Calidoscópio
Vol. 7, n. 3, p. 199-220, set/dez 2009

Leonel Figueiredo de Alencar
prof_leonel-lingcomp@yahoo.com.br

Produtividade morfológica e tecnologia do texto: aspectos da construção de um transdutor lexical do português capaz de analisar neologismos

Morphological productivity and text technology: Aspects of a lexical transducer of Portuguese capable of analyzing neologisms

RESUMO – Neste artigo, apresentamos o LEXPOR, protótipo de um componente morfológico do português capaz de segmentar e classificar os constituintes de derivados por meio da sufixação de *-ismo, -iano, -ês* e *-mente* bem como de derivados desses por prefixação com elementos de origem grega ou latina do tipo de *neo-*, *pseudo-*, *anti-* ou *ultra-*. Partimos do pressuposto de que uma representação das palavras complexas em termos de morfemas e categorias morfossintáticas não é só relevante na linguística de corpus, mas também em outras subáreas da tecnologia do texto, como a extração e a recuperação de informações. Este protótipo consiste de um transdutor lexical que modela o conjunto de palavras que se podem potencialmente construir usando esses afixos derivacionais. Esse transdutor foi compilado a partir de uma descrição da morfotática e das regras de alternância morfofonológicas e ortográficas desse fragmento do léxico, formalizada nas linguagens de programação de estados finitos xfst e lexc. A principal característica desse transdutor é a capacidade de realizar análises de neologismos construídos a partir de bases não lexicalizadas, tomadas de empréstimo de outras línguas. Como a utilização de antropônimos estrangeiros é uma das causas principais da extrema produtividade dos afixos derivacionais que focamos, o LEXPOR oferece uma arquitetura adequada para o desenvolvimento de um anotador automático de corpora do português capaz de preencher as lacunas de corpora como o CETENFolha e do analisador automático do projeto VISL. Em um como outro caso, as análises morfológicas de palavras complexas com os afixos derivacionais referidos frequentemente são insuficientemente detalhadas ou simplesmente incorretas.

Palavras-chave: derivação, sufixação, prefixação, autômatos, transdutores lexicais, morfologia de estados finitos, anotação automática de corpora, linguística computacional, linguística de corpus.

ABSTRACT – This paper presents LEXPOR, a prototype of a morphological component of Portuguese capable of segmenting and classifying the constituents of complex words resulting from suffixation *of -ismo*, *-iano*, *-ês* and *-mente* as well as from prefixing the words so derived with Greek or Latin prefixes such as *neo-*, *pseudo-*, *anti-*, or *ultra-*. We assume that a representation of complex words in terms of morphemes and morphosyntactic categories plays an important role not only in corpus linguistics, but also in other subfields of text technology, such as Information Extraction and Information Retrieval. This prototype consists of a lexical transducer modeling the set of words that can potentially be built using these derivational affixes. This transducer was compiled from a morphotactics and morphophonological description of this lexicon fragment as well as orthographic alternation rules formalized in the xfst and lexc finite-state programming languages. Its main feature is the ability to analyze neologisms built from non-lexicalized words borrowed from other languages. Since the use of foreign anthroponyms is one of the main causes of the extreme productivity of the derivational affixes we focus on, LEXPOR provides an adequate architecture for developing an automatic tagger for Portuguese, capable of overcoming the shortcomings of the CETENFolha corpus and of the parser for the VISL project. In both these cases, morphological analyses of complex words formed with the derivational affixes mentioned above are often either insufficiently detailed or simply incorrect.

Key words: derivation, suffixation, prefixation, automata, lexical transducers, finite-state morphology, automatic corpus annotation, corpus linguistics, computational linguistics.

Introdução

Segundo Lemnitzer e Wagner (2004, p. 246), sistemas de tecnologia da linguagem natural "necessitam de informações lexicais de forma muito mais abrangente e explícita do que usuários humanos de dicionários". Para esses dois autores, os recursos lexicais de um sistema desse tipo desempenham um papel equivalente ao do léxico mental de falantes humanos na produção e compreensão da linguagem.

Uma descrição de forma não só detalhada, mas também formalizada das estruturas lexicais constitui, portanto, pré-requisito para o processamento computacional de uma língua natural. Dicionários tradicionais como o

Aurélio ou o Houaiss, a exemplo do que se tem constatado em similares para outras línguas, não satisfazem essas duas exigências, pelo que se faz necessário desenvolver procedimentos que possibilitem a aquisição automática ou semi-automática de conhecimentos lexicais (Lemnitzer e Wagner, 2004, p. 246).

Entre esses procedimentos, destaca-se a anotação automática de corpora nos diferentes níveis de análise linguística. Com base nessas anotações, na situação ideal revisadas por humanos (Sasaki e Witt, 2004, p. 199), não só bases de dados lexicais, mas também analisadores (*parsers*) podem ser construídos automaticamente (Maier, 2007, p. 32). Corpora anotados automaticamente, por outro lado, têm se tornado cada vez mais indispensáveis à pesquisa não só em linguística descritiva, mas também teórica (Lemnitzer e Zinsmeister, 2006).

A anotação automática de corpora e a aquisição do conhecimento lexical constituem duas das subáreas da tecnologia do texto (*text technology*, *Text-technologie*), campo multidisciplinar voltado para o desenvolvimento de algoritmos para o processamento de textos enquanto dados semi-estruturados (Lobin e Lemnitzer, 2004, p. 1). Outras subáreas da tecnologia do texto são a extração de informações (*information extraction*, doravante IE) e a recuperação de informações (IR, do inglês *information retrieval*). Enquanto as duas primeiras subáreas integram também o campo da linguística computacional, as duas últimas constituem subdisciplinas da informática. Tanto a IE quanto a IR, porém, são campos de aplicação de várias tecnologias da linguagem natural, como a lematização, a etiquetação morfossintática de palavras e a análise sintática automática (Jurafsky e Martin, 2009; Krüger-Thielmann e Pailmans, 2004; Rehm, 2004).

O português, tanto em sua variedade europeia quanto brasileira, dispõe de vários corpora anotados morfossintática e/ou sintaticamente de forma automática. Esses corpora, contudo, não oferecem uma análise morfológica de palavras complexas como *patrioteirismo* ou *junguianismo*. Do mesmo modo, as principais ferramentas *on-line* de análise lexical do português também não dispõem da capacidade de segmentar neologismos desse tipo em seus morfemas constitutivos, classificando-os quanto ao tipo (raiz, sufixo etc.) e categoria lexical.

Neste artigo, apresentamos o LEXPOR, protótipo de um componente morfológico capaz de realizar análises, com esse nível de detalhamento, de derivados por meio da sufixação de *-ismo, -iano*, *-ês* e *-mente*[1] a partir de qualquer antropônimo bem como de derivados desses por prefixação com elementos de origem grega ou latina do tipo de *neo-*, *pseudo-*, *semi-*, *anti-*, *pós-* ou *sub-*. Este protótipo procura fazer jus à noção de léxico como subsistema da gramática mental onde também se verifica a criatividade linguística, a par do que ocorre na sintaxe (Anderson, 1992). Nessa visão, o léxico não se reduz a uma mera lista finita de itens, integrando, também, um componente morfológico, sob a forma de um sistema de regras que operam sobre esses itens para produzir um número potencialmente infinito de novos itens lexicais.

Implementado como um transdutor lexical, elaborado nas linguagens de programação de estados finitos xfst e lexc da Xerox (Beesley e Karttunen, 2003), o LEXPOR possui a capacidade de adivinhar e classificar como nome próprio a base de derivados *ad hoc* (passíveis, no entanto, de lexicalização) do tipo de *putinismo* (de Vladimir Putin, político russo) no Exemplo 1, extraído de artigo do jornalista Derek Bower, publicado no sítio UOL Notícias (2006) em 08/08/2006.[2]

Exemplo 1

> Robert Amsterdam [...] tornou-se um opositor veemente do "imperialismo energético da Rússia". Ele me disse no mês passado que os críticos alemães deste gasoduto andaram recebendo "ligações ameaçadoras durante a noite". Junto com mais gás russo virão outras "práticas de negócios" russas. "A Alemanha é o motor do *putinismo*", diz Amsterdam. "Nós temos diante de nós um imperialismo energético russo que nos é trazido por intermédio de agentes do setor financeiro da Alemanha".

Ao termo em itálico acima, o LEXPOR atribui a análise do Exemplo 2, caracterizando-o como substantivo (etiquetado como N) masculino singular derivado por meio da sufixação com -'*ism-* (o apóstrofo indica a sílaba tônica) a partir do nome próprio *Putin* (etiquetado como NPR). Por meio do programa *lookup*, que integra o pacote de ferramentas de estados finitos da Xerox, podemos utilizar o LEXPOR para anotar corpora automaticamente.

Exemplo 2

```
putin<NPR> 'ism<SUFF><N>o <Masc><Sg>
```

A formação de novas palavras com os sufixos *-ismo, -iano*, *-ês* e *-mente* a partir de antropônimos, com frequente

[1] Por mera conveniência, referimo-nos neste trabalho ao sufixo *-ismo* ou ao sufixo *-iano* etc., como se costuma tradicionalmente fazer (Monteiro, 1987, p. 150-161). No modelo da morfologia que subjaz à nossa análise, contudo, os sufixos são, na verdade, *-ism-* e *-ian-*, a vogal final constituindo um outro afixo (Schpak-Dolt, 1999, p. 84), como veremos mais adiante.

[2] Neste e nos demais exemplos extraídos da Internet e de corpora, destacamos em itálico a ocorrência em questão.

adjunção de prefixos de origem grega ou latina a esses derivados, é extremamente produtiva em português. No corpus CETENFolha[3] encontramos, entre muitos outros exemplos, os neologismos não dicionarizados *chedidismo* (do sobrenome *Chedid*, clã político paulista), *itamarismo* (relativo a Itamar Franco, ex-presidente da república), *spielberguiano* (referente a Steven Spielberg, cineasta norte-americano), *lacanês* (designativo da linguagem própria do psicanalista francês Jacques Lacan), *lawrencianamente* (do antropônimo *Lawrance*)[4] e *sub-hitchcockianismo* (do sobrenome de Alfred Hitchcock, cineasta britânico). Todos esses exemplos (e quaisquer outras criações lexicais análogas utilizando antropônimos não lexicalizados) são analisados corretamente pelo LEXPOR.

O sistema VISL[5], o mais robusto analisador automático do português disponível *on-line*, pelo contrário, não realiza uma decomposição morfológica das palavras complexas, classificando os diferentes componentes. O que é mais grave, com bastante frequência esse analisador oferece análises insatisfatórias de palavras complexas derivadas pela sufixação de *-iano*, *-ês*, *-ismo* e *-mente*, não as reduzindo, por exemplo, à base primitiva correta (ver análise do Exemplo 3, extraído do Exemplo 1).

Exemplo 3

A Alemanha é o motor do putinismo.
a [o] <artd> DET F S @>N
Alemanha [Alemanha] <civ> PROP F S @SUBJ>
é [ser] <fmc> V PR 3S IND VFIN @FMV
o [o] <artd> DET M S @>N
motor [motor] <mach> N M S @<SC
de [de] <sam-> PRP @N<
o [o] <artd> <-sam> DET M S @>N
putinismo [puta] <DERS> <DERS> <DERS> N M S @P<.

Nesse exemplo, podemos constatar, também, que são identificados três sufixos derivacionais em *putinismo*, quando, como veremos, se poderiam segmentar no máximo dois, admitindo por hipótese a forma-base *puta*.

Também derivados por meio de prefixos como *neo-*, a partir de formações por sufixação com *-iano*, *-ês* ou *-ismo*, não são analisados de forma adequada. Outra deficiência desse analisador é que, no caso de muitos derivados em *-iano*, apenas lematiza a palavra, sem detectar qualquer processo derivacional, como no caso, por exemplo, de *spielberguianos* (ver Exemplo 4).

Exemplo 4

Há muitos cineastas spielberguianos.
há [haver] <fmc> V PR 3S IND VFIN @FMV
muitos [muito] <quant> DET M P @>N
cineastas [cineasta] <Hprof> N M P @<ACC
spielberguianos [spielberguiano] ADJ M P @N<.

