EMANUEL SOUZA DE QUADROS

Competição morfológica e ilhas de confiabilidade na morfologia derivacional

Porto Alegre
2015

CIP - Catalogação na Publicação

Quadros, Emanuel Souza de
Competição morfológica e ilhas de confiabilidade na morfologia derivacional / Emanuel Souza de Quadros. -- 2015.
116 f.

Orientador: Luiz Carlos da Silva Schwindt.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Instituto de Letras, Programa de Pós-Graduação em Letras, Porto Alegre, BR-RS, 2015.

1. morfologia derivacional. 2. produtividade. 3. bloqueio. 4. competição morfológica. I. Schwindt, Luiz Carlos da Silva, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS
ÁREA DE CONCENTRAÇÃO: ESTUDOS DA LINGUAGEM
ESPECIALIDADE: TEORIA E ANÁLISE LINGUÍSTICA
LINHA DE PESQUISA: FONOLOGIA E MORFOLOGIA

Competição morfológica e ilhas de confiabilidade na morfologia derivacional

EMANUEL SOUZA DE QUADROS

Orientador: PROF. DR. LUIZ CARLOS SCHWINDT

Dissertação de Mestrado em Teoria e Análise Linguística, apresentada como requisito parcial para a obtenção do título de Mestre pelo Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul.

Porto Alegre

2015

Agradecimentos

Ao término deste trabalho, há muito que agradecer.

Agradeço ao professor Luiz Carlos Schwindt pela confiança, pela paciência e pela orientação ao longo de todos esses anos – que agora seriam dez, não fossem as descontinuidades da vida.

Agradeço aos demais professores das linhas de pesquisa Fonologia e Morfologia e Gramática, Semântica e Léxico (na nomenclatura da minha época) por todos os ensinamentos a que espero fazer alguma justiça. Em especial, a Gisela Collischonn, Marcos Goldnadel e Sergio Menuzzi, que, de muitas formas, ajudaram a construir meu pensamento sobre a Linguística.

Agradeço também aos amigos que me acompanharam em momentos diversos da trajetória acadêmica. Sobretudo a César Augusto González, Paulo Henrique Pappen e Tiago Martins (a galera do *Language Bar*); a Guilherme Duarte Garcia, pelas conversas decisivas; e a Tamara Melo, pela grande amizade e pelo apoio contínuo.

Sobre apoio, sou muito grato a minha família, por ter garantido as condições básicas para que eu perseguisse minhas aspirações. Sou grato também a Verônica Borsato, pelo suporte emocional, que é uma daquelas partes essenciais, ainda que invisíveis, de qualquer trabalho acadêmico; e pela curiosidade intelectual praticamente sem limites, que a fez ler e ouvir minhas ideias muitas vezes.

Por fim, agradeço ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo financiamento do meu desenvolvimento científico durante boa parte da graduação e do mestrado.

Resumo

No domínio da morfologia derivacional, é difícil encontrar padrões de formação de palavras que possam se aplicar a todas as bases que se encaixam em seus contextos de aplicação. Isso equivale a dizer que a produtividade de padrões derivacionais costuma ser limitada. Entre as causas dessa limitação, vemos que formações potenciais são frequentemente bloqueadas por itens lexicais já existentes; em outros casos, elas são suplantadas por expressões formadas por padrões derivacionais concorrentes. Este trabalho dedica-se a explorar tais situações de competição. Iniciamos pelo exame da ideia de produtividade e de como entender as diferenças quantitativas entre padrões rivais quanto a este aspecto. Fazemos, em seguida, uma discussão mais detida da competição morfológica e da noção central de bloqueio, contrapondo às teorias gramaticais de base lexical uma abordagem pragmática deste fenômeno. Por fim, apresentamos o modelo desenvolvido em Albright e Hayes (1999) e em trabalhos posteriores, que explora a ideia de que o grau de confiabilidade do emprego de padrões morfológicos em diferentes contextos fonológicos é um fator determinante da produtividade desses padrões, bem como da competição entre eles. Testamos este modelo utilizando dados dos sufixos *-ção* e *-mento*, que se encontram em competição há bastante tempo no português. Estes dados provêm do Dicionário Houaiss 3.0 e de um levantamento de textos de jornais e blogs, coletados com o auxílio de programas computacionais desenvolvivdos para este trabalho. Nossos resultados sugerem que a manutenção da produtividade de *-mento* ao longo da história, mesmo após *-ção* ter se tornado o padrão dominante de nominalização, foi escorada pela existência de contextos fonológicos em que *-mento* atinge um alto grau de confiabilidade. Dada a produtividade da primeira conjugação, foram particularmente importantes os contextos de aplicação de *-mento* encontrados entre palavras desta classe verbal. Com base nestas generalizações, mostramos como um modelo estatístico é capaz de prever, na maior parte dos casos, a escolha entre estes dois afixos diante de uma nova base verbal.

Palavras-chave: morfologia derivacional; produtividade; bloqueio; competição morfológica.

Abstract

In the field of derivational morphology, it is hard to find word formation patterns that may be applied to every base satisfying its context of application. This means that the productivity of derivational patterns is often limited. Among the causes of this limitation, we find that potential words are blocked by existing lexical items in many cases; in other cases, they are preempted by expressions formed by rival derivational patterns. This work devotes itself to exploring these instances of competition. We start by exploring the concept of productivity and by investigating how to understand quantitative differences between rival patterns in this respect. We then proceeed to a more detailed discussion of morphological competition and the fundamental notion of blocking, comparing a pragmatic approach to this phenomenon with lexicalist grammatical theories. Finally, we present the model of Albright e Hayes (1999) and later works, which explores the idea that the reliability of morphological patterns in different phonological contexts is a key determinant of the productivity of these patterns and the competition between them. We test this model on data formed by the suffixes *-ção* and *-mento*, which have been in competition for a long time in Portuguese. These data come from Dicionário Houaiss 3.0 and from a corpus created from newspapers and blogs with the help of software developed for this research. Our results suggest that the continued productivity of *-mento* throughout history, even after *-ção* had become the dominant nominalization pattern in the language, was supported by the existence of phonological contexts in which *-mento* reaches a high degree of reliability. Given the productivity of the first conjugation, contexts of application of *-mento* in words of this verbal class have shown to be especially important. We show that a statistical model equipped with these generalizations is able to predict the choice between these affixes in most cases.

Keywords: derivational morphology; productivity; blocking; morphological competition.

Sumário

Lista de tabelas

1 Introdução

Neste trabalho, apresentamos uma discussão sobre produtividade morfológica, na esteira de trabalhos anteriores que desenvolvemos a respeito deste tema (Quadros, 2009, 2011). Vista de modo geral, a produtividade é uma característica fundamental da linguagem humana que permite o emprego de um número limitado de elementos e mecanismos de combinação, para construir um número potencialmente infinito de expressões linguísticas. Na prática, essa potencialidade é limitada de diversas formas no domínio da morfologia, a começar pelo fato de ela operar sobre um número finito de itens lexicais, seja caracterizando suas formas em função do contexto morfossintático (morfologia flexional), seja dando origem a novos itens lexicais em função de necessidades comunicativas (morfologia derivacional).

Um dos fatores limitadores da produtividade de padrões morfológicos específicos é a existência (e o sucesso relativo) de padrões concorrentes que se apliquem sobre o mesmo conjunto de bases. Visto que palavras costumam ser memorizadas, elas também permanecem, com frequência, disponíveis para reutilização, o que diminui a necessidade de formação constante de novos itens lexicais. Assim, padrões morfológicos distintos, mas que se aplicam em um mesmo domínio lexical, acabam competindo por oportunidades limitadas de aplicação (Lindsay e Aronoff, 2013). Uma vez que uma nova palavra é gerada dentro de um desses padrões e se estabelece na língua, com uma determinada função, a probabilidade de que outra seja formada para suprir essa mesma função é severamente diminuída, um fenômeno conhecido como bloqueio (Aronoff, 1976).

Esta dissertação toma, então, como objeto a competição morfológica, vista como um fator determinante da produtividade de padrões lexicais. Partimos da premissa de que a escolha dos falantes entre padrões morfológicos rivais, na criação de uma palavra nova, não é aleatória; em vez disso, obedece a certas tendências, que podem se modificar ao longo da história. Em situações de criação, são concretizadas as intuições que falantes têm sobre a forma de palavras novas, incluindo a possibilidade de rejeitar construções agramaticais, mas também a de expressar preferências entre formas que seriam, em princípio, gramaticais. Sendo assim, é razoável supor que usuários de uma língua aprendem regularidades sobre a distribuição de padrões rivais, mesmo quando elas não são categóricas, à medida que internalizam o sistema morfológico desta língua.

Partindo desta conjectura, testamos neste trabalho o desempenho de um modelo de aprendizagem de regras morfofonológicas estocásticas, proposto por Albright e

Hayes (1999) e desenvolvido em trabalhos subsequentes. Este modelo explora a ideia de que, ao lado de regras gerais, como as tradicionalmente postuladas pela linguística gerativa, usuários de uma língua também aprendem generalizações mais específicas, com o mesmo *output* de uma regra geral, mas que podem ser mais robustas do que esta, por se referirem a contextos em que o mapeamento descrito pela regra é significativamente mais previsível – isto é, tem um número menor de exceções. Estas "ilhas de confiabilidade" seriam, assim, contextos em que usos produtivos das generalizações descritas pelas regras seriam mais esperados. A predição deste modelo é de que casos de competição morfológica devem ser sensíveis, em sua resolução, aos graus de confiabilidade que cada um dos padrões rivais demonstra nos contextos específicos em que a formação de palavras novas é necessária.

No Capítulo 2, apresentamos uma breve discussão da definição de produtividade morfológica, em seus aspectos qualitativos e quantitativos. Discutimos brevemente a medição da produtividade morfológica, buscando tornar mais precisas as afirmações do tipo "o padrão X é mais produtivo que o Y"; isso é exemplificado com uma comparação dos sufixos *-ção* e *-mento* em termos de seus níveis de produtividade. Em seguida, damos atenção à relação entre essa noção e a de gramática. Segue-se a isso o Capítulo 3, que desenvolve o tema da competição morfológica, em que figura, de forma central, o fenômeno do bloqueio, mencionado acima. Exploramos, neste capítulo, uma abordagem pragmática desse fenômeno, estendendo-a a situações de competição entre padrões morfológicos produtivos.

No Capítulo 4, detalhamos o modo como se deu a coleta de dados que se fez necessária tanto para a discussão anterior, sobre a produtividade desses afixos, quanto para o teste do modelo de aprendizagem de Albright e Hayes (1999), empreendido no Capítulo 5. Esse teste consiste em uma aplicação do modelo a dados do português, motivada por uma questão que surge quando olhamos para história da competição entre sufixos nominalizadores nesta língua: como *-mento* pôde se manter produtivo nos últimos séculos estando em competição com um rival consideravelmente mais frequente? Exploramos, neste ponto do trabalho, a hipótese de que a estabilidade de *-mento* tenha sido garantida pela existência de "ilhas de confiabilidade" como as que são depreendidas pelo aprendiz de Albright e Hayes (1999). Em seguida, apresentamos nossas considerações finais e algumas sugestões para estudos futuros.

2 Produtividade morfológica

2.1 O que é produtividade?

Produtividade, no domínio da morfologia, é entendida como o potencial de palavras novas serem formadas a partir de padrões morfológicos ativos em uma língua. Essa potencialidade é ambígua em pelo menos dois sentidos, que expressam duas dimensões importantes do que se entende por produtividade morfológica. Esse "potencial" pode ser entendido de forma qualitativa, como mera possibilidade, como a característica de algo que pode vir a ser dentro de uma língua. Por exemplo, em português, a afixação de *-eza* a adjetivos não derivados é possível (*caro* → *careza*), mas a mesma afixação não é possível no caso de adjetivos derivados (*comprável* → **comprabileza*, **compraveleza*).

"Potencial" também pode ser entendido de forma quantitativa. Neste caso, o que está em jogo é o quanto um determinado padrão é efetivamente utilizado na formação de palavras novas em uma língua. Assim, embora *-mento* e *-ção* estejam ambos disponíveis para a formação de substantivos a partir de verbos em português, pode ser o caso que um deles forme mais palavras do que o outro em um dado estágio da língua. Do ponto de vista qualitativo, ambos têm potencial para formar palavras novas. Porém, mesmo dentro de um mesmo domínio definido por restrições gramaticais, eles podem ter probabilidades distintas de participar de novas formações lexicais. É nesse sentido que Mark Aronoff afirma que "embora muitas coisas sejam possíveis na morfologia, algumas são mais possíveis do que outras" (Aronoff, 1976, p. 35).[1]

Essa distinção caracteriza duas preocupações distintas, mas complementares, nos estudos de morfologia. No entendimento qualitativo de produtividade, temos uma noção classificatória que separa padrões morfológicos entre aqueles que estão disponíveis para formar palavras novas e os que não estão. No caso dos que estão disponíveis, também é possível especificar, com maior ou menor detalhamento, seus domínios de atuação. Assim, como vimos, *-eza* está disponível para a formação de novas palavras, mas não atua em qualquer domínio, nem mesmo sobre qualquer

[1] Essa ambiguidade no entendimento de "produtividade" fez com que Corbin (1987) sugerisse deixar de lado essa noção "confusa e polissêmica" (p. 177); em vez dela, a autora sugere que se usem as noções de "regularidade" (*regularité*), "disponibilidade" (*disponibilité*) e "rentabilidade" (*rentabilité*), as duas últimas correspondendo, respectivamente, aos aspectos qualitativo e quantitativo da produtividade que discutimos acima. Nos valemos desses termos ao longo deste trabalho e da valiosa clarificação conceitual da autora, ainda que continuemos considerando útil falar em "produtividade" como uma noção complexa que envolve todas essas dimensões.

tipo de base adjetiva, mas especificamente apenas sobre adjetivos não derivados. Em contraste, *-idade* não tem essa restrição e possui um domínio de aplicação mais amplo; assim, *comprabilidade* é aceitável. Essa classificação, com as especificações mais finas decorrentes da análise dos contextos de disponibilidade de um padrão morfológico – "o escopo da regra" (Kastovsky, 1986) –, é o foco tradicional das teorias de gramática, já que nelas é preciso caracterizar o que faz parte de uma gramática e o que não faz. Assim, para algumas teorias, a afixação de *-eza*, por ser disponível no português brasileiro, corresponde a uma regra de formação de palavras distinta, definida por meio de um certo número de condições, ao passo que a presença de *-ebre* em *casebre* mereceria outro tratamento, não envolvendo uma regra desse tipo, por não se tratar de um afixo disponível para a formação de palavras no português.

Por outro lado, do ponto de vista quantitativo, há, em princípio, um número ilimitado de distinções entre padrões morfológicos quanto a suas "taxas de aplicação" (Kastovsky, 1986). Assim, se mesmo após uma caracterização exaustiva dos domínios de aplicação de *-ção* e *-mento*, observa-se que um desses afixos forma mais palavras do que o outro dentro de um mesmo domínio, temos uma diferença apenas do ponto de vista quantitativo. Essa diferença não é necessariamente importante para uma descrição gramatical preocupada apenas em definir as possibilidades qualitativas da língua, e não o modo como elas são postas em uso em situações reais. Por isso, diferenças quantitativas entre usos de padrões morfológicos não costumam receber mais do que menções breves em estudos voltados para a caracterização da competência linguística em uma língua.

Embora seja importante ter essas distinções em mente, o estudo da produtividade morfológica exige que ambas essas dimensões sejam seriamente consideradas. Um estudo quantitativo depende crucialmente de uma boa caracterização qualitativa e pode, nos melhores casos, informar as análises qualitativas, levando-as a refinamentos (Lüdeling e Evert, 2003). Os estudos qualitativos, por sua vez, são pouco informativos nos casos em que há diferenças consideráveis de uso que não sejam completamente explicadas por restrições gramaticais, ou quando essas restrições não se aplicam de forma categórica. Isso pode ser mero resultado de nossa falta de conhecimento sobre todos os fatores envolvidos em uma situação de fala; isto é, pode ser o caso que, dada uma descrição completa de todos esses fatores, poderíamos prever como cada falante preencheria cada lacuna lexical diante de uma necessidade comunicativa – por exemplo, qual seria a forma resultante em cada uma das escolhas entre *-ção* e *-mento*. Alternativamente, pode ser que uma tal descrição completa não seja suficiente para alcançar essa pretensão, pois pode ser o caso que diferenças probabilísticas entre padrões morfológicos sejam irredutíveis e

tenham valor funcional: por exemplo, utilizar um padrão menos provável pode servir para sinalizar pragmaticamente uma interpretação não canônica. Seja como for, só poderemos chegar a algum resultado em qualquer dessas direções após uma boa quantidade de trabalho quantitativo.

Possibilidade de uso, de um lado, e uso efetivo, de outro, são noções conceitualmente distintas, mas empiricamente interligadas. Para que um padrão morfológico possa ter usos concretos, é certamente necessário que ele esteja disponível na gramática da língua – exceto no caso de usos criativos da linguagem. Por outro lado, para que um padrão morfológico possa ser generalizado, isto é, para que esteja disponível para a formação de novas palavras, é necessário que já haja usos efetivos desse padrão que possam sustentar essa generalização – ou, pelo menos, um conjunto razoável de palavras que possam ser interpretadas como formadas por ele. Ademais, o uso frequente e não intencional de um padrão é o tipo de evidência mais clara de que ele está disponível na língua. Na ausência de observações desse tipo, a disponibilidade de um padrão poderia ser investigada apenas por métodos indiretos, como pela análise de intuições linguísticas. No entanto, na ausência de observações que evidenciem algum grau de uso sistemático de um padrão, seria difícil imaginar como um falante o generalizaria e lhe atribuiria um lugar na gramática durante a aquisição de um sistema morfológico.

Com essas observações, temos condições de colocar uma questão importante para os estudos sobre produtividade morfológica. Para que um padrão morfológico possa ser utilizado, ele precisa estar disponível; para continuar sendo disponível, ele precisa já ter sido utilizado. Então, como começa e como termina esse ciclo? Em outras palavras, quando um padrão passa a poder formar palavras dentro de uma língua e como ele pode perder esse potencial? Pode ser o caso que a presença de um desses aspectos da produtividade não seja suficiente para garantir a existência do outro. Isto é, além de estar disponível, um padrão morfológico pode necessitar de outras condições para se mostrar rentável na produção de palavras novas. Ou pode ser que o uso passado de um padrão morfológico não seja suficiente para que os falantes de um novo estágio da língua o interpretem como um padrão disponível para a formação de novas palavras. Uma parte importante dos estudos sobre produtividade morfológica deve ser, portanto, investigar essas condições adicionais.

Dessa discussão, resta claro que tanto noções qualitativas quanto noções quantitativas são importantes para caracterizar o fenômeno geral que nos interessa, e que cruza as fronteiras tradicionais entre competência e performance. Neste trabalho, portanto, o termo “produtivo” é reservado para padrões morfológicos que estejam disponíveis na língua e que sejam efetivamente utilizados na formação de palavras

novas, havendo a possibilidade de descrevermos graus de produtividade, de acordo com a rentabilidade relativa de cada padrão. Na seção seguinte, avaliamos a possibilidade de se quantificar de forma precisa essa noção de grau de produtividade, que se mostrou elusiva na história da morfologia. Nesse entendimento, deve ser possível responder, com "sim" ou "não", a pergunta "o padrão morfológico *m* é produtivo na língua L?" e responder, com alguma especificação gramatical, a pergunta "em que domínios da língua L o padrão morfológico *m* é produtivo?"; deve ser possível, ainda, oferecer uma resposta quantitativa à pergunta "em que medida o padrão morfológico *m* é produtivo na língua L?"

2.2 A análise quantitativa da produtividade morfológica

Vimos que muitos fatores estão envolvidos na configuração da produtividade de um padrão morfológico. Assim, uma dada escolha para o preenchimento de uma lacuna lexical é determinada por causas gramaticais e extragramaticais, e o agregado dessas escolhas, que caracteriza uma língua, é largamente inacessível a nossa introspecção, dada a sua complexidade. Para investigar de forma precisa essas questões, precisamos de um modelo quantitativo que nos permita formalizar nosso entendimento de produtividade e que possa dar suporte a análises qualitativas, isto é, ele deve ser linguisticamente significativo, e não apenas sumarizar estatísticas sobre a distribuição das palavras em um *corpus*.

Baayen (1992, p. 110) estabelece quatro exigências que qualquer medida deve satisfazer para que seja linguisticamente significativa no campo da produtividade morfológica:

1. "A medida deve fornecer um ranqueamento dos processos de formação de palavras que esteja em correspondência geral com um ranqueamento baseado em intuições linguísticas."

