UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE HUMANIDADES
DEPARTAMENTO DE LETRAS VERNÁCULAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

HÉLIO LEONAM BARROSO SILVA

EXPANSÃO DO MORPHOBR ATRAVÉS DA MODELAGEM COMPUTACIONAL DE PROCESSOS DE FORMAÇÃO DE PALAVRAS EM PORTUGUÊS

Fortaleza
2019

HÉLIO LEONAM BARROSO SILVA

EXPANSÃO DO MORPHOBR ATRAVÉS DA MODELAGEM COMPUTACIONAL DE PROCESSOS DE FORMAÇÃO DE PALAVRAS EM PORTUGUÊS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Linguística da Universidade Federal do Ceará, como requisito parcial para obtenção de título de Mestre em Linguística. Área de concentração: Linguística.

Orientador: Prof. Dr. Leonel Figueiredo de Alencar Araripe.

FORTALEZA

2019

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S58e Silva, Hélio Leonam Barroso.
Expansão do MorphoBr através da modelagem computacional de processos de formação de palavras em português / Hélio Leonam Barroso Silva. – 2019.
66 f. : il.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Humanidades, Programa de Pós-Graduação em Linguística, Fortaleza, 2019.
Orientação: Prof. Dr. Leonel Figueiredo de Alencar Araripe.

1. Morfologia de Estados Finitos. 2. Dicionário eletrônico. 3. Sufixação. 4. Processamento de Linguagem Natural. 5. Linguística Computacional. I. Título.

CDD 410

HÉLIO LEONAM BARROSO SILVA

EXPANSÃO DO MORPHOBR ATRAVÉS DA MODELAGEM COMPUTACIONAL DE PROCESSOS DE FORMAÇÃO DE PALAVRAS EM PORTUGUÊS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Linguística da Universidade Federal do Ceará, como requisito parcial para obtenção de título de Mestre em Linguística. Área de concentração: Linguística.

Aprovada em: ____ / ____ / ____

BANCA EXAMINADORA

_____
Prof. Dr. Leonel Figueiredo de Alencar Araripe (Orientador)
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Alexandre Rademaker
Escola de Matemática Aplicada da Fundação Getúlio Vargas (EMAp - FGV)

_____
Profa. Dra. Rosemeire Selma Monteiro-Plantin
Universidade Federal do Ceará (UFC)

_____
Profa. Dr. Gabriel de Ávila Othero
Universidade Federal do Rio Grande do Sul (UFRGS)

_____
Profa. Dr. Ronaldo Mangueira Lima Júnior
Universidade Federal do Ceará (UFC)

A minha mãe, Helita.

E a você, *Imzadi*.

AGRADECIMENTOS

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, pela concessão de bolsa de pesquisa que me auxiliou enquanto estive no mestrado.

A meu pai, pelo apoio, pelo sustento e pela educação que tive o privilégio de receber.

A você, minha mãe, por nunca ter desistido de cuidar da gente, de nos defender e proteger dos males da vida mesmo frente a tantos obstáculos. Obrigado pela força. Por ser guerreira. Por ser meu exemplo. É a você que tanto almejo dar ao menos um pouco de paz de espírito depois de décadas de luta.

A minhas irmãs, Florence e Narrimam, com quem compartilho morada desde que me entendo por gente e que tanto torcem por mim e a quem tanto quero orgulhar.

A meu avô, Manuel, que compartilhou seus últimos anos de vida com a gente. O *Lionam*, esse "minino tão direito", agora é mestre, vozinho!

A minha avó, Francisca, que finalmente agora descansa depois de uma vida de sofrimento.

A minha tia, France, que partiu prematuramente e nos deixou uma nova caçula, a Nicole.

Aos Rangers, Mayara, Victor, Priscila e Rogiellyson e, mais recentemente, Sávio, amigos com quem compartilhei esse percurso acadêmico e ao longo do qual me ajudaram de diversas formas.

A você, Mayara, a quem devo três vezes esse grau de mestre por três vezes ter me reerguido ao longo desse trajeto. Eu é que te agradeço por ter permanecido. Agradeço por ter acreditado quando eu mesmo não acreditava. Você que esteve comigo da reprovação à aprovação. Você que se preocupou depois que desisti de apresentar aquele seminário e que não se deixou vencer pela barreira de isolamento que ergui contra todos. Você que se fez presente na qualificação, em Seminários de Pesquisa, na defesa e nos ínterins, e que estará comigo novamente, agora na seleção de doutorado. Você é a que mais tem noção do quão foi difícil essa trajetória. Obrigado não somente pelo bem que você me fez e faz, obrigado também pelo

bem que você traz ao mundo. Você é minha referência, minha dupla, minha amiga. Obrigado pela confiança. Obrigado por se importar. Obrigado por me mostrar o caminho do otimismo. Obrigado por me ajudar a novamente me encontrar, a me lembrar do trajeto que um dia eu havia traçado pra mim e cujo mapa eu havia perdido. Esse agradecimento é pra você.

A meus colegas de CompLin, em especial a Daniel e Katiusha, com quem tenho compartilhado essa trajetória interdisciplinar entre humanidades e exatas.

A meu orientador Leonel, e aos professores Alexandre, Rosemeire e Pedro que participaram das bancas de mestrado ao longo desse processo e cujas recomendações me ajudaram a produzir minha dissertação.

À Liliana, com quem compartilhei mais de uma década da minha vida e que tanto me ajudou e torceu por mim.

A meus amigos de infância, Reginaldo e Luiz Carlos, a quem tive o prazer de reencontrar logo que fui aprovado no mestrado e de cuja companhia desfrutei mensalmente de lá pra cá em *happy hours* regados a *cervas*, tira-gostos, risos, confidências, lembranças e planos.

RESUMO

Neste trabalho, modelamos computacionalmente quatro processos de formação de palavras (PFP) do português usando a morfologia de estados finitos a fim de gerar novas entradas lexicais automaticamente, contribuindo com o grupo de pesquisa Computação e Linguagem Natural no desenvolvimento de recursos para o Processamento de Linguagem Natural (PLN) para a língua portuguesa. Dentre os numerosos desafios com que um sistema de PLN deve lidar ao processar textos em língua natural, vários deles têm a ver com o domínio lexical, que dialoga direta ou indiretamente com todos os outros níveis do sistema. Ter em mãos recursos lexicais bem estruturados e abrangentes influencia decisivamente na eficiência do sistema de PLN. O melhor recurso de que temos notícia é o MorphoBr (ALENCAR; RADEMAKER; CUCONATO, 2018), construído a partir da combinação, revisão e expansão de recursos análogos livremente disponíveis para o português, derivados, em sua maior parte, do Label-Lex (ELEUTÉRIO et al., 1995) e Unitex-PB (MUNIZ, 2004). Essa expansão ocorreu gerando-se automaticamente, por exemplo, formas do diminutivo de adjetivos e substantivos e formas flexionadas faltantes de vários verbos. Propomos a geração automática de entradas lexicais aproveitando as já existentes como bases de processos de formação de palavras por sufixação a fim de retroalimentar o conjunto de dados do MorphoBr. Os quatro PFPs selecionados foram os correspondentes aos sufixos *-vel*, *-idade*, *-izar* e *-mente*. Como apenas as classes morfossintáticas da base e do derivado não são suficientes para atestar a boa formação da palavra, levamos em conta as diversas restrições de cada PFP selecionado documentadas em Alves (2004), Basilio (1980, 1987, 1990, 2017), Cavalcante (1996), Maroneze (2005, 2011), Rocha (2008) e Villalva e Silvestre (2014). Em termos de quantidade relativa de lemas, alcançamos um aumento de 8,5% com *-vel*, de 9,5% com *-idade*, 9,8% com *-izar* e 12,9% com *-mente*. Em termos de quantidade absoluta de novas formas flexionadas, geramos 45.564 formas adjetivais, 16.962 advérbios, 24.978 formas substantivais e 833.560 formas verbais. Como uma primeira etapa para a modelagem dos PFPs relativos aos sufixos *-ção* e *-mento*, analisamos a concorrência dos dois sufixos por bases verbais de primeira conjugação.

Palavras-chave: Morfologia de estados finitos. Dicionário eletrônico. Sufixação. Processamento de Linguagem Natural. Linguística Computacional.

ABSTRACT

In this work, we computationally modeled four word-formation processes (WFP) of Portuguese using finite-state morphology in order to automatically generate new lexical entries contributing with the research group *Computação e Linguagem Natural* on the development of resources for the Natural Language Processing (NLP) for the Portuguese language. Among the numerous challenges a PLN system must deal with by processing texts in natural language, plenty of them has to do with the lexical aspect, which interacts directly and indirectly with all the other levels of the system. Having well-structured and comprehensive lexical resources in hand decisively influences the efficiency of the PLN system. The best resource we are aware of is MorphoBr (ALENCAR; RADEMAKER; CUCONATO, 2018), built from the combination, revision and expansion of freely available analogous resources of Portuguese derived mostly from Label-Lex (ELEUTÉRIO et al., 1995) and Unitex-PB (MUNIZ, 2004). This expansion occurred by automatically generating, for instance, diminutive forms of adjectives and nouns and missing inflected forms of verbs. We propose the automatic generation of lexical entries by taking advantage of the existing ones as base forms for the word-formation processes by suffixation in order to retrofeed MorphoBr's data set. The four WFP selected were the equivalent to the suffixes *-vel*, *-idade*, *-izar* e *-mente*. As the morphosyntactic classes of the base and the product only are not enough to ascertain the word's well-formedness, we take into account the various restrictions of every selected WFP documented in Alves (2004), Basilio (1980, 1987, 1990, 2017), Cavalcante (1996), Maroneze (2005, 2011), Rocha (2008) and Villalva & Silvestre (2014). In terms of relative lemma quantity, we reached an increase of 8,5% with *-vel*, of 9,5% with *-idade*, 9,8% with *-izar* and 12,9% with *-mente*. In terms of absolute inflectional forms quantity, we have generated 45,564 adjective forms, 16,962 adverbs, 24,978 noun forms, and 833,560 verb forms. As a first step towards the modeling of the WFP related to the suffixes *-ção* and *-mento*, we analyzed the competition between both suffixes for first conjugation verb bases.

Keywords: Finite-state morphology. Electronic dictionary. Suffixation. Natural Language Processing. Computational linguistics.

LISTA DE FIGURAS

Figura 1 – Como o léxico do português muda..... 14
Figura 2 – Representação da rede de estados finitos que analisa a palavra “básico”..... 21
Figura 3 – Representação da rede de estados finitos que analisa “básico” e “básicos”..... 22
Figura 4 – Representação da rede de estados finitos que analisa e reconhece “básico”..... 22
Figura 5 – Representação do transdutor que mapeia “básico” sobre “básico+A+M+Sg”.. 23
Figura 6 – Relação entre expressão regular, língua / relação e rede de estados finitos..... 28

LISTA DE TABELAS

Tabela 1 – Quantidade de formas flexionadas e de lemas..... 23
Tabela 2 – Produtividade Intraclasse Aparente dos PFPs selecionados..... 26
Tabela 3 – Produtividade Geral Aparente dos PFPs selecionados..... 26
Tabela 4 – Compilação das PLPs PLI e PLG de quatro PFPs..... 27
Tabela 5 – Representatividade de adjetivos complexos como base do sufixo -*idade*..... 31
Tabela 6 – Expansão lexical a partir da sufixação de -*idade* a adjetivos complexos..... 31
Tabela 7 – Adjetivos acentuados sufixados em -*ico* candidatos a base do sufixo -*idade*.. 32
Tabela 8 – Representatividade de adjetivos complexos como base do sufixo -*izar*..... 33
Tabela 9 – Expansão lexical a partir da sufixação em -*izar* de adjetivos complexos..... 34
Tabela 10 – Representatividade de adjetivos complexos como base do sufixo -*mente*..... 35
Tabela 11 – Expansão lexical a partir da sufixação em -mente de adjetivos complexos.... 35
Tabela 12 – Propriedade dos transdutores..... 53

SUMÁRIO

1 INTRODUÇÃO..... 10
2 REFERENCIAL TEÓRICO..... 13
2.1 Léxico..... 13
2.2 Expansão do Léxico..... 14
2.2.1 Adjetivalização..... 17
2.2.2 Substantivalização..... 18
2.2.3 Verbialização..... 18
2.2.4 Adverbialização..... 19
2.3 MorphoBr..... 19
2.4 Morfologia de Estados Finitos..... 21
2.5 Compilador xfst..... 28
3 METODOLOGIA..... 29
3.1 Extração dos lemas não hifenizados..... 29
3.2 Produtividade morfológica..... 32
3.3 Descrição e análise do léxico..... 35
3.3.1 Adjetivalização de bases verbais pela sufixação em -vel..... 36
3.3.2 Substantivalização de bases verbais pela sufixação em -ção..... 37
3.3.3 Substantivalização de bases verbais pela sufixação em -mento..... 38
3.3.4 Análise da concorrência entre -ção e -mento no MorphoBr..... 39
3.3.5 Substantivalização de bases adjetivais pela sufixação em -idade..... 39
3.3.6 Verbialização de bases adjetivais pela sufixação em -izar..... 42
3.3.7 Adverbialização de bases adjetivais pela sufixação em -mente..... 43
4 TRANSDUTORES..... 46
4.1 Transdutor adjetival..... 46
4.2 Transdutor substantival..... 49
4.3 Transdutor adverbial..... 50
4.4 Transdutor verbal..... 51
4.5 Transdutor resultante..... 57
5 CONSIDERAÇÕES FINAIS..... 59
REFERÊNCIAS..... 61

1 INTRODUÇÃO

O Processamento de Linguagem Natural, doravante PLN, é um ramo interdisciplinar fruto da interseção entre Ciência da Computação e Linguística. Dentre os numerosos desafios com que um sistema de PLN deve lidar ao processar textos em língua natural, vários deles têm a ver com o domínio lexical, que dialoga direta ou indiretamente com todos os outros níveis do sistema.

Para Villalva e Silvestre (2014, p. 28), ao léxico mental dos falantes "[...] compete garantir a boa comunicação entre as restantes partes da gramática (a morfologia, a sintaxe, a semântica, a fonologia)" e essa "boa comunicação" deve se refletir na modelagem computacional do léxico de uma língua.

Ter em mãos recursos lexicais bem estruturados e abrangentes influencia decisivamente na eficiência das tarefas, tanto mais básicas quanto mais complexas, executadas pelo sistema de PLN. Duran (2013, p. 867) afirma que "a qualidade da execução dessas tarefas [mais básicas] exerce impacto nas grandes tarefas de PLN, como tradução automática, sumarização mono e multidocumento, simplificação textual, sistemas de perguntas e respostas, análise de opiniões e sentimentos."

Com essa questão em mente, no âmbito dos esforços do grupo de pesquisa Computação e Linguagem Natural (CompLin[1]), pretendemos contribuir com o PLN do português ao aumentarmos a quantidade de entradas lexicais do MorphoBr[2], que representa o estado da arte em termos de recursos lexicais para o português. Atualmente, o recurso se encontra hospedado na plataforma Github[3] e conta com uma licença de software livre. "A atual versão do MorphoBr compreende as classes gramaticais mais numerosas, ou seja, adjetivo, advérbio, substantivo e verbo" (ALENCAR et al., 2018, p. 5, tradução nossa[4]), cada um em seu respectivo diretório reunindo pequenos arquivos de texto, de modo que computadores de memória RAM mais modesta possam carregá-los sem maiores problemas.