O analisador do projeto VISL é uma valiosa ferramenta para o tratamento computacional do português, permitindo a lematização e a anotação tanto morfossintática quanto sintática de corpora. Não oferece, contudo, análises com o grau necessário de refinamento exigido por outras áreas fundamentais da tecnologia do texto. Sob essa perspectiva, uma análise de palavras complexas no formato do Exemplo 2 é mais vantajosa. Como veremos na próxima seção, apenas corpora anotados com informações detalhadas sobre a estrutura das palavras podem constituir ponto de partida para análises automatizadas da estrutura morfológica do português. Por outro lado, no âmbito da IE/IR, uma análise morfológica como no Exemplo 2 permite que uma busca por *Putin* apresente também textos onde esse termo não ocorre, mas apenas palavras derivadas como *putinismo*.

No que segue, inicialmente fazemos um levantamento da situação da anotação automática de corpora do português, destacando as insuficiências dos corpora existentes quanto à análise morfológica de palavras complexas e as deficiências do analisador automático do projeto VISL. Esses problemas constituíram a motivação para a construção do LEXPOR. Na seção seguinte, realizamos uma análise linguística do fragmento da morfologia do português coberto pelo transdutor lexical. A penúltima seção, depois de delinear o modelo da Morfologia de Dois Níveis, mostra como o utilizamos na implementação computacional dessa análise em um transdutor de estados finitos, apresentando os resultados da aplicação desse transdutor a vários neologismos para os quais o analisador automático do projeto VISL oferece resultados insatisfatórios. A parte final do trabalho é destinada à conclusão.

Análise morfológica computacional do português: estado da arte

A investigação de fenômenos linguísticos nas variedades europeia e brasileira do português pode

[3] Este corpus está gratuitamente disponível para *download* no sítio da Linguateca (http://www.linguateca.pt/).
[4] Pelo contexto, provavelmente referente ao escritor inglês D.H. Lawrence.
[5] Esse analisador automático integra o projeto VISL (*Visual Interactive Syntax Learning*) da Universidade do Sul da Dinamarca, podendo ser acessado a partir da URL http://beta.visl.sdu.dk/visl/pt/parsing/automatic/.

fundamentar-se empiricamente em vários corpora anotados morfossintática e/ou sintaticamente. Esses corpora, contudo, não analisam morfologicamente as palavras complexas, decompondo-as em seus morfemas constitutivos, o que dificulta a pesquisa de fenômenos desse nível de análise.

O corpus CETENFolha, por exemplo, etiquetado automaticamente pelo PALAVRAS (Bick, 2000), a partir do qual se desenvolveu o analisador do projeto VISL, analisa a palavra *patrioteirismo* como sendo forma masculina singular do lema nominal PATRIOTEIRISMO:

Exemplo 5

<Sing> <a> Clóvis Rossi </s> <Pl> <Sing> São=Paulo=Ufanismo [São=Paulo=Ufanismo] PROP M S @SUBJ> e [e] <co-subj> KC @CO **patrioteirismo** [patrioteirismo] N M S @SUBJ> já [já] ADV @ADVL> são [ser] <fmc> V PR 3P IND VFIN @FMV duas [dois] <card> NUM F P @>N coisas [coisa] N F P @<SC que [que] <rel> SPEC M S @SUBJ> @#FS-N< enchem [encher] V PR 3P IND VFIN @FMV profundamente [profundamente] ADV @<ADVL as [o] <artd> DET F P @>N medidas [medida] N F P @<ACC

Uma criação lexical *ad hoc* como *chedidismo*, por sua vez, é analisada como substantivo masculino singular derivado por sufixação.

Exemplo 6

PT [PT] PROP M S @SUBJ> de [de] <sam-> PRP @N< a [o] <-sam> <artd> DET F S @>N região [região] N F S @P< de [de] PRP @N< Campinas [campina] N F P @P< se [se] <refl> PERS M/F 3S ACC @ACC> prepara [preparar] <fmc> V PR 3S IND VFIN @FMV para [para] PRP @<PIV combater [combater] V INF @IMV @#ICL-P< o [o] <artd> DET M S @>N $"« **chedidismo** [chedidismo] <DERS> N M S @<ACC $"» em [em] <sam-> PRP @N< a [o] <-sam> <artd> DET F S @>N região [região] N F S @P< de [de] PRP @N< Campinas [campina] N F P @P< $.

O tipo de anotação nos Exemplos 5 e 6 é suficiente do ponto de vista sintático, podendo ser utilizado, por exemplo, para a indução automática de regras de estruturação sintagmática do português. Léxicos de formas flexionadas também podem ser algoritmicamente construídos a partir dessas anotações. A partir de pares de formas e anotações como no Exemplo 7, não é difícil elaborar um programa que construa entradas lexicais de uma gramática independente de contexto baseada em estruturas de traços no formato do NLTK[6] (Bird *et al.*, 2009, p. 334), como no Exemplo 8, a qual pode ser utilizada num analisador (*parser*) sintático. No Exemplo 8, o atributo PRED tem como valor o predicado lógico-semântico expresso pelo verbo, como na LFG (Falk, 2001).[7]

Exemplo 7

enchem [encher] V PR 3P IND VFIN

Exemplo 8

V[TENSE=pres, PERS=3, NUM=pl, MOOD=ind, PRED='encher'] -> 'enchem'

Tanto do ponto de vista da investigação da estrutura morfológica das línguas naturais quanto da IE/IR, contudo, é necessária uma análise mais detalhada das palavras complexas do que a fornecida por corpora como o CETENFolha.

No que tange ao primeiro aspecto, um usuário do corpus que estivesse pesquisando, por exemplo, a iteração de processos derivacionais em português teria interesse em obter todas as palavras no corpus com dois ou mais sufixos. Para tanto, o corpus precisaria oferecer uma anotação que explicitasse a estrutura morfológica das palavras. No caso de *patrioteirismo*, essa análise, na forma mais simples possível, apenas segmentaria a palavra nos seus morfemas constitutivos, como o Exemplo 9 (Trost, 2004, p. 38).

Exemplo 9

patri+ot+eir+ism+o

Uma análise mais detalhada como o Exemplo 10, no formato do analisador automático do alemão SMOR (Schmid *et al.*, 2004), possibilitaria a realização de buscas mais sofisticadas no corpus. Essa análise constitui praticamente uma versão "lisa" (*flat*) de uma análise em termos

[6] O NLTK (abreviatura de *Natural Language Toolkit*) consiste, principalmente, de uma biblioteca em Python com várias ferramentas para o processamento computacional de línguas naturais, disponível *on-line* na URL http://www.nltk.org/.

[7] Na LFG, o atributo PRED inclui também a valência verbal, pelo que teríamos: 'encher <SUBJ OBJ>'. Graças à anotação sintática dos contextos precedente e subsequente, também a valência pode ser, com um algoritmo mais complicado, extraída automaticamente do CETENFolha.

de constituintes imediatos representada em um diagrama arbóreo como o da Figura 1, do tipo das análises que encontramos em trabalhos de morfologia de orientação estruturalista (ver, por exemplo, Schpak-Dolt, 1999, p. 74) ou gerativa (ver, por exemplo, Grewendorf *et al.*, 1989, p. 279). A diferença entre as duas análises é que, no formato do SMOR, uma palavra derivada como *patrioteirismo* é reduzida não à sua raiz (*pátri-*), mas à palavra primitiva da qual deriva (*pátria*).

Exemplo 10

```
pátria<N>ot<SUFF><A>eir<SUFF><A>'ism<S
UFF><N><Masc><Sing>
```

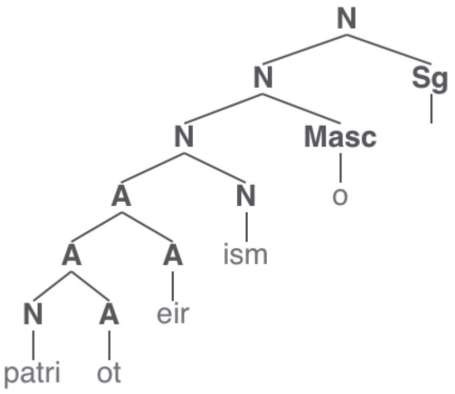

Figura 1: Representação arbórea de patrioteirismo.[8]
Figure 1: Tree representation of patrioteirismo.

Análises no formato do Exemplo 10, em que não só se segmentam, mas também se classificam os morfemas, fornecem informações extremamente úteis para pesquisas sobre a morfologia derivacional. No Exemplo 11, consideremos os exemplos anotados no formato do LEXPOR, em que se indicam as raízes, e não os lemas, das palavras primitivas das quais derivam, por meio de um ou sucessivos processos derivacionais, as palavras complexas etiquetadas. A sílaba tônica é indicada por meio de apóstrofo apenas no caso de palavras acentuadas como *lóbi* ou *árvore*. O sufixo *-ism-* é sempre tônico, mas só é graficamente acentuado em casos como *simultaneísmo*, em que é imediatamente precedido de vogal.

Exemplo 11

```
patrioteirismo/'patri<N>ot<SUFF><A>eir<SUFF
><A>'ism<SUFF><N><Masc><Sing>
antidirigismo/neo<PREF>dirig<V>'ism<SUFF>
<N><Masc><Sing>
purismos/pur<A>'ism<SUFF><N><Masc><Pl>
achismo/ach<V>'ism<SUFF><N><Masc><Sing>
lobismo/'lob<N>'ism<SUFF><N><Masc><Sing>
ultrapatriotismo/'patri<N>ot<SUFF><A>'ism<S
UFF><N><Masc><Sing>
arvorismo/'arvor<N>'ism<SUFF><N><Masc><Sing>
neofernandianismos/neo<PREF>fernand<NPR>
ian<SUFF><A>'ism<SUFF><N> <Masc><Pl>
subhitchcockianismo/sub<PREF>hitchcock<NPR
>ian<SUFF><A>'ism<SUFF><N> <Masc><Pl>
```

Utilizando expressões regulares, podemos extrair facilmente todos os exemplos de palavras com dois ou mais sufixos. O comando abaixo utiliza a ferramenta grep do sistema operacional UNIX para extrair todas as construções do Exemplo 11 (armazenadas num arquivo de nome *corpus*) que apresentam subcadeias (*substrings*) satisfazendo a expressão regular demarcada pelas aspas.

Exemplo 12

```
grep -E "(<SUFF>[\<\>[:alpha:]]+){2,}" corpus
patrioteirismo/'patri<N>ot<SUFF><A>eir<SUFF
><A>'ism<SUFF><N><Masc><Sing>
ultrapatriotismo/'patri<N>ot<SUFF><A>'ism<S
UFF><N><Masc><Sing>
neofernandianismos/neo<PREF>fernand<NPR>
ian<SUFF><A>'ism<SUFF><N> <Masc><Pl>
subhitchcockianismo/sub<PREF>hitchcock<NPR
>ian<SUFF><A>'ism<SUFF><N> <Masc><Sg>
```

Uma das questões fundamentais da morfologia é determinar a produtividade dos processos morfológicos. Como veremos mais adiante, o sufixo *-ismo* não restringe suas bases a uma única categoria lexical, admitindo tanto adjetivos quanto substantivos e verbos. Intuitivamente, porém, esse processo derivacional é muito mais produtivo a partir da primeira classe do que da segunda, enquanto, no caso da última classe, parece estarmos diante de um processo, quando muito, semiprodutivo. Como verificar essas intuições em corpora? Por meio de anotações no formato do LEXPOR, podemos extrair e calcular automaticamente a frequência das diferentes categorias sintáticas das bases de um sufixo. A seguinte sequência de comandos do UNIX exemplifica isso

[8] Esta representação foi construída por meio de ferramenta do NLTK (ver nota 6).

Exemplo 13

```
grep -Eo "<[[:alpha:]]+>ism<SUFF>" corpus |
sort | uniq -c
5 <A>ism<SUFF>
2 <N>ism<SUFF>
2 <V>ism<SUFF>
```

Pelo que sabemos, nenhum corpus do português oferece uma anotação da estrutura interna das palavras. A esse respeito em particular, a linguística de corpus do alemão está mais adiantada. Corpora da língua alemã como aqueles disponibilizados pelo sistema COSMAS II[9] permitem, por exemplo, obter todas as palavras derivadas a partir de uma dada palavra básica (*Grundwort*).