2. "A medida deve expressar 'a prontidão estatisticamente determinável com que um elemento entra em novas combinações.' Bolinger (1948)."

3. A medida deve ser sensível à existência de formas idiossincráticas no escopo de um padrão morfológico. Nas palavras do autor, "considerar as formações que são caracterizadas por propriedades idiossincráticas do ponto de vista formal ou semântico deve ter o efeito de diminuir o valor da medida de produtividade."

4. "A medida deve iluminar o fato empírico de que a produtividade não pode ser medida simplesmente em termos de frequência de tipos."

Vejamos, então, como o modelo apresentado em Baayen (1992) satisfaz essas condições. Este modelo parte da hipótese de que há uma correlação entre a frequência de *tokens* de um padrão morfológico e sua produtividade.[2] Se isso é verdade, a distribuição das frequências de palavras formadas por padrões produtivos deve ser significativamente diferente da de palavras formadas por padrões não produtivos.

Podemos entender uma distribuição de frequência, neste contexto, como um arranjo do número de ocorrências de cada uma das palavras em um *corpus*. Segue, por exemplo, a distribuição de frequência dos dez itens mais frequentes sufixados por *-mento* em um *corpus* composto por textos de jornais e de blogs, descrito com mais detalhes na Seção 4.1.

f_1	julgamento	4255
f_2	pagamento	3424
f_3	investimento	3316
f_4	atendimento	3151
f_5	desenvolvimento	2861
f_6	tratamento	2441
f_7	crescimento	2258
f_8	equipamento	2222
f_9	treinamento	1695
f_{10}	procedimento	1607

Tabela 1 – Dez primeiros itens da distribuição de frequência dos substantivos formados por *-mento* no corpus deste trabalho.

Essa distribuição pode ser organizada em grupos de frequência designados por um valor de r, de modo que, no grupo $r = 1$, entram todas as palavras que ocorrem apenas uma vez no *corpus* – chamadas de *hapax legomena*; no grupo $r = 2$, entram todas as que ocorrem apenas duas vezes; e assim por diante. No caso de *-mento*, considerando os *tokens* que instanciam esse afixo em nosso corpus, temos 77 itens no grupo $r = 1$, o mais numeroso; 74 no grupo $r = 2$; e quantidades cada vez menores de itens nos grupos seguintes, até o grupo $r = 4255$, que contém apenas uma palavra (*julgamento*). No Gráfico 1, abaixo, temos o número de palavras contidas nos primeiros 20 grupos de frequência de *-mento*.

[2] A frequência de *tokens* de um padrão morfológico é dada pela soma do número de atestações de cada um dos itens linguísticos por ele formados.

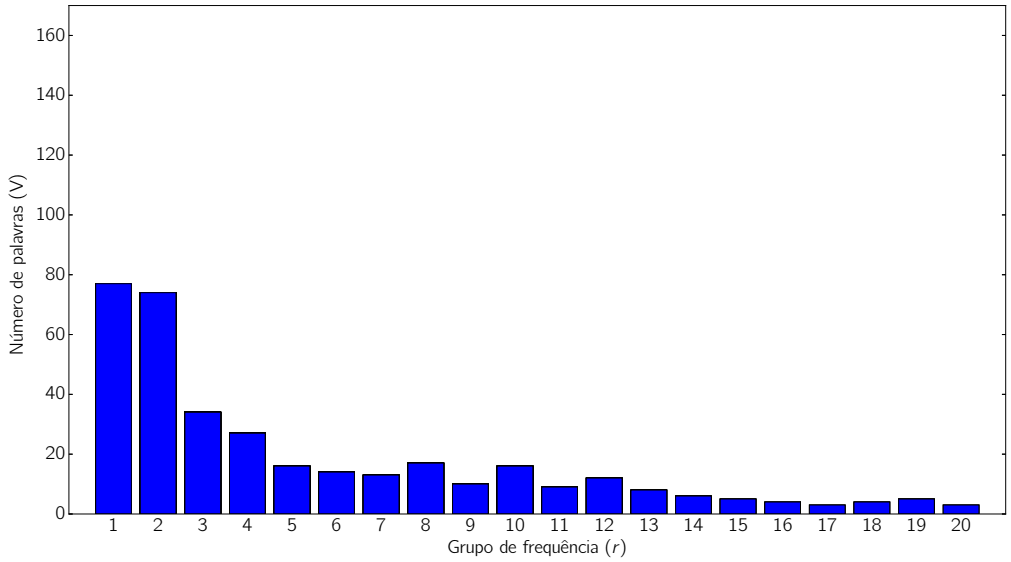

Gráfico 1: Distribuição de frequência agrupada dos substantivos deverbais em *-mento* ($N = 86653$).

Este gráfico apresenta uma tendência à esquerda, com um número significativo de palavras de baixa frequência.[3] Em uma amostra de tamanho suficiente, essa tendência reflete o enriquecimento do vocabulário, possivelmente por meio da formação de palavras novas. Sendo assim, amostras de padrões pouco (ou nada) produtivos não devem apresentar esse tipo de distribuição de forma tão acentuada.

Como exemplo disso, temos o padrão V-*ura*, que também caracteriza substantivos derivados de verbos, como *abertura* e *rachadura*. Este exemplo é interessante como contraste, pois, ao contrário de *-mento*, *-ura* parece ser apenas marginalmente produtivo no português atual. Assim, a distribuição de frequência agrupada deste afixo, no Gráfico 2, não apresenta uma tendência à esquerda. É interessante observar que todos os grupos de frequência visíveis neste gráfico apresentam menos palavras que os grupos apresentados no Gráfico 1. Isto não é consequência apenas de haver menos *tokens* em *-ura* dentro do corpus considerado. Antes, isto se deve ao fato de que a maior parte das palavras sufixadas por *-ura* se concentra em grupos de frequência caracterizados por valores de r maiores que 20, ou seja, são palavras de frequência relativamente alta, como podemos ver na Tabela 2.

[3] Para facilitar a visualização, o gráfico representa apenas grupos de frequência $r \leq 20$.

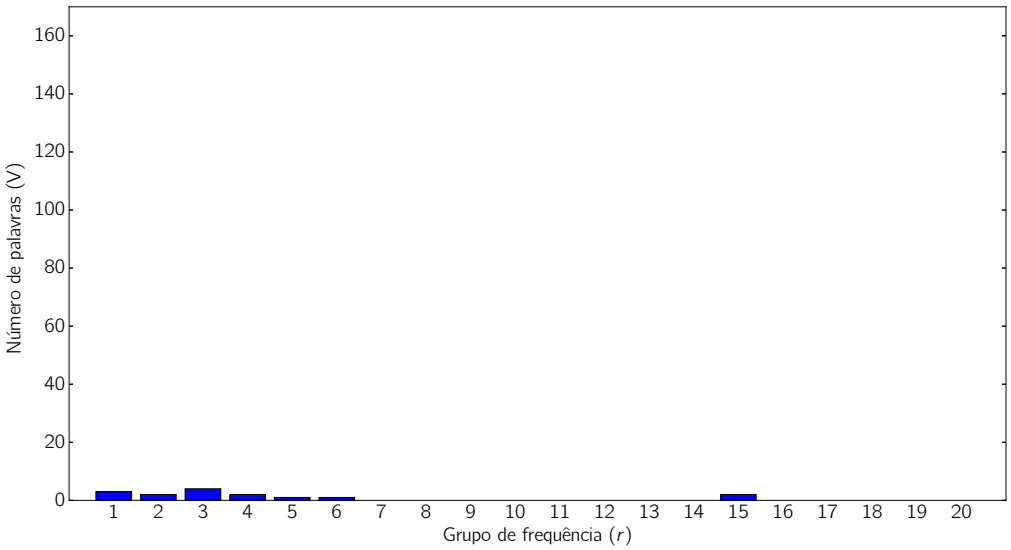

Gráfico 2: Distribuição de frequência agrupada dos substantivos deverbais em *-ura* ($N = 9205$).

r	Nº de palavras	*r*	Nº de palavras	*r*	Nº de palavras
1	3	48	1	249	1
2	2	53	1	340	1
3	4	62	1	828	1
4	2	71	1	1141	1
5	1	74	1	1481	1
6	1	117	1	1618	1
15	2	148	1	2317	1
23	1	156	1	-	-
26	1	171	1	-	-
34	1	180	1	-	-

Tabela 2 – Distribuição de frequência agrupada dos substantivos deverbais formados por *-ura* (N = 9205). Valores de *r* indicam grupos de frequência.

Considerando a definição de grupo de frequência (*r*) exposta anteriormente, esta tabela deve ser lida da seguinte forma: no grupo *r* = 1, temos três palavras que ocorrem apenas uma vez na amostra (*atadura*, *curvatura*, *gastura*); no grupo *r* = 2, temos duas palavras que ocorre duas vezes (*ligadura*, *abotoadura*); e assim por diante, até chegarmos ao grupo *r* = 2317, que contém uma palavra de frequência

relativamente alta (*abertura*), que ocorre 2317 vezes no corpus.

Para percebermos como se dá essa diferença de distribuição entre padrões produtivos, como X-*mento*, e padrões não produtivos, como X-*ura*, podemos imaginar uma leitura sequencial do *corpus*, em que vamos anotando cada palavra com o seu número de ocorrências. Inicialmente, todas as palavras que encontramos têm um número de ocorrências igual a 1. Porém, todas elas tendem a se repetir à medida que consideramos uma porção cada vez maior do *corpus*. Assim, elas passam a ocupar grupos de frequência caracterizados por valores de r cada vez mais altos. No caso de uma classe de palavras fechada (por exemplo, a dos pronomes pessoais do português), espera-se que todos os itens que a compõem se repitam várias vezes dentro de um *corpus* de tamanho razoável. Assim, o número de itens no grupo de frequência $r = 1$ deve rapidamente chegar a zero.

Os padrões morfológicos não produtivos caracterizam-se justamente por descreverem conjuntos fechados, isto é, que não podem ser atualizados com novos membros, a não ser por meio de recursos não morfológicos de enriquecimento lexical (empréstimos, criações lexicais intencionais, etc.). Por sua vez, padrões morfológicos produtivos definem conjuntos de palavras abertos. Nestes, espera-se que, ao lado da repetição de itens já estabelecidos, surjam novos itens, que, justamente por serem novos, tendem a ser pouco frequentes, pelo menos até que se tornem estabelecidos em uma comunidade linguística. Portanto, a expectativa é de que padrões morfológicos não produtivos sejam representados por muitos itens de alta frequência e por poucos itens de baixa frequência e de que padrões morfológicos produtivos sejam representados por uma proporção consideravelmente maior de itens de baixa frequência.

Além disso, devem-se encontrar distribuições de frequência diferentes mesmo entre os padrões morfológicos produtivos, sempre que eles se diferenciarem significativamente em seus níveis de produtividade. Isto é esperado porque quanto maior for a produtividade de um deles, maior deve ser a probabilidade de que seja encontrado um novo item (um *hapax legomenon*) instanciando este padrão na leitura sequencial do corpus. É o que podemos observar na comparação entre a distribuição de frequência agrupada dos nomes deverbais formados por *-ção*, apresentada no Gráfico 3, e a dos nomes deverbais formados por *-mento*, apresentada anteriormente no Gráfico 1.

Essas diferenças observadas entre as distribuições de frequência de padrões morfológicos – em particular, a observação de que classes morfológicas produtivas se caracterizam por terem muitos itens de baixa frequência –, estão na base das medidas de produtividade propostas por Baayen (1992). Detalhamos, abaixo, como

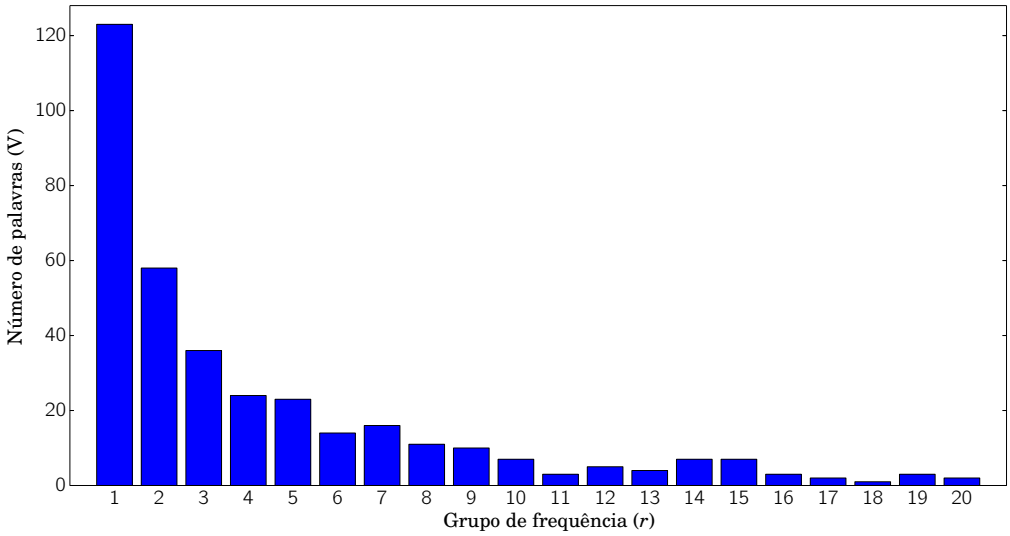

Gráfico 3: Distribuição de frequência agrupada dos substantivos deverbais formados por *-ção* (N = 291171).

derivar duas delas: o alcance de um padrão morfológico (que mede sua produtividade passada, observada até um dado momento) e um índice de produtividade no sentido estrito (que mede a velocidade de crescimento do vocabulário em um dado ponto e, por extensão, o quanto podemos esperar que ele cresça daí em diante).

Na leitura sequencial de um *corpus*, podemos observar o tamanho do vocabulário de um dado padrão morfológico em cada momento de amostragem. A cada vez que encontramos uma palavra instanciando esse padrão, o seu número total de *tokens* (N) aumenta. Esta palavra pode já ter aparecido antes ou não. Caso já tenha aparecido, não há crescimento de vocabulário, apenas reutilização de um item preexistente. Caso seja sua primeira ocorrência, temos crescimento de vocabulário, com uma palavra nova (em relação ao *corpus* considerado). O número de palavras únicas (ou tipos) para um tamanho determinado de amostra pode ser denotado por V ou, de forma mais explícita, $V(N)$, tendo em vista que o valor de V é sempre relativo ao tamanho da amostra. O crescimento do vocabulário pode ser visto graficamente ao representarmos o valor de $V(N)$ para cada momento da amostragem, como vemos no Gráfico 4.

Este gráfico mostra que o crescimento do vocabulário é bastante acentuado no caso dos padrões que consideramos produtivos, *-ção* e *-mento*; mais no caso do primeiro que no do segundo. Por outro lado, no caso do padrão em *-ura*, que é, no máximo, marginalmente produtivo em português, o crescimento do vocabulário é acentuado apenas no início da coleta, quando as palavras presentes no *corpus* ainda

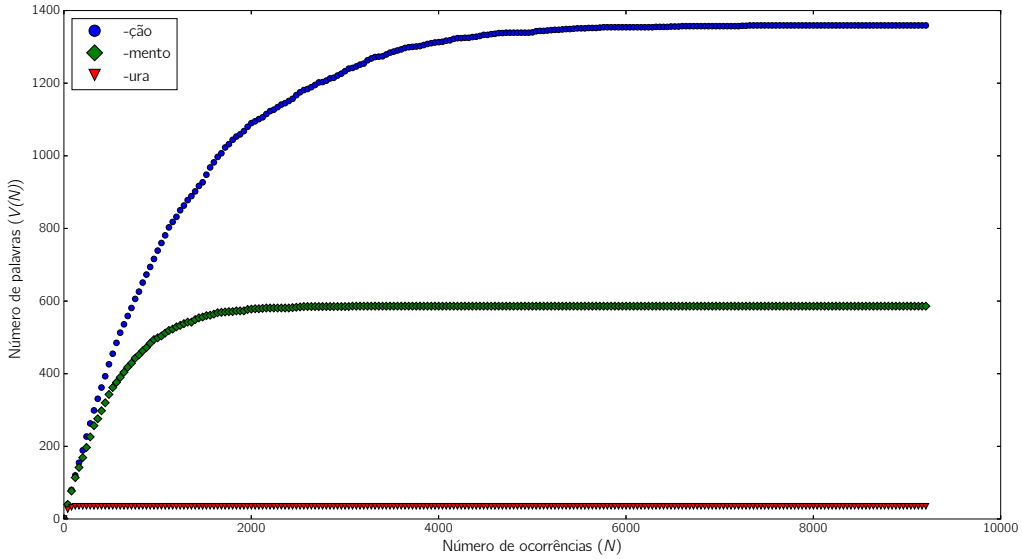

Gráfico 4: Crescimento do vocabulário dos sufixos nominalizadores *-ção*, *-mento* e *-ura* no corpus deste trabalho.

não tiveram muitas chances de se repetir e são, portanto, quase todas únicas.

A função *V*(*N*) é uma medida do alcance de um padrão morfológico, isto é, de quão generalizado é seu uso, em termos do número de palavras formadas por ele dentro de um *corpus* de certa extensão. Como tal, ela ainda não nos diz tudo o que gostaríamos de saber sobre produtividade. O problema é que não temos, até agora, uma boa caracterização da curva de crescimento de vocabulário. Podemos verificar o alcance de um padrão morfológico em qualquer ponto da amostragem (e traçar uma curva a partir dessas medições), mas nos resta saber se, a partir de qualquer um desses pontos, poderíamos prever se o alcance desse padrão aumentará ou não quando considerarmos uma porção maior do *corpus*.

Uma caracterização da curva de crescimento pode ser obtida a partir da proporção de *hapax legomena* no conjunto de palavras encontradas em uma amostra. Por exemplo, das primeiras 9.205 ocorrências de palavras com o sufixo *-ção* em nosso *corpus*, 162 são de *hapax legomena*; ou seja, aproximadamente 1,76% das palavras com *-ção* ocorrem uma única vez nesta amostra. No caso do sufixo *-mento*, com o mesmo número de ocorrências, temos 77 *hapax legomena*, o que representa cerca de 0,84% da amostra. No caso de *-ura*, por fim, em que o número total de *tokens* encontrados é 9.205, temos apenas 3 palavras que ocorrem apenas uma vez, correspondendo a cerca de 0,03% da amostra. Como já vimos anteriormente, os *hapax legomena* são indício de enriquecimento lexical, pois palavras que ocorrem uma única vez são itens recém adicionados ao vocabulário e que ainda não tiveram tempo de

se repetir, ou são genuinamente esporádicos. Calculando a proporção desses itens, obtemos, então, a taxa de crescimento do vocabulário em um determinado ponto da amostragem.

De forma geral, temos

$$\mathcal{P} = \frac{n_1}{N} \tag{2.1}$$

em que $\mathcal{P}$ denota a taxa de crescimento do vocabulário, n_1 denota o número de palavras que foram encontrados apenas uma vez no *corpus*, os *hapax legomena*, e N denota o número total de ocorrências (*tokens*) do padrão morfológico em questão. O valor de $\mathcal{P}$ é equivalente à inclinação da curva $V(N)$ no ponto N. A derivação detalhada dessa equação pode ser encontrada em Baayen (1992) e Baayen (2002).

É importante perceber que o índice $\mathcal{P}$ não é apenas uma estatística descritiva dos dados encontrados em um corpus. Pelo menos por hipótese, ele tem valor inferencial, expressando a probabilidade de que novos itens serão adicionados ao vocabulário caso a amostra seja ampliada. Assim, para Baayen (1992, p. 115), $\mathcal{P}$ "é a formalização quantitativa da noção linguística de produtividade morfológica", na medida em que expressa o grau de expectativa de que novas formações de um padrão morfológico serão encontradas ao observarmos porções maiores de texto.

No gráfico que segue, apresentamos a evolução das taxas de crescimento de vocabulário de *-ção*, *-mento* e *-ura* em nosso *corpus*, considerando incrementalmente as amostras de cada sufixo, até N = 9205. Observa-se que, para qualquer *N*, *-ção* apresenta uma taxa de crescimento vocabular mais alta que *-mento*, que, por sua vez, tem um crescimento vocabular bastante mais acentuado que *-ura*. Esses resultados estão de acordo com observações anteriores sobre os graus de produtividade dos sufixos nominalizadores do português – p. ex., Basilio (1996), Grodt (2009).