Os dois métodos de expansão do conteúdo dos recursos lexicais originais foram a identificação e preenchimento de lacunas (i) em formas verbais atreladas a pronomes clíticos e (ii) em diminutivos adjetivais e substantivais. Nossa contribuição se vale da morfologia derivacional, uma vez que expandimos ainda mais os recursos lexicais disponíveis para o

[1] O sítio eletrônico <http://complin.blogspot.com/> conta com informações sobre as atividades do grupo.
[2] "[...] um léxico de formas plenas construído a partir da combinação, revisão e expansão de recursos análogos livremente disponíveis para o português [...]". (ALENCAR et al., 2018, p. 3).
[3] Disponível em <https://github.com/LFG-PTBR/MorphoBr>.
[4] "The present version of MorphoBr comprises the most numerous word classes, namely, nouns, adjectives, adverbs, and verbs."

PLN do português através da modelagem de processos de formação de palavras, doravante, PFPs. Essa expansão foi quantificada a fim de avaliarmos nossa contribuição.

O presente trabalho toma como ponto de partida a versão atual do MorphoBr, apresentado e descrito em Alencar et al. (2018). Seguimos a linha de trabalhos como Santos et al. (2015), que modela o PFP por sufixação em *-ão* produzindo (i) nomes agentivos, (ii) adjetivos agentivos e (iii) nomes de ação. Alencar (2009) trata da criatividade linguística através da modelagem de quatro PFPs sufixais e seis PFPs prefixais, resultando na ferramenta LEXPOR, descrita como o

> [...] protótipo de um componente morfológico capaz de realizar análises [...] de derivados por meio da sufixação de *-ismo*, *-iano*, *-ês* e *-mente* 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-*. (ALENCAR, 2009, p. 200).

Assim como Alencar (2009), Santos et al. (2015) e Alencar et al. (2018), nós nos valemos da morfologia de estados finitos para modelar o produto de nosso trabalho. Há ao menos três boas razões para essa escolha: (i) sua flexibilidade única em comparação a outros programas algorítmicos tradicionais; (ii) sua eficiência computacional e consequente velocidade de processamento; e (iii) sua compressibilidade (BEESLEY; KARTTUNEN, 2013, p. 56)[5].

Apesar do avançado estágio dos estudos atuais de morfologia derivacional do português e da descrição teórica de variados PFPs na literatura especializada, falta ainda sua implementação computacional. Para mitigar parcialmente essa lacuna, propusemo-nos modelar seis PFPs. Selecionamos pelo menos um PFP de cada classe morfossintática derivada para modelar.

Rocha (2008, p. 123) afirma que os sufixos *-ção* e *-mento* são dois dos mais produtivos sufixos substantivalizadores do português. Alves (2004, p. 33) aponta como mais frequentes na formação de adjetivos os sufixos *-ista*, *-ano* e *-vel*, dos quais, escolhemos o último para modelar. Maroneze (2005, p. 160) aponta o sufixo *-izar* como um dos mais produtivos sufixos

[5] "First, the mathematical properties of finite-state networks are well understood, allowing us to manipulate and combine finite-state networks in ways that would be impossible using traditional algorithmic programs; there is a mathematical beauty to finite-state computing that translates into unparalleled flexibility. Second, finite-state networks are computationally efficient for tasks like natural-language morphological analysis, resulting in phenomenal processing speeds. Third, in most cases, finite-state networks can store a great deal of information in relatively little memory, and finite-state networks can be further compressed using commercial Xerox technology."

formadores de verbos do português. O sufixo *-mente* é, por excelência, o afixo formador de advérbios, por isso também foi selecionado em nossa amostra de PFPs.

As palavras a seguir são exemplos de derivados bem formados que poderiam constar no MorphoBr, mas que ainda não figuram no recurso lexical e que geramos a partir da modelagem e aplicação de três dos seis PFPs selecionados: “direcionalidade”, “dublável” e “separativizar”

Nossas questões de pesquisa foram (i) qual o aumento percentual do MorphoBr resultante da modelagem de cada PFP selecionado? e (ii) qual o aumento percentual geral do MorphoBr resultante da modelagem de todos os PFPs selecionados? Assumimos como hipótese (i) um aumento percentual individual entre 7,5% e 12,5% e (ii) um aumento percentual médio de 10,0%.

Ao longo deste trabalho, usaremos a palavra “terminação” para nos referirmos à sequência de letras finais de uma palavra, que pode servir como significante de um morfema ou de parte dele. As terminações serão aspeadas. Os morfemas em si serão representados por hífen seguido de seu significante em itálico, sem aspas. A terminação “ar”, por exemplo, refere-se à sequência final formada pelas letras “a” e “r”, presente tanto no final de adjetivos, na forma do sufixo *-ar*, a exemplo de “pendular”, como no final de verbos, na forma da concatenação de morfemas *-a-r*, como “ajudar”.

No capítulo 2, apresentamos o referencial teórico, deste trabalho. No capítulo 3, detalhamos a metodologia executada, explicamos as decisões tomadas e apresentamos o protótipo do nosso transdutor. No capítulo 4, listamos as contribuições alcançadas e descrevemos as lacunas que podem ser exploradas para trabalhos futuros.

2 REFERENCIAL TEÓRICO

Neste capítulo, apresentamos os conceitos que fundamentaram nossas escolhas metodológicas relativas a como expandimos os recursos lexicais do português.

2.1 Léxico

Dentre as várias possibilidades de definição de léxico presentes na literatura mais antiga ou mais recente, seguimos Villalva e Silvestre (2014, p. 19) ao focarmos em três interpretações: o léxico dos sistemas linguísticos, o léxico mental dos falantes e o léxico enquanto componente de um modelo de gramática.

O léxico dos sistemas linguísticos reúne as palavras em uso de cada falante de cada comunidade de fala da mesma língua tanto em uso na contemporaneidade, quanto documentadas, tanto na fala, quanto na escrita. (VILLALVA; SILVESTRE, 2014, p. 23). Por causa da impossibilidade prática de se alcançar diretamente o léxico de uma língua, e de abarcá-lo em sua totalidade, são necessários meios indiretos de acessá-lo e de se construir uma representação dele.

Na produção e na compreensão oral e escrita dos falantes, manifesta-se o léxico mental de cada um, em que se acumulam, se estruturam e se organizam os dados linguísticos a que o falante é exposto ao longo de sua história como membro daquela comunidade linguística. Entretanto, como o léxico mental varia tanto de indivíduo para indivíduo como ao longo do tempo, além de sofrer interferências em sua manifestação, principalmente na fala, mas também na escrita, ele não pode ser, em si, o objeto de uma Teoria Lexical. Em seu lugar, devemos visar o léxico como componente de um modelo de gramática.

Villalva e Silvestre (2014, p. 28) afirmam que:

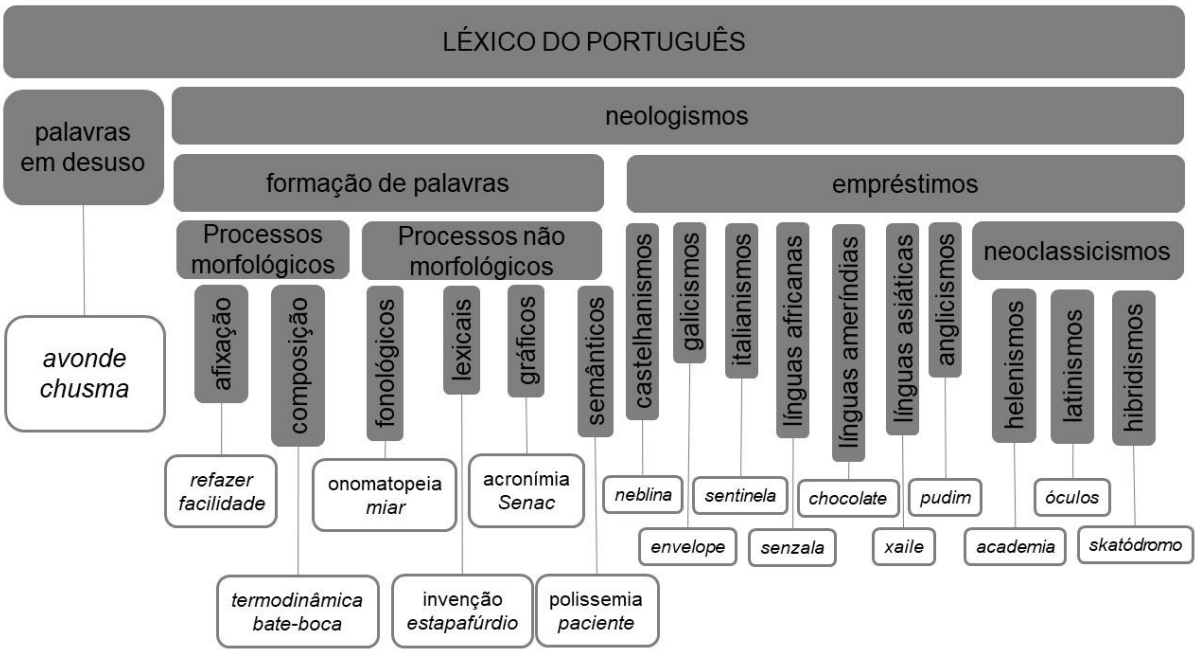
> “enquanto parte da gramática, o léxico pode ser visto como o lugar onde reside toda a informação que não é derivável, todas as propriedades idiossincráticas das línguas. É nesse papel que o léxico se distingue [...] da morfologia, a quem cabe a estruturação das palavras, igualmente formadas a partir da matéria-prima lexical”.

Essa afirmação ilustra o fato de que o léxico como componente da gramática é o ponto de partida e o ponto de chegada de qualquer trabalho que se proponha a expandir automaticamente um recurso lexical por meio de derivações, sendo a morfologia o território onde se deve transitar para alcançar essa expansão.

2.2 Expansão do léxico

A figura 1 serve para localizar, no cenário geral do léxico da língua portuguesa, o objeto deste trabalho, qual seja, processos morfológicos de formação de palavras dentro da morfologia derivacional sufixal em língua portuguesa.

Figura 1 - Como o léxico do português muda



Fonte: Villalva e Silvestre (2014, p. 72)

Uma forma de se construir um recurso lexical é a coleta de corpus. Esse tipo de procedimento aproveita a manifestação do léxico mental dos autores dos textos escritos usados como fonte para alimentar um dicionário eletrônico. A ocorrência da palavra no *corpus* indica sua existência na língua e justifica sua presença no dicionário eletrônico. Outra estratégia usada para alimentar o dicionário eletrônico é o aproveitamento de dicionários de referência. Essa estratégia é fundamentada no conceito de palavra atestada, que são as palavras para as quais se verificam algum tipo de registro de ocorrência.

Um contraponto aos dois procedimentos acima descritos vem de Villalva e Silvestre (2014, p. 84, grifo dos autores), ao argumentarem que "[...] ainda que ambos forneçam informações relevantes, nem o conceito de **palavra existente** nem o de **palavra atestada** [...] parecem ser os mais adequados quando pretendemos descrever o léxico dos falantes e, consequentemente, também o léxico da gramática".

Como o foco do presente trabalho se encontra exclusivamente no aspecto de expansão automática de um recurso lexical, deixamos de lado os conceitos de palavra existente e palavra atestada para aproveitarmos sobremaneira o conceito de palavra possível, explicado por Villalva e Silvestre (2014, p. 84, grifo dos autores):

> Resta o conceito de **palavra possível**, conceito que foi introduzido por Aronoff (1976) e que, recobrindo todas as potenciais ocorrências das palavras atestadas, acolhe também palavras que estão disponíveis para eventual uso dos falantes, por recurso a processos morfológicos de formação de palavras.

Uma abordagem semelhante é a de Rocha (2008, p. 44-45), que, ao tratar introdutoriamente de produtividade lexical, particiona o problema da existência de palavras em seis domínios:

(i) palavras impossíveis: “luzdor”, “gizdor”, “bonitodor”, “alegredor”;
(ii) palavras possíveis mas bloqueadas: “fabricador”, bloqueada por “fabricante”;
(iii) palavras possíveis e disponíveis: “cajuada”, “maracujada”, “atingimento”, “atingimento”, “efetuação”, “perduramento”, “espalhamento”;
(iv) palavras institucionalizadas, recém-criadas e não dicionarizadas: “faficheiro”, “fumódromo”, “malufar”, “crocodileiro”, “painelista”, “bicicleata”;
(v) palavras dicionarizadas em vias de entrar em desuso: arcaísmos, regionalismo e jargões;
(vi) palavras reais, em uso, dicionarizadas ou não.

Relevando qualquer distinção que possa haver entre o que o autor entende por “palavra institucionalizada”, “palavra real” e “palavra em uso”, é possível apontar que as Quatro oposições seguintes concorrem para estruturar a hierarquia descrita acima:

(1) palavras impossíveis x palavras possíveis;
(2) palavras disponíveis x palavras bloqueadas;
(3) palavras em desuso x palavras institucionalizadas / em uso / reais;
(4) palavras dicionarizadas x palavras não dicionarizadas.

As oposições que nos interessam compõem os pares (1) e (2), pois não nos propomos a traçar a fronteira do uso e do desuso (3) no léxico do português nem a identificar a entradas

lexicais dicionarizadas em nenhuma obra impressa. A nós nos interessam apenas as entradas lexicais presentes no MorphoBr. Palavras possíveis são formadas a partir de PFPs válidos da língua. Palavras bloqueadas, como “alisação”, são impedidas de se formar por já existir outra que supra a função requisitada, como “alisamento”, fenômeno explicado pelo princípio de economia linguística.

A validade do PFP é verificada na correta combinação de unidades lexicais. Na derivação sufixal, essa correta combinação é ditada pelo sufixo, que especifica a base e o produto em vários aspectos. Para além da especificação da classe sintática da base, é necessário especificar o tipo de base a que o sufixo se aplica. Essa especificação pode ser fonética, morfológica, sintática, semântica etc. (ROCHA, 2008, p. 37).

O aspecto fonético foi apenas indiretamente abordado através do aspecto grafemático. O aspecto morfológico foi o foco do nosso trabalho, pois analisamos a estrutura interna de palavras existentes para aproveitá-las como base para os processos morfológicos de formação palavras que modelamos. Não abordamos nem o aspecto sintático, que leva em consideração a estrutura argumental do vocábulo, nem o aspecto semântico, que subcategoriza o vocábulo num grupo semântico específico.

Em Villalva e Silvestre (2014, p. 106), os autores exemplificam as restrições do aspecto semântico: “[...] os sufixos derivacionais têm propriedades semânticas, distribuindo-se por grandes classes como a dos nomes-sujeito, a dos nomes de ação, a dos adjetivos relacionais ou dos verbos causativos”.

Um outro tipo de restrição tem a ver com a concorrência entre PFPs, que se manifesta, por exemplo, através “[d]os fenômenos de bloqueio, que impedem a ocorrência de uma palavra devido à existência de uma outra [e que estão] relacionados com o princípio geral de economia no funcionamento das línguas”. (VILLALVA; SILVESTRE, 2014, p. 145). Rocha (2008, p. 111) cita, por exemplo, os sufixos *-ção* e *-mento* como concorrentes por ambos “formarem substantivos abstratos a partir de verbos”. Esse fenômeno também foi levado em consideração na análise estrutural das palavras complexas.

Levamos em conta o que dizem Villalva e Silvestre (2014, p. 124) quanto à visível discrepância entre a primeira e as outras duas conjugações de modo que demos menos importância às bases verbais de segunda e terceira conjugação:

> [...] não existem verbos de flexão irregular na primeira conjugação - todos os verbos irregulares estão distribuídos pelas segunda e terceira conjugações. Por outro lado, todos os neologismos verbais (e.g. **clicar**, **teclar**, **printar**) se integram na primeira conjugação, o que traz, como contrapartida, a constatação de que as segunda e terceira conjugações são classes fechadas dentro da classe aberta dos verbos.

De posse dos dados resultantes dessa análise, fomos capazes de mensurar a "produtividade latente" de cada PFP modelado, que é a quantidade de derivados que aquele PFP tem potencial de formar utilizando-se apenas bases disponíveis no próprio dicionário em questão, no caso, o MorphoBr.