Por outro lado, enquanto o alemão dispõe do analisador automático SMOR, capaz de analisar formações completamente novas (i.e. não dicionarizadas nem presentes em corpora), decompondo-as em seus elementos constitutivos, a base de dados MorDebe[10] e o analisador morfológico do português do projeto VISL[11] são bastante limitados quanto a esse aspecto. A primeira não inclui uma palavra como *patrioteirismo*, uma vez que não implementa a produtividade morfológica na formação de palavras por derivação. Apenas palavras derivadas listadas em seu banco de dados, como *patrioteiro* e *patriotismo*, são contempladas.

A formação de substantivos a partir de outros substantivos ou de adjetivos por meio da sufixação com *-ismo* é um dos processos morfológicos mais produtivos do português. Com bastante frequência, deparamo-nos com formações completamente novas como *laranjismo* ou *mensalismo*, não registradas nem mesmo em dicionários *on-line* de atualização extremamente dinâmica, como o iDicionário Aulete.[12]

Exemplo 14

> Afinal, o Senado dirá ao país que tudo pode: falsificação, mentiras, *laranjismo*, sonegação, lobismo... (*Veja*, 2007).

Exemplo 15

> O petismo degradou-se em lulismo, nas esferas do Planalto, e degradou-se em *mensalismo*, aparelhismo e deboche no campo parlamentar (*Folha de São Paulo*, 2007).

Desse modo, como fonte de consulta a respeito da boa formação e estrutura das palavras do português, a base de dados MorDebe é inadequada pela falta de robustez. O analisador VISL, por sua vez, é robusto, uma vez que lematiza e analisa morfossintaticamente de forma correta neologismos com esse sufixo (ver Exemplo 16). Esse analisador, contudo, se por um lado determina a palavra primitiva da qual deriva, por prefixação e sufixação, uma forma como *neocafajestismos* (analisado como [cafajeste] <DERS> <DERP> N M P), não realiza uma análise morfológica de muitos derivados em *-ismo*, como podemos constatar no Exemplo 16.

Exemplo 16

```
Ufanismos e patrioteirismos me aborrecem muito.
ufanismos [ufanismo] <ism> N M P @SUBJ>
e [e] <co-subj> KC @CO
patrioteirismos [patrioteirismo] <ism> N M P @
SUBJ>
me [eu] PERS M/F 1S ACC @ACC>
aborrecem [aborrecer] <fmc> V PR 3P IND VFIN
@FMV
muito [muito] <quant> ADV @<ADVL.
```

Considerando as necessidades da pesquisa na área de morfologia derivacional com base em corpora, o analisador VISL é também inadequado porque não segmenta e classifica os morfemas de palavras como *shakespearianismo*, analisada como [Shakespeare] <DERS> <DERS> N M S. Essa análise informa apenas que se trata de um substantivo masculino singular derivado do lema *Shakespeare* por meio de dois sufixos derivacionais, mas não destaca os sufixos envolvidos. No entanto, no caso de derivados análogos a partir de nomes próprios que não constam na base de dados do sistema, como, por exemplo, *Jung*, a análise é bem menos satisfatória: dado o input *junguianismo*, o output é [junguiano] <DERS> N M S. Analogamente, *junguianos* é analisado como [junguiano] ADJ M P.

O sistema VISL, com bastante frequência, não reduz derivados desse tipo à base primitiva correta (ver análise do Exemplo 17, extraído do Exemplo 1), nem lida adequadamente com a interação entre os sufixos *-iano*, por um lado, e *-ismo* e *-mente*, por outro, como no caso, respectivamente, de *sartrianismo* e *sartrianamente*, ambos reduzidos a [sartriano] e não a [Sartre], sem que seja indicada, no segundo caso, incidência de um processo derivacional, como podemos ver no Exemplo 18.

[9] CosmasII disponível em http://www.ids-mannheim.de/cosmas2/.
[10] MorDebe disponível em http://www.iltec.pt/mordebe/.
[11] VISL disponível em http://beta.visl.sdu.dk/visl/pt/parsing/automatic/parse.php/.
[12] iDicionário Aulete disponível em http://aulete.uol.com.br/.

Exemplo 17

A Alemanha é o motor do putinismo.
a [o] <artd> DET F S @>N
Alemanha [Alemanha] <civ> PROP F S @SUBJ>
é [ser] <fmc> V PR 3S IND VFIN @FMV
o [o] <artd> DET M S @>N
motor [motor] <mach> N M S @<SC
de [de] <sam-> PRP @N<
o [o] <artd> <-sam> DET M S @>N
putinismo [puta] <DERS> <DERS> <DERS> N M S @P<.

Exemplo 18

O cineasta spielberguiano sartrianamente trocou o neo-itamarismo pelo subhitchcockianismo.
o [o] <artd> DET M S @>N
cineasta [cineasta] <Hprof> N M S @SUBJ>
spielberguiano [spielberguiano] ADJ M S @N<
sartrianamente [sartriano] ADV @ADVL>
trocou [trocar] <fmc> V PS 3S IND VFIN @FMV
o [o] <artd> DET M S @>N
neo-itamarismo [Itamar] <DERS> <DERS> N M S @<ACC
por [por] <sam-> PRP @<PIV
o [o] <artd> <-sam> DET M S @>N
subhitchcockianismo [hitchcockiano] <DERS> <DERP> N M S @P<.

Se, por um lado, o VISL acertadamente detecta o antropônimo do qual deriva o neologismo *neo-itamarismo* e identifica a derivação sufixal e prefixal em *subhitchcockianismo* (por meio, respectivamente, das etiquetas <DERS> e <DERP>), comete, por outro, vários erros graves. A análise de *putinismo* não é apenas absurda no contexto do artigo de Derek Bower, cuja temática central é a dependência da União Europeia em relação ao gás produzido na Rússia e como isso, por intermédio da Alemanha, o país mais rico da Europa e extremamente dependente desse recurso energético, tem fomentado uma prática política baseada na doutrina do Presidente Vladimir Putin. Do ponto de vista da morfologia do português, é completamente injustificável uma análise de *putinismo* como palavra derivada por meio de três sufixos a partir da base *puta*. No máximo, poderíamos segmentar os sufixos *-ino*, formador de adjetivos que indicam natureza, como em *cristalino* (Monteiro, 1987, p. 158), e *-ismo*: *put+a => put+in+a => put+in+ism+o*.

Em segundo lugar, o VISL, embora reduza corretamente *neo-itamarismo* a [Itamar], erroneamente analisa essa palavra complexa como resultado de uma dupla sufixação, ignorando, portanto, a natureza do prefixo *neo-*. Outro exemplo de derivado com esse prefixo para o qual o VISL não oferece uma análise satisfatória é *neo-molieriano*, analisado como ALT xxxeriano [neomolier] <DERS> ADJ M S.[13]

A análise automática do Exemplo 19 deixa claro que ao VISL não subjaz um modelo adequado da morfologia do português no que tange à prefixação com elementos de origem grega do tipo de *pseudo-* e *anti-*, extremamente produtivos. De fato, o primeiro derivado por prefixação com *pseudo-* e *anti-* é analisado como forma verbal da primeira pessoa do singular do presente do indicativo do lema [psi] – uma análise totalmente infundada. O derivado *antipseudo-sartrianismo* é equivocadamente reduzido à forma-base *psi-sartriano*, do qual resultaria por meio de um único processo derivacional, de natureza sufixal. A contribuição do prefixo *anti-* na formação dessa palavra complexa é, portanto, completamente ignorada.

Exemplo 19

O ensaísta denunciou o comportamento pseudo-antipseudo-intelectual do antipseudo-sartrianismo.
o [o] <artd> DET M S @>N
ensaísta [ensaísta] <Hprof> N M S @SUBJ>
denunciou [denunciar] <fmc> V PS 3S IND VFIN @FMV
o [o] <artd> DET M S @>N
comportamento [comportamento] <act> N M S @<ACC
pseudo-antipseudo [psi] <DERS> <DERS> <DERS> <DERP> <fmc> V PR 1S IND VFIN @FMV
-
intelectual [intelectual] <Hideo> N M S @S<
de [de] <sam-> PRP @N<
o [o] <artd> <-sam> DET M S @>N
antipseudo-sartrianismo [psi-sartriano] <DERS> N M S @P<.

Finalmente, outra deficiência do analisador VISL é não detectar nem a dupla sufixação nem o antropônimo do qual deriva *subhitchcockianismo*, i.e. *Hitchcock*, problema que se repete em *junguianamente*. Finalmente, de forma

[13] Segundo Bick (2000, p. 16), a etiqueta ALT assinala "orthographical changes introduced by the tagger's heuristics module – spelling/accent correction etc.". Quanto à notação xxx, é empregada para indicar tanto uma raiz hipotética, não encontrada no léxico do *parser*, quanto uma raiz que viola as regras de combinação morfológica.

bastante insatisfatória, apresenta, com frequência, como formas não derivadas, exemplos do tipo de *spielberguiano* (como vimos no Exemplo 4) ou *lacanês*, analisado como ALT xxxês [lacanês] <Hnat> <ling> N M S.

Sob a perspectiva da tecnologia do texto, há que se preferir uma análise de palavras complexas no formato do LEXPOR (ver Exemplo 20) ao formato do VISL (ver Exemplo 21).

Exemplo 20

```
sub<PREF>hitchcock<NPR>ian<SUFF><A>
'ism<SUFF><N><Masc><Pl>
```

Exemplo 21

```
subhitchcockianismo [hitchcockiano] <DERS>
<DERP> N M S
```

No formato de análise do Exemplo 20, não só se segmentam, mas também se classificam os morfemas constitutivos de palavras complexas, o que é especialmente valioso no caso de neologismos não dicionarizados como *subhitchcockianismo* ou *putinismo*. Como mostramos nos Exemplos 12 e 13, apenas corpora anotados com informações detalhadas sobre a estrutura das palavras podem contribuir de forma significativa para pesquisas empíricas em morfologia.

O formato do Exemplo 20 é mais adequado, também, para aplicações em IE e IR. Por um lado, uma das tarefas da IE é o reconhecimento de entidades nomeadas (NER, acrônimo de *named entity recognition*) e das relações que subsistem entre elas (Jurafsky e Martin, 2009, p. 768). No artigo sobre o papel da Rússia na questão energética europeia, de que transcrevemos trecho no Exemplo 1, não ocorre o sobrenome *Putin*. No entanto, esse texto trata desse estadista russo, fornecendo informação importante para a sua caracterização, na medida em que o Exemplo 22a é parafraseável como o Exemplo 22b. De fato, estabelece uma teia complexa de relações entre as entidades nomeadas *Alemanha*, *doutrina* e *Putin*. Essas relações só podem ser reconhecidas, contudo, a partir de uma análise morfológica nos moldes do Exemplo 2.

Exemplo 22

(a) A Alemanha é o motor do putinismo.
(b) A Alemanha é o motor da doutrina de Putin.

De modo análogo, uma análise morfológica como o Exemplo 2 permite, na área da IR, que uma busca por *Putin* apresente também textos onde esse termo não ocorre de forma isolada, mas em palavras derivadas que contribuam com informações sobre o objeto da pesquisa.

Aspectos da morfologia do português

Nesta seção, apresentamos as concepções sobre a estrutura nominal e a derivação prefixal e sufixal do português que subjazem ao LEXPOR.

Os nomes simples (i.e. não derivados e não compostos) do português, analogamente a outras línguas românicas, têm a estrutura esquematizada no Quadro 1, baseada, em linhas gerais, na análise de Schpak-Dolt (1999, p. 32-37) para o espanhol.[14] A flexão do plural é representada uniformemente por *s* também em casos como *ferozes*, pois pressupomos que essa forma resulta de regra fonológica que insere um *e* epentético no plural dos nomes com tema em consoante.

Em abordagens estruturalistas como Monteiro (1987), faz-se uma distinção entre vogal temática e desinência de gênero, como mostra a análise do Quadro 2 das formas *aventura*, *pura* e *puros*.

No entanto, vogal temática e desinência de gênero nunca aparecem separadas em português, configurando, portanto, em vez de morfemas distintos, o que Matthews (1974, p. 147) chama de exponência cumulativa. Trata-se de fenômeno típico das línguas flexionais como as línguas eslavas ou românicas, o qual consiste em que mais de uma propriedade morfossintática é expressa por um mesmo expoente (noção do modelo morfológico de Palavra e Paradigma *grosso modo* equivalente à de morfema). Nas línguas aglutinantes (como o turco e o finlandês), pelo contrário, as propriedades morfossintáticas são expressas por expoentes distintos.