É importante observar que a taxa de crescimento de vocabulário $\mathcal{P}$ é altamente dependente do tamanho da amostra, já que ela é uma função de *N* (conforme equação na página 19). Assim, como vimos, após observarmos 9.205 ocorrências de palavras com *-ção*, temos $\mathcal{P}$ = 0,0176 para este afixo. Contudo, ao considerarmos a totalidade das ocorrências com este sufixo (*N* = 292.171), obtemos $\mathcal{P}$ = 0,0005. De modo geral, quanto maior for o tamanho da amostra, menor tende a ser o valor de $\mathcal{P}$ de um mesmo padrão morfológico. No caso de padrões não produtivos, esse valor rapidamente tende a zero; no caso de padrões produtivos, o valor também tende a zero, mas de forma bem mais lenta. Esse resultado é esperado e reflete o fato de que o crescimento vocabular responde a necessidades comunicativas, de modo que, quanto maior for o número de palavras disponíveis, menor é a necessidade da inclusão de palavras novas em um vocabulário. Isso faz, todavia, com que $\mathcal{P}$ não possa

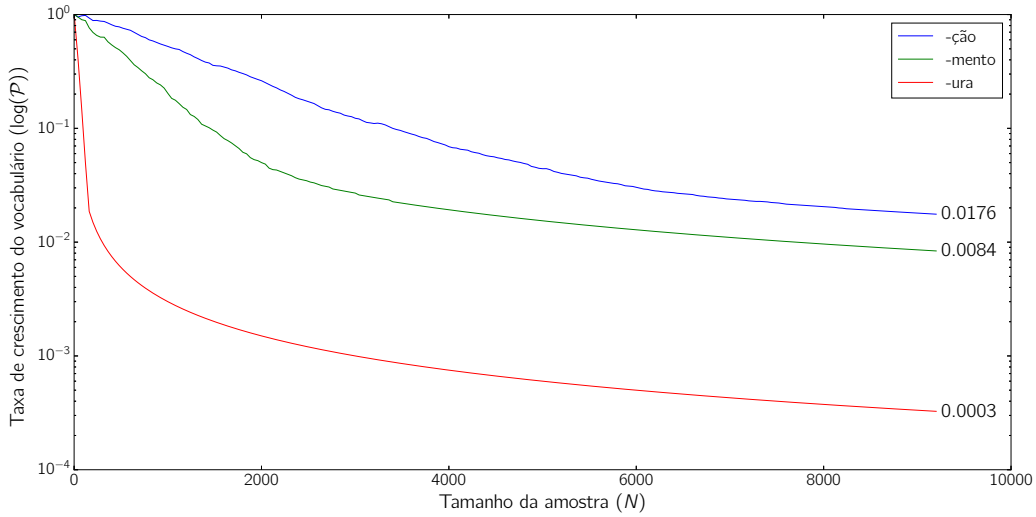

Gráfico 5: Evolução da taxa de crescimento do vocabulário dos sufixos nominalizadores *-ção*, *-mento* e *-ura*.

ser utilizado como um índice de produtividade absoluto. Em outras palavras, não é possível ranquear padrões morfológicos por esse índice sem referência a um *corpus* específico, e também não é possível ranqueá-los, mesmo dentro de um mesmo *corpus*, sem atentar para o fato de que o número total de ocorrências de palavras amostradas (N) pode ser bastante diferente para cada um desses padrões.[4]

Esse é um erro fácil de se cometer. Por exemplo, em Quadros (2011), afirmei que *-mento* apresentava uma taxa de crescimento ligeiramente maior do que *-ção* em um *corpus* composto por textos de blogs (0,0197 contra 0,0171, respectivamente). Entretanto, deve-se perceber que palavras em *-ção*, independentemente de questões de produtividade, ocorrem em maior número do que formas em *-mento*, o que significa que as amostras de cada um desses afixos não costumam ter o mesmo tamanho dentro de um mesmo corpus. No caso do *corpus* utilizado em Quadros (2011), a amostra de palavras com *-ção* continha 10.655 *tokens*, ao passo que a de palavras com *-mento* continha 5.472 *tokens*, não muito mais do que a metade do tamanho da amostra do afixo concorrente. Dada a dependência que o valor de $\mathcal{P}$ apresenta em relação ao número total de *tokens*, o resultado obtido deixa de ser surpreendente. Seria surpreendente se ambos os afixos tivessem o mesmo valor de N e, ainda assim, *-mento* tivesse apresentado uma taxa de crescimento maior do que *-ção*. Para fins de comparação, se recalcularmos o valor de $\mathcal{P}$ com apenas 5.472 das ocorrências de palavras com *-ção* no *corpus* de Quadros (2011), obtemos 0,0334 para este

[4] É por esse motivo que consideramos apenas 9.205 ocorrências para cada afixo no Gráfico 2.2, já que este é o número total de *tokens* contendo o sufixo *-ura*.

afixo (contra 0,0197 para *-mento*). Esse novo resultado mostra-se muito mais em concordância com nossas intuições a respeito das produtividades relativas de *-ção* e *-mento* e com os estudos anteriores sobre a expressão morfológica da nominalização no português.

O mesmo tipo de equívoco está na base das críticas de Van Marle (1992) ao modelo quantitativo de Baayen. Especificamente, Van Marle (1992) mostra como o ranqueamento de um conjunto de afixos do holandês, de acordo com os valores de $\mathcal{P}$ calculados por Baayen (1992), não corresponde a um ranqueamento obtido por meio de intuições linguísticas. No entanto, o tamanho da amostra (*N*) para cada um desses sufixos não é levado em conta pelo autor. A lição é a seguinte: $\mathcal{P}$ não nos oferece um índice por meio do qual se possam comparar padrões morfológicos de forma absoluta, independentemente de seus números totais de ocorrências. Uma comparação entre padrões morfológicos nesse modelo só pode ser feita de forma significativa, se forem consideradas amostras de mesmo tamanho para cada padrão, ou, ainda, se ela for feita com base na evolução da curva de crescimento de vocabulário, não em valores isolados de $\mathcal{P}$.

Agora temos condições de avaliar esse modelo quantitativo frente às exigências impostas por Baayen (1992) a qualquer medida de produtividade (ver página 12). Vimos que a taxa de crescimento de vocabulário ($\mathcal{P}$) expressa a prontidão com que um elemento entra em novas combinações na língua. Em cada momento da amostragem, esse valor nos oferece uma estimativa razoável da expectativa que podemos ter de que novas palavras com um determinado padrão morfológico sejam adicionadas ao vocabulário. Vimos também que, dado um número constante de N para um conjunto de padrões morfológicos sendo comparados, podemos oferecer um ranqueamento entre eles que, pelo menos no caso dos nominalizadores *-ção*, *-mento* e *-ura*, no português brasileiro, se mostra de acordo com nossas intuições.

A terceira exigência é de que a existência de palavras idiossincráticas tenha um efeito negativo no índice de produtividade de um padrão morfológico. A razão dessa expectativa é o fato de que formas idiossincráticas precisam ser memorizadas, já que, por definição, suas idiossincrasias não podem ser previstas pela caracterização geral do padrão morfológico. Sendo assim, o uso de uma dessas formas não conta como o emprego produtivo de um recurso morfológico, mas como a reutilização de um elemento armazenado na memória. O índice $\mathcal{P}$ é sensível à existência de formas idiossincráticas, pois, como já apontado por Aronoff (1976), elas costumam ser elementos de média ou alta frequência. Assim, sua presença aumenta o valor de N sem afetar o número de *hapax legomena*, o que resulta em diminuição na proporção de itens de baixa frequência e, consequentemente, no valor do índice $\mathcal{P}$.

Quanto à quarta exigência, vimos que o cálculo da taxa de crescimento de vocabulário não depende do número de tipos que instanciam um padrão morfológico. Para esta medida, interessa apenas o número de palavras de um padrão que ocorrem apenas uma vez em uma amostra. Em alguns casos, uma classe morfológica que contém um número bastante numeroso de palavras pode não ser mais produtiva. É o que vimos na Seção 2.3, sobre o sufixo *-ment*, do inglês. Embora esse afixo possua um vasto número de palavras já estabelecidas na língua, ele alcança índices bastante baixos de $\mathcal{P}$, em comparação com afixos produtivos do inglês Baayen (1992, p. 124).

2.3 Produtividade é uma questão gramatical?

Na medida em que falamos da disponibilidade de um padrão morfológico em termos qualitativos, não deve haver controvérsia sobre se produtividade, nesse sentido, é uma questão gramatical. Desde que se reconheça a realidade de algum tipo de gramática no domínio da morfologia, deve-se reconhecer a necessidade de que ela defina quais padrões morfológicos podem ser utilizados para criar palavras novas. A polêmica surge quando se entende produtividade de forma mais ampla, como fizemos na seção anterior, considerando que ela também apresenta uma dimensão quantitativa.

Assim, para alguns linguistas, a produtividade não é uma questão gramatical; ela seria, em vez disso, um resultado do modo como falantes reais fazem uso da gramática. Nesta posição, sustentada, por exemplo, por Di Sciullo e Williams (1987), a gramática é entendida como uma caracterização formal de quais são os objetos linguísticos possíveis dentro de uma língua. No caso da gramática da morfologia de uma língua, isso significa uma caracterização recursiva das palavras possíveis, o que envolve, entre outras coisas, o conjunto de restrições formais que determinam que bases podem figurar em cada padrão morfológico e quais são as propriedades fonológicas e morfossintáticas das palavras resultantes do uso de cada um desses padrões. Dito de outro modo, a gramática estabelece quais combinações são gramaticais, caso em que devem corresponder a palavras potenciais, e quais combinações são agramaticais, caso em que devem corresponder a palavras impossíveis. A gramática não teria, nessa perspectiva, nenhum compromisso de caracterizar o conjunto de palavras atestadas, já que ele é, em grande parte, determinado por circunstâncias acidentais do ponto de vista gramatical.

Bauer (2001) vê problemas nesse entendimento da produtividade. Para ele, a produtividade dos padrões morfológicos é uma parte importante da caracterização

da gramática de uma língua, e o uso que os falantes fazem desses padrões evidencia uma propriedade gramatical distinta. Vejamos dois exemplos trazidos pelo autor para esclarecer essa posição, e que podem servir de base para confrontarmos essas duas perspectivas. Trata-se da evolução do plural dos nomes terminados em *-al* do francês e da história do sufixo *-ment* no inglês. Para cada um desses casos, Bauer rejeita fatores não gramaticais que poderiam estar causalmente relacionados às mudanças envolvidas.

O caso do francês corresponde à perda de produtividade, por volta de 1600, do plural em *-aux*, que era característico dos nomes terminados em *-al*. Como resultado dessa perda, os nomes terminados em *-al* que entraram na língua a partir de então passaram a ter seu plural formado pelo padrão mais geral, com o sufixo *-s*. Assim, no francês atual, há dois conjuntos de nomes terminados em *-al*: aqueles que preservam o plural em *-aux*, como *cheval* 'cavalo' (pl. *chevaux* 'cavalos'); e aqueles com plural em *-s*, como *festival* 'festival' (pl. *festivals*).

Bauer (2001) sugere que, nesse caso, houve uma mudança na gramática do francês que não foi causada por nenhuma questão de frequência, já que, entre os nomes terminados em *-al*, o conjunto dos que se pluralizam em *-aux* ainda é mais numeroso que o dos que se pluralizam em *-als*. A sugestão de Bauer (2001) parece ser a de que, se houvesse algum efeito relevante de frequência na evolução do plural desse conjunto de palavras, ela deveria ser no sentido de aumentar a produtividade de *-aux*, não o contrário, já que esta é a terminação mais frequente dentro desse conjunto fonologicamente restrito de palavras.

> [...] esta mudança não pode ter tido ligação direta com o número de possíveis modelos, já que mesmo no século XVII havia mais nomes de uso comum que utilizavam o (então não produtivo) *-aux* do que os que utilizavam o (recém produtivo) *-als*. A mudança de produtividade parece ser independente da mudança de frequência. (Bauer, 2001, p. 8)[5]

O autor parece correto em considerar que houve uma mudança na gramática do francês: em um estágio da língua, havia um padrão morfológico produtivo; em outro estágio, ao que tudo indica, esse padrão não estava mais disponível para a criação de novas formas flexionais, e o seu espaço de atuação foi tomado por outro afixo. Esse é um tipo de mudança que teria de ser descrito gramaticalmente, mesmo em uma perspectiva como a de Di Sciullo e Williams (1987). O que não fica claro, porém, é o modo como Bauer (2001) conceptualiza o surgimento dessa mudança, já que parece

[5] "[...] this change cannot have been directly linked to the number of possible models for parallels, since even in the seventeenth century there were more common nouns which used the (then unproductive) *-aux* than used the (newly productive) *-als*. The change in productivity appears to be independent of the change in frequency."

considerá-la como puramente gramatical e não causada pelas frequências relativas das terminações. Para além da caracterização gramatical, é preciso entender por que novas gerações de usuários da língua tiveram uma interpretação diferente dos dados a partir dos quais a gramática é construída.

É importante perceber que, fora do conjunto restrito de nomes terminados em *-al*, o número de palavras formando seu plural com o sufixo *-s* já era muito maior na língua francesa – trata-se do padrão geral de formação de plural pelo menos desde o século XIV (Toynbee, 1896). Em outras palavras, olhando para a língua como um todo, o plural formado com o sufixo *-s* era o mais frequente, sendo instanciado pelo maior número de palavras. É plausível que gerações sucessivas de falantes adquirindo o francês tenham sido influenciadas pela frequência do padrão geral, que certamente era mais robusto do que o padrão em *-aux*. Por esse mesmo motivo, possivelmente, algumas variedades não padrão do francês moderno têm eliminado o plural em *-aux*, utilizando o sufixo *-s* para todos os nomes terminados em *-al* (Mayerthaler, 1977, apud BAUER, 2001); esta regularização já se observava no século XIX (Lodge, 2008) e pode ser vista hoje na fala de crianças adquirindo o francês (Hickmann, 1997). Assim, não parece correto afirmar que essa mudança na gramática do francês não esteve ligada ao número de modelos disponíveis que pudessem dar suporte à extensão do padrão em *-s*; a questão é que esses modelos não parecem ter sido buscados apenas no domínio dos nomes terminados em *-al*.

O outro caso discutido por Bauer é o do sufixo *-ment* do inglês. Com base em um levantamento de datações disponíveis no *Oxford English Dictionary*, o autor observa variação no número de palavras formadas por esse afixo ao longo da história da língua. Das 1100 palavras encontradas, quase metade delas foi formada no século XVI. Entre os séculos XVII e XVIII, houve uma grande queda no número de formações novas, e esse número voltou a crescer no início do século XIX, para diminuir novamente a partir de então, até chegar ao estágio atual, em que *-ment* parece ser apenas marginalmente produtivo. Bauer acrescenta que a perda de produtividade desse sufixo no século XVII não pode ter se devido a nenhuma perda de clareza semântica, visto que ele voltou a se tornar produtivo no século XIX; nem poderia ter se devido a uma falta de bases disponíveis para sufixação (embora, reconhecidamente, o autor afirme isso sem evidências para suportar sua asserção). Tudo o que seríamos capazes de observar nesse caso é uma mudança de produtividade. Ou seja, a mudança relevante na gramática do inglês seria a própria variação em produtividade, sem que ela fosse causada por fatores externos.

Em um estudo posterior, Lindsay e Aronoff (2013) mostram que, de fato, houve uma redução no número de bases verbais disponíveis para sufixação justamente no

período em que Bauer (2001) observa uma diminuição na produtividade de *-ment*. Lindsay e Aronoff (2013) contrastam o caso de *-ment*, formador de substantivos deverbais, com o de *-ity*, formador de substantivos deadjetivais. Ambos os sufixos entraram na língua inglesa pelo empréstimo de palavras do francês que continham os afixos *-ment* e *-ité*. Os empréstimos com *-ment* começaram já antes do século XIV, e esta terminação foi posteriormente generalizada, passando a ser produtiva dentro da língua inglesa. A diferença entre os dois sufixos é que *-ment* passou, gradualmente, a formar menos palavras a partir do século XVII (como apontara Bauer), ao passo que *-ity* manteve sua tendência de aumento de produtividade. Dois fatores explicam esse contraste, de acordo com Lindsay e Aronoff (2013). Primeiramente, ao contrário do que supunha Bauer, houve, de fato, um grande decréscimo na quantidade de verbos novos que entravam na língua inglesa no século XVII, ao passo que o número de adjetivos novos continuou sendo alto. Além disso, *-ment* já vinha ganhando um concorrente: *-ation*, generalizado a partir de empréstimos com o sufixo latino *-atio* e com o sufixo francês *-ation*. Justamente no período crítico em que o número de verbos novos estava em decréscimo, entravam na língua cinco vezes mais empréstimos com *-ation* do que com *-ment*. Com este suporte, *-ation* teve mais condições de se generalizar a partir do século XVII, ao passo que *-ment* foi gerando um número cada vez menor de derivados.

Na narrativa de Lindsay e Aronoff (2013), não parece ter entrado em jogo nenhuma mudança na gramática de *-ment*, no sentido estrito de Di Sciullo e Williams (1987). As mudanças na produtividade desse afixo ao longo dos séculos são explicadas pelas taxas de empréstimos e pela concorrência de um afixo que, por sua vez, também tinha sua produtividade afetada pela taxa de empréstimos. Bauer (2001) tem razão ao dizer que, na descrição de qualquer estágio da língua inglesa, as produtividades relativas de *-ment* e *-ation* teriam de ser devidamente descritas, bem como as porções da gramática referentes a cada um desses afixos, com suas restrições fonológicas, semânticas e morfossintáticas. No entanto, essas condições não eram suficientes para explicar a produtividade desses sufixos; por esse motivo, é lícito considerar que a produtividade é, em certa medida, independente da gramática, entendida no sentido estrito de Di Sciullo e Williams (1987). Tanto é que, em um outro estágio da língua, poderíamos ter, no domínio relevante, a mesma gramática, em que descreveríamos as mesmas condições para cada um dos afixos, mas com taxas de produtividade bastante diferentes – é exatamente essa situação que observamos na narrativa de Lindsay e Aronoff (2013) sobre a evolução do sufixo *-ment*.

Para Bauer (2001), a sua descrição da evolução do sufixo *-ment* implica que “a

produtividade é, por si só, uma parte importante de uma descrição linguística e que ela não pode ser necessariamente reduzida a outros fatores, como frequência, classe de *input* ou clareza de significado". Essa implicação é verdadeira mesmo quando consideramos as revisões advindas dos resultados de Lindsay e Aronoff (2013). No entanto, não parece possível derivar disso a conclusão de que a produtividade é parte da gramática, como uma propriedade primitiva associada a padrões morfológicos. Em vez disso, parece restar de nossa discussão que estamos diante de um fenômeno heterogêneo, que se apoia em fatores gramaticais, mas também, de forma crucial, em fatores extragramaticais – incluindo frequência, disponibilidade de bases, entre outros. A produtividade, nesse sentido, pode ser concebida como o resultado observável da interação desses fatores.

No próximo capítulo, nos concentraremos sobre a competição morfológica, já que, como vimos na discussão dos exemplos desta seção, a existência de expressões concorrentes é um dos grandes determinantes das possibilidades de uso de padrões morfológicos.

3 Competição e Bloqueio

Na discussão anterior sobre o plural do francês e sobre padrões de nominalização do inglês, esteve implícita a ideia de que um mesmo significado pode ser expresso por formas alternativas. Assim, para a expressão do plural de nomes terminados em *-al* no francês, falamos em duas terminações possíveis: *-als* e *-aux*. A primeira delas é mais geral, a segunda é lexicalmente restrita. Crucialmente, a escolha entre elas não é livre, mas governada por uma forte tendência de unicidade no preenchimento do paradigma de cada palavra, de forma que, normalmente, cada item lexical terá a categoria morfossintática relevante preenchida por apenas uma dessas alternativas. Assim, *festival* tem o plural *festivals*, e *cheval* tem o plural *chevaux*. A escolha de uma das formas na língua acarreta, em geral, a proibição da ocorrência da outra; assim, **festivaux* e **chevals* são formas agramaticais no francês padrão. Podemos dizer, portanto, que essas terminações estão em competição, pois o emprego de uma delas bloqueia o emprego da outra.