É importante frisar que o primeiro conjunto de derivados que propusemos por esse método, que chamamos de derivados de primeira geração, são potenciais bases para outros PFPs, formando o que chamamos de derivados de segunda geração. Todo adjetivo em *-vel* que propomos tem o potencial de servir de base para os sufixos *-idade*, *-izar* e *-mente*. Assim como todo verbo em *-izar* é candidato a base dos sufixos *-ção* e *-vel*. Os substantivos propostos não formam derivados de segunda geração, pois não modelamos nenhum PFP que tome bases substantivais. Os advérbios propostos também não são derivantes por causa da própria natureza não derivante dos advérbios em português.

2.2.1 Adjetivalização

O sufixo *-vel* toma o tema verbal do passado como base para formar um adjetivo paroxítono, que, pelas regras de ortografia do português, deve ser acentuado. O tema verbal do passado da primeira conjugação é o radical do verbo seguido da vogal temática "a" (5), que se acentua por ser o núcleo da sílaba tônica do paroxítono produzido. Já para as segunda e terceira conjugações, a vogal temática se neutraliza em "i", que também é acentuada.

(5) acessar / *acess-a* / acessável;
(6) conceber / *conceb-i* / concebível;
(7) conferir / *confer-i* / conferível;

Segundo Villalva e Silvestre (2014, p. 105), "este sufixo seleciona temas verbais de verbos que subcategorizam obrigatoriamente dois argumentos, um dos quais tem a função temática de tema [...]". Mais adiante, os autores afirmam que "cabe ao argumento que desempenha a função de tema na estrutura argumental do verbo assumir idêntica função na estrutura argumental do adjetivo, mas agora enquanto seu argumento externo." (VILLALVA; SILVESTRE, 2014, p. 109).

Em termos semânticos, o significado do adjetivo formado por adjunção do sufixo *-vel* a um tema verbal do passado é composicional. O adjetivo denota a possibilidade de se sofrer a ação do verbo que o deriva, por exemplo, "anotável" denota a possibilidade de ser anotado.

2.2.2 Substantivalização

Os sufixos *-ção* e *-mento* também tomam o tema verbal do passado como base, mas para formar substantivos abstratos. Como são sufixos tônicos, o primeiro forma oxítonas, e o segundo, paroxítonas; nenhum dos quais precisa ser acentuado. Ambos formam substantivos que denotam a ação ou o processo do verbo de que são derivados, por isso acabam por ser concorrentes entre si.

(8) anotar / *anot-a* / anotação;
(9) alçar / *alç-a* / alçamento;
(10) absolver / *absolv-i* / absolvição;
(11) derreter / *derret-i* / derretimento;
(12) partir / *part-i* / partição;
(13) aborrir / *aborr-i* / aborrimento.

O sufixo *-idade*, por sua vez, toma um radical adjetival como base para formar um substantivo abstrato que denota a qualidade relacionada ao adjetivo que o deriva. Um exemplo de radical adjetival presente no MorphoBr é “parental”, que deriva “parentalidade”, embora este substantivo bem formado não esteja presente no MorphoBr. Assim como os dois sufixos acima, este é tônico e forma paroxítonos não acentuados. Por não ter similaridade semântica com *-ção* ou *-mento*, acaba por não concorrer com eles na formação de substantivos.

2.2.3 Verbialização

O sufixo *-izar*, assim como *-idade*, também toma um radical adjetival, mas para formar um verbo causativo de primeira conjugação, embora também possa tomar um tema substantival. Um exemplo de verbo produzido a partir da adjunção desse sufixo é “parentalizar”, que não está presente no MorphoBr, embora seu primitivo “parental” esteja. Assim como os três sufixos da subseção anterior, *-izar* é tônico e forma oxítonos não acentuados. Verbos formados por esse sufixo denotam a mudança de estado.

2.2.4 Adverbialização

O sufixo *-mente* toma como base adjetivos flexionados no feminino singular para formar advérbios, o que, segundo (BASILIO, 2011), "vai contra a regra geral de que formas derivadas são construídas a partir do radical ou tema e não de formas já flexionadas". É talvez o PFP mais produtivo e menos restrito do português. Utilizaremos a descrição de Cavalcante (1996) para modelar esse PFP.

2.3 MorphoBr

De acordo com Alencar et al. (2018, p. 3, tradução nossa), o MorphoBr é "[...] um léxico de formas plenas construído a partir da combinação, revisão e expansão de recursos análogos livremente disponíveis para o português, derivados, em sua maior parte, do Label-Lex (ELEUTERIO et al., 1995) e Unitex-PB (MUNIZ, 2004)"[6].

A atual versão do Label-Lex é formada de três partes, LABEL-LEX-sw (formas flexionadas), LABEL-LEX-mw (expressões multi-palavras) e LABEL-LEX-gr (gramáticas), das quais apenas a primeira foi aproveitada pelo MorphoBr. García e Gamalo (2010) descrevem a integração do dicionário Label-Lex ao FreeLing (PADRÓ; STANISLOVSKY, 2012). García et al. (2014) expandem ainda mais o conteúdo do FreeLing, o que motivou a seguinte decisão metodológica na construção do MorphoBr: "Considerando todas as adições e melhorias à distribuição do Label-Lex feitas por Garcia e Gamalo, nós optamos por usar os arquivos FreeLing e Garcia (doravante GFL) em vez da distribuição Label-Lex" (ALENCAR et al., 2018, p. 4, tradução nossa)[7].

O DELAF (dicionário eletrônico de entradas simples flexionadas), que, junto dos dicionários DELAS (entradas simples não-flexionadas) e DELACF (entradas compostas flexionadas), compõem o conjunto de dicionários Unitex-PB, que foi construído a partir da base de dados lexicais Diadorim (GREGHI, 2002), selecionando, de dentro de seu conjunto de informações, apenas os lemas, as flexões e as etiquetas morfossintáticas.

[6] "[…] a full-form lexicon constructed from the combination, revision, and expansion of available free analog resources for Portuguese, mostly derived from Label-Lex (ELEUTERIO et al., 1995) and Unitex-PB (MUNIZ, 2004)".
[7] "Considering all the additions and improvements to the Label-Lex distribution made by Garcia and Gamalo, we opted to use the Garcia and FreeLing files (henceforth GFL) instead of the Label-Lex distribution."

A Diadorim foi criada a partir da unificação dos recursos lexicais no Núcleo Interinstitucional de Linguística Computacional (NILC[8]). Esses recursos foram o léxico do revisor gramatical ReGra (NUNES; OLIVERIA, 2000), o léxico do Thesaurus Eletrônico para o Português do Brasil, TeP, (DIAS-DA-SILVA et al., 2000) e o léxico da base de dados relacional do dicionário UNL-Português (OLIVEIRA et al., 2001).

As entradas lexicais do MorphoBr são formadas de quatro partes, que são a forma flexionada, o lema, a classe morfossintática e os traços morfossintáticos, além de dois separadores, os caracteres "\t", que é invisível e representamos por "•" e "+", como podemos ver nos seguintes exemplos:

(14) morfologias•morfologia+N+F+PL
(15) derivacionais•derivacional+A+F+PL
(16) contribuíram•contribuir+V+PRF+3+PL
(17) satisfatoriamente•satisfatoriamente+ADV

O exemplo (17) acima ilustra o fato de que os traços morfossintáticos não estão presentes em todas as classes. Além disso, cada entrada só admite uma classe, de modo que casos de homografia são registrados com entradas múltiplas, como nos exemplos (18) e (19) abaixo:

(18) casa casa+N+F+SG
(19) casa casar+V+PRS+3+SG

Outra característica do dicionário é que cada classe morfossintática tem uma determinada ordem de traços, como vemos abaixo. Os parênteses envolvendo a etiqueta "+NEG" significam opcionalidade.

Verbos: Tempo/Modo+Pessoa+Gênero+Número
Substantivos: Grau+Gênero+Número
Adjetivos: Grau+Gênero+Número
Advérbios: (+NEG)

[8] No sítio eletrônico http://www.nilc.icmc.usp.br/nilc/index.php, há mais informações sobre o núcleo.

De acordo com um dos arquivos[9] de sua documentação, o conjunto de etiquetas do MorphoBr é "baseado nas Regras de Glosagem de Lípsia[10] e outras abreviações de glosas morfológicas comumente adotadas na literatura linguística".

Os diretórios do MorphoBr que levamos em conta foram adjectives, adverbs, nouns e verbs. Em adjectives, encontram-se o arquivo extraído de GFL, adjs.gfl.dict, e os quatro extraídos do DELAF, a-c.delaf.dict, d-i.delaf.dict, j-p.delaf.dict e q-z.delaf.dict. Do mesmo modo, em nouns, temos o arquivo extraído de GFL, nouns.gfl.dict, e os extraídos do DELAF, a-c.delaf.dict, d-i.delaf.dict, j-p.delaf.dict e q-z.delaf.dict. Por sua vez, no diretório adverbs, há apenas os arquivos adverbs.delaf.dict e adverbs.gfl.dict, um extraído do DELAF, e o outro, do GFL.

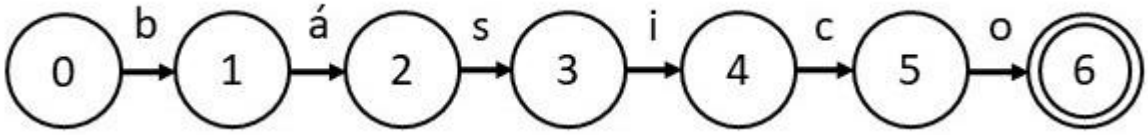
No diretório verbs[11], além do arquivo extraído de GFL, verbs.gfl.dict, há 50 arquivos com o conteúdo extraído do DELAF e nomeados de xaa.delaf.dict até xbx.delaf.dict. Os outros dois arquivos desse diretório, er-verbs-SBJF3SG.mbr.dict e IQ3s.mbr.dict, são resultado do preenchimento automático de lacunas na morfologia flexional que existia até então, não apresentando lemas adicionais e, por conseguinte, não sendo aproveitados em nossa expansão. O primeiro agrupa 728 formas de terceira pessoa do singular do futuro do subjuntivo de verbos de segunda conjugação. O segundo agrupa 25.423 formas de terceira pessoa do singular do imperfeito e do mais que perfeito do indicativo de verbos de primeira conjugação.

2.4 Morfologia de Estados Finitos[12]

Na figura 2, vemos a representação de um exemplo de rede de estados finitos. Essas redes são "máquinas abstratas que podem executar algumas tarefas linguísticas interessantes [...]" (BEESLEY; KARTTUNEN, 2013, p. 8). Os círculos representam os estados da rede. O círculo numerado com um zero representa o estado inicial. O círculo duplo representa o estado final. Numa rede de estados finitos, há apenas um estado inicial, mas pode haver zero, um ou mais estados finais. As setas, ou arcos, representam as transições entre os estados. As transições são acionadas através de sinais de entrada. Esses sinais podem ser equivalentes, por exemplo, a caracteres.

[9] Disponível em https://github.com/LFG-PTBR/MorphoBr/blob/master/TAGSET.
[10] Disponível em https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf.
[11] O subdiretório clitics, pertencente a verbs, foi ignorado por conter apenas arquivos com formas verbais unidas a pronomes clíticos.
[12] A estrutura desta seção foi inspirada no primeiro capítulo de Beesley e Karttunen (2013).

Figura 2 - Representação da rede de estados finitos que analisa a palavra “básico”

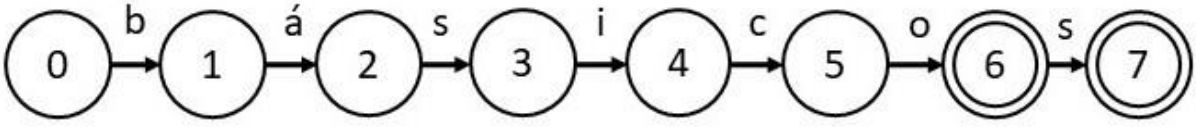


No caso da rede da figura 2, há sete estados e seis arcos. O primeiro estado só é atingido se a rede receber o sinal equivalente ao caractere “b”. Do mesmo modo, o estado dois só é alcançado após a rede receber o sinal equivalente ao caractere “á”. Seguindo esse raciocínio, se a rede receber a sequência de caracteres equivalente aos da palavra “básico”, a rede será percorrida desde o estado inicial até o estado final e essa palavra será reconhecida. Se, a qualquer momento, o caractere recebido não corresponder a uma transição entre aquele estado e o estado seguinte, a sequência completa é rejeitada.

Cada sequência de caracteres que a rede reconhece é considerada uma palavra da língua formal codificada por ela. No caso do nosso exemplo, a única palavra reconhecida pela rede é “básico”, ou seja, a língua formal codificada pela rede é formada pela única palavra “básico”.

Se quisermos uma rede que reconheça tanto a forma “básico”, quanto a forma “básicos”, precisamos da rede representada na figura 3, que tem oito estados, sendo dois deles finais, e sete arcos. O conjunto de caracteres aceitos pela rede é chamado de alfabeto. Observe que, embora a rede da figura 3 tenha um arco e um estado a mais que a da figura 2, ambas têm o conjunto de caracteres “b”, “á”, “s”, “i”, “c” e “o” como alfabeto.

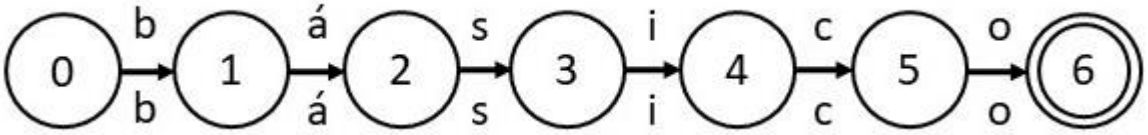
Figura 3 - Representação da rede de estados finitos que analisa “básico” e “básicos”



Até aqui, as redes tomadas de exemplo se comportaram apenas como analisadores de palavras, mas essas máquinas abstratas também são capazes de gerar palavras. No caso dos dois exemplos, a rede tem apenas um nível, o que quer dizer que cada caractere é consumido ou rejeitado e não há nenhum retorno mais elaborado. Entretanto, cada palavra analisada por uma rede pode ser retornada por ela se for de dois níveis.

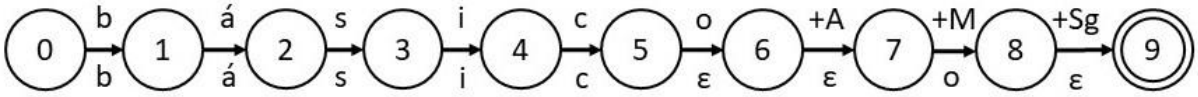
Na figura 4, temos a representação de uma rede de estados finitos de dois níveis, cada nível sendo equivalente a uma língua formal, uma língua superior e uma língua inferior. Nesse caso, as duas línguas são idênticas, pois são formadas pela mesma palavra “básico”. Segundo Beesley e Karttunen (2013, p. 47), “usamos o termo rede de estados finitos para cobrir tanto autômatos simples, que codificam uma língua regular, quanto transdutores, que codificam relações regulares”[13], ou seja, redes de um nível são chamadas autômatos, e redes de dois níveis são chamadas transdutores.

Figura 4 - Representação da rede de estados finitos que analisa e reconhece “básico”



Não há obrigatoriedade de que as palavras retornadas sejam idênticas às analisadas. É possível introduzir informações morfossintáticas a uma das línguas de modo que o transdutor funcione como um etiquetador morfossintático. Um analisador morfossintático relaciona pares de palavras da forma “básico+A+M+Sg” e “básico”. Os símbolos multicaractere “+A”, “+M”, “+Sg” são etiquetas morfossintáticas que são concatenadas ao lema da palavra analisada “básico”. Na figura 5, temos um transdutor que relaciona esse par caractere por caractere.