Conforme o Quadro 1, o morfema *-a* de um substantivo como *aventura* acumula, como é típico nas línguas flexionais, duas funções: (i) indicar o gênero feminino do substantivo e (ii) classificá-lo no paradigma flexional dos substantivos terminados em *-a*. Uma análise nos moldes do Quadro 2, pelo contrário, força o enquadramento da estrutura nominal do português num modelo mais adequado a línguas aglutinantes.

Enquanto o gênero do substantivo é inerente à sua raiz (ou ao seu radical), a raiz adjetival (ou o radical no caso de adjetivos complexos) é desprovida de traço de gênero, que é determinado pela desinência que serve também de morfema classificador. As desinências *-o* e *-a* determinam o gênero masculino ou feminino ao adjungir-se a radicais de adjetivos do tipo de *puro*. No

[14] Para Schpak-Dolt (1999), a vogal átona final dos substantivos constitui um morfema classificador; nos adjetivos, essa vogal, para ele, é flexão de gênero.

Quadro 1: Estrutura nominal do português.
Chart 1: Portuguese noun structure.

Forma	Raiz	Flexão de gênero/classificador	Flexão de número
detalhes	detalh	e	s
aventura	aventur	a	∅
puros	pur	o	s
rebelde	rebeld	e	∅
ferozes	feroz	∅	s

Quadro 2: Estrutura nominal do português segundo Monteiro (1987).
Chart 2: Portuguese noun structure according to Monteiro (1987).

Raiz	Vogal temática	Desinência de gênero	Desinência de número
aventur	a	∅	∅
pur	∅	a	∅
pur	o	∅	s

caso de adjetivos do tipo de *rebelde* e *celta*, as desinências *-e* e *-a*, respectivamente, são ambíguas quanto ao gênero, o que ocorre também com a desinência zero em adjetivos do tipo de *feroz*. Consequentemente, as formas desses adjetivos são sistematicamente ambíguas.

Contrariamente ao ponto de vista amplamente difundido em análises morfológicas do português e do espanhol[15], vocábulos como *menina* e *aluna* não são considerados formas flexionadas no feminino dos lemas *menino* e *aluno*, respectivamente, mas substantivos derivados por sufixação zero a partir dos correspondentes masculinos, analogamente a substantivos como *profetisa* e *heroína*, derivados, respectivamente, por meio dos sufixos (não mais produtivos) *-isa* e *-ina*. A classe de substantivos passíveis de alimentar esse processo derivacional caracteriza-se pelo traço "moção" (Schpak-Dolt, 1999, p. 35-36), i.e. podem ser "movidos" por derivação de um gênero para outro.

Como dissemos, as vogais átonas finais dos nomes funcionam tanto como flexão de gênero quanto como classificador. Com base nessas vogais, podemos estabelecer classes nominais, como, por exemplo, substantivos masculinos em *-a* (*problema*), substantivos femininos em *-e* (*ponte*), adjetivos em *-o* ou *-a*, adjetivos uniformes em *-e* (*ilustre*) etc. Outro parâmetro classificatório é a propriedade da "moção", i.e. a capacidade de substantivos masculinos como *aluno* de sofrerem mudança de gênero. Em português, há também nomes que, no singular, não apresentam vogal átona final, constituindo a classe dos nomes de tema em consoante (*futebol*, *mulher*, *feroz*)[16]. Também integram essa classe alguns nomes terminados em *-ão*, os quais se desdobram em várias subclasses, segundo a forma do feminino singular e a do masculino plural (ver Quadro 3).[17] O símbolo N representa, conforme Câmara Jr. (1987, p. 58-59, 95), arquifonema que pode realizar-se como consoante nasal de uma sílaba seguinte (por exemplo, em *cidadanismo*) ou subsistir como travamento nasal, estendendo a nasalidade à vogal ou ditongo precedente (por exemplo, em *cidadão*).

A vogal *-e* (ou *-i*) da terminação do plural dos temas em consoante resulta de regra de epêntese. As formas de superfície dos nomes em *-ão* também são resultado da operação de regras fonológicas. Na próxima seção, trataremos dessas regras, mostrando como implementá-las computacionalmente num transdutor de estados finitos.

Como vimos na seção anterior, o sufixo *-ismo* é extremamente produtivo. Uma das razões para essa produtividade é a diversidade categorial das bases. De fato, esse sufixo adjunge-se não só a adjetivos e substantivos, mas também a verbos, do que damos alguns exemplos consignados no iAulete:

Exemplo 23

dirigir → dirigismo, achar → achismo, transformar → transformismo

[15] Um exemplo em português é Monteiro (1987, p. 67). Exemplos em espanhol são citados por Schpak-Dolt (1999).
[16] Não consideramos aqui os temas em vogal tônica. Por outro lado, em nossa análise dos nomes do tipo de futebol ou feroz, divergimos de Câmara Jr. (1987, p. 86, 95), que postula, nesses casos, um tema teórico em *-e*.
[17] Em conformidade com Câmara Jr. (1987, p. 95), postulamos um tema teórico em -o para os nomes em *-ão* que fazem o plural em *-ãos*. Diferentemente desse autor, porém, analisamos os nomes do tipo de *campeão* e *alemão* como membros da classe dos temas em consoante.

A razão maior para a produtividade do sufixo, contudo, é a possibilidade de utilizar como base não só itens listados no léxico, como nos Exemplos 23 e 24, mas também o produto de processos morfológicos extremamente produtivos, como a derivação (ver Exemplo 25), a composição (ver Exemplo 26) e a acrossemia (ver Exemplo 27).[18]

Exemplo 24

laranjismo ← laranja, lobismo ← lóbi

Exemplo 25

salesianismo ← salesiano ← Sales

Exemplo 26

novo-riquismo ← novo-rico, quarto-mundismo ← Quarto Mundo

Exemplo 27

petismo ← PT (Partido do Trabalhadores), uspismo ← USP (Universidade de São Paulo), vipismo ← VIP (*very important person*), cebrapianismo ← cebrapiano ← CEBRAP (Centro Brasileiro de Análise e Planejamento)

Outra fonte inesgotável de bases para a derivação por meio de *-ismo* são os antropônimos, devido à produtividade dos processos de criação de novos itens desse tipo (*Itamar*, de que deriva *itamarismo*), como no caso dos hipocorísticos (*lulismo* ← *Lula* ← *Luís*, *janguismo* ← *Jango* ← *João Goulart*, *lalauísmo* ← *Lalau* ← *Nicolau*) e sobrenomes estrangeiros (*junguismo* ← *Jung*).[19] Pelo fato de esse processo derivacional poder ser alimentado com o resultado de outros processos por si mesmos altamente produtivos, podemos dizer, sem exagero, que o potencial de criação de neologismos em *-ismo* é infinito.

O sufixo *-iano* normalmente é analisado como alomorfe de *-ano* (Cunha e Cintra, 1985, p. 98; Monteiro, 1987, p. 152). Uma análise dos derivados a partir desses dois elementos no CETENFolha, porém, sugere fortemente que se trata de dois sufixos distintos, hipótese que implementamos no LEXPOR, como veremos na próxima seção. Conforme o Quadro 4, *-iano* é adjungido tipicamente a antropônimos, para formar adjetivos parafraseáveis como "referente a, próprio ou característico de X", onde X é a base do derivado. O sufixo *-ano*, por sua vez, adjunge-se a substantivos para produzir adjetivos parafraseáveis como "procedente de X", onde X geralmente designa um lugar.

O Quadro 4 mostra claramente que *-iano* e *-ano* ocorrem nos mesmos contextos fonológicos. Por exemplo, tanto a raiz de *Édipo* quanto a de *Sergipe* terminam em oclusiva bilabial surda, mas, enquanto a primeira é sufixada com *-iano*, à segunda se adjunge *-ano*. Não se trata, porém, de variação livre entre alomorfes, uma vez que **edipano* e **sergipiano* são agramaticais e os dois sufixos diferem quanto às propriedades selecionais e à funcionalidade.

No CETENFolha, encontramos exemplos que não se enquadram na distribuição exemplificada no Quadro 4, mas que constituem, em grande parte, exceções apenas aparentes. No primeiro caso, temos exemplos como *morumbiano* e *ucraniano*, em que a raiz nominal termina em /i/. Casos como *oxfordiano*, *hollywoodiano*, *liliputiano*, *uspiano* ou *febiano* (de *FEB*, Força Brasileira Expedicionária) explicam-se pela presença desse fonema em posição final na pronúncia das bases. Diferentemente de *-iano*, o sufixo *-ano* adjunge-se não à raiz, mas ao tema de nomes em *-e* ou *-es*, com mudança da vogal final para

Quadro 3: Tipos de nomes em *-ão* em português.
Chart 3: Types of nouns ending in *-ão* in Portuguese.

Singular masculino		Feminino	Plural masculino	Tema
(i)	cidadão	cidadã	cidadãos	cidadaNo
(ii)	campeão	campeã	campeões	campeoN
(iii)	alemão	alemã	alemães	alemaN
(iv)	leão	leoa	leões	leoN

[18] Segundo Monteiro (1987, p. 175-176, 185), a acrossemia é um processo de formação de palavras a partir de uma sigla (por ex. PT, USP e VIP) ou de sílabas de dois ou mais vocábulos (por exemplo, o antropônimo *Jomar*, derivado de João e Maria), resultando na criação de um novo vocábulo com "autonomia de significante". Desse modo, um vocábulo acrossêmico como USP pronuncia-se "uspe" e não "uessepê".
[19] Sobre a formação de antropônimos e hipocorísticos em português, consulte-se Monteiro (1987, p. 184-196).

Quadro 4: Distribuição dos sufixos *-iano* e *-ano* em exemplos do CETENFolha.
Chart 4: Distribution of the suffixes *-iano* and *-ano* in examples from the CETENFolha corpus.

Derivado a partir de antropônimo*		**Derivado a partir de topônimo, nome comum de lugar ou de instituição**	
Édipo	edipiano	Sergipe	sergipano
Villa-Lobos	vilalobiano	Piracicaba	piracicabano
Lombroso	lombrosiano	diocese	diocesano
Rosa	rosiano		
Mozart	mozartiano	Esparta	espartano
Ford	fordiano	Tibet	tibetano
Machado	machadiano	Minessota	minessotano
Eduardo	eduardiano		
Lacan	lacaniano		
Kafka	kafkiano	Tijuca	tijucano
Spielberg	spielberguiano		
Sherlock	sherloquiano	Franca	francano
Artur	arturiano		
César	cesariano	Itabira	itabirano
Guerra	guerriano**	serra	serrano

* O sufixo *-iano* também ocorre com teônimos: *venusiano* ← *Vênus*.
** Exemplo extraído de Loureiro (2009). Os demais exemplos são do CETENFolha.

i: *Iraque* → *iraquiano*, *Cabo Verde* → *cabo-verdiano*, *Açores* → *açoriano* etc. Casos como *pessoano*, *galileano* e *mallarmeano* sugerem que o fonema /i/ do sufixo *-iano* sofre aférese diante de vogal final da raiz nominal: *pesso+ian+o* → *pesso+an+o*.[20]

As verdadeiras exceções explicam-se pela lexicalização. De um lado, temos itens dicionarizados como *elizabetano*, *franciscano*, *maometano* e *luterano*, derivados de antropônimos por sufixação com *-ano*.[21] De outro, temos derivados de topônimos em que se segmenta o sufixo *-iano*, como *caucasiano* ← *Cáucaso*, *equatoriano* ← *Equador*, *iraniano* ← *Irã* e *veneziano* ← *Veneza*, também lexicalizados.

Como *-iano*, segundo nossa análise, adjunge-se a antropônimos e novos itens desse tipo são constantemente criados ou introduzidos no português a partir de outras línguas, esse sufixo, a exemplo de *-ismo*, também tem uma produtividade inesgotável.

Tradicionalmente, o sufixo *-ês* adjunge-se a nomes designativos de lugar (*montanhês* ← *montanha*, *francês* ← *França*), expressando origem, em concorrência com o sufixo *-ano*. Cunha e Cintra (1985, p. 98) e Monteiro (1987, p. 156) consideram-no variação de *-ense*. Um levantamento no corpus CETENFolha evidencia uma especialização semântica desse sufixo no português brasileiro contemporâneo, o qual passa a designar jargão, dialeto ou congênere próprio da entidade designada pela base, que pode ser um antropônimo (*lacanês* ← *Lacan*) ou um nome de instituição (*sorbonnês* ← *Sorbonne*, *ibeemês* ← *ibeeme* ← *IBM*), de profissão (*psicologuês*, *sociologuês*), de atividade humana (*politiquês*, *futebolês*) etc. Também adjetivos pátrios constituem bases para o sufixo *-ês* nessa acepção mais recente (*gauchês*, *paulistês*, *baianês*).