Essa situação de bloqueio na morfologia flexional recebeu diversos tratamentos na literatura. Minimamente, assume-se algum princípio que garanta que haja apenas uma forma de exponência para cada célula de um paradigma morfológico (por exemplo, o princípio de *Unicidade* em Wunderlich (1996)), com condições que estabeleçam como essa escolha é feita pela gramática, como o Princípio de Elsewhere (p. ex. Kiparsky (1982)). Em termos informais, este princípio estabelece que entre duas regras que poderiam se aplicar em um mesmo contexto, a regra que tiver a descrição estrutural mais específica é a que tem precedência, aplicando-se em detrimento da regra mais geral. É assim que no exemplo do plural do francês, mencionado acima, a regra que forma o plural em *-aux* tem precedência sobre a regra que formaria o plural em *-als*, no caso da palavra *cheval*. Isso acontece porque a terminação *-aux* é reservada a um conjunto fechado de palavras do francês, sendo, portanto, mais restrita em sua aplicação do que a regra geral, que forma plurais em *-s*.

No domínio da morfologia derivacional, no entanto, essas considerações não dão conta das situações de competição e bloqueio. Em primeiro lugar, pode ocorrer bloqueio mesmo quando não há concorrência entre regras. É este o caso em um dos exemplos explorados por Aronoff (1976), em que a formação **gloriosity*, a partir de *glorious*, é bloqueada por *glory*. Kiparsky (1983) afirma que "não há, evidentemente, como estender [o Princípio de Elsewhere] de modo que ele faça com que a mera existência de *glory* bloqueie a adição de *-ity* a *glorious*". Essa extensão não é possível porque *glory* não é produto da aplicação de uma regra de formação de palavras a

glorious. Assim, não há duas regras em competição que pudessem estar sujeitas a esse princípio. Outra razão para se concluir que o bloqueio de **gloriosity* por *glory* não pode ser resolvido por algum princípio de competição entre regras é o fato de que *gloriousness* é uma formação possível em inglês. Caso o Princípio de Elsewhere estivesse em jogo aqui, a formação de *gloriousness* também deveria ser obstruída pela existência de *glory*.

Para Aronoff (1976), **gloriosity* é bloqueada por um princípio que impede a listagem de uma palavra no léxico sempre que houver um sinônimo contendo o mesmo radical - no caso, *glory*. Assim, não é necessário que haja interação ou competição entre regras. Basta que o léxico possua algum tipo de estrutura que permita identificar posições lexicais a serem preenchidas por uma e apenas uma palavra. Por exemplo, a posição lexical correspondente ao substantivo abstrato que designa "a qualidade de ser *glorious*". Pode-se traçar aqui um paralelo com o papel da noção de paradigma na morfologia flexional, embora isso não seja feito explicitamente por Aronoff (1976).[6] A palavra *gloriousness* escaparia a esse princípio por não precisar ser listada: *-ness* é um sufixo altamente produtivo no inglês, capaz de ser adicionado a virtualmente qualquer adjetivo; assim, na ausência de outras motivações, nenhum novo substantivo em *ness* precisaria ser listado no léxico. Substantivos em *-ity*, por outro lado, não podem ser formados a partir de qualquer adjetivo do inglês; portanto, cada nova formação precisaria ser listada, de acordo com Aronoff (1976), estando assim sujeita à atuação do princípio de bloqueio.

Essa explicação, porém, encontra problemas. Mesmo alguns dos exemplos de Aronoff (1976) contrariam sua teoria de bloqueio, como aponta Kiparsky (1983). É o caso das palavras formadas pelo sufixo produtivo *-(c)y*, que impedem a formação de palavras em *-ness* com a mesma função:

(1) decency *decentness
aberrancy *aberrantness
profligacy *profligateness

Essa situação é inesperada diante da grande produtividade de *-ness* e da consequente expectativa de que palavras formadas por esse sufixo não precisem ser listadas. Essa expectativa é ainda maior no caso de construções frasais, dada sua produtividade virtualmente ilimitada. Assim, é ainda mais problemático para a teoria de bloqueio de Aronoff (1976) o fato de que pode haver bloqueio de construções frasais por itens lexicais (Poser, 1992). Exemplo disso é a restrição à formação de

[6] Para uma teoria que desenvolve essa ideia de paradigmas derivacionais de forma explícita, ver Miyagawa (1981).

comparativos frasais no inglês, como **more good*, quando há um comparativo lexical equivalente, *better*.

Kiparsky (1983) não discute o bloqueio de construções frasais, mas propõe, para os demais casos, que se trate a questão no nível semântico, propondo o princípio Evite Sinonímia:

(2) "A saída de uma regra lexical não pode ser um sinônimo de um item lexical existente." (Kiparsky, 1983, p. 13)

Essa formulação cobre os casos de bloqueio previstos por Aronoff (1976), mas não é restrita a palavras com o mesmo radical. Isso é importante, porque também há bloqueio em pares como *thief* - **stealer*, em que não há relação morfológica entre as duas palavras, mas a ocorrência da segunda é geralmente bloqueada pela existência da primeira. Assim, esta proposta parece ser mais promissora que as anteriores, mas precisaria, ainda, ser revista para acomodar as construções frasais envolvidas no tipo de bloqueio apontado por Poser (1992), pois, neste caso, não se trata de uma proibição sobre a "saída de uma regra lexical" - pelo menos não no entendimento de regra lexical de teorias como a de Kiparsky (1982).

Outro problema do princípio expresso em (3) é que ele parece ser forte demais. Horn (1984) observa que alguns pares de expressões podem coexistir, para um mesmo usuário da língua, ainda que haja equivalência semântica entre os elementos desses pares, trazendo exemplos como *icebox* - *refrigerator* e *synonymy* - *synonymity*. No corpus utilizado nesta pesquisa, é possível encontrar, em um mesmo artigo de jornal, as palavras *interceptação* e *intercepção*, com significado equivalente. Da mesma forma, em uma busca na internet, é possível encontrar *coercivo* e *coercitivo* coexistindo, com o mesmo significado, até dentro de um mesmo texto. Contraexemplos como estes fortalecem a concessão feita por Kiparsky (1982) ao sugerir que a proposição em (3) seria, de fato, um princípio curioso e que talvez seja "mais correto vê-la como uma estratégia de aprendizagem da língua do que como uma restrição formal da gramática".

3.1 Morfologia e aquisição de vocabulário

A ideia de que a existência de bloqueio na morfologia esteja ligada a uma estratégia de aprendizagem ecoa a sugestão de Dowty (1979) de que os mecanismos morfológicos de formação de palavras têm como função principal servir como "um auxílio na aquisição de vocabulário adicional". Na prática, isso significa que as regras morfológicas e as regras de interpretação semântica associadas a elas garantem que uma pessoa possa inferir algumas informações relevantes sobre palavras novas com que

tenha contato, desde que elas sejam formadas de acordo com os padrões morfológicos da língua. Assim, diante de *paulistização*, qualquer falante de português brasileiro que consiga depreender a base *paulista* é capaz de inferir, minimamente, que se trata de um substantivo abstrato, que provavelmente denota uma "mudança de estado" em que algo se torna "paulista" em alguma dimensão.[7] Da mesma forma, um falante que queira expressar essa noção em algum contexto, pode se valer dos sufixos *-izar* e *-ção* e confiar que as regras de interpretação semântica associadas a eles darão ao interlocutor boas pistas para a interpretação do enunciado.

Seguindo este raciocínio, é razoável supor que o próprio emprego de um mecanismo de formação de palavras sinalize ao interlocutor a necessidade de expressar um conteúdo que não poderia ser expresso por meio de algum item lexical já existente. Ou seja, parte do que é comunicado com o uso de uma palavra nova, como *paulistização*, é a necessidade de se referir a uma ideia que ainda não encontrava expressão no léxico. Quando essa necessidade é espúria, isto é, quando já existe uma palavra para expressar a ideia pretendida, a formação nova pode ser descartada como anômala. Dentro desse entendimento, são compreensivas as ocorrências de bloqueio parcial, em que uma nova formação é bloqueada apenas nos contextos em que teria o mesmo significado de uma palavra existente, mas não quando pode se referir a uma noção distinta. Por exemplo, para a maioria dos usuários do português brasileiro, possivelmente, a palavra *falador* em "os (?)faladores de português brasileiro" seria anômala em um contexto neutro, como em uma dissertação de linguística, mas bastante aceitável em um contexto mais pejorativo, como "o menino era chato e falador".

Um tipo de proposta que parece dar conta dessa concepção da função da morfologia é a abordagem pragmática do problema do bloqueio, que remonta, pelo menos, ao trabalho de McCawley (1978). Nessa perspectiva, algumas situações de bloqueio são analisadas à luz dos princípios conversacionais propostos por Grice (1975). Vejamos, por exemplo, a discussão feita por Horn (1984) dos seguintes dados trazidos por McCawley (1978), sobre a distribuição de causativos lexicais e perifrásticos, que pode ser vista como um exemplo de bloqueio parcial, em que uma expressão é bloqueada na interpretação canônica, mas pode ocorrer com outro sentido:

(3) a. Black Bart killed the sheriff.
'Black Bart matou o xerife.'

[7] Exemplo presente no corpus: *As federações de futebol de Rio Grande do Sul, Minas Gerais, Bahia e Rio chegaram a ensaiar uma oposição contra o que chamavam de "paulistização da CBF"*. (http://www1.folha.uol.com.br/fsp/esporte/129417-influente-na-cbf-futebol-paulista-patina.shtml)

b. Black Bart caused the sheriff to die.
'Black Bart causou a morte do xerife.'

Nesse exemplo, o causativo lexical *kill*, em (3a), é restrito a situações de causação prototípica – direta, não mediada, por meio de ação física. Para Horn (1984), essa interpretação pode ser derivada por meio de um princípio de economia centrado no falante, expresso em (4b). Assumindo que um enunciador da frase (3a) obedece esse princípio conversacional e, assim, utiliza a forma menos marcada para chegar à interpretação canônica, pode-se derivar a interpretação de causação prototípica; essa interpretação tende, então, a ser convencionalmente associada ao causativo lexical. O uso da construção mais marcada, em (3b), por outro lado, implica uma interpretação não prototípica, via Princípio Q, em (4a). Assumindo que um enunciador da frase (3b) obedece esse princípio, e sabendo que ele escolheu não utilizar a expressão menos marcada, que levaria à interpretação prototípica, pode-se inferir que esta interpretação não é compatível com o seu conhecimento da situação ou com suas intenções comunicativas – portanto, tem-se a implicação de que a situação não marcada não é a pretendida; por exemplo, talvez "Bart, tendo socado algodão na arma do xerife, tenha feito com que o disparo saísse pela culatra; ou tenha providenciado para que escorpiões fossem colocados no quarto do xerife (que sabidamente tem um coração fraco), etc".

(4) a. Princípio Q (centrado no ouvinte)
FAÇA UMA CONTRIBUIÇÃO SUFICIENTE
DIGA O QUANTO PUDER (considerando R)

b. Princípio R (centrado no falante)
FAÇA UMA CONTRIBUIÇÃO NECESSÁRIA
NÃO DIGA MAIS DO QUE VOCÊ PRECISA (considerando Q)

(Horn, 1984, p. 13)

Uma vantagem óbvia dessa abordagem é que ela deriva efeitos de bloqueio total e parcial de princípios que são, por hipótese, geralmente válidos na comunicação humana, não de condições específicas do léxico.

Poser (1992), contudo, coloca três problemas para esse tipo de abordagem. O primeiro deles se refere mais diretamente à formulação de McCawley (1978), em que uma forma A bloqueia a forma B, de mesmo significado, caso a forma A envolva um "esforço menor" do que a forma B – tendo, por exemplo, menos material fonológico. O problema é que, em muitos casos, as formas em competição não se diferenciam, aparentemente, em termos de esforço necessário para produzi-las: em português,

por exemplo, *estive* (flexão do verbo *estar*) não é mais simples do que **estei* (cf. *testar* - *testei*) em termos de quantidade de material fonológico ou morfológico; ainda assim, a primeira dessas formas bloqueia a outra. O exemplo dado por Poser (1992) é do bloqueio de **oxes* por *oxen*.

O segundo problema colocado por Poser (1992) para as abordagens pragmáticas é o fato de que, em casos típicos de bloqueio morfológico, costuma haver julgamentos fortes de agramaticalidade, que parecem ser diferentes dos julgamentos de aceitabilidade dependentes de contexto dos casos típicos da literatura pragmática.

Por fim, segundo Poser (1992), a vantagem da abordagem pragmática de não restringir efeitos de bloqueio a itens lexicais, e de, portanto, capturar relações desse tipo existentes entre palavras e estruturas frasais, carrega o ônus de prever bloqueio em casos em que ele não ocorre. Por exemplo, assim como há bloqueio entre comparativo lexical e perifrástico em (5), esperar-se-ia que houvesse o mesmo efeito entre as sentenças em (6).

(5) a. John is smarter than Tom.
    b. * John is more smart than Tom.

(6) a. John's intelligence exceeds Tom's.
    b. John has more intelligence than Tom.
    c. John has greater intelligence than Tom.

O primeiro problema centra-se na questão de "menor esforço" ou de complexidade. Os exemplos de Poser (1992), contudo, não precisam levar à conclusão de que essas noções não estão envolvidas no bloqueio morfológico. Em vez disso, pode ser o caso que, em morfologia, complexidade não possa ser definida simplesmente em função do número de morfemas ou de segmentos fonológicos de uma expressão. Ao tratar dessa questão, Horn (1984) fala em termos de "marcado" e "não marcado", o que, tradicionalmente, envolve outras dimensões de análise, além da contagem de elementos. No caso de flexões irregulares, como *oxen* x **oxes* (exemplo de Poser (1992)), as formas podem se diferenciar pelo fato de uma delas poder ser recuperada pronta da memória, enquanto a outra precisaria ser formada por algum processo morfológico regular. A diferença entre esses dois mecanismos de expressão morfológica certamente tem implicações para uma avaliação plena da ideia de "menor esforço". De fato, as evidências psicolinguísticas acerca da competição entre formas regulares e irregulares mostram que ela é influenciada pela relativa facilidade de acesso a itens armazenados na memória (em comparação com o acesso a processos produtivos). Assim, padrões regulares podem se aplicar sempre que formas memorizadas (possivelmente irregulares) forem inexistentes ou inacessíveis, como

se pode ver na tendência de formas irregulares de baixa frequência (memorizadas com menos facilidade) sofrerem mais "erros" de produção e serem eventualmente substituídas por formas regulares ao longo da história de uma língua (Prasada e Pinker, 1993). Quando uma forma irregular está acessível na memória, contudo, a estratégia de produção mais simples, não marcada, é simplesmente recuperá-la antes que qualquer processo produtivo possa se aplicar.

Quanto ao segundo problema, é importante notar que o exemplo de julgamento forte de agramaticalidade (e de "caso típico de bloqueio") citado por Poser (1992) vem da morfologia flexional: *men* - **mans*. Este é comparado pelo autor com um caso de bloqueio entre um item lexical e uma construção sintática, que não envolve um julgamento tão forte quanto no primeiro caso; o exemplo é o de *pink* bloqueando parcialmente o sintagma (?)*pale red*. O contraste entre esses tipos de construção é importante, pois a flexão parece estar, por razões independentes, sujeita a princípios que não se aplicam diretamente a outros tipos de morfologia ou a construções sintáticas. Além de ela ser geralmente obrigatória, a expressão de um dado conjunto de traços flexionais costuma ser categórica (não apresentar opcionalidade), como regra geral. Teorias morfológicas costumam ter mecanismos gramaticais para dar conta desse fato; por exemplo, por meio da ideia de paradigmas flexionais, em que cada célula deve ser preenchida por uma e apenas uma forma, ou por meio de princípios de ordenamento, como o Princípio de Elsewhere, que garante que apenas uma operação morfológica se aplique em um dado contexto, etc. Isso significa que, mesmo que não assumíssemos uma abordagem pragmática do bloqueio morfológico, ainda precisaríamos de mecanismos distintos para dar conta dos julgamentos fortes de agramaticalidade que se observam na morfologia flexional e que, muitas vezes, não se observam na morfologia derivacional, em casos em que se esperaria bloqueio. Assim, supor que existe uma tendência geral de bloqueio regida por princípios pragmáticos é compatível com a existência dessas restrições adicionais às quais a morfologia flexional é sujeita.

Ademais, uma abordagem pragmática das situações de bloqueio parcial, como a de (?)*pale red*, pressupõe que haja uma porção de significado restante a ser ocupada pela forma produtiva que escapa ao bloqueio total; é isso que observamos no caso de (3a-3b), em que o causativo perifrástico é bloqueado na leitura de causação prototípica, mas é aceitável em outras interpretações. No caso da morfologia flexional, por ela ser responsável pela exponência de traços morfossintáticos definidos estruturalmente, é normalmente difícil imaginar que porção de significado restaria disponível para a forma regular (e.g. **mans*) quando já existe uma forma irregular

estabelecida para a mesma função. [8]

O terceiro problema levantado por Poser (1992) é mais sério para uma abordagem pragmática do bloqueio, e já fora levantado por Horn (1984) e Horn (1978), ao comentar que nem sempre é fácil determinar quando uma expressão conta como alternativa a outra. No caso de (5), temos uma boa intuição de que estamos diante de duas expressões comparativas equivalentes, *smarter* e *more smart*. No caso de (6), por outro lado, podemos encontrar diferenças de estrutura informacional que motivariam uma distinção entre, pelo menos, o primeiro dos exemplos e os demais, com implicações sobre os contextos em que esperamos encontrar cada um desses casos: no primeiro exemplo, o tópico é "a inteligência de John"; nos demais, o tópico é "John".

Para dar conta desse problema, Poser (1992) sugere que o fenômeno do bloqueio é restrito ao domínio das categorias que tipicamente se manifestam por mecanismos morfológicos, mesmo quando elas são expressas por construções frasais em uma língua particular. O autor conjectura que esse domínio esteja restrito ao dos sintagmas que contêm apenas projeções de nível zero, na Teoria X-Barra. Por exemplo, a forma perifrástica do comparativo do inglês (*more intelligent*, **more smart*) seria "presumivelmente [...] de tipo A^1, contendo apenas categorias de tipo ADV^0 e A^0" (Poser, 1992, p. 127); portanto, estaria no domínio em que o bloqueio é atuante. O mesmo não poderia ser dito das estruturas que são comparadas nos exemplos em (6).

Ainda é uma questão em aberto se o tipo de domínio proposto por Poser (1992) restringe corretamente os tipos de bloqueio atestados. A proposta parece não dar conta, por exemplo, do bloqueio parcial observado em (3a-3b), em que o causativo perifrástico envolve, presumivelmente, um VP interno. Seja como for, uma vez reconhecida a existência de bloqueio entre formas lexicais e certas construções frasais, resta para qualquer teoria desse fenômeno a tarefa não trivial de definir o domínio em que esse tipo de interação pode ocorrer. Com a consciência dessa necessidade de refinamento teórico, não nos concentraremos sobre este ponto neste trabalho. Em vez disso, voltamo-nos brevemente ao problema da noção de complexidade envolvida na abordagem pragmática do bloqueio.

Como vimos anteriormente, a ideia de que formas mais complexas podem ser bloqueadas por formas que exigem "menor esforço", na formulação de McCawley (1978), não pode depender meramente da contagem do número de elementos lin-

[8] Quanto a *pale red*, a abordagem pragmática só prevê que esse sintagma não possa, normalmente, se referir à mesma porção do espectro de cores a que se refere *pink*; mas ele pode certamente se referir a qualquer outra dessas porções que seja importante diferenciar de *pink* e *red* em um contexto relevante. O ponto crucial é que *pale green* e *pale blue*, por exemplo, não são restritos da mesma forma.

guísticos envolvidos. No caso de formas irregulares armazenadas no léxico, não é difícil derivar sua vantagem em relação a formas regulares a partir dos princípios conversacionais em (4). Do ponto de vista do falante, é desnecessário gerar uma forma nova ("dizer mais"), a não ser que se queira sinalizar ao interlocutor que a interpretação pretendida não é a canônica. Do ponto de vista do ouvinte, o uso de uma forma diferente daquela que já é consagrada na língua licencia a inferência de que a interpretação deve ser outra. Caso não haja uma interpretação disponível que possa ser atribuída à forma nova nos contextos relevantes, o enunciado pode ser anômalo. Esse mecanismo é aplicável mesmo a casos de bloqueio que não envolvem nenhuma diferença em termos de número de operações ou de elementos morfológicos. Blutner (1998) cita, a esse respeito, um exemplo envolvendo o fenômeno de *grinding*, por meio do qual nomes contáveis podem ter uma leitura massiva, denotando a coisa de que são feitos, como em (7), em que *fish* não denota mais um objeto animal discreto, mas uma substância que pode ser consumida. Essa operação semântica normalmente falha no caso de animais como *pig* e *cow*; por hipótese, isso acontece porque a língua inglesa já possui palavras com esse sentido especializado, *pork* e *beef*, que bloqueiam a operação de *grinding*.