Figura 5 - Representação do transdutor de estados finitos que mapeia a palavra “básico” sobre a forma etiquetada “básico+A+M+Sg”



Por causa da flagrante disparidade de quantidade de caracteres, nove num caso e seis no outro, necessariamente sobram três caracteres para serem pareados. Transdutores de estados finitos têm a opção de parear caracteres com o caractere vazio, tradicionalmente representado por “ε” e que pode ser usado como equivalente ao morfema zero.

[13] “We use the term FINITE-STATE NETWORK to cover both simple automata that encode a regular language and transducers that encode a regular relation.”

Esse exemplo de transdutor lexical mapeia apenas um par, mas a mesma tecnologia é suficiente para criar desde um transdutor que analise e gere os pares referentes às quatro flexões do lema "básico" listadas em (16) até um recurso lexical com milhões de entradas lexicais de várias classes morfossintáticas de uma língua natural, como é o caso do MorphoBr.

(16) básico+A+M+Sg:básico
básico+A+F+Sg:básica
básico+A+M+Pl:básicos
básico+A+F+Pl:básicas

Beesley e Karttunen (2013, p. XVI) apontam os dois problemas centrais da morfologia como sendo (i) formação de palavra, morfotática e morfossintaxe e (ii) alternância fonológica e ortográfica. A morfotática lida com a correta combinação de morfemas para a geração de palavras válidas na língua, aspecto de suma importância na modelagem de PFPs. A morfossintaxe, por sua vez, lida também com a função sintática da palavra. As alternâncias em (ii) lidam com as acomodações que os morfemas sofrem ao se combinarem uns com os outros, como os dois fenômenos que abordamos neste trabalho: a alomorfia *-bil* / *-vel* e a perda do diacrítico de adjetivos sufixados por *-ico* ao receberem sufixos como *-idade* e *-izar*.

Nós nos valemos da morfologia de estados finitos para modelar os PFPs selecionados por ao menos três bons motivos: (i) sua flexibilidade única em comparação a outros programas algorítmicos tradicionais; (ii) sua eficiência computacional e consequente velocidade de processamento; e (iii) sua compressibilidade (BEESLEY; KARTTUNEN, 2013, p. 56)[14]. Alencar et al. (2018, p. 13) afirma o seguinte sobre essa tecnologia:

> A morfologia de estados finitos é o paradigma padrão para a construção de modelos computacionais baseados em regras de processos flexionais e de formação de palavras. Essa abordagem tem dois pontos fortes. Primeiramente, processos de formação de palavra podem ser formalizados de modo a espelhar estreitamente descrições linguísticas. Assim, não é necessário reinventar a roda ao implementar um fenômeno morfológico já descrito em detalhes na literatura linguística. Tudo o que se precisa fazer é traduzir a descrição de uma língua natural para uma especificação formal. Em segundo lugar, essa especificação formal pode ser

[14] "First, the mathematical properties of finite-state networks are well understood, allowing us to manipulate and combine finite-state networks in ways that would be impossible using traditional algorithmic programs; there is a mathematical beauty to finite-state computing that translates into unparalleled flexibility. Second, finite-state networks are computationally efficient for tasks like natural-language morphological analysis, resulting in phenomenal processing speeds. Third, in most cases, finite-state networks can store a great deal of information in relatively little memory, and finite-state networks can be further compressed using commercial Xerox technology."

> compilada em um FST usando software livre e open source como o Foma (HULDEN, 2009).[15]

Nossa modelagem se deu na forma de um transdutor de estados finitos, que é uma representação computacional em rede que mapeia palavras de uma língua formal de superfície sobre palavras de uma língua formal profunda. A aplicação linguística se dá traduzindo as formas flexionadas das palavras da língua natural no formato da língua formal superficial e reservando a língua formal profunda para seu equivalente etiquetado morfossintaticamente.

Dentre as três grandes vantagens de se utilizar esse formalismo, a que se destaca é a flexibilidade, uma vez que o mesmo transdutor de estados finitos pode ser percorrido nos dois sentidos, no da geração e no da análise. O da geração relaciona a forma profunda, etiquetada, com a forma superficial, flexionada. A análise representa o caminho inverso. Essa capacidade contribui com maior simplicidade para o funcionamento de um sistema de PLN.

Na subseção 3.3.5 e 3.3.6, descrevemos o caso do prefixo adjetivalizante *-ico*, que, junto de outros cinco prefixos adjetivalizantes, tomamos como indicador dos adjetivos complexos que serviram de recorte para as bases a receber o sufixo substantivalizante *-idade* e o sufixo verbalizante *-izar*. O sufixo *-ico* é postônico e aparece em proparoxítonas. Na sufixação por *-idade* e por *-izar*, é necessário o apagamento do diacrítico. Tecnicamente, o que ocorre é a substituição do caractere vocálico combinado ao diacrítico pela versão do caractere vocálico sem nenhum diacrítico.

Dentre os adjetivos sufixados por *-ico*, vamos usar o vocábulo "básico" para ilustrar o procedimento. Esse item lexical serve de primitivo no PFP que produz o derivado "basicidade". O transdutor que criamos para relacionar esse primitivo e esse derivado (i) toma "básico" como entrada e (ii) concatena "^idade" para formar "básico^idade". O caractere "^" é tradicionalmente usado para marcar fronteiras de morfema. Em seguida, (iii) o caractere "á" é substituído pelo caractere "a", gerando "basico^idade". Então, (iv) o caractere "o", identificado como antecedendo o caractere "^", é apagado. Por fim, (v) o marcador de fronteira de morfema "^" é apagado gerando "basicidade".

[15] "Finite-state morphology is the standard paradigm for the construction of rule-based computational models of inflectional and word-formation processes. This approach has two strengths. First, morphological processes can be formalized in a way that closely mirrors linguistic descriptions. Thus, one does not have to reinvent the wheel when implementing a certain morphological phenomenon already described in detail in the linguistic literature. All one needs to do is to translate the description from a natural language into a formal specification. Second, this formal specification can be compiled into an FST using free, open source software, e.g. Foma (HULDEN, 2009)."

(i) básico
(ii) básico^idade
(iii) basico^idade
(iv) basic^idade
(v) basicidade

Cada etapa desse processamento tem a forma de um transdutor de estados finitos, que é um autômato de dois níveis. Cada palavra de um nível é mapeada sobre uma palavra do outro nível. Assim, uma representação mais precisa das cinco etapas desse processamento é uma sequência de pares de palavras, uma da língua superior e outra da língua inferior. Esses pares também podem ser representados na forma "básico:básico", em que a palavra à esquerda de ":" é uma palavra da língua superior, e a palavra à direita é uma palavra da língua inferior.

(i) básico
básico

(ii) básico
basico^idade

(iii) basico^idade
basic^idade

(iv) basic^idade
basicidade

(v) basicidade
basicidade

As formas que nos interessam são "básico" e "basicidade". As outras três, "básico^idade", "basico^idade", "basic^idade", são consideradas formas intermediárias. O último passo para criar o transdutor final se dá através da operação de composição entre os cinco transdutores. Nessa operação, as formas intermediárias colapsam. Esse colapso ocorre em cascata. O compilador de estados finitos xfst (BEESLEY; KARTTUNEN, 2003), que é a ferramenta utilizada para operar e manipular os autômatos e transdutores, identifica as quatro

coincidências entre a língua inferior de um transdutor e a língua superior do seguinte, apontadas em (17) e elimina uma a uma até ficarmos apenas com um transdutor que relaciona os pares do tipo “básico:basicidade”.

(17) língua inferior de (i) e língua superior de (ii): “básico”
língua inferior de (ii) e língua superior de (iii): “basico^idade”
língua inferior de (iii) e língua superior de (iv): “basic^idade”
língua inferior de (iv) e língua superior de (v): “basicidade”

A operação de composição, descrita e exemplificada acima, é a maior vantagem da tecnologia de estados finitos, uma vez que contribui diretamente com cada um dos três pontos fortes dessa tecnologia apontados por Beesley e Karttunen (2013, p. 56), a flexibilidade, a eficiência e a compressibilidade. Contribui com a flexibilidade por causa da possibilidade de se modularizar a modelagem da morfologia das línguas. Contribui com a compressibilidade por causa da possibilidade de se tomar um número ilimitado de transdutores e transformá-los em um equivalente.

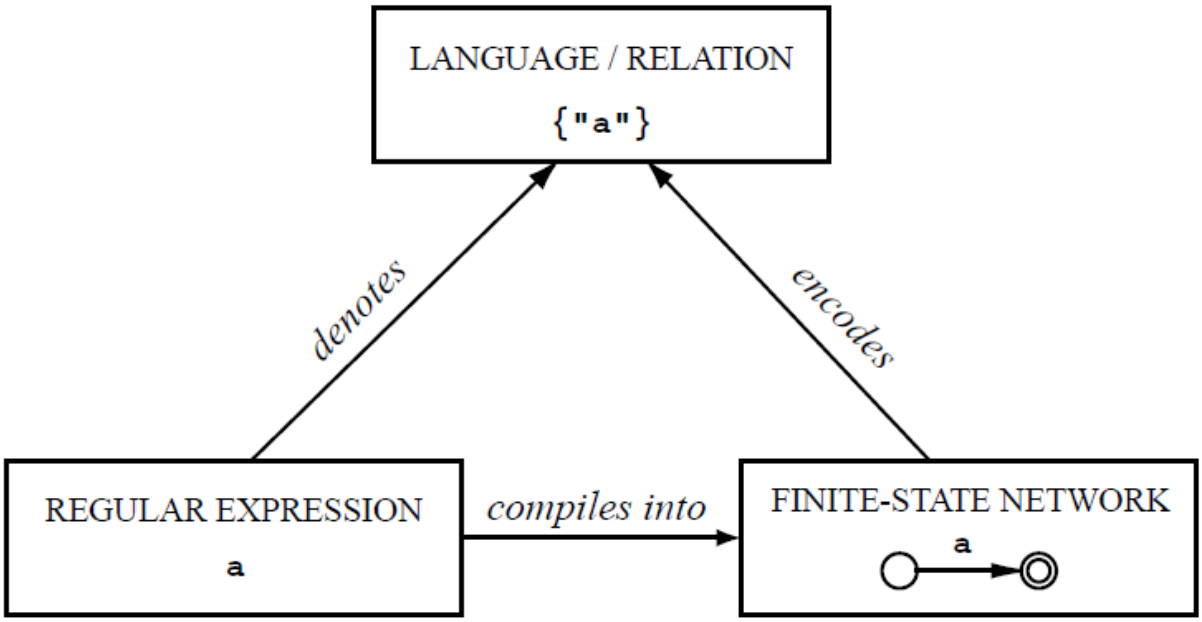
Beesley e Karttunen (2013, p. 30, tradução nossa) descrevem a operação de composição do ponto de vista de cada par de palavras da seguinte forma:

> Composição é uma operação sobre duas relações, e o resultado é uma relação. Se uma relação contém o par ordenado <x, y> a outra relação contém o par ordenado <y, z>, a relação resultante da composição <x, y> e <y, z>, nessa ordem, conterá o par ordenado <x, z>.[16]

Na figura 6, reproduzimos o diagrama de Beesley e Karttunen (2013, p. 46) que ilustra a relação entre línguas / relações, redes de estados finitos e expressões regulares. Uma expressão regular *denota* uma língua ou uma relação entre línguas. Uma expressão regular pode ser *compilada* em uma rede de estados finitos. Uma rede de estados finitos *codifica* uma língua ou uma relação entre línguas.

[16] “Composition is an operation on two relations, and the result is a new relation. If one relation contains the ordered pair <x, y> and the other relation contains the ordered pair <y, z>, the relation resulting from composing <x, y> and <y, z>, in that order, will contain the ordered pair <x, z>.”

Figura 6 – Relação entre expressão regular, língua / relação e rede de estados finitos



2.5 Compilador xfst

O compilador xfst trabalha com dois formalismos, lexc e xfst, que estão respectivamente ligados à morfotática e às alternâncias morfológicas, fonológicas, fonéticas e ortográficas. Cada um segue uma sintaxe específica. As regras de substituição e de contexto traduzidas no formalismo xfst são aplicadas ao léxico estruturado no formalismo lexc. O compilador armazena e opera com transdutores organizados numa pilha, em inglês *stack.*

As funções que utilizamos na construção do transdutor foram *source*, *read lexc*, *compose*, *regex*, *define*, *turn stack* e *save stack*. O comando *read lexc* serve para indicar ao compilador o arquivo em formato “.lexc” a ser lido e carregado na pilha. Já o comando *source* serve para indicar ao compilador o arquivo em formato “.xfst” a ser lido e carregado na pilha. O comando *turn stack* serve para o compilador inverter a ordem dos elementos armazenados na pilha. O comando *save stack* serve para o compilador salvar as redes armazenadas na pilha em um arquivo binário e de extensão “.fst” ou “.fsm”. Esse arquivo binário não legível por humanos, mas ocupa bem pouco espaço e pode ser recarregado na pilha por meio do comando *load*. O comando *compose* serve para o compilador executar a operação de composição em todos os transdutores armazenados na pilha. O comando *regex* serve para indicar ao xfst que compile uma expressão regular e carregue o transdutor equivalente. O comando *define* serve para nomear uma expressão regular.

3. METODOLOGIA

Este trabalho foi executado em duas etapas. Primeiramente, foi feita a investigação morfológica do conteúdo do MorphoBr utilizando-se um editor de texto e o ambiente de desenvolvimento interativo da linguagem de programação Python[17]. Essa combinação de ferramentas, aliadas ao uso de expressões regulares, manipulação de arquivos e cadeias de caracteres permitiu verificação de hipóteses, tomada de decisões e quantificação de dados. A segunda etapa foi a escrita dos arquivos que servem de fonte para a compilação do transdutor de estados finitos.

3.1 Extração dos lemas não hifenizados

Por o MorphoBr não conter arquivos prontos com os lemas adjetivais, adverbiais, substantivais e verbais, escrevemos um programa em linguagem de programação Python para extrai-los dos arquivos disponíveis. Durante o processo de extração de lemas, no entanto, encontramos 3.296 palavras hifenizadas – 836 adjetivos, 2.392 substantivos, 63 verbos e 5 advérbios. Foi necessário decidir manter ou descartar completa ou parcialmente esse conjunto de palavras hifenizadas, uma vez que a modelagem desse tipo de vocábulo poderia se mostrar muito complexa.

Segundo Antunes (2004, p. 41), na língua portuguesa, o hífen é usado em seis casos: translineação[18], divisão silábica, clíticos, sufixação, prefixação e composição por justaposição. Os casos de translineação e divisão silábica fogem do escopo deste trabalho. É seguro dizer que nenhuma das 3.296 palavras hifenizadas é formada por clítico, uma vez que, dentro do MorphoBr, esse tipo de vocábulo está reunido em arquivos que foram ignorados pelos procedimentos aqui adotados. Quanto à sufixação, os casos descritos em Antunes (2014, p. 73)[19] estão grafados sem hífen no MorphoBr. Concluímos, portanto, que as palavras hifenizadas eram casos ou de prefixação ou de justaposição.

Como exemplos de palavras hifenizadas derivadas por prefixação presentes no MorphoBr, temos os adjetivos “ab-reativo”, “ab-rogável” e “ad-retal”; os substantivos “ab-reação”, “ab-rogamento” e “ab-rogação”; e os verbos “co-administrar” e “ob-rogar”. Como

[17] Disponível em https://www.python.org/downloads/.
[18] Translineação refere-se à “partição dos vocábulos no final da linha” (ANTUNES, 2004, p. 41).
[19] “[...] nas formações por sufixação apenas se emprega o hífen nas palavras terminadas em sufixo de origem tupi-guarani que apresentam formas adjetivas, como ‘açu, guaçu e mirim’, quando o primeiro termo acaba em vogal acentuada graficamente ou a ortoépia exige a distinção gráfica dos dois elementos.”

tais casos não são numerosos, decidimos não os diferenciar dos hifenizados compostos por justaposição.