Por admitir como bases também antropônimos estrangeiros, o potencial de criação de novas palavras com esse sufixo é grande, a exemplo de *-iano*, embora não se constatem no CETENFolha tantos derivados quanto com esse último sufixo. Os Exemplos 28 e 29, coletados na Internet, evidenciam a produtividade desse sufixo na formação de termos designativos da linguagem própria de um indivíduo ou classe de indivíduos. No segundo exemplo, constatamos que um derivado em *-ês* pode constituir a base de um processo de derivação prefixal.

[20] No CETENFolha, ocorre a grafia não padrão *sartreano* (derivado de Sartre), ao lado da forma padrão *sartriano*.
[21] É possível que o sufixo *-ano* em pelo menos alguns desses exemplos se explique por empréstimo de outras línguas (por exemplo, *elizabetano* do inglês *elizabethan*). A única exceção desse tipo encontrada no CETENFolha que não se explica pela lexicalização é *espinosano*.

Exemplo 28

Neste sentido, incorporar Kafka é incorporar uma dicção específica da modernidade, o *kafkês* é uma língua nova [...] (Martinez, 2006).

Exemplo 29

O colunista da Folha de S. Paulo José Simão é autor de verbetes hilários que nos ajudam a entender o cenário político brasileiro. Desde que criou o *lulês*, o *tucanês* e o *antitucanês* [...] (View, 2007).

Enquanto os sufixos *-ismo*, *-iano* e *-ês* se adjungem a radicais, *-mente* é sufixado a formas de adjetivos flexionados no feminino (Schpak-Dolt, 1999, p. 74), no caso de temas em vogal, e a formas de adjetivos no masculino, nos demais casos, como mostram as derivações do Exemplo 30.[22]

Exemplo 30

(a) Kant → kantiano → kantianismo
(b) Kant → kantismo
(c) kantiana → kantianamente
(d) célebre → celebremente
(e) feroz → ferozmente
(f) francês → francesmente

O potencial de criação lexical por meio da sufixação com *-mente*, como no caso dos sufixos *-ismo* e *-iano*, é igualmente infinito, uma vez que se pode irrestritamente alimentar do produto da sufixação por meio de *-iano*.

Como na sintaxe, também na morfologia há recursividade (Trommer, 2004, p. 217). Em português, no âmbito da formação de palavras, constatamos essa propriedade na prefixação. Formações com dois prefixos não são incomuns, como mostram os neologismos do Exemplo 31, todos extraídos, por meio do Google, de textos em português na WWW. A natureza recursiva da prefixação em português evidencia-se na possibilidade de adjungir ainda mais um prefixo a construções com dois ou mesmo três prefixos, como nos derivados do Exemplo 32, também colhidos na Internet. Nos dois grupos de derivados, mantivemos as idiossincrasias no uso do hífen e do espaço em branco dos exemplos originais, as quais vão de encontro às normas ortográficas do português. Conforme Cunha e Cintra (1985, p. 66-67), deveríamos grafar, por exemplo, *antipseudoliterários*, *pseudo-antipseudo-intelectual* e *hipersupermegaultrapoderoso*.

Exemplo 31

anti-pseudo-cientistas, antipseudo-literários, anti-neo-fascismo, anti-neo-liberal, neo-neo-realismo

Exemplo 32

pseudo-anti-pseudo-intelectual, hiper-super-mega-feliz, hiper-super-mega-ultra poderoso

Em português, duas características gerais opõem a derivação por prefixação à derivação por sufixação. No plano morfossintático, prefixos não alteram as propriedades gramaticais das suas bases, ao passo que sufixos tipicamente o fazem. O sufixo *-ismo*, por exemplo, transforma adjetivos e verbos em substantivos. Mesmo quando a base é um substantivo feminino, o produto da derivação por meio desse sufixo é um substantivo masculino. No plano fonológico, enquanto o acento das bases é preservado nos derivados por prefixação, a maioria dos sufixos são tônicos, resultando num deslocamento do acento. Os derivados dos Exemplos de 33 a 36 exemplificam esses contrastes.[23]

Exemplo 33

árvore (N) (Fem.) → arvorismo (N) (Masc.) → semi-arvorismo (N) (Masc.)

Exemplo 34

Sartre (N) → sartriano (A) → pseudo-sartriano (A)

Exemplo 35

França (N) → francês (A) → francesismo (N) → neofrancesismo (N)

[22] Segundo Cunha e Cintra (1985, p. 101), a discrepância entre formas como *lindamente* e *francesmente* (**francesamente* na norma culta) explica-se pelo fato de todos os adjetivos em *-ês* terem sido uniformes num estágio anterior da língua, propriedade que, à exceção de poucos exemplos como *pedrês* e *montês*, desapareceu do português atual.
[23] Um exemplo de sufixo átono é *-vel* (cf. por ex. *realizável e vendível*).

Exemplo 36

simultâneo (A) → simultaneísmo (N) → anti-simultaneísmo (N)

LEXPOR: um transdutor lexical robusto do português

Nesta seção, delineamos inicialmente a Morfologia de Dois Níveis, o modelo da estrutura das palavras que normalmente subjaz aos analisadores morfológicos implementados como transdutores de estados finitos. Veremos que esse tipo de abordagem é especialmente útil no caso da morfologia produtiva, dada a impossibilidade de listar todas as formas que podem ser criadas. Em seguida, mostramos como, na construção do LEXPOR, exploramos as duas dimensões do modelo, apresentando, em um primeiro momento, um autômato que implementa os aspectos morfotáticos do fragmento da morfologia do português da seção anterior. Em um segundo momento, explicamos como a utilização de diacríticos que realizam operações sobre as propriedades morfossintáticas dos morfemas, combinada com a implementação, em um transdutor, de regras de alternâncias ortográficas e morfofonológicas, permite corrigir a hipergeração desse autômato. Finalmente, testamos o LEXPPOR na análise de vários neologismos para os quais o analisador automático do projeto VISL não oferece resultados satisfatórios.

Podemos implementar computacionalmente, de modo eficiente, a morfologia flexional de uma língua como o inglês por meio de uma simples listagem das diferentes formas e suas propriedades. De fato, em inglês, um paradigma verbal tem no máximo cinco elementos diferentes. Nessa língua, adjetivos têm no máximo três formas distintas e substantivos, no máximo duas. No caso de uma língua como o português europeu, com paradigmas verbais com mais de 60 formas, esse método é bem menos atraente. No caso de uma língua como o finlandês, porém, em que, segundo Arppe (2001), cada substantivo tem mais de 1850 formas, cada adjetivo, mais de 6000, e cada verbo, cerca de 20.000 formas, essa abordagem é completamente impraticável.

No que diz respeito à morfologia derivacional, tanto do português (brasileiro ou europeu) quanto de línguas de morfologia flexional relativamente pobre como o inglês, a listagem de formas é igualmente inviável, não só devido à imensa quantidade de itens a serem listados, mas também, sobretudo, devido à produtividade morfológica.[24]

O modelo da Morfologia de Dois Níveis, não por acaso desenvolvido pelo linguista computacional finlandês Kimmo Koskeniemmi (Karttunen e Beesley, 2005), permite modelar de forma elegante e eficiente tanto a morfologia flexional quanto a formação de palavras por derivação ou composição. Esse modelo caracteriza-se principalmente pela distinção, no âmbito do processamento das palavras, entre um nível lexical e um nível superficial, como mostramos nos Exemplos 37 e 38 para a forma ambígua *caça*, utilizando o formato proposto por Beesley e Karttunen (2003). Do ponto de vista computacional, um dos atrativos principais do modelo é a sua bidirecionalidade. Por conta disso, um mesmo programa pode ser usado para a análise (produzindo, por exemplo como componente de um analisador sintático, os Exemplos 37a e 38a a partir da forma *caça*) ou para a geração (gerando, por exemplo, como componente de um sistema de tradução automática, a forma *caça* a partir do Exemplo 37a).

Exemplo 37

(a) **Lexical**: caça+Subst+Fem+Sg
(b) **Superficial**: caça

Exemplo 38

(a) **Lexical**: caçar+Verbo+PresInd+3P+Sg
(b) **Superficial**: caça

A descrição da morfologia de uma língua, conforme esse modelo, desdobra-se em dois componentes:

Exemplo 39

Componentes da Morfologia de Dois Níveis
I. **Morfotática**: *montanh+ism+o*, *montanh+ês*
ultra+anti+neo+fernand+ian+a+s
II. **Alternâncias ortográficas e morfofonológicas**:
franç+ês+as > francesas
árvor+ism+o >arvorismo

O primeiro componente descreve as combinações permitidas de morfemas. Por exemplo, elementos do tipo de *ultra*, *anti* e *neo* podem prefixar-se, em qualquer ordem, a um nome como a forma adjetival *fernandianas*, do que damos apenas três exemplos:

Exemplo 40

(a) ultra-antineofernandianas
(b) antineo-ultrafernandianas
(c) neo-ultra-antifernandianas

[24] Jurafsky e Martin (2009, p. 84) contrapõem a simplicidade da morfologia flexional do inglês à complexidade de sua morfologia derivacional.

A forma *fernandianas*, por sua vez derivada a partir do antropônimo *Fernando* por meio da sufixação de *-ian-* e das flexões de gênero e número, não permite, contudo, adjunção de um sufixo como *-mente* (ver Exemplo 41a), passível, sim, de adjungir-se à forma feminina do adjetivo (ver Exemplo 41b) ou a uma forma dele derivada por prefixação (ver Exemplo 41c).

Exemplo 41

(a) *fernandianasmente
(b) fernandianamente
(c) neo-ultra-antifernandianamente

O sufixo *-ism-* pode adjungir-se tanto a formas simples quanto a formas derivadas por meio de *-ian-* e *-ês-*. Formas derivadas por meio de *-ism-*, contudo, não admitem outros sufixos derivacionais.[25]

Exemplo 42

(a) fernandismo
(b) fernandianismo, francesismo
(c) *francesismomente

No segundo componente, formulam-se as regras responsáveis pelas alternâncias ortográficas e morfofonológicas que resultam de ajustes nas formas dos morfemas ao se concatenarem. No primeiro caso, temos, por exemplo, uma regra que substitui *ç* por *c* antes de vogal não posterior e outra que substitui *ê* por *e* se seguido de exatamente uma sílaba.[26] Enquanto essas duas regras são meramente ortográficas, em casos como *árvor+ism+o >arvorismo* temos, concomitantemente com a mudança na grafia, uma alteração fonológica, que é o deslocamento da sílaba tônica da raiz nominal para o sufixo derivacional.

A teoria linguística, através da fonologia gerativa de Chomsky e Halle na obra clássica *The Sound Pattern of English* (Chomsky e Halle, 1968), propôs, há quatro décadas, um método eficiente para lidar com essas alternâncias, por meio da utilização de regras de reescrita sensíveis ao contexto. Um exemplo de regra desse tipo, extraído de Kenstowicz (1994, p. 21), está no Exemplo 43. Essa regra especifica que vogais posteriores devem ser substituídas por vogais não posteriores, se precedidas de vogais não posteriores seguidas de zero, uma ou mais consoantes. Por exemplo, na língua chamorro, temos alternância, produzida por essa regra, entre *lagU* "norte" e *sæn lægU* "em direção ao norte" (Kenstowicz, 1994, p. 18).

Exemplo 43

$$\begin{bmatrix} -\text{cons} \\ +\text{back} \end{bmatrix} \rightarrow [-\text{back}] \; / \; \begin{bmatrix} -\text{cons} \\ -\text{back} \end{bmatrix} \; C_o \; \_\_\_$$

As regras de alternância têm o formato geral do Exemplo 44. Segundo essa regra, devemos substituir uma representação A por uma representação B toda vez que A estiver precedido de C e seguido de D. As representações C e D podem ser vazias.