(7) I ate fish.

'Eu comi peixe'

(8) a. I ate pork/?pig.

'Eu comi (carne de) porco.'

b. I ate beef/?cow.

'Eu comi (carne de) vaca.'

Seguindo essa perspectiva, parece-nos que uma interpretação promissora da noção de "menor esforço" em uma abordagem pragmática do bloqueio (e da competição morfológica) é vê-la como uma tendência conservadora, que promove o uso de mecanismos e expressões já existentes em usos não marcados; e que deixa disponível para contextos inesperados o uso de expressões novas ou marcadas. Vista dessa forma, essa tendência é um fator limitador da produtividade de padrões morfológicos, pois prevê que a formação de palavras novas seja bastante limitada pelo léxico já existente. De forma especulativa, podemos supor que ela também governa a competição entre estruturas produtivas, mesmo na ausência de expressões lexicalizadas que possam bloqueá-las. Dada uma nova necessidade comunicativa, usuários de uma língua se veem, frequentemente, diante da escolha entre mecanismos linguísticos distintos que poderiam ser utilizados para suprir essa necessidade. A abordagem pragmática de bloqueio discutida acima, se estendida para esses casos, prevê que

essa escolha deve ser governada pelo grau de convencionalidade dos mecanismos rivais no contexto específico em que a escolha se coloca. Esse grau de convencionalidade, por sua vez, pode ser definido pelos usos prévios desses mecanismos no contexto em questão.

Na seção seguinte, exploraremos um modelo de aprendizagem e de competição morfofonológica baseado na ideia de que padrões produtivos (ou semiprodutivos) de uma língua podem ser mais ou menos prováveis em cada contexto de aplicação, de acordo com o sucesso que obtiveram nesses contextos em usos prévios encontrados pelo aprendiz. Essa abordagem surgiu no contexto do debate sobre a competição entre formas regulares e irregulares (centrada, sobretudo, na morfologia do passado do inglês, p. ex. Albright e Hayes (2002)), com o fim de capturar os efeitos de frequência observados nesse domínio, sem referência direta às preocupações deste trabalho. Contudo, propomos que a noção de confiabilidade advinda desse debate é um fator importante para capturar a ideia de "menor esforço" (ou complexidade, ou marcação) que tem sido explorada até aqui.

3.2 Ilhas de confiabilidade na competição morfológica

Na seção anterior, o foco da discussão foram casos de competição entre novas construções e formas preexistentes. Vimos que esse tipo de competição pode ser um fator limitador da produtividade de padrões morfológicos, visto que palavras novas costumam ser bloqueadas quando suas funções já são preenchidas por outros itens lexicais, o que limita o potencial criativo da morfologia derivacional. Na investigação da produtividade relativa dos padrões morfológicos, também nos interessa entender um outro tipo de competição: a que se dá entre possibilidades distintas de formação de palavras, mesmo na ausência de qualquer item preexistente para exercer a mesma função.

Um exemplo desse tipo de competição é a que vimos na Seção 2.3, entre os afixos *-ment* e *-ation* do inglês. Por algum tempo, aparentemente, falantes dessa língua tiveram uma escolha entre esses dois sufixos na formação de substantivos abstratos a partir de verbos. Diante de uma base verbal nova na língua, não relacionada, portanto, a nenhuma nominalização preexistente, as alternativas eram uma derivação em *-ment* e uma em *-ation*.[9] Como vimos, de acordo com Lindsay e Aronoff (2013), essa competição foi resolvida em favor de *-ation*, com a resultante perda de produti-

[9] Evidentemente, esta é uma simplificação, pois à época em que tanto *-ation* quanto *-ment* eram produtivos no inglês, também havia outros padrões de nominalização, de base germânica, de modo que a competição era provavelmente mais complexa. Ainda assim, tratava-se de meios alternativos de formação de palavras e não de uma alternativa entre produtividade e uso de itens lexicais já existentes.

vidade de *-ment*, graças ao grande número de empréstimos contendo o sufixo *-ation* em períodos anteriores da língua, que forneceram uma grande base de generalização para que falantes posteriores interpretassem este afixo como a escolha mais segura de nominalização deverbal.

Entretanto, a extinção de um dos padrões rivais é apenas uma das opções de resolução de uma situação de competição morfológica. Outra possibilidade é a de que a língua se organize de forma a comportar a coexistência desses padrões. Por exemplo, Lindsay e Aronoff (2013) discutem o caso de *-ic* e *-ical*, do inglês, ambos formadores de adjetivos. Assim como *-ment* e *-ation*, *-ic* entrou na língua inglesa por meio da reanálise de empréstimos do francês. *-ical*, por sua vez, é produto de uma amalgamação resultante do amplo uso do sufixo *-al* com palavras técnicas terminadas em *-ic*, como em *mathematical* and *poetical*. Embora sejam sinônimos, *-ic* e *-ical* ainda são ambos produtivos na língua atual. Em uma busca realizada por meio da Google Search API, os autores encontraram uma preferência 7,84 vezes maior para a afixação de *-ic*, sugerindo que este é o sufixo dominante deste par.[10]

Para explicar a manutenção de *-ical* frente a um rival mais produtivo, os autores investigaram as terminações dos radicais aos quais esses afixos se juntam, com o objetivo de encontrar contextos em que houvesse favorecimento de um ou outro formativo. A única subregularidade encontrada refere-se ao conjunto de palavras terminadas em *-olog-*, em que *-ical* ocorre com uma frequência 6,42 vezes maior do que *-ic*, praticamente invertendo a regularidade encontrada em favor de *-ic* no restante do léxico. Como se trata de um contexto frequente, os autores sugerem que essa distribuição tenha dado suporte à manutenção da produtividade de *-ical*.

Outro exemplo discutido por Lindsay e Aronoff (2013), e com mais detalhes por Plag (2000), é o da competição entre os verbalizadores *-ize* e *-ify*. Utilizando o mesmo método de busca pelo número estimado de atestações através da Google Search API, Lindsay e Aronoff (2013) observam que *-ize* é, em geral, mais generalizado, com uma razão de aproximadamente 5:1, mas que, no subconjunto de bases monossilábicas, essa mesma razão se dá em favor de *-ify*. Novamente, essa distribuição quase complementar parece ter garantido a coexistência desses dois afixos sinônimos na língua.

[10] Não é claro a partir deste levantamento o quão larga seria a vantagem de *-ic* sobre bases novas, que é o que mais nos interessa no estudo da produtividade. Uma possível limitação do estudo de Lindsay e Aronoff (2013) a esse respeito é o fato de que a busca por atestações de palavras formadas por esses afixos foi feita a partir de bases extraídas de itens já dicionarizados com algum desses sufixos. Na descrição dos autores, "[...] we identified all words ending in either -ic or -ical (or both) in Webster's 2nd International Dictionary and stripped off the suffixes to produce 11,966 unique stems. [...] we then executed automated queries for each stem and suffix combination (e.g. biolog- + -ic,-ical)" (Lindsay e Aronoff, 2013, p. 11). É possível que esse procedimento não nos responda sobre qual seria a escolha dos falantes no caso de bases que ainda não possuem uma forma correspondente, e institucionalizada, em *-ic* ou *-ical*.

A ideia geral que podemos extrair desses exemplos é de que a estabilidade de um padrão morfológico pode ser garantida, mesmo diante de um rival mais generalizado, caso novas gerações de usuários da língua sejam capazes de encontrar nichos em que o emprego de um padrão morfológico é suficientemente previsível. Em princípio, esses nichos podem se formar em torno de qualquer dimensão linguística – Lindsay e Aronoff (2013) trazem exemplos formados por critérios fonológicos, morfológicos, semânticos e pragmáticos. Como consequência dessa ideia, podemos supor que a transmissão de um sistema morfológico não envolve apenas o aprendizado de regras extremamente gerais e de alternâncias restritas a itens lexicais. Entre esses dois extremos, parece haver diversos tipos de subregularidades às quais falantes parecem ser sensíveis e que estão envolvidas na evolução das línguas, seja favorecendo mudanças, seja garantindo a manutenção de certos padrões. Por isso, voltaremos nossa atenção para um modelo de aprendizagem que parece ser capaz de capturar algumas propriedades desejáveis desse processo de transmissão linguística, como a sensibilidade a subgeneralizações e a capacidade de explorá-las produtivamente.

Em uma série de trabalhos, Albright (2002) e Albright e Hayes (1999, 2002, 2003) apresentam um modelo de aprendizagem de regras morfológicas e fonológicas que visa a emular as intuições dos falantes sobre as regularidades e subregularidades envolvidas na produção de novas formas linguísticas. Esses autores apresentam uma alternativa intermediária no debate entre teorias puramente associativas da morfologia, que assumem um único mecanismo para a geração de formas irregulares e regulares, e teorias que assumem uma dissociação entre uso da memória para a produção de formas irregulares e de regras gerais para as regulares. O *Minimal Generalization Learner* (doravante, MGL) de Albright e Hayes (1999, 2002, 2003) também assume que formas regulares são geradas por regras; porém, o modelo postula múltiplas regras estocásticas, com diferentes níveis de generalidade, capazes de dar conta de efeitos de similaridade mesmo na produção de formas regulares.

O MGL é um aprendiz artificial que toma como input pares de formas relacionadas e aprende, a partir disso, um conjunto de regras capazes de generalizar os mapeamentos encontrados entre essas formas. Posteriormente, o algoritmo pode aplicar essas regras produtivamente, sobre bases novas. O algoritmo de aprendizagem funciona de forma iterativa, encontrando, para cada par de formas de treinamento, uma regra bastante específica que descreve a mudança estrutural envolvida. Por exemplo, comparando o par ironizar - ironizasãw, o modelo encontra a seguinte regra altamente específica: ɾ → sãw / #ironiza ____ #. Essa comparação começa pela busca da maior sequência compartilhada pelas duas formas, da esquerda para a direita; no caso, #ironiza. Em seguida, o algoritmo busca a maior sequência compartilhada

da direita para a esquerda, que, no caso, é o marcador de fronteira de palavra, #. O material remanescente em cada uma das formas compõe a mudança estrutural: ɾ → sãw̃.

Após acumular um grande número de regras específicas relacionando cada par de formas de treinamento, o algoritmo compara as que possuem a mesma mudança estrutural, buscando depreender contextos mais gerais em que essa mudança ocorre. A comparação é feita por um procedimento semelhante ao descrito acima. Assim, olhando para os contextos das regras em (9a), o modelo compara as sequências adjacentes ao local da mudança, encontrando a maior sequência compartilhada à esquerda: oniza. Caso haja segmentos não compartilhados mais à esquerda, no caso [ɾ] e [k], eles são comparados entre si, a fim de se verificar se contêm traços fonológicos compartilhados. O restante dos segmentos não compartilhados é reduzido a uma variável X. Esse procedimento é repetido para a sequência à direita do local da mudança, que, neste caso, é a borda da palavra: #. Como resultado, neste exemplo, o algoritmo obtém a regra mais geral em (9b), que engloba os contextos das regras em (9a).

(9) a. ɾ → sãw̃ / #iɾoniza ____#
ɾ → sãw̃ / #pɾekoniza ____#

b. $\text{ɾ} \rightarrow \text{sãw̃} \;/\; \#\,\text{X} \begin{bmatrix} -\text{silábico} \\ -\text{nasal} \\ -\text{labial} \\ -\text{lateral} \end{bmatrix} \text{oniza} \;\_\_\_\_\#$

Posteriormente, outras regras específicas são comparadas com as mais gerais, como (9b), gerando regras cada vez mais abrangentes. Essas iterações se dão de forma conservadora, no sentido de que o modelo obtém, a cada passo, a regra mais específica possível que contenha ambos os contextos comparados. Por isso, ele é chamado de *minimal generalization learner*, ou aprendiz de generalização mínima.

Outra característica importante do modelo, que o diferencia de outros procedimentos de indução de regras, como o apresentado por Yang (2005), é o fato de que ele não descarta regras que já tenham sido aprendidas, mesmo quando seus contextos formam subconjuntos dos de regras mais gerais depreendidas posteriormente. Desse modo, o resultado de todo o procedimento é uma vasta lista de regras, de vários graus de generalidade; evidentemente, a maioria delas, como a expressa em (9b), não é o tipo de regra que seria tipicamente formulada por um linguista analisando os mesmos dados. Uma consequência dessa grande quantidade de regras no modelo é que diversas delas podem corresponder ao mesmo output. Como veremos adiante, isso é crucial para que o MGL seja capaz de capturar intuições sobre sub-

generalizações mesmo no caso de padrões bastante regulares. Outra consequência do modo como o aprendiz organiza as regras no modelo é que um mesmo input pode corresponder a mais de uma possibilidade de output. Isso acontece sempre que um mesmo contexto é englobado, no modelo, por regras que preveem mudanças estruturais distintas. De fato, essa é uma situação comum no modelo, exceto em casos em que há distribuição complementar entre operações. Esta característica permite ao MGL dar conta do fato de que pessoas podem aceitar mais de uma possibilidade de realização de uma forma nova, tendo preferências gradientes entre elas. Para dar conta dessa gradiência, o modelo atribui escores de boa-formação a cada um de seus possíveis outputs, por meio do cálculo do valor de confiabilidade das regras utilizadas para gerá-los.

O valor de confiabilidade de uma regra é calculado a partir de duas informações: o número de formas da lista de treinamento que satisfazem o contexto de aplicação da regra, isto é, o seu *escopo*; e o seu número de *acertos*, isto é, a quantidade de formas no escopo da regra sobre as quais a sua aplicação geraria um output correto, conforme definido pela própria lista de treinamento. Intuitivamente, o valor de confiabilidade deve expressar o grau de certeza do aprendiz de que a regra seria capaz de gerar outputs corretos quando aplicada a novos contextos. Por isso, ele é expresso neste modelo pela razão entre o número de acertos e o de formas no escopo da regra. Por exemplo, nos dados de treinamento de Albright e Hayes (2003), a operação bastante geral que adicionaria [d] a qualquer verbo do inglês para formar seu passado simples alcança um valor de confiabilidade de 0,949 (correspondente a 4.034 acertos em um universo de 4.253 palavras no escopo da regra).

Em seguida, o valor assim obtido é ajustado, para dar conta do fato de que o grau de certeza sobre a confiabilidade de uma regra deve ser proporcional ao número de casos observados no seu escopo. Por exemplo, temos mais certeza de que uma regra tem 100% de sucesso quando a observamos se aplicar em 46 de 46 casos do que quando ela se aplica em 2 de apenas 2 casos. Para penalizar regras baseadas em poucos dados, o aprendiz utiliza o limite inferior de um intervalo de confiança.[11] O

[11] Este limite (π_L) é calculado da seguinte forma, seguindo Mikheev (1997):

$$\pi_L = \hat{p}_i^* - z_{(1-\alpha)/2} \times \sqrt{\frac{\hat{p}_i^*(1-\hat{p}_i^*)}{n}}$$

Nesta fórmula, o valor de z, relativo a um determinado coeficiente de confiança α, é encontrado por uma busca em uma tabela da distribuição t de Student, e o de $\hat{p}_i^*$ é a razão e entre o número de acertos e o escopo da regra, com adição de valores mínimos para evitar zeros no numerador ou no denominador:

$$\hat{p}_i^* = \frac{x_i + 0.5}{n_i + 1.0}$$

grau de ajuste é obtido a partir do coeficiente de confiança α, em que 0.5 < α < 1. Quanto maior é esse coeficiente, maior é a penalização da estimativa inicial da confiabilidade de uma regra. Assim, por exemplo, com α = 0.75, nossos dois exemplos de 100% de aplicação resultariam nos valores de confiabilidade ajustados de 0.979 e 0.57, respectivamente. A interpretação desses valores, dado um coeficiente de confiança, é a seguinte: podemos ter α (e.g. 75%) de confiança de que o escore da regra A não seria menor do que x (e.g. 0.979 ou 0.57) caso tivéssemos uma quantidade arbitrariamente maior de dados de treinamento.[12]

Outra diferença deste modelo em relação a teorias linguísticas tradicionais está na escolha das melhores regras para se gerar um determinado conjunto de outputs. Em Chomsky e Halle (1968), por exemplo, e em teorias derivadas, ao identificar duas possíveis regras, A e B, que descrevem a mesma mudança estrutural, mas têm seus contextos de aplicação em uma relação de inclusão, em que o contexto de B é um subconjunto do contexto de A, o analista se vê obrigado considerar a regra mais geral A como sendo a correta. Em contraste, o aprendiz artificial de Albright e Hayes (2002) pode dar mais peso a uma regra menos geral, desde que ela obtenha um escore de confiabilidade mais alto, isto é, desde ela se aplique de forma menos excepcional do que a regra mais geral.

Essa abordagem do aprendizado de generalizações morfológicas e fonológicas é motivada por resultados experimentais que demonstram preferências gradientes entre formas de palavra possíveis, influenciadas pela existência de subgeneralizações no léxico. Em experimentos de avaliação e produção de pseudopalavras, Albright e Hayes (2003) observaram que participantes preferem formas do passado simples do inglês que se conformem ao que Albright (2002) chama de ”ilhas de confiabilidade”, definidas como ”subgeneralizações sobre contextos fonológicos em que um processo morfológico é especialmente robusto” (Albright, 2002, p. 2). Crucialmente, esse efeito foi encontrado tanto para formas irregulares quanto para as que seguem o padrão regular. Esse resultado se coloca em contraste, portanto, com teorias que postulam uma dissociação forte entre os mecanismos de produção desses dois tipos de formas (e.g. Pinker (1998)), em que as irregulares seriam armazenadas e recuperadas prontas da memória, ao passo que as regulares seriam geradas por uma regra simples e extremamente geral. A predição destas teorias é de que, nos casos em que um padrão irregular é esporadicamente estendido a palavra novas, podem ser observados efeitos de similaridade, em que a possibilidade de uso produtivo do padrão irregular ocorre em analogia com formas preexistentes, disponíveis na memória – assim, *splung* ocorre como passado da pseudopalavra *spling* em da-

[12] O valor exato de α é um parâmetro do modelo, e pode ser selecionado de acordo com o melhor ajuste aos dados.

dos experimentais, no molde de *swung*, *strung*, *wrung*, *stung*, *slung*, *flung* e *clung* (Albright e Hayes, 2003; Bybee e Moder, 1983; Prasada e Pinker, 1993). Por outro lado, no caso do uso produtivo de padrões regulares, a aplicação da regra geral seria automática, sem fazer referência à composição do léxico e, portanto, sem efeitos de similaridade. Com base em seus resultados, Albright e Hayes (2003) defendem que a produção de formas regulares também está sujeita a efeitos contextuais que não podem ser modelados por apenas uma regra geral.

Outra característica do MGL, que é corroborada pelos resultados de Albright e Hayes (2003), é que mesmo as formas irregulares, ainda que memorizadas, podem corresponder neste modelo a regras nos casos em que o algoritmo for capaz de detectar generalizações entre elas. Isto é necessário porque, quando são suficientemente robustas, essas subregularidades podem ser estendidas a novas palavras, o que, de acordo com os resultados obtidos pelos autores, não ocorre por meio de analogia irrestrita a formas preexistentes, mas por meio de similaridades estruturais do tipo que é capturado por regras linguísticas.

Podem-se ver no MGL características de um modelo analógico, na medida em que a probabilidade de uso de uma forma nova não depende apenas da existência de uma regra capaz de gerá-la, mas também do grau de suporte que ela encontra em itens lexicais preexistentes que contenham essa generalização. Contudo, como apontam Albright e Hayes (2003), essa proposta também se distancia de modelos analógicos comuns ao não propor comparações diretas entre formas linguísticas com base em similaridades de qualquer tipo. Em vez disso, elas são relacionadas por meio de regras definidas por características estruturais. Assim, *spiff* e *push* podem ser relacionadas nesse modelo pelo contexto [+contínuo, -voz] ____ #, que define um conjunto de outras formas relacionadas a essas palavras, de forma sistemática, pela característica de terminarem em uma consoante fricativa desvozeada. Modelos puramente analógicos, por outro lado, não precisam se restringir a esse tipo de similaridade estruturada, de modo que essas mesmas palavras poderiam servir de base analógica para verbos hipotéticos como *spooshen* ou *puv*. Portanto, ao se basear em regras definidas em termos estruturais, o MGL carrega a afirmação empírica de que o aprendizado de generalizações morfológicas e fonológicas por seres humanos não utiliza todo o poder disponível a modelos analógicos. Nos testes realizados pelos autores, comparando o MGL a uma implementação de modelo analógico, este último obteve resultados piores na predição de dados experimentais do passado do inglês, justamente por tender a encontrar nos dados de treinamento relações que não correspondem a generalizações feitas por aprendizes reais.