Como não pretendíamos tratar de palavras formadas por composição – nem como primitivas, nem como derivadas dos processos abordados neste trabalho –, selecionamos dois critérios: um para tratar da composição por aglutinação e outro para tratar da composição por justaposição.

Optamos por não diferenciar as palavras compostas por aglutinação das palavras simples, não sendo necessário descartar aquelas. Essa decisão se mostra como a mais sensata por causa da complexidade na identificação desses compostos. Quanto às formadas por justaposição, identificadas pela ocorrência do hífen, optamos por filtrar todos os 3.296 adjetivos, substantivos, verbos e advérbios[20] hifenizados.

A tabela 1 apresenta a comparação quantitativa entre formas flexionadas presentes no MorphoBr, lemas extraídos automaticamente e lemas não hifenizados aproveitados na construção dos arquivos “adjectives.lemas”, “nouns.lemas”, “verbs.lemas” e “adverbs.lemas”, que serviram de ponto de partida para a modelagem dos PFPs.

Tabela 1 - Quantidade de formas flexionadas e de lemas

Classe	**Formas Flexionadas**	**Lemas**	**Lemas Não Hifenizados**
Adjetivo	190.669	44.074	43.238
Substantivo	175.753	71.225	68.833
Verbo	1.380.555	15.256	15.193
Advérbio	4.223	4.185	4.180
Total	**1.752.952**	**134.740**	**131.444**

Em português, o advérbio é uma classe invariável, ou seja, não varia nem em gênero nem em número, mas a tabela 1 mostra que o número de formas adverbiais flexionadas é diferente do número de lemas adverbiais. Ora, além da forma superlativa, como “cedíssimo”, já se atesta o uso de formas diminutivas como “cedinho”, em que se emprega o sufixo mais frequente *-inho*, ou mesmo “cedito”, em que se emprega um sufixo menos frequente, *-ito*. A classe de advérbios não manifesta concordância, mas flexiona em grau. Daí advém a

[20] Independentemente da quantidade de advérbios hifenizados, todos poderiam ser mantidos pelo fato de os advérbios não serem primitivos de nenhum dos processos abordados neste trabalho, não exigindo, de nossa parte, qualquer descrição e análise adicional. Entretanto, para podermos aplicar o mesmo recorte nas quatro classes morfossintáticas, optamos por também descartar os cinco advérbios hifenizados.

disparidade entre a quantidade de formas adverbiais, 4.223, e lemas adverbiais, 4.185, o que produz uma diferença de 38 entradas lexicais, descritas a seguir.

Estes 14 superlativos sintéticos, formados pelo sufixo *-íssimo*, estão ausentes no DELAF e presentes no GFL: “brevíssimo”, “cedíssimo”, “certíssimo”, “depressíssimo”, “devagaríssimo”, “juntíssimo”, “longíssimo”, “malíssimo”, “muitíssimo”, “pertíssimo”, “pianíssimo”, “pouquíssimo”, “tantíssimo”, “tardíssimo”.

Estes 13 diminutivos, 12 em *-inho* e 1 em *-inha*, estão ausentes no DELAF e presentes no GFL: “baixinho”, “cedinho”, “certinho”, “depressinha”, “devagarinho”, “juntinho”, “loguinho”, “longinho”, “mansinho”, “pertinho”, “pianinho”, “pouquinho”, “nadinha”. Estes 6 diminutivos, 5 em *-ito* e 1 em *-ita* estão ausentes no DELAF e presentes no GFL: “cedito”, “devagarito”, “longito”, “pertito”, “pouquito”, “depressita”.

Diferentemente das 33 formas acima, os 3 advérbios de negação a seguir estão presentes tanto do DELAF, quanto no GFL, mas apenas no GFL estão acompanhados da etiqueta “+NEG”: “jamais”, “nada” e “não”.

A forma plena do advérbio “eis” está presente apenas no DELAF, enquanto a forma reduzida “ei” está presente apenas no GFL. A forma reduzida aparece nas combinações de “eis” com pronomes oblíquos átonos de terceira pessoa, singular e plural, como “ei-lo”, “ei-la”, “ei-los” e “ei-las”.

Por último, identificamos uma inconsistência entre as lematizações da forma comparativa “pior” no DELAF e no GFL. Os advérbios “bem”, “melhor”, “mal” e “pior” estão presentes no DELAF, e seus lemas são, respectivamente “bem”, “melhor”, “mal” e “pior”. No entanto, dos quatro advérbios listados, o único presente no GFL, “pior”, foi lematizado como “mal”.

Duas alternativas se apresentam para mitigarmos essa pequena inconsistência no MorphoBr: (i) manter o padrão do DELAF, considerando “melhor” e “pior” como advérbios independentes de “bem” e “mal” e descartando manualmente a entrada lexical presente no GFL; ou (ii) adotando o padrão do GFL, descartando a entrada (18) e substituindo (19) por (20).

(18) pior pior+ADV
(19) melhor melhor+ADV
(20) melhor bem+ADV

3.2 Produtividade morfológica

A produtividade de um PFP mensura seus derivados em termos de quantidade (produtividade bruta) ou em termos de porcentagem (produtividade relativa). A produtividade relativa pode tomar como universo o léxico completo ou uma parte desse léxico. Julgamos útil elencar dois subléxicos como universos amostrais: (i) o de todos os vocábulos não hifenizados oriundos de todas as classes morfossintáticas disponíveis no léxico original e (ii) o subléxico de uma classe morfossintática específica. A produtividade relativa que toma (i) como universo é chamada produtividade global. A produtividade relativa que toma (ii) como universo é chamada produtividade intraclasse.

Achamos útil também opor produtividade *aparente* e produtividade *real*. Quanto mais precisa e refinada for a modelagem do PFP em questão, mais a produtividade aparente se aproxima da produtividade real. No atual estágio do MorphoBr, ao analisarmos a produtividade *aparente* de cada PFP no contexto da classe de palavra do derivado e no contexto de todos os lemas não hifenizados extraídos do MorphoBr, chegamos aos números das tabelas 2 e 3.

Referimo-nos a produtividade aparente, pois a simples ocorrência de uma sequência de grafemas finais homógrafos a um sufixo não garante que a palavra seja derivada do PFP relativo àquele sufixo, embora sirva como um primeiro filtro. A palavra “aumento”, por exemplo, termina em “mento”, mas não é sufixada por *-mento* já que “au” não termina nem em “a”, nem em “i”, não podendo ser um tema verbal nem de primeira, nem de segunda, nem de terceira conjugação. Falsos-positivos como esse foram incluídos no cálculo da produtividade aparente.

A produtividade *intraclasse* aparente, doravante PIA (tabela 2), mensura, no contexto da classe morfossintática do derivado, o percentual de candidatos a produtos do PFP em questão. Os percentuais foram calculados dividindo-se o número de ocorrências potenciais pelo número de lemas não hifenizados pertencentes àquela classe. Já a produtividade *global* aparente, doravante PGA (tabela 3), considera como universo a união das quatro classes presentes no MorphoBr, ou seja, adjetivos, advérbios, substantivos e verbos.

É importante frisar que os valores reunidos nas tabelas 2 e 3 referem-se apenas aos dados já presentes no MorphoBr. Apresentamos nossas contribuições, que propomos incluir no MorphoBr, ao longo da seção 3.3. A utilidade de se trabalhar com a produtividade aparente é o de fornecer uma estimativa de base para a quantidade máxima de ocorrências daquele PFP. Esse valor é superestimado por causa dos falsos-positivos incluídos no cálculo. Só

depois de encontrar essa estimativa é que iniciamos uma investigação desses falsos-negativos de modo a refinar nossos resultados.

Tabela 2 - Produtividade Intraclasse Aparente dos PFPs selecionados

Natureza da Base	**Natureza do Derivado**	**Sufixo**	**Ocorrências Potenciais**	**Entradas da mesma classe**	**PIA**
Verbo	Adjetivo	*-vel*	3.461	43.238	8,0%
	Substantivo	*-ção*	3.909	68.833	5,7%
		-mento	2.052	68.833	3,0%
		-idade	1.755	68.833	2,5%
Adjetivo	Verbo	*-izar*	1.000	15.193	6,6%
	Advérbio	*-mente*	3.864	4.180	92%

Tabela 3 - Produtividade Geral Aparente dos PFPs selecionados

Natureza da Base	**Natureza do Derivado**	**Sufixo**	**Ocorrências Potenciais**	**Entradas do MorphoBr**	**PGA**
Verbo	Adjetivo	*-vel*	3.461	131.444	2,6%
	Substantivo	*-ção*	3.909		3,0%
		-mento	2.052		1,6%
		-idade	1.755		1,3%
Adjetivo	Verbo	*-izar*	1.000		0,8%
	Advérbio	*-mente*	3.864		3,2%

Em suma, a produtividade de um PFP pode ser bruta ou relativa, aparente ou real. Dentre as relativas, selecionamos a intraclasse e a global. Todos os tipos de produtividade apresentados até aqui servem para descrever um recurso lexical em um determinado estado. Entretanto, como usamos subléxicos do MorphoBr como base para produzir novas entradas lexicais e expandir o recurso lexical original – essencialmente retroalimentando-o – achamos útil introduzir o conceito de produtividade *latente*.

Na seção 2.2, descrevemos a produtividade latente como “a quantidade de derivados que aquele PFP tem potencial de formar utilizando-se apenas bases disponíveis no próprio dicionário em questão”. A título de ilustração, consideremos os seguintes conjuntos de

adjetivos (21) e verbos (22) como componentes de um microrrecurso lexical, que chamaremos de MicroMorphoBr.

(21) alegre, real, municipal, parental
(22) comprar, realizar, municipalizar, entrar

Ao aplicarmos o PFP verbializante deadjetival de sufixo *-izar* aos dois conjuntos, podemos identificar quatro subconjuntos:

(i) adjetivos que não servem de primitivo: alegre;
(ii) verbos que não servem de produto: comprar, entrar;
(iii) adjetivos que servem de primitivo: real, municipal, parental; e
(iv) verbos que servem de produto: realizar, municipalizar.

Entretanto, ao parearmos os elementos dos subconjuntos (iii) e (iv) da seguinte forma "real / realizar" e "municipal / municipalizar", notamos que o adjetivo "parental" não possui seu derivado verbal listado no recurso lexical, embora "parentalizar" seja perfeitamente bem formado. Caso expandíssemos o microrrecurso lexical incluindo o verbo "parentalizar", produziríamos o MicroMorphoBr Expandido. Essa potencialidade pode ser mensurada de modo absoluto ao se afirmar que a produtividade latente *bruta* do PFP verbializante deadjetival de sufixo *-izar* vinculada ao MicroMorphoBr tem valor um.

Seguindo a tipologia de produtividades já apresentada, a produtividade latente *relativa* pode ser dividida em *intraclasse* e *global*, doravante PLI e PLG respectivamente. No caso do exemplo em questão, a PLI é calculada dividindo-se a produtividade latente bruta (um) pela quantidade de entradas lexicais (verbos) já presentes no recurso lexical original (MicroMorphoBr), ou seja, quatro. Assim, a PLI = 1 / 4 = 0,25 = 25%. Do mesmo modo, a PLG é calculada dividindo-se a mesma produtividade latente bruta (um) pela quantidade de entradas lexicais totais (verbos e adjetivos) já presentes no recurso lexical original (MicroMorphoBr), ou seja, oito. Assim, PLG = 1 / 8 = 0,125 = 12,5%.

Entretanto, como dito anteriormente, outras produtividades relativas podem ser definidas. Assim, vamos definir a produtividade latente *processual*, doravante PLP, como aquela que toma o subléxico dos produtos do PFP em questão como referência. No caso deste exemplo, a PLP é calculada dividindo-se a produtividade latente bruta (um) pela quantidade de produtos do PFP (*-izar*) em questão já presentes no recurso lexical original

(MicroMorphoBr), ou seja, dois (“realizar” e “municipalizar”). Assim, PLP = 1 / 2 = 0,5 = 50%.

3.3 Descrição e Análise do Léxico

Ao longo dessa subseção, descrevemos em detalhes os procedimentos que seguimos para chegar às produtividades latentes apresentadas na tabela 4. Apenas os PFPs relativos aos sufixos *-vel, -idade, -izar* e *-mente* figuram na tabela. Quanto ao sufixos *-ção* e *-mento*, não foram geradas novas entradas lexicais. Apenas analisamos como os dois sufixos competem entre si dentro do estado atual do MorphoBr identificando os verbos que funcionam como primitivos dos dois PFPs, os verbos que funcionam como primitivos de apenas um PFP e os verbos que não servem como primitivos nem de um nem de outro PFP.

Tabela 4 - Compilação das PLPs, PLIs e PLGs de quatro PFPs

PFP	**PLP**	**PLI**	**PLG**
-vel	329%	26,3%	8,5%
-idade	709%	18,1%	9,5%
-izar	1.282%	84,4%	9,8%
-mente	439%	406%	12,9%

Ao longo deste trabalho, usamos o termo “cruzar” para nos referirmos à operação de extração do conjunto diferença. A operação “diferença”, quando aplicada a dois conjuntos A e B, nos dá o conjunto de elementos pertencentes ao conjunto A mas não pertencentes ao conjunto B. Por exemplo, ao cruzarmos o conjunto de verbos do MicroMorphoBr Expandido (23) com o conjunto de verbos do MicroMorphoBr (24), obtemos o conjunto diferença abaixo (25):

(23) comprar, realizar, municipalizar, entrar
(24) comprar, realizar, municipalizar, entrar, parentalizar
(25) parentalizar

É essa operação que, nas seções 3.3.1, 3.3.5, 3.3.6 e 3.3.7, nos permite identificar as entradas lexicais que propomos serem incluídas no MorphoBr.

3.3.1 Adjetivalização de bases verbais pela sufixação em -vel

Tomemos, primeiramente, os adjetivos terminados em "vel". Como não há etiquetas morfológicas que explicitem onde há, de fato, a ocorrência desse sufixo, uma busca simples por adjetivos terminados pela sequência de letras "v", "e" e "l" superestima a quantidade real de ocorrências do sufixo em 3.461. Essa busca simples funciona como um primeiro filtro, que descrevemos como produtividade aparente bruta.

Um segundo filtro, mais preciso, leva em consideração uma das restrições de ocorrência do sufixo *-vel*. Como apenas sufixos *-vel* precedidos das vogais temáticas tônicas "á" e "í" indicam reais ocorrências do sufixo, podemos restringir a busca e concluir que há 2.748 lemas adjetivais terminados por "ável" (candidatos a derivados de bases verbais de primeira conjugação) e 681 terminados por "ível" (candidatos a derivados de bases verbais de segunda e terceira conjugações).

Somados, corrigimos a produtividade aparente bruta para 3.429 adjetivos, o que faz a PIA cair de 8,0% para 7,9% e a PGA a se manter em 2,6%.

As 32 ocorrências restantes são falsos positivos, verificados no dicionário online Priberam[21], que descartamos e descrevemos a seguir divididos em três casos.

(i) há sete ocorrências de "-avel" e uma de "-uvel" em que o acento parece estar ausente: "abagulhavel", "conferenciavel", "destrinçavel", "ponderavel", "tachavel", "tacheavel", "tenteavel" e "insoluvel"[22];

(ii) há dez ocorrências de "úvel", oito de "-óvel" e duas de "-ével", que são não servem de primitivos para o PFP em questão : "dextrivolúvel", "dissolúvel", "hidrossolúvel", "indissolúvel", "insolúvel", "irresolúvel", "lipossolúvel", "resolúvel", "solúvel", "volúvel", "automóvel", "electromóvel", "eletromóvel", "hipomóvel", "imóvel", "locomóvel", "móvel", "semimóvel", "delével", "indelével";

(iii) há os quatro adjetivos "javel", "novel", "revel", "cascavel", que já têm "vel" como parte de seus radicais.