Exemplo 44

$$A \rightarrow B \; / \; C \_\_\_ D$$

No Morfologia de Dois Níveis, a morfotática é modelada por meio de subléxicos e classes de continuação. Por exemplo, prefixos como *ultra*, *anti* e *neo* constituem um subléxico que tem como classe de continuação raízes nominais como *franç-*, *kafk-*, *bob-*, *pur-* etc. As raízes nominais, por sua vez, constituem subléxico que tem como classe de continuação sufixos do tipo de *-ian-* ou *-ês*. Esses constituem subléxico que tem como classe de continuação sufixos como *-ism-*, os quais têm como classe de continuação morfemas como *-a* ou *-o* (que, como vimos, acumulam as funções de classificador e flexão de gênero). Se incluirmos morfemas zero em alguns subléxicos e compusermos o autômato resultante (como veremos mais abaixo) com regras de alternância apropriadas, conseguiremos gerar, por meio dessa descrição, palavras como as do Exemplo 45, entre outras.

Exemplo 45

francesismo, *kafkês*, *pseudokafkiana*, *pseudokafkianismo*, *pura*, *neobobo*[27], *neobobismo*[28]

[25] Uma exceção poderia ser o diminutivo: *francesismin*ho ou *francesismozinho.*
[26] Para simplificar as regras de acentuação, analisamos palavras como *fêmea* como proparoxítonas (Cunha e Cintra, 1985, p. 70). Como o encontro vocálico final, sob essa perspectiva, constitui duas sílabas (embora seja normalmente pronunciado como ditongo crescente), a palavra subtrai-se à regra que substitui *ê* por *e* em casos como *franç+ês+as > francesas.*
[27] Termo utilizado por Professor Luizinho, citado pela revista *Veja* (2003).
[28] Neologismo cunhado pelo ex-presidente Fernando Henrique Cardoso (*in Veja*, 2002).

O grande avanço da Morfologia de Dois Níveis, no âmbito do processamento automático no nível da palavra, deve-se à descoberta de que esse modelo pode ser implementado como um transdutor de estados finitos (Karttunen e Beesley, 2005). Nesse caso, as regras de alternância, formuladas na fonologia gerativa como regras de uma gramática sensível ao contexto, podem ser implementadas num único transdutor desse tipo. Um transdutor representando as combinações possíveis entre morfemas pode ser composto com um transdutor que modela as alternâncias, resultando num único transdutor que pode ser utilizado tanto para a análise como para a geração. O ganho de eficiência com isso é enorme, uma vez que autômatos e transdutores de estados finitos são processados de forma muito mais rápida que gramáticas sensíveis ao contexto (ou mesmo que gramáticas independentes de contexto).

No Exemplo 46, formalizamos, na linguagem de estados finitos lexc da Xerox, uma descrição em termos da Morfologia de Dois Níveis de um fragmento da morfologia do português que inclui não só construções vocabulares nos moldes do Exemplo 45 (mas sem as alternâncias ortográficas e morfofonológicas), no singular e no plural, mas também derivados em *-mente*.

Exemplo 46

```
LEXICON Pref
< Pref >        Pref ;
                Roots ;
LEXICON Roots
< Root >        Suff1 ;
LEXICON Suff1
< Suff1 >       Suff2 ;
                Suff2 ;
LEXICON Suff2
< Suff2 >       Gen ;
                Gen ;
LEXICON Gen
< Gen >         Suff3 ;
LEXICON Suff3
< Suff3 >       # ;
                Num ;
LEXICON Num
< Num >         # ;
```

Na linguagem lexc, um subléxico obedece ao esquema do Exemplo 47. A classe de continuação é um outro subléxico ou o final de palavra, simbolizado pela classe de continuação #.

Exemplo 47

```
LEXICON Nome
Morfema      ClasseDeContinuação ;
```

No Exemplo 46, os elementos entre parênteses angulados (<>) constituem expressões regulares definindo cada subléxico, conforme o Quadro 5, onde o épsilon, que, na teoria das línguas formais, simboliza a cadeia vazia (*empty string*), representa o morfema zero do singular.

Quadro 5: Subléxicos de um fragmento da morfologia do português.
Chart 5: Sublexicons of a morphology fragment of Portuguese.

Expressão regular	Subléxico
Pref	*ultra*, *anti, neo*
Root	*franç*, *kafk*, *bob*, *pur*
Suff1	*ian*, *ês*
Suff2	*ism*
Gen	*a*, *o*
Suff3	*mente*
Num	*s*, ε

O fragmento descrito no Exemplo 46, equivalente à expressão regular do Exemplo 48 na metalinguagem da Xerox, é compilado no autômato de estados finitos (FSA) da Figura 2.

Exemplo 48

```
Pref* Root (Suff1) (Suff2) [Gen Suff3 | Gen Num];
```

Nesse FSA, 0 é o estado inicial e 5, o estado final. Os estados estão ligados por arcos, cada um anotado com um ou mais rótulos (*labels*), constituindo transições que, na aplicação da FSA no reconhecimento de uma cadeia, vão sendo percorridas à medida em que se processa símbolo por símbolo a cadeia. Uma dessas transições é o par ordenado <<0,Root>,1>, que pode ser interpretado como uma instrução para passar ao estado 1 se, no estado 0, for encontrado na cadeia o símbolo Root.

Um FSA é uma codificação da língua regular constituída pelas cadeias (*strings*) que se podem formar (ou que se podem reconhecer) percorrendo cada um dos caminhos do estado inicial ao final, concatenando, um após o outro, os rótulos das transições entre os estados. Uma dessas cadeias é, por exemplo, o Exemplo 49, que representa a estrutura da palavra do Exemplo 41c e corresponde ao caminho constituído pelos arcos 0-0, 0-0, 0-0, 0-1, 1-3, 3-2, 2-5.

Exemplo 49

```
Pref Pref Pref Root Suff1 Gen Suff3
```

No arco 2-5, há dois rótulos: Num e Suff3. Isso implica duas transições diferentes: <<2,Num>,5> e <<2,Suff3>,5>. Desse modo, em vez do Exemplo 49, também pode ser formada uma cadeia como o Exemplo 50, que representa a estrutura das palavras do Exemplo 40.

Exemplo 50

Pref Pref Pref Root Suff1 Gen Num

A recursividade da prefixação está visualmente expressa na Figura 2 pelo fato de o arco rotulado como "Pref" partir do estado 0 e retornar a esse mesmo estado, num percurso cíclico. A opcionalidade da prefixação está representada pela possibilidade de formar uma palavra partindo do estado 0 diretamente ao estado 1, sem dar volta alguma no ciclo responsável pela prefixação.

Para gerar não estruturas em termos de rótulos abstratos como Root Gen Num, mas palavras da língua portuguesa como *bobos* ou *puras*, que instanciam esse esquema, basta incluir na formalização do Exemplo 46 as definições do Exemplo 51, que expandem esses rótulos em subconjuntos de morfemas.

Exemplo 51

```
Definitions
Pref = {neo} | {ultra} | {anti} ;
Root = {franç} | {kafk} | {bob} | {pur} ;
Suff1 = {ês} |{ian} ;
Suff2 =  {ism} ;
Gen = o | a | 0 ;
Suff3 = {mente} ;
Num = (s) ;
```

O FSA, compilado a partir do Exemplo 46 com as definições do Exemplo 51, codifica língua regular que inclui palavras gramaticalmente bem-formadas do português, como *antineobobismo*. No entanto, esse FSA reconhece e gera muitas palavras agramaticais em português, como, por exemplo, as seguintes:

Exemplo 52

*franços

Exemplo 53

*françêsisma

De uma maneira geral, a hipergeração explica-se pela violação de dois tipos diferentes de princípios, relacionados aos dois componentes da Morfologia de Dois Níveis. A má formação dos Exemplos 52 e 53 decorre, por um lado, da violação de princípios morfotáticos do português. No primeiro caso, a raiz *franç-* exige a flexão de gênero (e classificador) *-a*. No segundo caso, o sufixo *-ism-* forma substantivos masculinos em *-o*. Por outro lado, o Exemplo 53 também viola regras ortográficas e morfofonológicas da língua.

Na linguagem de estados finitos do programa xfst da Xerox, o primeiro tipo de problema pode ser sanado por meio de um recurso chamado *flag diacritics*, que permite implementar traços morfossintáticos em termos de estruturas de atributos e valores. No LEXPOR, a raiz *franç-*, por exemplo, é concatenada aos diacríticos "@U.GEND.FEM@" e "@U.TH.A@", onde U representa a operação de unificação de traços. O primeiro diacrítico codifica a seguinte instrução: "atribua a GÊNERO o valor FEMININO, caso isso não contradiga atribuição anterior de valor". O segundo diacrítico envolve o atributo TEMA e o valor A, i.e. a raiz em questão exige o classificador *-a*. O morfema *-o*, por sua vez, é concatenado a "@U.GEND.MASC@" e "@U.TH.O@". Ao concatenar a raiz *franç-* com a flexão *-o*, o algoritmo de aplicação do FSA na geração ou reconhecimento de palavras realiza uma operação de unificação entre as estruturas de traços [GEND=FEM, TH=A] e [GEND=MASC, TH=O], a

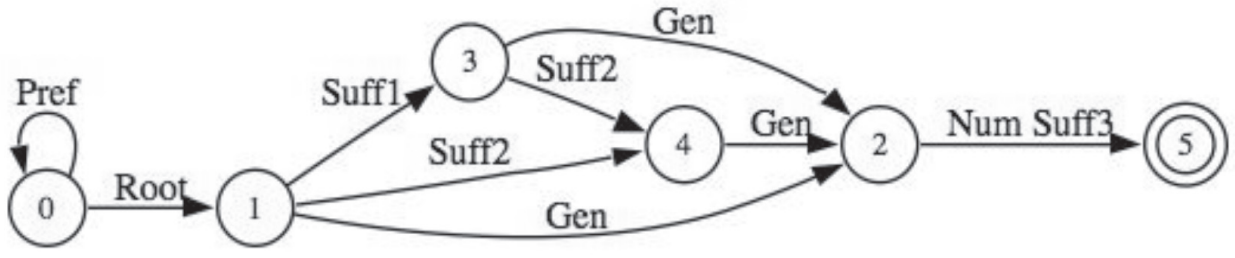

Figura 2: Autômato de estados finitos representando fragmento da morfologia do português.
Figure 2: Finite-state automaton representing a morphology fragment of Portuguese.

qual, naturalmente, fracassa, devido à incompatibilidade de valores dos atributos GEND e TH. Esse caminho é, portanto, ignorado pelo algoritmo.

O LEXPOR se vale de várias outras operações com estruturas de traços disponíveis no xfst. Extremamente útil, por exemplo, é a operação simbolizada por C (representando mnemonicamente o verbo *clear* 'limpar'), utilizada para "esvaziar" um atributo. Na representação de sufixos formadores de adjetivos como *-ian-*, incluímos os diacríticos "@C.GEND@" e "@C.TH@". Isso garante que a partir de uma raiz como *brasil-*, concatenada a "@U.GEND.MASC@" e "@U.TH.C@" (onde TH.C representa tema em consoante), se possam construir formas tanto masculinas quanto femininas de adjetivos por meio da concatenação com as flexões *-a* e *-o* (no caso, *brasiliano* e *brasilianas*). A capacidade de um sufixo como *-ism-* de produzir derivados com gênero e classe flexional discrepantes dos da base é modelada por meio dos diacríticos "@P.GEND.MASC@" e "@P.TH.O@", onde P simboliza operação de fixação de valor de um atributo, independentemente de valor que possa ter sido anteriormente fixado.

O outro tipo de problema do FSA definido nos Exemplos 46 e 51 resulta da não observância de regras ortográficas e morfofonológicas do português. Na variação do Exemplo 53 em 54, a morfotática está correta. No entanto, o exemplo é agramatical por apresentar cedilha antes de vogal não posterior, violando convenção ortográfica do idioma, e pela posição do acento na antepenúltima e não na penúltima sílaba, o que já constitui violação de princípio fonológico da língua. Como vimos, o sufixo *-ismo* implica, como quase sempre na sufixação em português, o deslocamento do acento.