Em resumo, o MGL fornece um modelo de aprendizagem que consegue dar conta

de resultados experimentais importantes sobre o aprendizado da morfologia. Por utilizar regras com variáveis, ele é capaz de gerar outputs corretos mesmo no caso de formas desconhecidas que não apresentam grande similaridade com formas preexistentes, o que tradicionalmente é uma vantagem dos modelos que empregam regras gerais sobre os modelos associativos, que dependem da existência de expressões similares no léxico para a determinação da forma de expressões produtivamente geradas. Por outro lado, ao computar regras de diversos níveis de generalidade, e ao diferenciá-las de acordo com graus de confiabilidade, o modelo é capaz de capturar o papel exercido por generalizações específicas na organização da língua, um tipo de dado que sempre motivou modelos associativos da morfologia. Interessantemente, o MGL é capaz de aprender subgeneralizações mesmo no caso de padrões bastante gerais e regulares.

Na Seção 5.1, demonstraremos uma aplicação desse modelo ao português em um domínio empírico distinto do que é comumente considerado nesse tipo de estudo: a morfologia derivacional. Uma diferença importante desse domínio é que, diferentemente da morfologia flexional, raramente se pode encontrar nele algum padrão morfológico regular que funcione como default, aplicando-se de forma geral a qualquer palavra nova na categoria relevante. Antes desse teste, no capítulo seguinte, esclarecemos como se deu a coleta de dados que dá suporte a este estudo.

4 Coleta de dados

Análises quantitativas da produtividade morfológica dependem do uso de *corpora* de grande extensão. Para os propósitos de nossa pesquisa, tendo em vista a discussão da Seção 2.2, é minimamente necessário que a extensão do corpus permita que os itens lexicais se diferenciem significativamente entre os de alta e os de baixa frequência. Como vimos, espera-se que os produtos de morfologia produtiva distribuam-se, predominantemente, entre os itens de baixa frequência, ao passo que as formas não analisáveis e as que são fruto de morfologia não mais produtiva tendam a se concentrar em níveis mais altos de frequência.

Essa exigência torna pouco prático qualquer tipo de coleta manual de textos. Assim, este capítulo detalha os procedimentos metodológicos que desenvolvemos com o fim de possibilitar a coleta automática de uma grande quantidade de dados. Apresentamos também alguns problemas que esse tipo de coleta enfrenta e o modo como buscamos solucioná-los. O código-fonte documentado de todos os scripts utilizados nesta pesquisa encontra-se disponível nos apêndices deste trabalho e no repositório virtual `http://github.com/shoeki/ling`.

4.1 Seleção dos textos

Nossos dados foram extraídos de textos escritos disponíveis na internet, a partir de dois tipos de fontes: jornais e blogs, conforme a tabela abaixo.[13]

Fonte	Nº de artigos	Tokens	Período
Folha (jornal)	19.837	6.939.846	2013-2014
Correio do Povo (jornal)	6.908	1.919.491	2013-2014
Lola (blog)	3.509	3.050.703	1998-2014
Sakamoto (blog)	923	580.101	2006-2014
Cisco (blog)	2.604	423.240	2003-2014

Tabela 3 – Número de artigos, número de tokens e período compreendido pela coleta, para cada fonte.

[13] Endereços dos blogs pesquisados:
`http://escrevalolaescreva.blogspot.com.br/`
`http://blogdosakamoto.blogosfera.uol.com.br/`
`http://ciscocosta.com/filisteu/`

Embora todos esses textos sejam representativos da língua escrita culta, pois são retirados de jornais de grande circulação ou de blogs produzidos por escritores com ensino superior completo, há uma diferença de registro entre esses dois meios. Blogs costumam ser mais informais e ter pouco ou nenhum controle editorial. A seleção de textos de registros diferentes nos parece ser uma consideração metodológica importante em vista das considerações feitas anteriormente sobre a necessidade de que itens de baixa frequência sejam identificáveis no corpus. Tendo isso em mente, devemos evitar situações em que a raridade de uma construção seja mero artefato da seleção dos textos que compõem o corpus. Pode ser o caso, por exemplo, que uma construção como *bateção* seja rara dentro de um corpus simplesmente por ele ser composto por textos de um registro formal; em outros usos da língua, essa construção pode ser mais comum. Da mesma forma, construções formais podem aparecer com raridade em registros informais e serem, por isso, confundidas com expressões novas, no conjunto de itens de baixa frequência.

Apesar dessas considerações, não buscamos neste estudo nos aprofundar no estudo da influência de diferentes registros, ou de outras variáveis sociolinguísticas, sobre a produtividade morfológica. Fazemos, na discussão que segue, algumas alusões breves a essas influências, mas um estudo adequado desses fatores exige que enriqueçamos nosso corpus com uma maior variedade de textos estratificados de acordo com as categorias relevantes. De qualquer sorte, na construção do nosso banco de dados, tomamos o cuidado de manter informações sobre as fontes de cada texto, de modo a facilitar estudos futuros que abordem essas questões.[14]

4.1.1 Coleta

Na coleta dos textos que compuseram nosso corpus, utilizamos a plataforma Scrapy versão 0.14.1 (http://scrapy.org), que possibilita a criação de scripts de extração automática de dados de páginas da internet, utilizando a linguagem de programação Python. De cada um dos sites, foram coletados o título, o corpo e a data de publicação de cada artigo encontrado. No caso dos blogs, todos os artigos publicados até o momento da coleta foram incluídos. No caso dos jornais, buscamos incluir um número variado de seções, a partir das quais o script é capaz de encontrar os artigos disponíveis em cada site. Os textos foram salvos em formato JSON para processamento posterior. Segue, como exemplo, um trecho de um artigo coletado, no formato utilizado.

[14] Do ponto de vista técnico, o uso de um banco de dados SQLite nos permite estender o corpus posteriormente, de acordo com a necessidade, com anotações em diversos níveis de análise.

```
1 {
2     "hash": 1954311185683236400,
3     "url": "http://blogdosakamoto.blogosfera.uol.com.br/2012/10/30/troco-um-feriado-
          cristao-por-uma-pausa-pelo-calor/",
4     "aut": "Leonardo Sakamoto",
5     "titulo": [
6         "Troco um feriado cristão por uma pausa pelo calor"
7     ],
8     "corpo": [
9         " ",
10        "Durante as eleições, discutiu-se tanto sobre Deus e o Diabo na capital
              paulista que o chão se abriu e o capeta montou uma sucursal por aqui.
              Está quente, muito quente.",
11        "Sugiro trocar um dos vários feriados cristãos deste país laico pela
              possibilidade do poder público decretar uma parada obrigatória em dias
              irritantemente quentes e/ou poluídos. Um feriado religioso faz bem à
              alma de dezenas de milhares de fieis dedicados. A ideia que apresento
              faria um bem enorme ao corpo dos mais de 11 milhões de moradores de um
              município como São Paulo, os que crêem e os que não.",
12
13                ...
14
15    ],
16    "data": [
17        "30/10/2012"
18    ]
19 }
```

A entrada “hash” que aparece na linha 2, acima, é um número computado a partir do texto do artigo, por meio de uma implementação em Python do algoritmo simhash, desenvolvido por Charikar (2002). Esse número foi útil para permitir uma detecção eficiente de artigos duplicados na coleta. Eventualmente, os scripts de extração automática de dados construídos a partir do Scrapy podem retornar textos duplicados, pois o mesmo artigo pode aparecer mais de uma vez dentro de um site, sob endereços diferentes; daí a necessidade de essas repetições serem detectadas e eliminadas. Entretanto, a comparação entre textos em larga escala é bastante custosa computacionalmente. Algoritmos hash permitem contornar essa dificuldade por meio do cálculo de um número com base no conteúdo de cada texto, que pode ser rapidamente comparado com os demais. A vantagem específica do simhash é que o número de cada texto não é necessariamente único. Textos muito similares podem receber valores iguais ou similares; desse modo, mesmo que dois artigos, com endereços diferentes, diferenciem-se na formatação, na capitalização, ou mesmo em parâmetros textuais pequenos, é possível detectá-los como sendo instâncias do mesmo artigo.

4.2 Processamento dos *corpora*

4.2.1 Tokenização

O primeiro passo no processamento de cada texto foi a tokenização, que consiste em separá-lo em palavras, considerando cada ocorrência de uma palavra como um *token* distinto. Para os fins deste trabalho, a tokenização é aplicada apenas ao conteúdo da chave 'corpo' de cada artigo. A chave 'titulo' é ignorada, seguindo a opinião de Bauer (2001), no que diz respeito ao uso de *corpora* em estudos de produtividade morfológica, que considera que essa parte de um artigo costuma ser pensada de forma a atrair a atenção dos leitores e, portanto, está mais propensa a conter formações criativas que não necessariamente caracterizam a noção intuitiva de produtividade morfológica como criação lexical não intencional por meio de recursos morfológicos.

O procedimento de tokenização utilizado neste trabalho (linhas 68-70 do Apêndice A.1) deixa todas as palavras em letras minúsculas, a fim de não superestimar o número de palavras únicas devido a diferenças de capitalização, e divide os textos em sequências de caracteres alfanuméricos que não sejam separados por espaços, permitindo a existência de hífens, para capturar palavras prefixadas e compostas grafadas com hífen.

Uma vez identificadas essas sequências, elas são submetidas a uma função removedora de sufixos adaptada do algoritmo RSLP (Removedor de Sufixos da Língua Portuguesa) proposto por Orengo e Huyck (2001). Originalmente, esse algoritmo remove de uma palavra todos os seus sufixos, flexionais ou derivacionais. Para esta coleta de dados, no entanto, o algoritmo foi modificado, de modo a retirar de uma palavra apenas seus sufixos flexionais, visto que não buscamos depreender a raiz primária das palavras, mas apenas identificar variantes flexionais de um lexema como pertencentes ao mesmo tipo. Nosso objetivo com esse algoritmo também não foi o de chegar a uma análise morfológica linguisticamente correta em todos os casos, mas simplesmente o de resolver de forma automática os casos que eram de interesse para esta pesquisa. Por exemplo, "destruição" e "destruições" deveriam sempre ser reconhecidas como variantes do tipo *destruição*. A implementação desse algoritmo encontra-se no Apêndice A.2, na classe RSLPStemmer.

Não apenas as variantes flexionais deveriam ser consideradas como pertencentes ao mesmo tipo, mas também as variantes ortográficas. Assim, "destrução" também deveria ser considerada como um *token* de *destruição*. Portanto, o script tokenizador consulta uma lista de substituições criada para esse fim. A lista, em formato JSON contém entradas do seguinte tipo:

```
"armazanamento": "armazenamento",
"desncontentamento": "descontentamento",
"atentimento": "atendimento",
"impalamento": "empalamento",
"pró-armamento": "armamento",
"prontoatendimento": "atendimento",
"ecodesenvolvimento": "desenvolvimento",
```

Sempre que uma das formas à esquerda é encontrada no corpus, o script a considera como sendo um *token* da expressão à direita.

Pode-se notar nas entradas acima que alguns casos de prefixação e de composição, como *pró-armamento* e *prontoatendimento*, também foram considerados como tokens de outras palavras. Isso pode parecer contraintuitivo, já que uma ocorrência de *pró-armamento* em um texto é normalmente reconhecida como uma instância da palavra *pró-armamento*, não da palavra *armamento*, por um usuário de língua portuguesa. Todavia, a relação tipo-*token* que nos interessa neste trabalho não é esta. Em vez dela, nos interessa a relação que existe entre o tipo *armamento* e seus tokens. Ou seja, entre o objeto morfológico formado pela junção da base *armar* com o sufixo *mento* e suas ocorrências, mesmo que elas se manifestem dentro de objetos morfológicos maiores, formados por afixações ou composições subsequentes. Considerar *pró-armamento* como um tipo independente teria como resultado a contabilização de mais um provável *hapax legomenon* na categoria X-mento. Esse resultado é indesejado, pois superestimaria a proporção de *hapax legomena* para esse afixo, sem que isso representasse sua afixação a uma nova base; em vez, disso, *pró-armamento* é muito mais um indício de produtividade do prefixo *pró-* que do sufixo *mento*.

Note-se que um cuidado metodológico sempre presente neste trabalho é o de buscar não superestimar nem o número de palavras que instanciam um padrão morfológico, nem o número de *hapax legomena* entre essas palavras. É para isso que serve o controle das variantes ortográficas e flexionais, e das afixações ou composições externas às que caracterizam a classe morfológica estudada. Na ausência desse cuidado, não seria possível dizer que um resultado obtido no cálculo do índice $\mathcal{P}$ reflete a produtividade de um padrão morfológico, pois ele poderia ser causado pela interveniência desses outros fatores.

Por fim, cada um dos tokens é salvo em um banco de dados SQLite, juntamente com uma referência ao texto em que foi encontrado (por meio do valor hash), à sua posição nesse texto, e ao tipo que instancia, conforme definição apresentada no Apêndice A.3.

O passo seguinte na coleta de dados consiste em coletar as palavras que exemplificam o padrão morfológico de interesse. Isto é feito por meio do script apresentado

no A.4, que toma como input o nome de um afixo e retorna uma lista de palavras que o contêm. Para fazer essa busca, o algoritmo encontra a expressão regular correspondente ao sufixo desejado no arquivo tools.py (Apêndice A.5) e compara cada palavra do banco de dados com essa expressão. As palavras encontradas são organizadas em uma lista de frequência e salvas em um arquivo de texto.

4.3 Revisão da coleta

O processamento automático de textos escritos dificilmente pode ser suficiente para uma boa análise linguística. Como mostram Evert e Lüdeling (2001), em um estudo de sufixos do alemão, os métodos estatísticos disponíveis para a análise da produtividade são altamente sensíveis a diversos fatores, como a existência de erros ortográficos, de palavras que acidentalmente terminem com uma sequência estudada (e.g. *alimento* em um estudo sobre *-mento*), etc. Diante de problemas desse tipo, não há métodos de processamento automático que sejam sofisticados a ponto de tornar dispensável qualquer correção manual dos dados.

Nos casos de variantes gráficas, como *jugalmento* para *julgamento*, ou *rankeamento* para *ranqueamento*, devemos contar cada um desses pares como uma palavra só. Caso ignorássemos esse problema, o número de *hapax legomena* seria bastante superestimado, visto que ocorrências específicas de grafias divergentes costumam ter frequências muito baixas em um corpus (*jugalmento*, por exemplo, poderia ocorrer apenas uma vez, mesmo dentro de um *corpus* extenso).

No caso de palavras que apenas coincidentemente apresentam as sequências gráficas características dos padrões morfológicos estudados (e.g. *aumento*), o risco maior é de que se infle o número total de tokens para um padrão morfológico, contabilizando palavras de alta frequência que não o instanciam de fato. Isso teria como resultado diminuir indevidamente o índice de produtividade $\mathcal{P}$ para esse padrão morfológico, visto que o número total de tokens é o denominador no cálculo desse índice.

Como mencionado na seção anterior, utilizamos listas de substituição para dar conta das variantes gráficas e de listas de exclusão para dar conta das palavras que não são analisáveis pelo padrão estudado. No entanto, a própria necessidade de essas listas existirem mostra que o problema não pôde ser resolvido de forma automática. Para alimentar essas listas de substituição, foram necessárias inspeções manuais das listas de palavras de cada afixo. Em alguns casos, foi necessário observar a palavra em seu contexto, a fim de verificar se realmente se tratava de um erro de grafia ou de uma palavra distinta. Por exemplo, *secretamento* poderia ser

tanto um nome derivado do verbo *secretar* quanto uma ocorrência do advérbio *secretamente*, grafada de forma incorreta. Neste caso, a segunda alternativa se mostrou verdadeira.

Os casos que exigiram mais atenção foram os que envolveram decisões sobre a estrutura morfológica de palavras, quando a transparência de um formativo não é clara, ou quando é preciso identificar a ordem em que múltiplas afixações se aplicaram em uma formação. Utilizamos, como critério básico para exclusão das palavras obtidas pelo procedimento descrito na seção anterior, a inexistência da base como uma palavra independente e semanticamente relacionada. Houve tolerância em relação a irregularidades fonológicas quando a relação semântica entre a base e o derivado era clara (e.g. *eleição*).

No caso de múltiplas afixações, consideramos apenas os casos em que nos parece claro que o afixo buscado encontra-se na última camada de derivação (e.g.). Em contraste, há palavras como *pró-traição*, em que, claramente, o prefixo tem escopo sobre *traição*, não sobre o verbo de base, *trair*. Nesses casos, optamos por utilizar a lista de substituições, definindo regras de reescrita como *pró-traição* $\rightarrow$ *traição*. Assim, este *token* foi contabilizado como uma instância do tipo *traição*. A alternativa de considerá-lo como instância de um novo tipo, *pró-traição*, como discutido anteriormente, inflaria o número de *hapax legomena* do sufixo *-ção* de forma espúria.

Encerrando esta seção, é importante mencionarmos que a resolução desses casos é feita de forma particularizada, e ainda que se sigam os critérios expostos, nunca é possível saber se conseguimos eliminar todos os vieses que podem influenciar a seleção de um conjunto de dados por um analista. Se não podemos aspirar a essa certeza, podemos ao menos buscar o maior grau possível de transparência. Tendo isso em mente, os códigos-fonte de todos os scripts, as listas de substituições e as listas de vocabulário utilizadas nesta pesquisa podem ser encontradas nos apêndices deste trabalho e/ou no seguinte repositório virtual: `https://github.com/shoeki/ling`.

5 Mudança e estabilidade na produtividade morfológica: *-ção* x *-mento*

Na Seção 2.3, vimos que o sufixo *-ment* do inglês deixou de ser produtivo nessa língua, perdendo espaço para o sufixo *-ation* a partir do século XVII. Essa perda gradual de produtividade levou à situação atual, em que o sufixo *-ment* não é mais utilizado (senão marginalmente) na formação de palavras novas do inglês. Ao longo dos séculos, o que se observa entre esses dois afixos é uma situação de competição pelo nicho de formação de substantivos abstratos a partir de verbos. Essa competição teve como resultado uma mudança na língua inglesa que pode ser descrita em termos de produtividade: um afixo que era produtivo deixou, gradualmente, de sê-lo, ao passo que outro afixo, que era pouco produtivo, passou a ser o padrão dominante de nominalização. Vimos também que, contrariamente à opinião de Bauer (2001), essa mudança pode ser explicada, pelo menos em parte, pelo contexto em que essa competição se desenrolou. Especificamente, Lindsay e Aronoff (2013) sugerem que o grande número de empréstimos em *-ation* introduzidos na língua antes do século XVII garantiu a esse afixo uma ampla base de suporte para generalização, o que se mostrou crucial em um período em que o número de verbos novos na língua, passíveis de sofrer nominalização, era escasso.

Os sufixos cognatos de *-ment* e *-ation* no português, *-mento* e *ção*, também se encontram em competição há séculos. Ambos podem formar substantivos abstratos a partir de verbos, tendo contextos de aplicação, em grande parte, coincidentes. Isso significa que, para muitos verbos, poderíamos esperar tanto um substantivo em *-ção* quanto um em *-mento*, e, às vezes, podemos encontrar as duas opções atestadas na língua. É assim, por exemplo, que falantes do sul do Brasil costumam falar em *alagamento*, ao passo que, em variedades da região Norte, costuma-se falar em *alagação*, com o mesmo sentido, a partir do verbo *alagar*.

(10) O Rio dos Sinos transbordou e já atinge a pista lateral da BR-116 entre Sapucaia do Sul e Esteio. O **alagamento** acontece nos dois sentidos. (`http://gaucha.clicrbs.com.br/rs/noticia-aberta/alagamento-em-pista-lateral-provoca-congestionamento-de-3-km-na-br-116-entre-sapucaia-e-esteio-9994.html` - Acesso: 08/10/2014)

(11) “Sendo janeiro um mês muito chuvoso e considerando também que começa a prevalecer o sistema característico de concentração de umidade na atmosfera, afetando diretamente o Estado, é praticamente certo que vai haver uma

> **alagação** nos próximos dias", disse Alejandro. (`http://www.ufac.br/portal/news/segundo-pesquisador-da-ufac-chuvas-de-janeiro-sinalizam-para-enchente-do-rio-acre` - Acesso: 08/10/2014)

É interessante notar, a respeito desse exemplo, que as duas formas, *alagamento* e *alagação*, estão disponíveis para os falantes das duas variedades do português, e ambas as frases são plenamente aceitáveis para os dois grupos de falantes. A diferença entre as duas variedades, quanto a essa questão, está apenas em qual das alternativas de nominalização se institucionalizou em cada comunidade linguística. Esse exemplo indica que a competição entre esses dois padrões morfológicos não é completamente decidida por fatores linguísticos, de modo que falantes de localidades geográficas distintas podem selecionar opções diferentes entre as disponibilizadas pela gramática. Além disso, mesmo dentro de uma única variedade linguística é possível encontrar pares coexistentes; Sandmann (1988) cita, entre outros, *indiciação*/*indiciamento* e *formigamento*/*formigação*, este último encontrado em seu corpus de textos jornalísticos.