[21] Disponível no site https://dicionario.priberam.org/.

[22] Apenas as contrapartes acentuadas de "abagulhavel", "conferenciavel", "tacheavel" e "tenteavel" não presentes no MorphoBr.

Dentre os 15.193 verbos do arquivo “verbs.lemas”, extraímos todos os 13.538 verbos da primeira conjugação, gerando o arquivo “v1.lemas”. De “v1.lemas”, substituímos a terminação “ar” por “ável”, gerando o arquivo “v1_vel.lemas” com 13.538 adjetivos terminados em “ável”.

Ao cruzar os 13.538 adjetivos de “v1_vel.lemas” com os 43.238 adjetivos não hifenizados de “adjectives.lemas”, identificamos 11.391 novos adjetivos em “-ável”, o que equivale: (i) a um aumento de 415% em relação aos 2.748 adjetivos formados por sufixação em *-vel* e derivados de verbos de primeira conjugação; (ii) a um aumento de 329% em relação aos 3.461 adjetivos formados por sufixação em *-vel*; (iii) a um aumento de 26,3% em relação a todos os 43.238 adjetivos do MorphoBr; e (iv) a um aumento de 8,5% em relação a todas as 131.444 entradas lexicais do MorphoBr. A partir dos valores (ii), (iii) e (iv), concluímos que a PLP, a PLI e a PLG desse PFP são 329%, 26,3% e 8,5% respectivamente.

3.3.2 Substantivalização de bases verbais pela sufixação em -ção

No caso dos 3.909 substantivos terminados com as letras “ção”, apenas 3.178 terminam em “ação” (candidatos a derivados de bases verbais de primeira conjugação) e 269 em “ição” (candidatos a derivados de bases verbais de segunda e terceira conjugações), que totalizam 3.447. Ao dividirmos esse número pelos 68.833 substantivos não-hifenizados, obtemos uma PIA de 5,0%. Dividindo-se os mesmos 3.447 pelas 131.444 entradas lexicais do MorphoBr, obtemos uma PGA de 2,6%.

As 462 ocorrências restantes, descritas nos três próximos parágrafos, reúnem substantivos terminados em “eção”, “oção”, “ução”, “nção”, “cção”, “lção”, “pção”, “rção” e “scção”, e, por não seguirem a regularidade que nos propomos a explorar, não foram aproveitados no processamento.

Há 98 ocorrências de “eção”, dentre os quais há derivados de verbos em “etar” e “igir” ou da deleção do “c” das terminações “ecção”. Há 21 ocorrências de “oção”[23], das quais 4 derivam de verbos em “otar”, 8, de verbos em “mover” e 9 são lexicalizados. Há 81 ocorrências de “ução”, das quais 28 são candidatos a derivados de verbos em “duzir”, 17

[23] São elas: adoção, readoção, promoção, despromoção, autopromoção, comoção, demoção, locomoção, remoção, emoção, premoção, tremoção, carroção, devoção, indevoção, loção, noção, moção, prenoção, pescoção, poção.

terminam em “lução”, 9 terminam em “strução”, 5 derivam de “locução”, restando 20 lexicalizadas ou formadas por prefixação ou composição[24].

Há 75 ocorrências de “nção”, das quais 6 terminam em “ânção” e são lexicalizadas; 31 terminam em “enção” e são ou derivadas de verbos em “entar” ou “ter” ou lexicalizadas; 8 terminam em “inção” e são ou derivadas de verbos em “tinguir” ou lexicalizadas; 26 terminam em “unção”, das quais 9 terminam em “junção”, 6, em “função”, 4 derivam de “pungir”, 7 terminam em “sunção”, restando 4 lexicalizadas: “bênção”, “monção”, “contramonção” e “unção”.

Há 126 ocorrências de “cção”, sendo 44 de “acção”, 64 de “ecção”, 12 de “icção”, 2 de “ocção” e 3 de “ucção”, restando “mincção” como variante de “micção”. Há 3 ocorrências de “lção”: “alção”, “balção” e “calção”. Há 34 ocorrências de “pção”, dos quais 15 terminam em “epção”, 10 terminam em “rupção” e são candidatos a derivados de verbos em “romper”, 6 terminam em “mpção”, e 3, em “opção”, sendo derivados de verbos em “optar”. Há 25 ocorrências de “rção”, das quais 4 terminam em “arção” e são lexicalizadas[25], 7 terminam em “erção” e 11 terminam em “orção”. A única ocorrência de “sção” refere-se à palavra “imisção”.

3.3.3 Substantivalização de bases verbais pela sufixação em -mento

Seguindo o raciocínio aplicado no caso de “ção”, dos 2.052 substantivos terminados com a sequência “mento”, 1.597 terminam em “amento” – candidatos a derivados de bases verbais de primeira conjugação – e 411, em “imento” – candidatos a derivados de bases verbais de segunda e terceira conjugações, que totalizam 2.008. Ao dividirmos esse número pelos 68.833 substantivos não-hifenizados, obtemos uma PIA de 2,9%. Dividindo-se os mesmos 2.008 pelas 131.444 entradas lexicais do MorphoBr, obtemos uma PGA de 1,5%.

Das 44 ocorrências restantes, a maioria é de substantivos ~~primitivos,~~ que já contêm “mento” no próprio tema. Podemos dividir as 44 ocorrências em três grupos:

(i) 35 primitivos: argumento, armento, aumento, cemento, comento, complemento, decremento, documento, elemento, emolumento, escarmento, excremento, fermento, fomento, fragmento, frumento, implemento, incremento, indumento, instrumento, jumento, lomento, memento, mento,

[24] São elas: carapução, caução, consecução, corrução, precaução, desprecaução, imprecaução, sução, lipossução, exsução, execução, inexecução, eletrexecução, electrexecução, loução, nução, oução, persecução, prossecução, retoução.
[25] São elas: arção, camarção, escarção e garção.

momento, monumento, pigmento, recremento, sarmento, segmento, suplemento, tegumento, tomento, tormento;

(ii) 7 derivados por prefixação: adimplemento, inadimplemento, auripigmento, contradocumento, hiperdocumento, radielemento, radioelemento;

(iii) 2 derivados por sufixação: asmento, ciumento.

3.3.4 Análise da concorrência entre -ção e -mento

Como descrito nas subseções 3.3.2 e 3.3.3, os sufixos *-ção* e *-mento* tomam radicais verbais para formar substantivos. A título de investigação, tomamos apenas os verbos de primeira conjugação, derivamos os substantivos através dos dois PFPs, cruzamos cada lista com os substantivos já presentes no MorphoBr e interpretamos os resultados.

Podemos dividir os verbos da primeira conjugação em quatro casos: (i) verbos que não têm derivados em *-ção* nem em *-mento*; (ii) 1.194 verbos não têm derivados em *-ção*, mas têm em *-mento*; (iii) 2.775 verbos têm derivados em *-ção*, mas não têm em *-mento*; e (iv) 403 verbos têm derivados em *-ção* e em *-mento*.

3.3.5 Substantivalização de bases adjetivais pela sufixação em -idade[26]

No caso dos 1.761 substantivos terminados em "idade", devemos levar em consideração o tema do adjetivo de base. Segundo Villalva e Silvestre (2014, p. 132), há cinco classes formais de adjetivos: (i) os de tema em *-o*, (ii) os de tema em *-a*, (iii) os de tema em *-e*, (iv) os de tema consonantal e (v) os atemáticos, exemplificados por "alto", "baixa", "livre", "normal" e "simples", respectivamente.

A substantivalização de adjetivos com tema vocálico – casos i, ii e iii – suprime a vogal temática e toma apenas o radical adjetival como base. Já com adjetivos de tema consonantal – caso iv –, há apenas a adjunção do sufixo *-idade*. Não trataremos de bases adjetivais atemáticas – caso v – neste projeto.

A tabela 5 ilustra um dado que encontramos ao analisar o conteúdo do MorphoBr e que consideramos pertinente, que é o fato de que derivados de um subconjunto de adjetivos complexos correspondem a, aproximadamente, três quartos – 74,2% – dos substantivos em

[26] Existem 80 ocorrências de *-dade* não precedidos por -i no arquivo "nouns.lemas".

"idade". Essa porcentagem justifica nossa escolha metodológica de focar nos adjetivos formados pelos seis sufixos da tabela 5 como bases para o sufixo *-idade*.

Tabela 5 - Representatividade de adjetivos complexos como base do sufixo *-idade*

Sufixo da Base Adjetival	**Terminação Pós-Substantivalização**	**Número de Ocorrências**	**Porcentagem de Ocorrências**
-al	*-alidade*	262	14,9%
-ar	*-aridade*	69	3,9%
-bil (*-vel*)	*-bilidade*	511	29,0%
-ico	*-icidade*	148	8,4%
-ivo	*-ividade*	164	9,3%
-oso	*-osidade*	152	8,6%
Total		**1.306**	**74,2%**

Aproveitando essa constatação, dentre os 43.238 adjetivos do arquivo "adjectives.lemas", extraímos 14.532 adjetivos formados pelos sufixos *-al*, *-ar*, *-ável*, *-ico*, *-ivo* e *-oso*, que quantificamos na tabela 6. Os 14.532 substantivos formados a partir da sufixação de *-idade* desses 14.532 adjetivos-base correspondem a 825% dos 1.761 já presentes no MorphoBr.

Tabela 6 – Mapeamento de expansão a partir da sufixação em *-idade* de adjetivos complexos

Sufixo da Base Adjetival	**Terminação Pós-Substantivalização**	**Adjetivos-Base no MorphoBr**	**Adjetivos-Base Latentes**
-al	*-alidade*	262	1.808
-ar	*-aridade*	69	582
-abil (*-ável*)	*-abilidade*	289	2.748
-ico	*-icidade*	148	7.209
-ivo	*-ividade*	164	983
-oso	*-osidade*	152	1.202
Total		**1.084**	**14.532**

Esses seis sufixos podem ser divididos em quatro grupos baseados no modo como interagem com o sufixo *-idade*. Essa divisão se dá como resultado da confluência de dois aspectos: a classe temática a que pertence o adjetivo base e a tonicidade do sufixo. Os sufixos

-al e *-ar* recebem *-idade* por simples adjunção, sem nenhuma acomodação gráfica ou morfológica, como mostram os pares de exemplos (26) e (27).

(26) “real”, “realidade”
(27) “polar”, “polaridade”

O sufixo *-vel* recebe *-idade* sofrendo alomorfia para *-bil*, e a vogal temática que o precede perde o acento tônico para o sufixo *-idade*, como vemos em (28).

(28) “arável” e “arabilidade”

O sufixos *-ico*, *-ivo* e *-oso*, poderiam formar o terceiro e último grupo a partir dos seis sufixos elencados, pois todos os três recebem *-idade* após perda da desinência de gênero. Entretanto, diferentemente dos outros cinco sufixos, o sufixo *-ico* é postônico e aparece, necessariamente, em proparoxítonas, que sempre são acentuadas. Como podemos ver na tabela 7, todas as combinações de vogais e diacríticos válidas na língua portuguesa aparecem nesse contexto.

Tabela 7 – Mapeamento dos 7.371 adjetivos acentuados sufixados em *-ico* candidatos a base do sufixo *-idade*

Vogal Acentuada	**Ocorrências no MorphoBr**	**Exemplo de Base em *-ico***	**Exemplo de Derivado em *-idade***
á	1.421	aeron**á**utico	aeron**a**uticidade
â	250	din**â**mico	din**a**micidade
é	1.452	ciberné**t**ico	ciberne**t**icidade
ê	392	d**ê**itico	d**e**iticidade
í	973	cr**í**tico	cr**i**ticidade
ó	1.735	t**ó**xico	t**o**xicidade
ô	654	t**ô**nico	t**o**nicidade
ú	114	l**ú**dico	l**u**dicidade

Do ponto de vista teórico, a perda de acentuação entre base e derivado no PFP descrito acima deve ocorrer ou antes ou depois da sufixação de *-idade*. Em termos de modelagem, ambas alternativas apresentam os mesmos resultados e a mesma simplicidade. A perda da acentuação foi modelada como uma regra no transdutor de estados finitos.

Ao cruzar os 13.448 substantivos em *-idade* de adjIDADE.lemas com os 68.833 substantivos de "nouns.lemas", identificamos 12.489 novos substantivos em *-idade*, o que equivale: (i) a um aumento de 709% em relação aos 1.761 substantivos formados por sufixação em *-idade*; (ii) a um aumento de 18,1% em relação a todos os 68.833 substantivos não hifenizados do MorphoBr; e (iii) a um aumento de 9,5% em relação a todas as 131.444 entradas lexicais não hifenizadas do MorphoBr. A partir dos valores (i), (ii) e (iii), concluímos que a PLP, a PLI e a PLG desse PFP são 709%, 18,1% e 9,5% respectivamente.

3.3.6 Verbialização de bases adjetivais pela sufixação em -izar

Seguindo a mesma lógica do PFP acima, o sufixo *-izar*, que também toma bases adjetivais, tem numerosas formações a partir de adjetivos complexos. Na tabela 8, vemos que seis sufixos servem de base para quase 40% dos 1.000 verbos terminados nesse sufixo. Limitar-nos-emos, também, a esses seis sufixos para modelarmos esse PFP.

Tabela 8 - Representatividade de adjetivos complexos como base do sufixo *-izar*

Sufixo da Base Adjetival	**Terminação Pós-Verbialização**	**Número de Ocorrências**	**Porcentagem de Ocorrências**
-al	*-alizar*	204	20,4%
-ar	*-arizar*	47	4,7%
-bil (*-vel*)	*-bilizar*	46	4,6%
-ico	*-icizar*	9	0,9%
-ivo	*-ivizar*	5	0,5%
-ano	*-anizar*	67	6,7%
Total		**378**	**37,8%**

Dentre os adjetivos do arquivo "adjectives.lemas", extraímos os 14.126 adjetivos terminados em *-al*, *-ar*, *-ável*, *-ico*, *-ivo* e *-ano*, que quantificamos na tabela 9. O comportamento dos cinco primeiros sufixos da lista com *-izar* e com *-idade* é idêntico. O sufixo *-ano* se comporta como *-ivo* e *-oso*, recebendo *-idade* após perder a desinência de gênero *-o*.

Tabela 9 - Mapeamento de expansão a partir da sufixação em *-izar* de adjetivos complexos

Sufixo da Base Adjetival	**Terminação Pós-Verbialização**	**Adjetivos-Base no MorphoBr**	**Adjetivos-Base Latentes**
-al	*-alizar*	204	1.808
-ar	*-arizar*	47	582
-abil (*-ável*)	*-abilizar*	46	2.748
-ico	*-icizar*	9	7.209
-ivo	*-ivizar*	5	983
-ano	*-anizar*	67	796
Total		**378**	**14.126**

Ao cruzar os 13.042 verbos em *-izar* de "adjIZAR.lemas" com os 15.193 verbos de "verbs.lemas", identificamos 12.824 novos verbos em *-izar*, o que equivale a um aumento de (i) 1.282% em relação aos 1.000 verbos formados por sufixação em *-izar*, (ii) a um aumento de 84,4% em relação aos 15.193 verbos não hifenizados (iii) e a um de 9,8% em relação a todas as 131.444 entradas lexicais não hifenizadas do MorphoBr. A partir dos valores (i), (ii) e (iii), concluímos que a PLP, a PLI e a PLG desse PFP são 1.282%, 84,4% e 9,8% respectivamente.