Exemplo 54

*françêsismo

Para modelar esse tipo de fenômeno, o LEXPOR utiliza a solução padrão no âmbito da morfologia de estados finitos, que é a composição de transdutores. No FSA definido nos Exemplos 46 e 51, como em todo autômato de estados finitos, apenas uma face de cada arco é rotulada. O FSA codifica, portanto, uma língua regular, que resulta de operações de concatenação desses símbolos. A grande vantagem da tecnologia de estados finitos para o tratamento computacional das línguas naturais no nível da palavra advém da possibilidade de utilizar transdutores de estados finitos na modelação de fenômenos morfológicos complexos. Um transdutor de estados finitos (doravante FST) codifica uma relação regular, i.e. um conjunto de pares ordenados constituídos a partir de elementos de duas línguas regulares, chamadas, respectivamente, de língua superior (*upper language*) e língua inferior (*lower language*). Um autômato como o da Figura 2 representa apenas a dimensão fonológica ou ortográfica dos morfemas. Transdutores como o que definimos em (55), pelo contrário, podem representar cada morfema como associação entre representação fonológica ou ortográfica e propriedades morfossintáticas. Por exemplo, a flexão de gênero feminino é representada como a relação denotada pela expressão regular [a %<Fem%>:0], que é compilada pelo xfst no transdutor da Figura 3.

Exemplo 55

```
Definitions
Pref = {neo} | {ultra} | {anti} ;
Adj =  {bob} | {pur} ;
Gen = o %<Masc%>:0 | a %<Fem%>:0 ;
Num = s %<Pl%>:0 |  %<Sg%>:0 ;

LEXICON Pref
< Pref %<PREF%>:0 >          Pref ;
                  Adjs ;

LEXICON Adjs
< Adj %<A%>:0 >      Gen ;

LEXICON Gen
< Gen >        Num ;

LEXICON Num
< Num >                  # ;
```

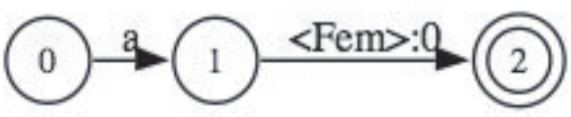

Figura 3: Transdutor representando a flexão de gênero feminino.
Figure 3: Transducer representing the feminine gender inflection.

No Exemplo 56, transcrevemos parte de sessão interativa do xfst em que se aplica, por meio do comando *up* ('para cima' em inglês), o transdutor definido no Exemplo 55 na análise das palavras *neobobas* e *ultraneobobo*. Esse comando exibe as representações lexicais, localizadas nos arcos na parte de cima (ou à esquerda), correspondentes às formas de superfície, situadas abaixo (ou à direita).

Exemplo 56

```
xfst[1]: up neobobas
neo<PREF>bob<A>a<Fem>s<Pl>
xfst[1]: up ultraneobobo
ultra<PREF>neo<PREF>bob<A>o<Masc><Sg>
```

As regras responsáveis pelas alternâncias ortográficas e morfofonológicas também são implementadas como transdutores. No Exemplo 57, temos expressão regular que denota relação onde qualquer par ordenado que apresente, como primeiro elemento, uma cadeia com *ç* antes de *ê* está associado a uma cadeia com *c* nessa mesma posição. Exemplos de pares ordenados que integram essa relação são <*françês, francês*> e <*francês, francês*>, onde o primeiro elemento integra a língua superior da relação. O par <*francês, françês*>, ao contrário, não pertence a essa relação.

Exemplo 57

```
ç -> c || _ ê
```

A Figura 4 exibe o transdutor que resulta da compilação da expressão do Exemplo 57 pelo xfst. O rótulo *?* representa qualquer símbolo exceto aqueles que constituem o alfabeto do transdutor. Desse modo, na aplicação para cima do transdutor a uma palavra como *francês*, o algoritmo ou percorre apenas a transição cíclica que parte do estado 0, retornando *francês*, ou percorre esse ciclo somente até a letra *n*, passando, então, para o estado 2 e, deste, para o estado 0, onde dá uma volta final no ciclo, retornando *françês* .

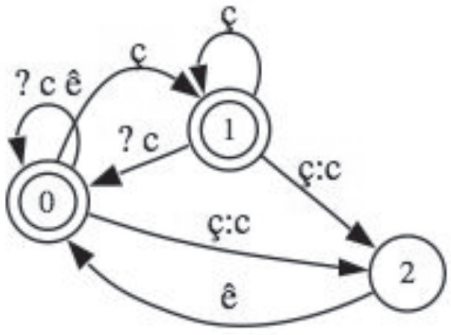

Figura 4: Transdutor que modela regra de substituição de *c* por *ç* antes de *ê*.
Figure 4: Transducer modeling rule that substitutes *c* for *ç* before *ê*.

O LEXPOR constitui um transdutor que resulta da composição entre um transdutor do tipo do Exemplo 55, mas dezenas de vezes mais abrangente, modelando os fenômenos morfológicos descritos anteriormente e representados esquematicamente no FSA da Figura 2, com um transdutor que implementa mais de 30 regras de alternância ortográfica ou morfofonológica do português. No Quadro 6, exemplificamos as principais alternâncias modeladas no LEXPOR. Para facilitar a identificação do tipo de alteração envolvido, indicamos por meio de "+" a segmentação morfêmica das palavras ou de suas formas subjacentes (indicadas por meio de "*"). O símbolo î representa um *i* assilábico.

Vejamos agora, por meio da transcrição de uma sessão interativa do xfst em ambiente UNIX, alguns exemplos de aplicação do LEXPOR na análise de palavras lexicalizadas e de neologismos:

Exemplo 58

```
UNIX> xfst
Copyright © Palo Alto Research Center 2001-2009
PARC Finite-State Tool, version 2.10.43 (lib-
cfsm-2.14.4) (svn 29656)

Type "help" to list all commands available or "help
help" for further help.

xfst[0]: load lexpor.fst
Opening input file 'lexpor.fst'
0%>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
>>100%
July 04, 2009 16:10:14 GMT
Closing input file 'lexpor.fst'
xfst[1]: up
apply up> irmão
irmaN<N>o<Masc><Sg>
apply up> irmãs
irmaN<N><SUFF><N>a<Fem>s<Pl>
apply up> irmanismo
irmanism<FRG>o<Masc><Sg>
irmani<FRG>'ism<SUFF><N>o<Masc><Sg>
irman<FRG>'ism<SUFF><N>o<Masc><Sg>
irmaN<N>'ism<SUFF><N>o<Masc><Sg>
apply up> kafkês
kafk<FRG>ês<SUFF><A><Masc><Sg>
apply up> francês
franc<FRG>ês<SUFF><A><Masc><Sg>
franç<NPR>ês<SUFF><A><Masc><Sg>
apply up> neo-ultrafrancesismo
neo-<PREF>ultra<PREF>francesism<FRG>o<
Masc><Sg>
neo-<PREF>ultra<PREF>francesi<FRG>'ism<S
UFF><N>o<Masc><Sg>
neo-<PREF>ultra<PREF>frances<FRG>'ism<S
UFF><N>o<Masc><Sg>
neo-<PREF>ultra<PREF>franc<FRG>ês<SUFF
><A>'ism<SUFF><N>o<Masc><Sg>
neo-<PREF>ultra<PREF>franç<NPR>ês<SUFF
><A>'ism<SUFF><N>o<Masc><Sg>
```

Como se evidencia no caso de *irmão*, o LEXPOR, quando aplicado "para cima" a uma palavra do português, retorna uma análise morfêmica em que a

Quadro 6: As principais alternâncias modeladas no LEXPOR.
Chart 6: The main alternations modeled by LEXPOR.

Tipo de alternância	Exemplos
Ortográfica	Franç+a => franc+ês franc+ês => franc+es+a mong+e => monj+a Jung => jungu+ismo
Morfofonológica	utopi+a => utop+ismo ação (< *açîoN) => acion+ismo *açîoN+s => ações irmão (<=*irmaN+o) => irmã (<=*irmaN+a) irmão (<=*irmaN+o) => irmãos (<=*irmaN+o+s) irmã (<=*irmaN+a) => irmãs (< *irmaN+a+s) *irmaN+'ism+o => irmanismo balão (<=*baloN) => balões (<=*baloN+s) *baloN+'ism+o => balonismo árvor+e => arvor+ismo simultâne+o => simultane+ísmo real => reais (*real+s)

representação de morfemas individuais não necessariamente coincide com a sua forma de superfície. De fato, na esteira de Câmara Jr. (1987, p. 90), representamos a raiz de *irmão* <=*irmaN+o como /irmaN/. Diferentemente desse autor, porém, não analisamos o substantivo feminino correspondente como resultado de processo flexional consistindo na supressão da vogal temática *-o* da forma masculina, mas por meio de sufixo derivacional zero (representado no Exemplo 58 por uma cadeia vazia) que transforma um substantivo masculino em substantivo de gênero feminino. Na análise do LEXPOR, a flexão de gênero em *irmãs* é a vogal átona *-a*, que sofre contração com a nasal (Monteiro, 1987, p. 68):

Exemplo 59

*irmaN+a+s => *irmã+a+s => irmãs

O neologismo *irmanismo* (não dicionarizado no iAulete) recebe quatro análises diferentes no LEXPOR. Na primeira análise, essa palavra é considerada antropônimo masculino singular com a raiz de proveniência estrangeira *irmanism-*, rotulada como <FRG> (do inglês *foreign*). As duas análises seguintes, mais plausíveis, tratam a palavra como derivada, respectivamente, das raízes estrangeiras *irmani-* e *irman-*, por meio da sufixação com *-ismo*. Finalmente, na quarta e última análise, a palavra é analisada como derivada do substantivo *irmão* (mais precisamente, da raiz /irmaN/).

O neologismo *kafkês* é analisado como derivado por meio de sufixação de *-ês*, que forma adjetivos a partir de substantivos, à raiz de antropônimo estrangeiro *kafk-*. Postulamos que o uso desse tipo de derivado em posição típica de substantivo, em exemplos análogos ao Exemplo 28, é licenciado na sintaxe.

A palavra *francês* é considerada ambígua pelo LEXPOR, sendo analisada como derivada por sufixação de *-ês* ou à raiz *franc-* de um antropônimo de origem estrangeira (por ex. *France*, do escritor Anatole France), a exemplo de *kafkês*, ou à raiz de nome próprio *franç-* (de *França*). É claro que dificilmente algum falante utilizaria *francês* no primeiro sentido, dada a homonímia com *francês* derivado de *França*.

As cinco análises de *neo-ultrafrancesismo* explicam-se pelas diferentes interpretações teoricamente possíveis das cadeias *francesism-* (como em *irmanismo*) e *francês*. A última dessas análises evidencia exemplarmente o papel dos sufixos derivacionais na determinação das propriedades morfossintáticas dos derivados. Em um derivado com mais de um sufixo derivacional, o último à direita é o que determina a categorial lexical e as demais propriedades do derivado. Enquanto *franç-* é substantivo feminino e *francês*, adjetivo, *francesismo* é substantivo masculino.

À exceção de *kafkês*, todos os derivados do Exemplo 58 têm, em pelo menos uma análise, base que consta do léxico de raízes do LEXPOR. Vejamos agora como o transdutor analisa derivado que não se enquadra nesse caso:

Exemplo 60

```
apply up> antineopseudo-ultramerkelianas
anti<PREF>neo<PREF>pseudo-<PREF>ultra<P
REF>merkelian<FRG>a<Fem>s<Pl>
anti<PREF>neo<PREF>pseudo-<PREF>ultra<PR
EF>merkeli<FRG>ian<SUFF><A>a<Fem>s<Pl>
anti<PREF>neo<PREF>pseudo-<PREF>ultra<PR
EF>merkel<FRG>ian<SUFF><A>a<Fem>s<Pl>
```

A terceira análise do Exemplo 60 remete diretamente à chanceler alemã Angela Merkel, embora o elemento *merkel-* não conste do inventário de raízes do LEXPOR. Se abstrairmos desse sobrenome, porém, as outras duas análises de *antineopseudo-ultramerkelianas* parecem também plausíveis, na medida em que as respectivas bases são sobrenomes estrangeiros possíveis. Por um lado, *Merkeli* é o nome de uma cidade na Letônia. Por outro, *Merkelian* tem a terminação típica dos sobrenomes armênios. Na terceira análise do Exemplo 61, o LEXPOR remete ao sobrenome de origem armênia da atriz brasileira Aracy Balabanian.

Exemplo 61

```
apply up> balabanianismo
balabanianism<FRG>o<Masc><Sg>
balabaniani<FRG>'ism<SUFF><N>o<Masc><
Sg>
balabanian<FRG>'ism<SUFF><N>o<Masc><
Sg>
balabani<FRG>ian<SUFF><A>'ism<SUFF><N
>o<Masc><Sg>
balaban<FRG>ian<SUFF><A>'ism<SUFF><N
>o<Masc><Sg>
```

Na parte final da sessão do xfst, aplicamos o LEXPOR a mais dois exemplos de derivados a partir de uma base não listada, mas que é adivinhada pelo transdutor.