Diferentemente do que ocorre no caso de seus cognatos no inglês, os sufixos *-ção* e *-mento* continuam sendo ambos produtivos no português. Porém, assim como no inglês, diversos estudos sugerem que *-ção* é consideravelmente mais produtivo que *-mento* (Basilio, 2008; Rocha, 1999; Silveira, 2015). Este resultado é corroborado por um levantamento em nosso corpus, como podemos ver na Tabela 4.

Sufixo	Palavras (V)	*Hapax Legomena* (n_1)	Índice de produtividade $\mathcal{P}$
-ção	1.359	162	0,00187
-mento	586	77	0,00089

Tabela 4 – Produtividade dos sufixos *-ção* e *-mento* (tamanho da amostra de cada afixo: 86.653 *tokens*).

Como vimos na Seção 2.2, o índice $\mathcal{P}$ expressa a probabilidade de que novas atestações de um sufixo sejam de palavras novas, isto é, que não foram observadas até então durante a amostragem ($\mathcal{P} = n_1/N$). Assim, após 86.653 *tokens* contendo o sufixo *-mento* terem sido encontrados no corpus (correspondentes a 586 palavras distintas), a probabilidade de que uma nova forma contendo esse sufixo seja um *hapax legomena* é de 0,089%. Com o mesmo número de *tokens* observados, a probabilidade de que uma nova observação do sufixo *-ção* corresponda a um *hapax legomenon* é mais de duas vezes maior: 0,187%.

Contudo, ainda que se mostre menos produtivo, *-mento* continua sendo uma fonte estável de formação de novas palavras, diferentemente de seu cognato no in-

glês. Essa situação se manteve mesmo após *-ção* ter se tornado o afixo nominalizador predominante no português no século XVII. No Gráfico 6, podemos ver o número de novas atestações de palavras com esses sufixos em cada período da história da língua, de acordo com as datações disponíveis no Dicionário Houaiss (versão eletrônica 3.0).[15]

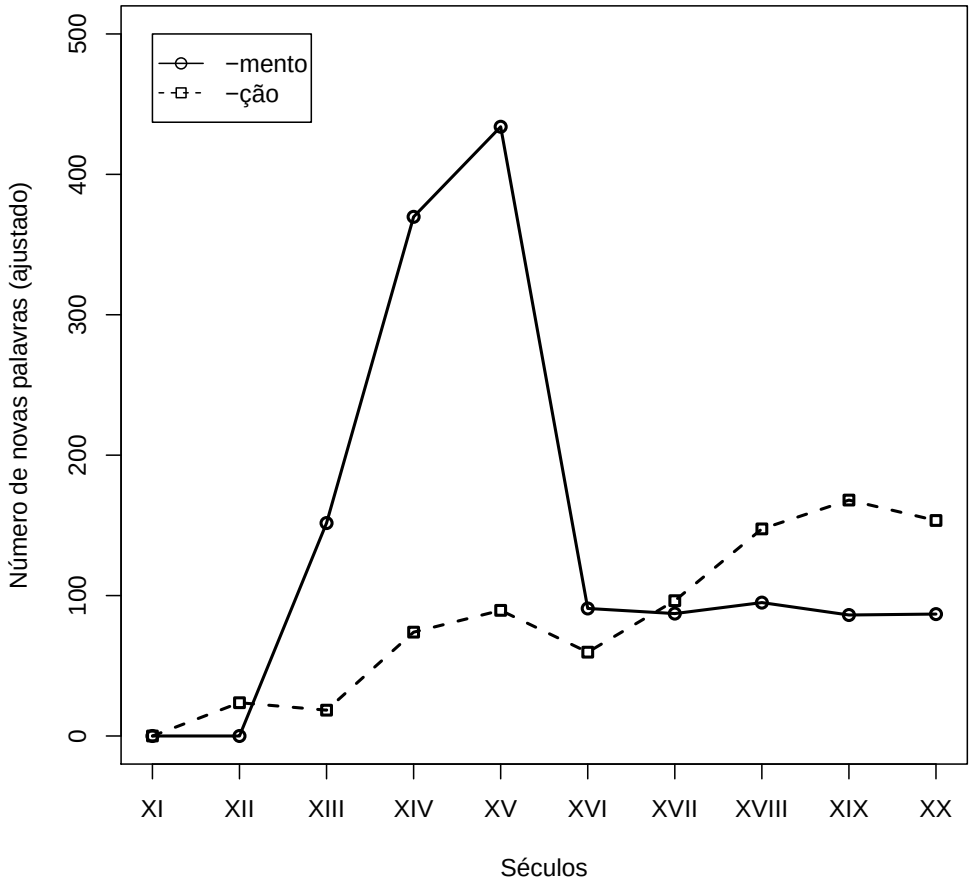

Gráfico 6: Número de palavras derivadas com *-mento* e *-ção* na língua portuguesa, por século.

O número de novas palavras por século foi ajustado, a fim de dar conta da variação no número geral de atestações por período no dicionário. Por exemplo, o século XIX contém o maior número bruto de novas palavras atestadas com os sufixos *-ção* e *-mento*, tanto entre os empréstimos quanto entre as palavras formadas dentro da língua portuguesa. Porém, o século XIX também é aquele que tem o maior número de palavras novas atestadas (N = 27735) de forma geral. Portanto, consideramos no

[15] Incluímos apenas palavras que o dicionário indica terem sido formadas no português pela adição desses sufixos, excluindo, portanto, palavras que não foram registradas como derivadas, que apenas terminam com as sequências gráficas <ção> ou <mento>, bem como palavras herdadas do latim ou empréstimos posteriores.

gráfico acima um número ajustado de atestações, de acordo com a seguinte fórmula: N_{ajustado} = (número de atestações com o sufixo no período / número total de atestações no período) $\times$ 10^4. O mesmo tipo de ajuste é utilizado por Lindsay e Aronoff (2013).

Vemos no Gráfico 6 que o sufixo *-ção* passou a ser a forma predominante de nominalização do português pelo menos a partir do século XVII. Assim como na história contada por Lindsay e Aronoff (2013) sobre a generalização de *-ation* no inglês, é possível que *-ção* tenha encontrado suporte para sua expansão na grande base de empréstimos que entraram no português em séculos anteriores e que puderam ser reanalisados, por gerações posteriores, como palavras formadas por este afixo. Como podemos ver no Gráfico 7, o número de empréstimos contendo *-ção* que adentravam o léxico do português era consideravelmente maior do que o de empréstimos com *-mento*, uma situação que se manteve por um longo período.[16]

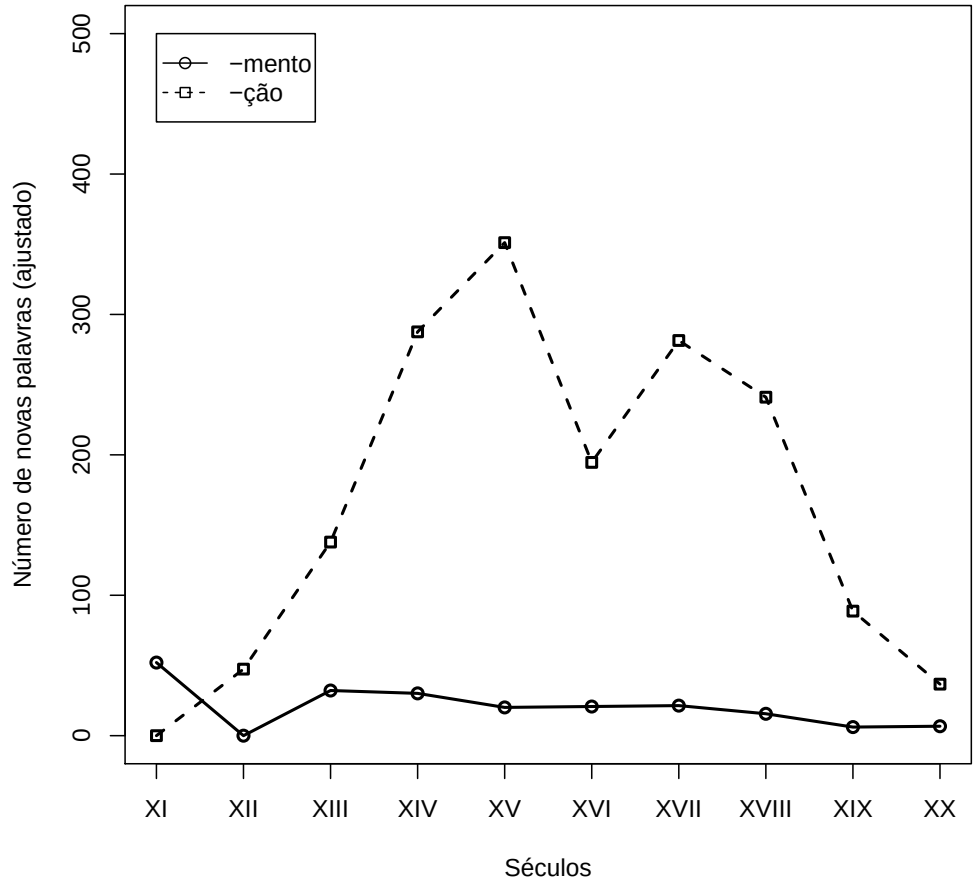

Gráfico 7: Número de empréstimos contendo *-ção* e *-mento*, por século.

Essa explicação, por si só, não é suficiente para a história do português, pois o

[16] O número de empréstimos para cada período também foi computado a partir das datações e indicações etimológicas do Dicionário Houaiss.

padrão de nominalização em *-mento*, de acordo com os dados do dicionário Houaiss, também tinha um grande suporte de formações já estabelecidas no século XVII, graças a sua grande produtividade em períodos anteriores. Como podemos ver no Gráfico 6, *-mento* dera origem, antes do século XVI, a um número de palavras derivadas ligeiramente maior até do que o de empréstimos com *-ção*. Entre outros fatores que podem ter estado em jogo para garantir a predominância de *-ção* está um contexto sociolinguístico favorável, dado o prestígio com que contavam os empréstimos de origem latina entre os falantes de português da época (Teyssier, 1982).

Esses fatores merecem uma investigação mais aprofundada em outros trabalhos. Por ora, nos concentraremos na seguinte questão: por que, mesmo com o predomínio de *-ção* a partir do século XVII, *-mento* continuou sendo um sufixo produtivo no português? Note-se que este afixo mantém certa estabilidade na língua, com uma taxa de novas atestações mais ou menos constante desde que *-ção* se tornou o padrão de nominalização predominante. Note-se, ainda, que essa situação contrasta com a da língua inglesa, em que o sufixo cognato *-ment* teve sua produtividade gradualmente reduzida, até perdê-la por completo, diante da ascensão de *-ation*.

Essa é uma versão específica de uma questão geral sobre a competição morfológica, qual seja, em que condições um padrão de formação de palavras pode sobreviver em uma língua, estando em concorrência com um padrão mais geral? Poderia ser o caso que as línguas simplesmente tolerassem padrões morfológicos com a mesma função, e que falantes pudessem escolher livremente entre eles na formação de uma nova palavra. Entretanto, como vimos no Capítulo 3, há uma forte tendência de que haja bloqueio nesses casos, ainda que pares de sinônimos possam existir ocasionalmente. Dessa forma, o uso de um padrão morfológico tende a ser restringido pelo uso prévio de algum padrão concorrente. Essa tendência pode levar à perda de produtividade de um dos rivais e, até, ao seu desaparecimento. Sendo assim, faz-se necessário explicar as situações em que há estabilidade dos padrões em competição.

5.1 Ilhas de confiabilidade no léxico do português

Vimos anteriormente que antes de perder seu status como principal padrão de nominalização em português, o sufixo *-mento* já havia dado origem a um grande número de formações. Assim como uma vasta base de empréstimos parece ter dado suporte à generalização de *-ção* na língua, podemos imaginar que a base de derivados em *-mento* tenha também oferecido algum suporte a usos futuros desse afixo. Nossa hipótese a esse respeito é de que gerações subsequentes puderam encontrar,

nesse conjunto de formações antigas, nichos lexicais em que a afixação de *-mento* era particularmente predominante. Lindsay e Aronoff (2013) mostram que esta é uma opção para a resolução de situações de competição em sistemas morfológicos, além da possibilidade de perda completa de produtividade de um dos padrões rivais. Um dos casos estudados pelos autores, em que se criou uma distribuição quase complementar como resolução de uma rivalidade, é o do verbalizador *-ify*, do inglês, que sobreviveu diante do afixo mais produtivo *-ize*, encontrando um nicho definido fonologicamente, em torno de radicais monossilábicos (e.g. *tube* 'tubo' - *tubify*, **tubize* 'tubificar'; *random* 'aleatório' - **randomify*, *randomize* 'aleatorizar').[17]

A fim de investigar a existência de nichos que tenham favorecido a afixação de *-mento*, garantindo sua estabilidade, empregamos o procedimento de indução de regras introduzido na Seção 3.2. Como vimos, trata-se de um aprendiz artificial que busca um conjunto de regras capazes de dar conta do mapeamento entre formas relacionadas presentes em uma lista de treinamento. O conjunto de regras depreendido pelo aprendiz diferencia-se do tipo de gramática tradicionalmente assumido nas teorias linguísticas por conter, normalmente, mais de uma regra capaz de mapear duas formas relacionadas. A cada uma dessas regras é atribuído um valor de confiabilidade, definido pela razão entre o número de casos em que a regra se aplica de fato e o número de casos em que ela poderia se aplicar em princípio. Isso garante que regras bastante específicas possam ser preferidas às mais gerais capazes de dar conta de um mesmo mapeamento, desde que as regras específicas sejam significativamente mais robustas em sua aplicação. Com isso, esperamos detectar nichos fonológicos em que esses afixos sejam particularmente dominantes.[18]

A importância de se utilizar um algoritmo de aprendizagem nessa detecção decorre da hipótese de que a distribuição desses afixos no português atual foi determinada, em grande parte, pelos procedimentos empregados na aquisição do sistema morfológico do português pelas últimas gerações de usuários da língua. De forma mais específica, ao empregarmos o *Minimal Generalization Learner* (MGL) de Albright e Hayes (1999), consideramos a hipótese de que esses procedimentos

[17] Como *-ize* e *-ify* possuem significados equivalentes e uma distribuição (quase) complementar, Plag (2000) chega a considerá-los como alomorfes em uma relação supletiva atualmente.

[18] Sendo esta análise apenas um exercício de aplicação do modelo de Albright e Hayes (1999) à morfologia derivacional, não pretendemos realizar aqui uma análise apreciável dos condicionamentos prosódicos, morfossintáticos e semânticos a que esses afixos podem estar sujeitos. Além de não ser este nosso objetivo, a consideração desses fatores envolveria algumas complicações técnicas, ainda que não insuperáveis, pois a implementação disponível do MGL opera apenas sobre representações segmentais. De qualquer sorte, como a gama de significados disponível a esses sufixos é basicamente a mesma, nos parece razoável, para um estudo quantitativo preliminar, assumir que esses padrões concorrentes são sinônimos e avaliar a hipótese de que suas distribuições podem ser previstas, em grande medida, por parâmetros segmentais; ainda que essa hipótese vá necessitar, possivelmente, de revisões futuras.

sejam sensíveis a subregularidades presentes no léxico, e de que estas, quando suficientemente robustas, podem ter preferência sobre regras mais gerais. Se esta suposição estiver correta, a gramática obtida pelo MGL deve ser capaz de prever a forma de palavras novas, com base nessas subregularidades.

5.1.1 Procedimentos metodológicos

Nosso primeiro passo na aplicação do MGL aos dados de nominalização do português foi uma primeira rodada do aprendiz sobre uma lista de treinamento contendo 1.919 nominalizações com suas respectivas bases verbais. Essa lista é composta pelos nomes afixados por *-mento* ou *-ção* atestados pelo Dicionário Houaiss 3.0, excluindo empréstimos. Os pares de verbo e nominalização assim obtidos foram transcritos por um script de conversão grafema-fonema desenvolvido para este trabalho (Apêndice A.6), a que também se seguiu uma checagem manual. Cada segmento dos dados transcritos é associado pelo MGL a uma matriz de traços fonológicos.[19]

A partir dessa lista, o aprendiz é capaz de induzir um conjunto de regras que descrevem contextos fonológicos em que as operações de adição de *-mento* ou *-ção* são aplicáveis, além de computar, para cada uma delas, a razão entre o número de acertos e o de bases que se encaixam no escopo da regra. Os contextos em que essa razão é especialmente alta são chamados de "ilhas de confiabilidade", pois é neles que uma regra se mostra especialmente previsível, com menos exceções do que as regras mais gerais.

O modelo formado por esse conjunto de regras, com seus valores de confiabilidade, pode então ser aplicado a bases que não estavam disponíveis nos dados de treinamento, a fim de verificar o quão bem as escolhas feitas pelo modelo correspondem às de falantes reais. Com isso, podemos testar a hipótese de que, ao escolherem entre *-mento* e *-ção*, falantes de português se valem das ilhas de confiabilidade depreendidas por esse aprendiz artificial.

Os dados de teste provêm do corpus descrito na Seção 4.1. Deste corpus, coletamos palavras nominalizadas contendo os afixos *-ção* e *-mento* que não constam da versão do Dicionário Houaiss considerada – e que, portanto, não fizeram parte dos dados de treinamento do aprendiz artificial. Todas as palavras passaram por correção ortográfica automática, com posterior checagem manual.

[19] O objetivo principal desta transcrição foi eliminar as inconsistências da relação grafema-fonema do português. Dessa forma, não nos comprometemos com propostas específicas sobre o inventário fonológico do português e realizamos uma transcrição relativamente superficial, mantendo, por exemplo, semivogais e vogais nasais como segmentos simples. Exemplos de formas transcritas a partir deste script podem se encontrados no Apêndice B.3, que contém um dos outputs da rodada do MGL.

Além das palavras que já constavam da lista de treinamento, também foram excluídas as palavras prefixadas cuja base verbal primária fizesse parte dessa lista, como *desatualização* (pois *atualização* está no Dicionário Houaiss, e, portanto, *atualizar* está na lista de treinamento), visto que nosso teste consiste em avaliar o desempenho do MGL diante de bases que não haviam sido encontradas antes por esse aprendiz. Além disso, foram excluídas palavras prefixadas cuja base também fosse atestada no corpus; por exemplo, *desautomatização* foi excluída, porque *automatização* também foi encontrada no corpus. O motivo desta exclusão é que a base verbal primária *automatizar* já é incluída uma vez na lista de teste, tornando redundante incluir *desautomatizar*, pois ambas têm uma composição fonológica idêntica nas proximidades do sufixo nominalizador. Além disso, esses verbos têm o mesmo núcleo em sua estrutura morfológica, já que prefixos, no português, não exercem esta função. Como a seleção morfológica costuma ser governada pelo núcleo da base, é esperado que esses verbos selecionem o mesmo sufixo. Para fins de análise estatística, contudo, é importante que as observações das escolhas feitas pelos falantes sejam independentes umas das outras.

Após essas exclusões, computamos as bases verbais das nominalizações encontradas, utilizando uma adaptação do algoritmo de stemização desenvolvido para o português por Orengo e Huyck (2001). Na implementação destes autores, o algoritmo encontra, para cada palavra, um radical, que não precisa ser ele mesmo uma palavra, por meio da retirada de todos os sufixos. Para *comprometimento*, por exemplo, o output da stemização seria *compromet*. Em nosso trabalho, no entanto, interessa-nos encontrar o verbo que serve de base para cada nominalização. Por isso, modificamos o algoritmo de modo a manter intacta a vogal que precede o sufixo, já que, com essa informação, nossa versão do programa consegue prever a terminação correta da base verbal na maioria dos casos. O resultado dessa etapa, contudo, exigiu uma checagem manual, visto que nem sempre é possível determinar se o verbo base é de segunda ou de terceira conjugação com base na vogal que precede o sufixo da forma nominalizada. Isso é demonstrado pelo exemplo de *comprometimento*, cuja base verbal é de segunda conjugação, embora a vogal que precede o sufixo seja [i] (cf. *preterimento*, cuja base verbal é de terceira conjugação).