3.3.7 Adverbialização de bases adjetivais pela sufixação em -mente

Por fim, dentre os 4.180 advérbios não hifenizados do MorphoBr, identificamos 3.864 advérbios terminados em "mente". Seguindo a mesma estratégia adotada para os sufixos *-idade* e *-izar*, elencamos dez sufixos adjetivais como marcadores de adjetivos-base para a sufixação de *-mente*, que, no total, representam quase metade dos advérbios já presentes no MorphoBr e que detalhamos na tabela 10. Na tabela 11, quantificamos a contribuição que podemos oferecer expandindo o MorphoBr ao gerar esses advérbios adicionais.

Tabela 10 - Representatividade de adjetivos complexos como base do sufixo *-mente*

Sufixo da Base Adjetival	**Terminação Pós-Adverbialização**	**Número de Ocorrências**	**Porcentagem de Ocorrências**

-al	*-almente*	377	9,8%
-ar	*-armente*	55	1,4%
-bil (*-vel*)	*-bilmente*	5	0,1%
-ico	*-icamente*	530	13,7%
-ivo	*-ivamente*	254	6,6%
-ano	*-anamente*	33	0,9%
-dor	*-doramente*	55	1,4%
-ista	*-istamente*	75	1,9%
-ante	*-antemente*	150	3,9%
-oso	*-osamente*	285	7,4%
Total		**1.819**	**47,1%**

Tabela 11 - Mapeamento de expansão a partir da sufixação em *-mente* de adjetivos complexos

Sufixo da Base Adjetival	**Terminação Pós-Adverbialização**	**Adjetivos-Base no MorphoBr**	**Adjetivos-Base Latentes**
-al	*-almente*	377	1.808
-ar	*-armente*	55	582
-vel	*-velmente*	5	2.748
-ico	*-icamente*	530	7.209
-ivo	*-ivamente*	254	983
-ano	*-anamente*	33	796
-dor	*-doramente*	55	2.418
-ista	*-istamente*	75	887
-ante	*-antemente*	150	1.338
-oso	*-osamente*	285	1.202
Total		**1.819**	**19.971**

Podemos dividir o comportamento dos dez sufixos acima em cinco grupos de acordo com suas interações com o sufixo *-mente*. Os sufixos *-al*, *-ar*, *-ista* e *-ante* recebem *-mente* por simples adjunção. O sufixo *-vel*, também recebe *-mente* por simples adjunção, mas sofre perda do acento gráfico. O sufixo *-ico* perde a vogal temática antes da adjunção de *-mente* além de seu vocábulo perder o acento gráfico. Como visto em 3.3.5, essa perda de acento gráfico é mais complexa que a de *-ável* por haver oito casos diferentes. Os sufixos *-ivo*, *-ano* e *-oso*

perdem a vogal temática “o” antes de receberem *-mente*. Por fim, o sufixo *-dor* ganha a vogal temática “a” antes de receber *-mente*.

Após cruzar os 18.887 advérbios de “adjMENTE.lemas” com os advérbios de “adverbs.lemas”, encontramos 16.962 novos advérbios em *-mente*, o que equivale (i) a um aumento de 439% em relação aos 3.864 advérbios formados por sufixação em *-mente*, (ii) a um de 406% em relação aos 4.180 advérbios e (iii) a um de 12,9% em relação a todas as entradas não hifenizadas do MorphoBr. De (i), (ii) e (iii), concluímos que os valores de PLP, PLI e PLG relativos a esse PFP são 439%, 406% e 12,9%.

4 TRANSDUTORES

Construímos quatro transdutores, cada um para dar conta dos derivados de uma das quatro classes morfossintáticas a que pertencem os derivados produzidos. A arquitetura dos transdutores segue o molde da arquitetura dos transdutores do MorphoBr.

4.1 Transdutor adjetival

Cada um dos quatro é composto de quatro partes, como podemos ver pelos arquivos manipulados pelo xfst para produzir as flexões dos novos adjetivos: (29) o arquivo “build-fst.xfst”, que reúne os comandos para gerenciar os demais componentes; (30) o arquivo nAdj.lemas com os lemas que servem de base para o PFP e cujo conteúdo é ilustrado pela palavra “desnudável”; (31) o arquivo nAdj.lexc que modela a morfotática; e (32) o arquivo “regras.xfst”, que modela as regras de substituição, inserção e apagamento necessárias.

(29)
```
source regras.xfst
read lexc nAdj.lexc
compose
save stack nAdj.fst
```

(30)
```
desnudável
```

(31)
```
Multichar_Symbols +A +M +F +SG +PL

Lexicon Root
<@txt"nAdj.lemas">    Adjetivo;

Lexicon Adjetivo
+A:0        Gênero;

Lexicon Gênero
+M:0        Número;
+F:0        Número;
```

Lexicon Número
+SG:0 #;
+PL:^s #;

(32)
clear
define MB "^" ;
regex [l -> i || _ MB] ;
define DeleteMB [MB -> 0] ;
regex DeleteMB ;
turn stack
compose net

Assim como fizemos com o par “básico:basicidade” na seção 2.3, vejamos por que formas intermediárias a palavra “desnudável” passa. O comando *<@txt“nAdj.lemas”>* de dentro do arquivo “nAdj.lexc” lê o arquivo “nAdj.lemas” e carrega 12.489 adjetivos, dentre os quais está “desnudável”. A partir daí, o compilador será direcionado de bloco em bloco concatenando partes de palavra e etiquetas até concluir todos os pares de palavras. Essas etiquetas devem ser declaradas na primeira linha do arquivo, logo após o comando “Multichar_Symbols”[27], de forma que o compilador não as considere como caractere comuns.

A palavra “Adjetivo” à direita do comando *<@txt“nAdj.lemas”>* indica para o compilador que se deve continuar para o bloco nomeado “Adjetivo”, formando o par em (33). A mesma coisa acontece com o bloco “Gênero”, formando os dois pares em (34), e com o bloco final “Número”, formando os quatro pares em (35). O compilador só para de concatenar quando encontra o caractere “#”.

(33) desnudável+A:desnudável

(34) desnudável+A+M:desnudável
desnudável+A+F:desnudável

(35) desnudável+A+M+SG:desnudável
desnudável+A+F+SG:desnudável

[27] Abreviação da expressão de língua inglesa *multicharacter symbols*, que significa “símbolos multicarateres”.

desnudável+A+M+PL:desnudável^s
desnudável+A+F+PL:desnudável^s

A ordem de blocos que o compilador segue não é influenciada pela ordem em que eles aparecem no arquivo. O compilador, por padrão, procura o bloco de nome “Root” para iniciar o processo de concatenação. A partir daí, cada linha indicará ao compilador que bloco buscar em seguida.

Os pares de palavras em (34) representam o estado em que os dados do transdutor se encontram quando apenas o arquivo “lexc” é compilado. Nesse ponto, as formas masculina e feminina do singular estão prontas, mas as do plural, não. Ainda é necessário substituir o caractere “l” pelo caractere “i” e apagar a fronteira de morfema “^” nas formas masculina e feminina do plural. Disso se encarrega o arquivo “regras.xfst” em (32) através da regra de substituição, que produz os pares em (36), e da regra de apagamento, que produz os pares em (37).

(36) desnudável+A+M+SG:desnudável
desnudável+A+F+SG:desnudável
desnudável+A+M+PL:desnudávei^s
desnudável+A+F+PL:desnudávei^s

(37) desnudável+A+M+SG:desnudável
desnudável+A+F+SG:desnudável
desnudável+A+M+PL:desnudáveis
desnudável+A+F+PL:desnudáveis

Na seção 2.5 explicamos as funções e os comandos utilizados para construir nossos transdutores. Explicaremos agora cada linha do arquivo “regras.xfst” apresentado em (32). O comando *clear* serve para limpar a pilha. O comando *define MB “^” ;* serve para atribuir à variável MB o valor “^”. A regra de substituição tem a forma do comando *regex [l -> i || _ MB]*. Nesse comando, o caractere (ou a sequência de caracteres) à esquerda do operador “->” é substituído pelo caractere (ou sequência de caracteres) à direita do operador. O local exato da troca é marcado pelo caractere *underscore* “_”, e o contexto em que a substituição deve ocorrer é expresso pelo que está presente à esquerda e/ou à direita do *underscore*. No caso em

questão, a substituição ocorre antes no contexto de precedência à variável MB, ou seja, o caractere “^”.

A regra de apagamento tem a forma do comando *define DeleteMB [MB -> 0]*. Esse comando atribui à variável “DeleteMB” a regra de substituição do caractere “^” (conteúdo da variável MB) pelo caractere vazio, representado por “0”, essencialmente apagando o caractere “^”. Em seguida, o comando *regex DeleteMB* interpreta o conteúdo de DeleteMB como uma expressão regular, a traduz na forma de um autômato e a adiciona à pilha. O comando *turn stack* inverte a pilha de modo que os autômatos fiquem na ordem correta para a operação de composição, executada pelo comando seguinte, *compose net*.

4.2 Transdutor substantival

No caso dos novos substantivos, o procedimento é mais simples por não haver necessidade de nenhuma regra de substituição, inserção ou apagamento, logo não ser necessário um arquivo xfst, apenas o arquivo “build-fst.xfst” (38), o arquivo “nSubs.lemas”, ilustrado pela palavra “cremosidade” (39), e o arquivo “nSubs.lexc” (40). Também não foi necessário utilizar o marcador de fronteira de morfema.

(38)
```
read lexc nSubs.lexc
save stack nSubs.fst
```

(39)
```
cremosidade
```

(40)
```
Multichar_Symbols +N +F +SG +PL

Lexicon Root
<@txt"nSubs.lemas">  Substantivo;

Lexicon Substantivo
+N:0        Gênero;

Lexicon Gênero
+F:0        Número;
```

```
Lexicon Número
+SG:0       #;
+PL:^s      #;
```

Assim como fizemos em (33), (34) e (35), listamos as etapas pelas quais a palavra “cremosidade” passa na compilação do arquivo “lexc”. O par em (41) é resultado da aplicação do bloco “Substantivo”. O par em (42) é resultado da aplicação do bloco “Gênero”. Já os pares em (43) resultam da aplicação do bloco “Número”.

(41) cremosidade+N:cremosidade

(42) cremosidade+N+F:cremosidade

(43) cremosidade+N+F+SG:cremosidade
cremosidade+N+F+PL:cremosidades

4.3 Transdutor adverbial

No caso dos advérbios, o procedimento é mais simples ainda, pois não há flexão de gênero nem de número. O arquivo “build-fst.xfst” (44) é mantido, mas o arquivo xfst não é necessário, e o arquivo “nAdv.lexc” (45) só serve para adicionar a etiqueta morfossintática “+Adv”. Usamos a palavra “cremosamente” para ilustrar o arquivo “nAdv.lemas” (46).

(44)
```
read lexc nAdv.lexc
save stack nAdv.fst
```

(45) cremosamente

(46)
```
Multichar_Symbols +ADV

Lexicon Root
<@txt"nAdv.lemas">   Advérbio;

Lexicon Advérbio
```

```
+ADV:0    #;
```

4.4 Transdutor verbal

Já o caso dos verbos é diferente. A morfologia verbal do português é rica, e o transdutor de estados finitos deve dar conta das 64 formas verbais de cada um dos 12.824 verbos propostos. Em (47), temos o arquivo “build-fst.xfst”. Em (48), temos o verbo “textualizar”, que usamos aqui para ilustrar o conteúdo do arquivo “nVerbs.lemas”. Em (49), temos o arquivo “nVerb.lexc”. O compilador ignora tudo o que é escrito depois do caractere “!”. Esse recurso serve para escrita de comentários no corpo do código. Em (50), temos o arquivos “regras.xfst”.

(47)
```
source regras.xfst
read lexc nVerb.lexc
compose net
save stack nVerb.fst
```

(48)
```
textualizar
```

(49)
```
Multichar_Symbols +V +PRS +PRF +IMPF +PQP +FUT +COND +SUBJR
+SUBJP +SUBJF +IMP +INF +GRD +PTPASS +SG +PL +1 +2 +3

Lexicon Root
<@txt"nVerb.lemas">  Verbo;

Lexicon Verbo
+V:0 Flexão;

Lexicon Flexão
! indicativo
+PRS:0     PresInd;     ! presente
+PRF:0     PrfInd;      ! pretérito perfeito
+IMPF:0    ImpfInd;     ! pretérito imperfeito
+PQP:0     PqpInd;      ! pretérito mais-que-perfeito
```

```
+FUT:0      FutInd;      ! future do presente
+COND:0  CondInd;    ! condicional / futuro do pretérito

! subjuntivo
+SUBJR:0  PresSubj;
+SUBJP:0  ImpfSubj;
+SUBJF:0  FutSubj;

! imperativo
+IMP:0      Imp;

! infinitivo
+INF:^ar    #;

! gerúndio
+GRD:^ando      #;

! particípios
+PTPASS:^ad     Gênero;

Lexicon PresInd
+1+SG:^o          #;
+2+SG:^as         #;
+3+SG:^a          #;
+1+PL:^amos     #;
+2+PL:^ais        #;
+3+PL:^am         #;

Lexicon PrfInd
+1+SG:^ei         #;
+2+SG:^aste       #;
+3+SG:^ou         #;
+1+PL:^amos     #;
+2+PL:^astes      #;
```

```
+3+PL:^aram      #;

Lexicon ImpfInd
+1+SG:^ava       #;
+2+SG:^avas      #;
+3+SG:^ava       #;
+1+PL:^ávamos    #;
+2+PL:^áveis     #;
+3+PL:^avam      #;

Lexicon PqpInd
+1+SG:^ara       #;
+2+SG:^aras      #;
+3+SG:^ara       #;
+1+PL:^áramos    #;
+2+PL:^áreis     #;
+3+PL:^aram      #;

Lexicon FutInd
+1+SG:^arei      #;
+2+SG:^arás      #;
+3+SG:^ará       #;
+1+PL:^aremos    #;
+2+PL:^areis     #;
+3+PL:^arão      #;

Lexicon CondInd
+1+SG:^aria      #;
+2+SG:^arias     #;
+3+SG:^aria      #;
+1+PL:^aríamos   #;
+2+PL:^aríeis    #;
+3+PL:^ariam     #;
```

```
Lexicon PresSubj
+1+SG:^e        #;
+2+SG:^es       #;
+3+SG:^e        #;
+1+PL:^emos     #;
+2+PL:^eis      #;
+3+PL:^em       #;

Lexicon ImpfSubj
+1+SG:^asse     #;
+2+SG:^asses    #;
+3+SG:^asse     #;
+1+PL:^ássemos  #;
+2+PL:^asseis   #;
+3+PL:^assem    #;

Lexicon FutSubj
+1+SG:^ar       #;
+2+SG:^ares     #;
+3+SG:^ar       #;
+1+PL:^armos    #;
+2+PL:^ardes    #;
+3+PL:^arem     #;

Lexicon Imp
+2+SG:^a        #;
+3+SG:^e        #;
+1+PL:^emos     #;
+2+PL:^ais      #;
+3+PL:^em       #;

Lexicon Gênero
+M:^o       Número;
+F:^a       Número;
```

```
Lexicon Número
+SG:0       #;
+PL:^s      #;
```

(50)
```
define MB "^" ;
regex [ r -> 0 || a _ MB ];
regex [ a -> 0 || z _ MB ];
define DeleteMB [ MB -> 0 ];
regex DeleteMB;
turn stack
compose net
```

Assim como fizemos de (33) a (35) e de (41) a (43), mostramos abaixo como evoluem os pares de palavras ao longo do processo de concatenação executado na compilação do arquivo “nVerb.lexc”. Em (51), temos o par resultante da aplicação do bloco “Verbo”. Em (52), temos os treze pares resultantes da aplicação do bloco “Flexão”, o que já ilustra a explosão de formas verbais. Com exceção dos pares “textualizar+V+INF:textualizar^ar” e “textualizar+V+GRD:textualizar^ndo”, aos quais não se concatena mais nenhuma sequência de caracteres, todos os outros pares direcionam o compilador para outros blocos.