Exemplo 62

```
apply up> ultramedvedevesmente
ultra<PREF>medvedev<FRG>ês<SUFF><A><
Masc>'mente<SUFF><Adv>
apply up> antimedvedevianamente
anti<PREF>medvedevi<FRG>ian<SUFF><A>a
<Fem>'mente<SUFF><Adv>
anti<PREF>medvedev<FRG>ian<SUFF><A>a<
Fem>'mente<SUFF><Adv>
apply up> END;
xfst[1]: quit
bye.
UNIX>
```

Nas duas análises, o LEXPOR reduz os derivados à raiz *medvedev-*, que coincide com o sobrenome do atual presidente da Rússia, Dmitri Medvedev, permitindo que, num levantamento de informações sobre esse político, no âmbito de um sistema de IR ou de IE, documentos com esses neologismos também sejam levados em conta.

Conclusão

Neste artigo, partimos do pressuposto de que, no âmbito da tecnologia do texto, análises morfológicas que não apenas lematizam e etiquetam morfossintaticamente palavras complexas, mas também segmentam e classificam os morfemas desempenham um papel importante em várias aplicações, como a anotação automática de corpora e a extração ou recuperação de informações.

Corpora eletrônicos abrangentes como o CETENFolha e textos disponíveis na Internet evidenciam que a morfologia derivacional do português, tanto na prefixação quanto na sufixação, é extremamente produtiva. Sob esse aspecto, destacam-se especialmente os prefixos de origem grega ou latina do tipo de *neo-*, *pseudo-*, *semi-*, *anti-*, *pós-* ou *sub-* e os sufixos *-ismo, -iano*, *-ês* e *-mente*, com um potencial inesgotável para criação de novas palavras. Essa produtividade deve-se, em grande parte, à recursividade, no caso da prefixação, e à possibilidade de utilização, como bases para esses processos morfológicos, dos produtos de vários outros processos ou de empréstimos de outras línguas (principalmente antropônimos).

Não obstante a importância da análise morfológica automática em termos de segmentação e classificação dos constituintes de palavras complexas não dicionarizadas, tanto um corpus da dimensão do CETENFolha quanto um analisador automático do português amplamente utilizado como o do projeto VISL deixam essa lacuna em aberto. Partindo das deficiências desse analisador, que apresenta, com frequência, resultados insatisfatórios na análise de neologismos com os prefixos e sufixos mencionados, construímos o LEXPOR, um transdutor lexical elaborado nas linguagens de programação de estados finitos xfst e lexc da Xerox. Concebido como um protótipo para a construção de um transdutor lexical do português de ampla cobertura, capaz de ser aplicado na anotação automática de corpora, o LEXPOR modela apenas um fragmento da morfologia do português. No entanto, é capaz de realizar análises plausíveis para um número potencialmente infinito de palavras completamente novas, incluindo derivados de

qualquer antropônimo de origem estrangeira, como, por exemplo, *pseudo-ultramerkelianas* ou *enzensbergeresmente*, aos quais o LEXPOR atribui, respectivamente, as representações pseudo-<PREF>ultra<PREF>merkel <FRG>ian<SUFF><A>a<Fem>s<Pl> e enzensberger<F RG>ês<SUFF><A><Masc>'mente <SUFF><Adv>. De fato, nessas duas representações, as duas palavras complexas são reduzidas às raízes de antropônimos *merkel-* e *enzensberger-*, que remetem aos sobrenomes da chanceler alemã Angela Merkel e do poeta alemão Hans Magnus Enzensberger. O LEXPOR alcança esse resultado não por incorporar uma vasta lista de antropônimos, mas pela capacidade de "adivinhar" como antropônimo de origem estrangeira todo elemento que não integre o léxico de elementos nativos. O analisador do projeto VISL, pelo contrário, não reduz esses derivados às suas bases primitivas, ignorando quase todos os afixos derivacionais desses neologismos.

Enquanto protótipo, o LEXPOR, na fase de desenvolvimento que constituiu objeto deste trabalho, não visou, como ressaltamos, a abarcar uma vasta porção do léxico do português e de sua morfologia produtiva, mas à profundidade das análises. Para os próximos meses, planeja-se, por um lado, criar uma interface *on-line*, acessável pela WWW, em que usuários possam testar o analisador e fornecer *feedback* sobre seu funcionamento, ao mesmo tempo em que se procurará, por outro lado, expandir a sua cobertura, ampliando o leque de prefixos e sufixos e as classes de bases às quais se adjungem.

Referências

ANDERSON, S. R. 1992. *A-morphous morphology*. Cambridge, Cambridge University Press, 434 p.

ARPPE, A. 2001. Focal points in frequency profiles: how some word forms in a paradigm are more significant than others in Finnish. *In*: CONFERENCE ON COMPUTATIONAL LEXICOGRAPHY AND CORPUS RESEARCH, 6, Birmingham, 2001. Disponível em: http://www.ling.helsinki.fi/~aarppe/Publications/COMPLEX-01.rtf. Acesso em: 29/06/2009.

BEESLEY, K.R.; KARTTUNEN, L. 2003. *Finite state morphology*. Stanford, CSLI Publications, 510 p.

BICK, E. 2000. *The parsing system "Palavras"* : *automatic grammatical analysis of Portuguese in a Constraint Grammar framework*. Århus, Dinamarca. Tese de PhD. University of Århus, 505 p. Disponível em: beta.visl.sdu.dk/pdf/PLP20-amilo.ps.pdf Acesso em: 27/10/2009.

BIRD, S.; KLEIN, E.; LOPER, E. 2009. *Natural language processing with Python*: *analyzing text with the Natural Language Toolkit*. Sebastopol, O'Reilly. 502 p.

CÂMARA JR., J.M. 1987. *Estrutura da língua portuguesa*. 17ª ed., Petrópolis, Vozes, 124 p.

CHOMSKY, N.; HALLE, M. 1968. *The sound pattern of English*. New York, Harper & Row, 484 p.

CUNHA, C.; CINTRA, L. 1985. *Nova gramática do português contemporâneo*. Rio de Janeiro, Nova Fronteira, 714 p.

FALK, Y. N. 2001. *Lexical-functional grammar*: *an introduction to parallel constraint-based syntax*. Stanford, CSLI Publications, 237 p.

FOLHA DE SÃO PAULO. 2007. Disponível em: http://www1.folha.uol.com.br/fsp/opiniao/ fz1301200701.htm. Acesso em: 16/06/2009.

GREWENDORF, G.; HAMM, F.; STERNEFELD, W. 1989. *Sprachliches Wissen*: *eine Einführung in moderne Theorien der grammatischen Beschreibung*. 3ª ed., Frankfurt am Main, Suhrkamp, 467 p.

JURAFSKY, D.; MARTIN, J.H. 2009. *Speech and language processing*: *an introduction to natural language processing, computational linguistics, and speech recognition*. 2ª ed., London, Pearson International, 1024 p.

KARTTUNEN, L.; BEESLEY, K.R. 2005. Twenty-five years of finite-state morphology. *In:* A. ARPPE; L. CARLSON; K. LINDÉN; J. PIITULAINEN; M. SUOMINEN; M. VAINIO; H. WESTERLUND; A. YLI-JYRÄ (eds.), *Inquiries into words, constraints and contexts*: *Festschrift for Kimmo Koskenniemi on his 60th Birthday*. Stanford, CSLI, p. 71-83.

KENSTOWICZ, M. 1994. *Phonology in generative grammar*. Cambridge, Blackwell, 704 p.

KRÜGER-THIELMANN, K.; PAIJMANS, H. 2004. Informationserschließung. *In*: H. LOBIN; L. LEMNITZER (eds.), *Texttechnologie*: *Perspektiven und Anwendungen*. Tübingen, Stauffenburg, p. 353-378.

LEMNITZER, L.; WAGNER, A. 2004. Akquisition lexikalischen Wissens. *In*: H. LOBIN; L. LEMNITZER (eds.), *Texttechnologie*: *Perspektiven und Anwendungen*. Tübingen, Stauffenburg, p. 245-266.

LEMNITZER, L.; ZINSMEISTER, H. 2006. *Korpuslinguistik*: *eine Einführung*. Tübingen, Narr, 220 p.

LOBIN, H.; LEMNITZER, L. (eds.). 2004. *Texttechnologie*: *Perspektiven und Anwendungen*. Tübingen, Stauffenburg, 487 p.

LOUREIRO, I. R. 2009. A figura feminina em *Lucía Miranda*, de Rosa Guerra. *In*: C.S. de BARROS; E.T.R. AMARAL; E.A. VIEIRA; S. ROJO (orgs.), *Anais do V Congresso Brasileiro de Hispanistas e I Congresso Internacional da Associação Brasileira de Hispanistas*. Belo Horizonte, Faculdade de Letras da UFMG. Disponível em: http://www.letras.ufmg.br/espanhol/Anais/anais_paginas_%20 1005-1501/A%20figura%20feminina.pdf. Acesso em: 31/08/2009.

MAIER, W. 2007. NeGra und Tuba-D/Z: ein Vergleich. 2007. *In*: G. REHM; A. WITT; L. LEMNITZER (eds.), *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen*. Tübingen, Narr, p. 29-38.

MARTINEZ, L. 2006. Murilo Mendes, o imbele no campoconcentração. *Boletim de pesquisa — NELIC*, **8/9**. Disponível em: http://www.cce.ufsc.br/~nelic/boletim8-9/leonilmartinez.htm. Acesso em: 07/07/2009.

MATTHEWS, P.H. 1974. *Morphology*: *an introduction to the theory of word-structure*. Cambridge, Cambridge University Press, 243 p.

MONTEIRO, J.L. 1987. *Morfologia portuguesa*. 2ª ed., Fortaleza, EDUFC, 218 p.

REHM, G. 2004. Texttechnologie und das Wide World Web. *In*: H. LOBIN; L. LEMNITZER (eds.). *Texttechnologie*: *Perspektiven und Anwendungen*. Tübingen, Stauffenburg, p. 433-464.

SASAKI, F.; WITT, A. 2004. Linguistische Korpora. *In*: H. LOBIN; L. LEMNITZER, L. (eds.), *Texttechnologie*: *Perspektiven und Anwendungen*. Tübingen, Stauffenburg, p. 195-216.

SCHMID, H.; FITSCHEN, A.; HEID, U. 2004. SMOR: A German computational morphology covering derivation, composition, and inflection. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 4, Lisboa, 2004. *Proceedings*... Lisboa, 1263-1266.

SCHPAK-DOLT, N. 1999. *Einführung in die Morphologie des Spanischen*. Tübingen, Niemeyer, 140 p.

TROMMER, J. 2004. Morphologie. *In*: K.-U. CARSTENSEN; C. EBERT; C. ENDRISS; S. JEKAT; R. KLABUNDE; H. LANGER (eds.), *Computerlinguistik und Sprachtechnologie*: *eine Einführung*. Heidelberg, Spektrum Akademischer Verlag, p. 190-217.

TROST, H. 2004. Morphology. *In*: R. MITKOV (ed.), *The Oxford handbook of computational linguistics*. Oxford, Oxford University Press, p. 25-47.

UOL NOTÍCIAS. 2006. Disponível em: http://noticias.uol.com.br/midiaglobal/prospect/2006/08/08/ult2678u42.jhtm. Acesso em: 22/05/2009.

VEJA. 2007. Disponível em: http://veja.abril.com.br/120907/andre_petry.shtml. Acesso em: 24/04/2008.
VEJA. 2003. Disponível em http://veja.abril.com.br/250603/vejaessa.html. Acesso em: 31/08/2009.
VEJA. 2002. Disponível em: http://veja.abril.com.br/especiais/anos-fhc/melhores-frases-63749.shtml. Acesso em: 31/08/2009.
VIEW. 2007. Disponível em: http://topview.indexet.info/revistas/2007/10/10/2241/agenda-zeh-simao.html. Acesso em: 31/08/2009.

Submissão: 08/07/2009
Aceite: 30/10/2009

Leonel Figueiredo de Alencar
UFC - Departamento de Letras Estrangeiras
Av. da Universidade, 2683, Benfica
60020-181, Fortaleza, CE, Brasil