(51)
```
textualizar+V:textualizar
```

(52)
```
textualizar+V+PRS:textualizar
textualizar+V+PRF:textualizar
textualizar+V+IMPF:textualizar
textualizar+V+PQP:textualizar
textualizar+V+FUT:textualizar
textualizar+V+COND:textualizar
textualizar+V+SUBJR:textualizar
textualizar+V+SUBJP:textualizar
textualizar+V+SUBJF:textualizar
textualizar+V+IMP:textualizar
textualizar+V+INF:textualizar^ar
```

textualizar+V+GRD:textualizar^ando
textualizar+V+PTPASS:textualizar^d

Por questão de simplicidade e economia, elencamos em (53) apenas três dos treze pares de (52) para ilustrar o resto do processo de concatenação.

(53) textualizar+V+PRS:textualizar
⋮
textualizar+V+SUBJR:textualizar
⋮
textualizar+V+PTPASS:textualizar^d

Em (54), (55), (56) e (57), respectivamente, temos os pares resultantes da aplicação dos blocos “PresInd”, “PresSubj”, “Gênero” e “Número”.

(54) textualizar+V+PRS+1+SG:textualizar^o
textualizar+V+PRS+2+SG:textualizar^as
textualizar+V+PRS+3+SG:textualizar^a
textualizar+V+PRS+1+PL:textualizar^amos
textualizar+V+PRS+2+PL:textualizar^ais
textualizar+V+PRS+3+PL:textualizar^am

(55) textualizar+V+SUBJR+1+SG:textualizar^e
textualizar+V+SUBJR+2+SG:textualizar^es
textualizar+V+SUBJR+3+SG:textualizar^e
textualizar+V+SUBJR+1+PL:textualizar^emos
textualizar+V+SUBJR+2+PL:textualizar^eis
textualizar+V+SUBJR+3+PL:textualizar^em

(56) textualizar+V+PTPASS+M:textualizar^d^o
textualizar+V+PTPASS+F:textualizar^d^a

(57) textualizar+V+PTPASS+M+SG:textualizar^d^o
textualizar+V+PTPASS+F+SG:textualizar^d^a

textualizar+V+PTPASS+M+PL:textualizar^d^o^s
textualizar+V+PTPASS+F+PL:textualizar^d^a^s

Os 16 pares listados em (54), (55) e (57) ilustram a estrutura dos dados no fim da compilação do arquivo “nVerbs.lexc”. As duas regras de apagamento formalizadas no arquivo “regras.xfst” dão conta de apagar a vogal temática “a” junto da desinência de infinitivo “r” e de apagar o marcador de fronteira de morfema “^”, produzindo a lista de pares em (58).

(58) textualizar+V+PRS+1+SG:textualizo
textualizar+V+PRS+2+SG:textualizas
textualizar+V+PRS+3+SG:textualiza
textualizar+V+PRS+1+PL:textualizamos
textualizar+V+PRS+2+PL:textualizais
textualizar+V+PRS+3+PL:textualizam
⋮
textualizar+V+SUBJR+1+SG:textualize
textualizar+V+SUBJR+2+SG:textualizes
textualizar+V+SUBJR+3+SG:textualize
textualizar+V+SUBJR+1+PL:textualizemos
textualizar+V+SUBJR+2+PL:textualizeis
textualizar+V+SUBJR+3+PL:textualizem
⋮
textualizar+V+PTPASS+M+SG:textualizado
textualizar+V+PTPASS+F+SG:textualizada
textualizar+V+PTPASS+M+PL:textualizados
textualizar+V+PTPASS+F+PL:textualizadas

4.5 Transdutor resultante

Na tabela 12, vemos as propriedades dos quatro transdutores:

Tabela 12 - Propriedades dos transdutores

Transdutor	Tamanho	Estados	Arcos	Caminhos
nAdj.fst	254,2 kB	6.076	16.202	45.564
nAdv.fst	418,9 kB	13.019	26.762	16.962
nSubs.fst	325,2 kB	9.850	20.767	24.978
nVerb.fst	335,2 kB	10.155	21.380	833.560

Por meio da execução do arquivo “build-fst.xfst” em (59), os quatro arquivos binários salvos ao fim da execução de cada arquivo “build-fst.xfst” foram carregados na pilha por meio do comando *load* e serviram de argumento para a operação de união, gerando assim um único transdutor com os 11.391 novos adjetivos, 16.962 novos advérbios, 12.489 novos substantivos e os 12.824 novos verbos.

(59) load nAdj.fst
load nAdv.fst
load nSubs.fst
load nVerb.fst
union net
save stack mbr-exp.fst

De posse do binário “mbr.exp.fst”, geramos uma lista com as palavras da língua inferior, que nomeamos “lower.txt”, como em (60), e que foi analisada pelo autômato através da ferramenta lookup com um comando em (61), gerando os pares de palavras como em (62). Esse comando, aplicado ao terminal do Linux, lê o conteúdo do arquivo “lower.txt” linha por linha e consulta no dicionário em formato de autômato finito a forma flexionada. Todas as consultas são salvas no mesmo arquivo “mbr.exp.dict”. Esse procedimento foi aplicado aos transdutores individuais de forma a criar dicionários eletrônicos separados por classe morfossintática, que nomeamos “Adj.dict”, “Adv.dict”, “Subs.dict” e “Verbs.dict”.

(60) lufáveis
(61) cat lower.txt | lookup mbr.exp.fst > mbr.exp.dict
(62) lufável • lufável+A+M+SG

5 CONSIDERAÇÕES FINAIS

Neste trabalho, expandimos o recurso lexical mais abrangente do português, o MorphoBr, através da modelagem computacional de quatro PFPs por sufixação. Primeiramente, apresentamos os PFPs que nos propusemos modelar, depois apresentamos a estrutura do MorphoBr. Em seguida, apresentamos a morfologia de estados finitos, formato que estruturou nosso produto, e o compilador xfst, ferramenta com a qual construímos nossos transdutores.

Dos dados do MorphoBr, extraímos os lemas não hifenizados. Esse recorte foi necessário, por dois motivos. Primeiramente, foi necessário extrair os lemas, pois os primitivos e derivados dos processos de formação se manifestam como lemas e não como formas flexionadas. Em segundo lugar, optamos por não tratar os lemas hifenizados por uma questão de simplificação do nosso objeto de estudo. Trabalhos futuros poderão tratar desses casos.

Para podermos mensurar a expansão do recurso lexical, trabalhamos com o conceito de produtividade morfológica. Definimos subconceitos para descrever de modo mais preciso tanto um recurso lexical em um determinado estado, quanto a expansão de um recurso lexical através de sua retroalimentação por um PFP.

A partir dos vários tipos de produtividade definidos, pudemos analisar o estado atual do MorphoBr em termos de primitivos e derivados dos PFPs em questão e quantificar uma estimativa para nossas contribuições. A partir dos resultados compilados na tabela 4, podemos concluir que as hipóteses levantadas foram confirmadas ou levemente superadas. A produtividade latente global dos PFPs relativos aos sufixos *-vel*, *-idade* e *-izar*, respectivamente 8,5%, 9,5% e 9,8%, se encontram dentro da faixa de 7,5% a 12,5% prevista na hipótese original. A produtividade latente global do PFP relativo ao sufixo *-mente* excedeu o intervalo ficando em 12,9%.

A análise da concorrência entre sufixos *-ção* e *-mento* serviu como uma primeira etapa para a modelagem de novas entradas lexicais em trabalhos futuros. Nesses trabalhos futuros podem investigadas que tipos de seleção de bases fazem cada um dos dois sufixos de modo a modelar os dois PFPs com precisão.

Todos os resultados encontrados neste estudo, subléxicos analisados e transdutores construídos foram documentados e armazenados em diretórios do repositório do projeto na

plataforma GitHub[28] e distribuídos sob a licença GNU General Public License Version 3[29]. Há espaço ainda para muitos refinamentos na modelagem dos PFPs abordados aqui. Outros PFPs, tanto por sufixação, quanto por prefixação, serão modelados também seguindo a metodologia aqui aplicada. Há também a possibilidade de se comparar as acurácias do MorphoBr e do MorphoBr expandido contra um mesmo *corpus*, o que não foi feito neste trabalho.

[28] Disponível em https://github.com/heliolbs/MorphoBrExpansion.
[29] Cópia disponível em https://github.com/heliolbs/MorphoBrExpansion/blob/master/gpl-3.0.txt

REFERÊNCIAS

ALENCAR, L. F. de. Produtividade morfológica e tecnologia do texto: aspectos da construção de um transdutor lexical do português capaz de analisar neologismos. **Calidoscópio (UNISINOS)**, v. 7, p. 199-220, 2009.

ALENCAR, L. F. de et al. JMorpher: a Finite-State Morphological Parser in Java for Android. In: BAPTISTA, J. et al. (Eds.). **Computational Processing of the Portuguese Language**. 11th International Conference, PROPOR 2014. São Carlos/SP, Brazil, October 6-8, 2014. Proceedings. Series Lecture Notes in Computer Science / Subseries Lecture Notes in Artificial Intelligence. Berlin; Heidelberg: Springer, 2014.

ALENCAR, L. F. de; RADEMAKER, A.; CUCONATO, B. **MorphoBr – resources for morphological analysis of Portuguese**. 2018. Disponível em: <https://github.com/LFG-PTBR/MorphoBr>. Acesso em: 11 nov. 2018.

ALENCAR, L. F.; CUCONATO, B; RADEMAKER, A. MorphoBr: an open source large-coverage full-form lexicon for morphological analysis of Portuguese. **Texto Livre: Linguagem e Tecnologia**, [S.l.], v. 11, n. 3, p. 1-25, dez. 2018. ISSN 1983-3652. Disponível em: <http://www.periodicos.letras.ufmg.br/index.php/textolivre/article/view/14294>. Acesso em: 26 dezembro 2019. doi: http://dx.doi.org/10.17851/1983-3652.11.3.1-25.

ALVES, I. M. **Neologismo**. Criação Lexical. 2. ed. São Paulo: Ática, 2004.

ANTUNES, C. C. **Um estudo das regras de uso do hífen, segundo o acordo ortográfico de 1990**. 2014. Dissertação. Mestrado em Língua Portuguesa, Pontifícia Universidade Católica de São Paulo, São Paulo, 2014.

BASILIO, M. **Estruturas Lexicais do Português**. Petrópolis: Vozes, 1980.

_____. **Teoria Lexical**. São Paulo: Ática, 1987.

_____. Produtividade e função do processo de formação de palavras do português. **Anais do Congresso Internacional da Associação de Lingüísticas e Filologia da América Latina**. Campinas: ALFAL, 9 (1): 1-9, 1990.

_____. Morfológica e Castilhamente: um Estudo das Construções X-mente no Português do Brasil. **DELTA**, São Paulo, v. 14, n. spe, p. 17-28, 1998. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S010244501998000300003&lng=en_&nrm=iso>. Acesso em: 26 jul. 2017. doi: http://dx.doi.org/10.1590/S0102-44501998000300003.

_____. **Formação e classes de palavras no português do Brasil**. 3. ed. São Paulo: Contexto, 2011.

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

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

CAVALCANTE, F. T. A produtividade lexical do sufixo -mente na língua portuguesa. **Revista de Letras**, v. 18., n. 2, jul/dez, 1996.

DIAS-DA-SILVA, B. C.; MORAES, H. R. Construção de um thesaurus eletrônico para o português do Brasil. **Processamento Computacional do Português Escrito e Falado (PROPOR)**, v. 4, p. 1-10, 2000.

DURAN, M. S. A importância dos recursos lexicais para o processamento automático do português. **Estudos Linguísticos**, São Paulo, v. 42, n. 2, p. 866-877, 2013. Disponível em: <http://www.gel.org.br/estudoslinguisticos/volumes/42/el42_v2_maio-ago_t23.pdf>. Acesso em 8 nov. 2018.

ELEUTÉRIO, S. et al. A system of electronic dictionaries of portuguese. **Lingvisticae Investigationes**, v. 19, n. 1, p. 57-82, 1995. Disponível em: <http://label.ist.utl.pt/publications/docs/Eleuterio_et_al_95.pdf>. Acesso em: 18 set. 2018.

GARCIA, M.; GAMALLO, P. Análise morfossintáctica para português europeu e galego: Problemas, soluções e avaliação. **Linguamática**, Braga, v. 2, n. 2, p. 59-67, 2010. Disponível em: <http://linguamatica.com/index.php/linguamatica/article/view/56>. Acesso em: 11 abr. 2019.

GREGHI, J. G. **Uma base de dados lexicais para o português do brasil**. 2002. Dissertação – Instituto de Ciências Matemáticas de São Carlos, Universidade de São Paulo, São Carlos 2002.

HULDEN, M. Foma: a finite-state compiler and library. In: **Conference of the European Chapter of the Association for Computational Linguistics**, 12, 2009, Athens. *Proceedings*... [S.l.]: Association for Computational Linguistics, 2009. p. 29-32. Disponivel em: <http://www.aclweb.org/anthology/E09-2008>. Acesso em: 20 jan. 2019.

KAPLAN, R. M.; BRESNAN, J. **The Mental Representation of Grammatical Relations** ed. Joan Bresnan: Cambridge, 1982.

MARONEZE, B. O. **Um estudo da nominalização no Português do Brasil com base em unidades lexicais neológicas**. 2005. Dissertação (Mestrado em Filologia e Língua Portuguesa). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2005.

_____. **Um estudo da mudança de classe gramatical em unidades lexicais neológicas**. 2011. 198 f. Tese (Doutorado em Letras) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2011.

MARTINS, R. T. et al. Linguistic issues in the development of ReGra: a Grammar Checker for Brazilian Portuguese. **Natural Language Engineering**, Volume 4 (Part 4 December 1998): p287-307; Cambridge University Press.

MUNIZ, M. C. M. **A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB**. 2004. Dissertação. Instituto de Ciências Matemáticas de São Carlos, Universidade de São Paulo, São Carlos, 2004.

NUNES, M. G. V.; OLIVEIRA Jr., O. N. O processo de desenvolvimento do Revisor Gramatical ReGra. **Anais do XXVII SEMISH** (XX Congresso Nacional da Sociedade Brasileira de Computação), Volume 1, p.6 (resumo). Artigo Completo na Versão em CD-Rom. PUC-PR, Curitiba, julho 2000.

OLIVEIRA JÚNIOR, O. N. et al. **O uso de interlíngua para comunicação via Internet: O Projeto UNL/Brasil**. Série de Relatórios do NILC. NILC-TR-01-3, julho 2001, 14p.

PADRÓ, L.; STANILOVSKY, E. FreeLing 3.0: Towards Wider Multilinguality. In: **Proceedings of the Language Resources and Evaluation Conference**, 8., 2012, Istanbul, Anais… Istambul: ELRA, 2012. Disponível em: <http://nlp.lsi.upc.edu/publications/papers/padro12.pdf >. Acesso em: 22 nov. 2018.

ROCHA, L. C. A. **Estruturas Morfológicas do Português**. São Paulo: WMF Martins Fontes, 2008.

SANTOS, A. F.; TORRES, C. E. A.; SILVA, H. L. B. Sobre a construção de um recurso léxico de elementos nominais agentivos e de ação para o processamento computacional do Português Brasileiro. In: **Domínios de Linguagem**. [S.l.], v. 9, n. 3, p. 137-155, jul/set 2015. Disponível em: < http://www.seer.ufu.br/index.php/dominiosdelinguagem/article/view/28963/16975>. Acesso em: 26 jan. 2016. doi: https://doi.org/10.14393/DL19-v9n3a2015-8.

VILLALVA, A.; SILVESTRE, J. P. **Introdução ao estudo do léxico: descrição e análise do Português**. Petrópolis, RJ: Vozes, 2014, 247 p.