Finalmente, essas bases foram fonemicamente transcritas, compondo uma lista de teste a ser submetida ao modelo depreendido pelo MGL a partir da lista de treinamento. Neste ponto, o MGL age como um mecanismo de produção, tentando encontrar, para cada base, uma regra que gere um output em *-mento* e outra que gere um output em *-ção*. As regras selecionadas são as que possuem o maior escore de confiabilidade dentre as que têm sua descrição estrutural compatível com a base

testada.

5.1.2 A gramática prevista pelo MGL

O aprendiz foi capaz de induzir 8.453 regras a partir da lista de treinamento, sendo a maioria delas, é claro, generalizações muito específicas e/ou pouco robustas. No entanto, algumas regras bastante interessantes foram encontradas, com altos valores de confiabilidade, o que é relevante para os objetivos de nossa investigação. Na Tabela 5, podemos ver as regras mais robustas encontradas pelo modelo (com valores de confiabilidade maiores que 0,75) para a geração de formas em *-mento*. Reportamos nesta tabela apenas regras robustas que se mostraram relevantes para pelo menos uma das bases da lista de teste. Para cada regra apresentada, trazemos um exemplo de predição do modelo sobre uma base da lista de teste na primeira coluna. O símbolo ☹ indica outputs que foram previstos pelo modelo, mas divergem do que foi atestado no corpus.

O fato de o aprendiz ter induzido regras com escore bastante alto a partir dos dados de treinamento confere suporte inicial a nossa hipótese de que há ilhas de confiabilidade, isto é, subgeneralizações bastante robustas no léxico do português, para a produção dessas formas nominalizadas. A mais confiável entre essas regras, que responde pela formação de palavras como *pertencimento*, descreve uma generalização já mencionada na literatura, ainda que de forma mais restrita. Rocha (1999) comenta que bases terminadas em *-ecer* tendem a se combinar com o afixo *-mento*. Interessantemente, o autor reconhece que não se trata apenas de seleção morfológica, pois bases em que a sequência *ecer* não é afixal também estão sujeitas a essa combinação. O que podemos ver em nossos dados, contudo, é que essa generalização também não parece ser restrita a essa sequência, como mostra o caso de *pertencimento*. Na lista de treinamento, a afixação de *mento* também se mostrou bastante previsível sobre qualquer raiz de segunda conjugação terminada em [ʃ] (e.g. *enchimento*) ou [x] (e.g. *socorrimento*); por isso, a regra depreendida pelo MGL abrange estes contextos.

Também podemos ver na Tabela 5 outras subgeneralizações quase tão confiáveis quanto a primeira. Talvez as mais importantes delas, para nosso propósito de entender a estabilidade de *-mento*, sejam as regras relacionando verbos da primeira conjugação a este padrão de nominalização (marcadas em cinza). Elas são importantes, porque a maioria dos novos verbos do português são de primeira conjugação; a segunda e a terceira são apenas marginalmente produtivas, dependendo da prefixação de verbos já existentes para sua renovação, ou da formação de verbos em *-ecer*, no caso da segunda. Assim, o sufixo *-mento* poderia não ter tido um fluxo de

bases novas para promover sua estabilidade, caso fosse restrito a verbos de segunda e terceira conjugação. As ilhas de confiabilidade para este sufixo entre verbos de primeira conjugação garantem-lhe fontes mais seguras de novas nominalizações, fora do universo fechado de raízes das demais conjugações.

Exemplo	Regra[20]	Acertos/Escopo	Conf.
pertencimento	eɾ → imẽto / [X {ʃ, s, x} ____]$_N$	84/85	.959
carvoejamento	ɾ → mẽto / [X {a, e, o, ã, ẽ, õ} ʒa ____]$_N$	26/26	.936
☹ fervimento	eɾ → imẽto / [X [1-2]abertura, +solt. ret., −nasal ____]$_N$	115/119	.934
erguimento	eɾ → imẽto / [X −silábico, −nasal, −labial, −arr. ____]$_N$	140/146	.928
engessamento	ɾ → mẽto / [X {e, o} sa ____]$_N$	14/14	.882
☹ zoamento	ɾ → mẽto / [X {ʎ, ʃ, ʒ, l, s, x, z} oa ____]$_N$	13/13	.873
guinchamento	ɾ → mẽto / [X {ʃ, ʒ} a ____]$_N$	46/49	.869
amarelamento	ɾ → mẽto / [X {tʃ, t, ʃ, ʒ, d, ɾ, s, t, z} ela ____]$_N$	11/11	.849
tensionamento	ɾ → mẽto / [X {a, e, i, j, l, ʎ, ɾ} ona ____]$_N$	17/18	.810
embaralhamento	ɾ → mẽto / [X {ʎ, ʃ, ʒ, ŋ} a ____]$_N$	94/104	.804
encoleiramento	ɾ → mẽto / [X [-nasal] eira ____]$_N$	8/8	.791
patrolamento	ɾ → mẽto / [X −nasal, −arredondado, −estridente, −lateral ola ____]$_N$	7/7	.760

Tabela 5 – Ilhas de confiabilidade robustas (> .75) para a produção de nominalizações em *-mento*.

Podemos perceber que a maioria dessas regras favorecedoras da ocorrência de *-mento* na primeira conjugação aplica-se após consoantes com um traço de coronalidade. Como o próprio sufixo *-ção* apresenta esse traço, a evitação da sequência resultante pode ser, ou pode ter sido, uma tendência atuante na escolha do sufixo nominalizador, ainda que não haja uma restrição absoluta a sequências de sílabas iniciadas por coronais na língua portuguesa.

A regra que prevê a geração de *tensionamento* é interessante, porque grande parte das palavras que dão suporte a essa generalização na lista de treinamento contêm a sequência *ão* em sua história derivacional. Por exemplo, *relação* → *relacionar* → *relacionamento*. Seria difícil atribuir a confiabilidade dessa regra a uma evitação sincrônica, foneticamente motivada, da adição de *-ção* neste caso, sendo que *ão* já não se superficializa no verbo que serve de base para a nominalização. Em vez disso, podemos supor que essa ilha de confiabilidade tenha se formado com base em formações de períodos da língua em que a evitação de *-ção* neste contexto podia ter motivação fonética mais transparente, tendo em vista a forma arcaica desse sufixo, *-ção*.

5.1.3 Comparação com o corpus

Na subseção anterior, vimos que o MGL foi capaz de encontrar um bom número de contextos fonológicos em que o uso de *-mento* é predominante, o que dá suporte inicial a nossa hipótese sobre a estabilidade desse sufixo. Como a composição do léxico da língua é, em parte, reflexo da história de aplicação de sua morfologia derivacional, o fato de um sufixo nominalizador ter concentrações estatisticamente significativas em determinados contextos sugere uma tendência atuante ao longo da história da língua nesse domínio. Em outras palavras, esse resultado sugere que, na transmissão do sistema morfológico do português, os falantes da língua foram sensíveis a essas subregularidades.

Ainda precisamos saber, contudo, se falantes de português são capazes de explorar essas regularidades na formação de novas palavras. Em outras palavras, devemos testar se elas não são apenas fatos distribucionais dos dados de treinamento, mas tendências com realidade psicológica que podem servir como guias aos falantes na resolução de situações de competição morfológica. Isso nos leva aos resultados da aplicação do modelo sobre os dados de teste.

A rodada do MGL sobre as bases não dicionarizadas extraídas do corpus coloca em destaque três ilhas de confiabilidade que não estiveram entre as mais robustas

[20] Na representação do contexto das regras desta tabela, utilizamos símbolos segmentais ou traços fonológicos de acordo com o que se mostrou mais conveniente em termos de economia de espaço em cada caso.

na análise da lista de treinamento. Elas merecem comentário, entretanto, porque foram responsáveis pelo maior número de previsões corretas de palavras da lista de teste. Juntas, as três regras expressas na Tabela 6 responderam por 24 das 86 previsões corretas feitas pelo MGL, ao passo que outras 42 regras responderam pelas previsões restantes.

Exemplo	Regra	Acertos/Escopo	Conf.
propagandeamento	ɾ → mẽto / [X {e, o} a ____]$_N$	83/124	.598
prestigiamento	ɾ → mẽto / [X {ʎ, e, i, o, u, w, j} a ____]$_N$	147/242	.555
empoderamento	ɾ → mẽto / [X {ʎ, ʃ, ʒ, ɾ} a ____]$_N$	144/241	.521

Tabela 6 – Ilhas de confiabilidade que tiveram mais sucesso na previsão de formas em *-mento*.

A primeira delas não fez nenhuma predição incorreta e foi a que previu o maior número de palavras da lista de teste (dez): *baleamento*, *bloqueamento*, *coqueamento*, *escamoteamento*, *esfaqueamento*, *jateamento*, *pareamento*, *pisoteamento*, *propagandeamento* e *ranqueamento*. Em grande parte dos casos, as bases abrangidas por esta ilha são formadas pelo verbalizador *-ear*, como em *propaganda* → *propagandear*, mas a generalização por ela descrita é, aparentemente, mais ampla, envolvendo também bases que não parecem ser derivadas, como *bloquear*, *escamotear* e *pisotear*. Na lista de treinamento, é possível ver que essa ilha de confiabilidade encontra suporte também em formas como *abalroamento*, *coroamento*, *escoamento*, etc. Infelizmente, porém, não foi possível encontrar nominalizações não dicionarizadas formadas a partir de verbos terminados em oaɾ no corpus, o que não nos permite testar se elas seguiriam a escolha de sufixo predita por esta regra.

O grande sucesso preditivo dessa ilha de confiabilidade nos faz pensar que seu valor de confiabilidade pode estar sendo subestimado pelo modelo. Isso pode se dever ao fato de termos composto nossa lista de treinamento a partir de um dicionário geral, que, como tal, contém muitas palavras que não fazem parte do léxico corrente dos falantes. Assim, pode ser que grande parte dos contraexemplos a essa generalização encontrados pelo MGL nessa lista não sejam palavras acessíveis a aprendizes reais, como *balneação*, *caseação* e *manuseação*; dessa forma, o valor de confiabilidade da regra para aprendizes reais pode ser maior do que o estimado pelo MGL. No caso de alguns dos contraexemplos a essa regra que constam do dicionário e, portanto, da lista de treinamento, é possível encontrar, de fato, formações em *-mento* de uso mais corrente (a julgar pelos resultados do mecanismo de busca do Google);

é o caso, por exemplo, de *abotoamento*, *alheamento*, *branqueamento*, *caseamento*, *delineamento* e *manuseamento*, que são bastante mais frequentes do que as formas em *-ção* listadas, em cada caso, no dicionário.

A segunda regra listada na Tabela 6 prevê corretamente a forma de oito palavras de nosso corpus, todas elas envolvendo a sequência [ia]: *acumpliciamento*, *diligenciamento*, *fatiamento*, *justiciamento*, *prestigiamento*, *referenciamento*, *silenciamento* e *taxiamento*. Houve também duas predições incorretas: ☹*instanciamento* e ☹*remediamento*, casos em que a forma atestada no corpus contém o sufixo *-ção*. Outras sequências descritas pela regra podem ser encontradas, por exemplo, em *embrulhamento*, *vozeamento*, *atordoamento* e *apaziguamento*. Podemos ver que esta regra define, em sua descrição estrutural, um conjunto de palavras que inclui o que é definido pela regra anterior – trata-se de uma generalização que inclui as bases terminadas em [ear] e [oar], juntamente com bases terminadas em [ʎar], [iar] e [uar]. Com a maior abrangência, a regra alcança um escore de confiabilidade um pouco mais baixo. Porém, como o MGL considera, para esta ilha de confiabilidade, os mesmos contraexemplos que discutimos acima, para a primeira regra, podemos supor que aqui também estamos diante de uma generalização que tem seu escore de confiabilidade subestimado pelo MGL graças a contraexemplos dicionarizados que podem não fazer parte do léxico corrente de aprendizes reais.

A terceira regra da Tabela 6 teve seis predições corretas: *aparamento*, *destemperamento*, *empoderamento*, *enamoramento*, *regramento* e *tratoramento*; e uma predição incorreta: ☹*oneramento*. Podemos ver que todas as bases da lista de teste que foram associadas pelo MGL a esta ilha de confiabilidade têm uma raiz terminada em [ɾ]. As outras três possibilidades previstas pela descrição estrutural da regra são atestadas no corpus, mas o MGL não aplica esta regra a elas, pois há outras regras mais confiáveis englobando esses contextos e que, portanto, têm precedência; é o caso das regras que produzem *guinchamento* e *embaralhamento* na Tabela 5.

Como vimos anteriormente, para cada base da lista de teste, o MGL propõe possíveis formas de output (em nosso caso, uma com *-mento* e outra com *-ção*) e atribui a elas escores de boa formação, de acordo com o valor de confiabilidade das regras responsáveis por gerá-las. Ao avaliar a adequação do modelo com as formações do corpus, contamos uma situação de concordância entre o modelo e os dados empíricos (uma predição correta) sempre que a forma atestada no corpus também tiver sido a que recebeu o maior escore no modelo, em comparação com a alternativa contendo o sufixo rival. De outro modo, temos discordância entre o modelo e os dados. A hipótese nula nessa comparação é de que a taxa de concordância não ultrapassa 50%; isto é, de que dada uma escolha do modelo, não teríamos razão para esperar que ela

seja ou não a forma atestada no corpus. Isso deve ser observado caso a escolha dos falantes entre *-mento* e *-ção* seja aleatória, ou caso ela seja determinada somente por fatores não capturados pelas ilhas de confiabilidade encontradas pelo MGL.

Encontramos, no teste, 87% de concordância entre as formas preditas pelo modelo e as que foram encontradas no corpus. Essa proporção mostra-se significativa em um teste binomial exato (p < 0.001, intervalo de confiança (α = 95%): 81,3 a 91,5%). Vemos na Tabela 7 que houve aproximadamente a mesma taxa de concordância para outputs com *-ção* e com *-mento*, sugerindo que falantes respeitam ilhas de confiabilidade para ambos os afixos.

	Discordância	Concordância	Teste binomial exato
-mento	6	43	p < .001, I.C. (α = 95%): 75,2% a 95,4%
-ção	18	117	p < .001, I.C. (α = 95%): 79,7% a 91,9%

Tabela 7 – Número de concordâncias e discordâncias entre as predições do modelo e os dados empíricos.

No entanto, um olhar mais atento para os dados (expostos no Apêndice B.3) revela que o trabalho do modelo pode ter sido fácil demais ao prever as formas em *-ção*, devido ao grande número de bases verbais na lista de teste que terminam no sufixo *-izar*, e também a um número menor de bases terminadas no sufixo *-ificar*. Ambos estes sufixos reconhecidamente selecionam o nominalizador *-ção*. Portanto, não é surpreendente que o modelo tenha sido capaz de induzir regras altamente confiáveis contendo essas sequências, e que elas tenham obtido concordância com os dados empíricos. De fato, isso serve como validação do MGL, mostrando que ele é capaz de aprender um exemplo claro de seleção da morfologia do português, o que é um critério mínimo de razoabilidade para um algoritmo de aprendizagem neste domínio.

Por outro lado, por se tratar de um caso relativamente claro de seleção morfológica, que responde por mais da metade dos nossos dados de teste, a inclusão desses dados não nos ensina nada de novo sobre ilhas de confiabilidade na língua, e ainda pode ser responsável por inflar a taxa de concordância entre o modelo e o corpus. Por isso, excluímos essas bases da lista de teste. Com isso, restaram apenas 36 predições com *-ção* no modelo. Destas, 19 (52,7%) estão de acordo com o que foi atestado no corpus. Agora, no entanto, não é possível descartar a hipótese nula, de que não há relação entre as predições do modelo e os dados empíricos, no caso do sufixo *-ção* (teste binomial exato: p = 0,8679; I.C. (α = 95%): 35,5% a 69,6%.). No caso do sufixo *-mento*, os resultados não se alteraram, já que todas as bases retiradas eram de

dados para os quais o modelo previa afixação de *-ção*.

Nas tabelas 8 e 9, correspondentes aos sufixos *-mento* e *-ção*, respectivamente, vemos os casos de discordância entre o modelo e os dados, isto é, de palavras que receberam um escore mais alto do que suas concorrentes no MGL, mas não foram atestadas no corpus. A terceira coluna das tabelas apresenta a diferença entre os escores destas palavras e os de suas rivais atestadas.

Preferência do modelo	Escore	Diferença de escore
fervimento	0,934	0,622
zoamento	0,873	0,488
protocolamento	0,760	0,201
instamento	0,731	0,172
repactuamento	0,616	0,093
instanciamento	0,555	0,021

Tabela 8 – Preferências do modelo contendo o sufixo *-mento* que não foram atestadas no corpus.

A primeira observação que podemos fazer a respeito desses dados é a existência de alguns itens com escore bastante alto que não foram atestados no corpus. Entre eles, estão *fervimento* e *zoamento*. No primeiro caso, temos uma palavra que é corrente na língua, ainda que não apareça no corpus; a forma atestada nos dados é *ferveção*, que não é, contudo, uma alternativa legítima a *fervimento* para expressar a nominalização do verbo *ferver*, no sentido canônico de ebulição. *Zoamento*, por outro lado, não parece mesmo ser uma forma corrente, ainda que se encontrem algumas ocorrências suas em uma busca na internet. A forma atestada para esta base foi *zoação*, que existe ao lado das nominalizações já dicionarizadas, *zoada* e *zoeira*. Como previsto na discussão sobre bloqueio parcial da Seção 3.1, essas formas em *-ção* (com escore de confiabilidade baixo) conseguem escapar do bloqueio das formas atestadas porque possuem significados não canônicos. No caso de *ferveção*, há um sentido metafórico, em que uma festa agitada, por exemplo, pode estar "em ebulição"; e em ambos os casos, há um componente de significado frequentativo/iterativo, com a implicação de que não se trata de um evento simples de "ferver" ou "zoar", mas de um evento composto de diversas instâncias, possivelmente com diversos participantes.

Também podemos notar nas tabelas 8 e 9 que a maioria dos outputs não atestados do MGL recebeu escores medianos, que não se diferenciam significativamente dos que foram obtidos por formas alternativas atestadas com o sufixo rival. Essa tendência é mais visível no caso da Tabela 9, referente aos outputs com sufixo *-ção*. A diferença de escores entre o melhor output e sua alternativa, de acordo com o

Preferência do modelo	Escore	Diferença de escore
cercação	0,738	0,325
empoderação	0,719	0,198
destemperação	0,719	0,198
apenação	0,665	0,106
fatiação	0,641	0,086
descarnação	0,630	0,154
prestigiação	0,574	0,019
tratoração	0,573	0,052
enamoração	0,573	0,052
abrigação	0,573	0,038
outorgação	0,573	0,038
valetação	0,571	0,047
aparação	0,568	0,047
emparedação	0,566	0,077
capotação	0,561	0,039
agendação	0,547	0,058
regração	0,544	0,022

Tabela 9 – Preferências do modelo contendo o sufixo *-ção* que não foram atestadas no corpus.

modelo, é dada na terceira coluna dessas tabelas. Trata-se, portanto, de casos em que o modelo não decide inequivocamente entre as formas concorrentes, porque ambas são cobertas por generalizações pouco robustas de cada um dos afixos rivais. O fato de esses casos serem numerosos na Tabela 9 é interessante por duas razões. Primeiramente, outputs cujo escore se diferencia do de seu rival por menos de 0,1 representam 12 das 17 formas em *-ção* previstas pelo MGL que não foram atestadas no corpus. Se elas fossem desconsideradas, ou seja, se considerássemos apenas as ilhas de confiabilidade razoavelmente robustas (que acarretam diferenças maiores que 0,1), o modelo voltaria a obter sucesso significativo em prever quando uma forma em *-ção* pode ser atestada, acertando em 79,2% dos casos nesta amostra (teste binomial exato: p = 0,006611; CI (α = 95%): 57,8% a 92,9%).

O outro motivo por que os resultados da Tabela 9 são interessantes é que eles sugerem que, na ausência de algum motivo para que *-ção* ou *-mento* sejam escolhidos para a produção de uma forma nominalizada – quando não há, por exemplo, uma ilha de confiabilidade robusta decidindo a competição –, a escolha atestada tende a ser em *-mento*. Esta conclusão é desencorajada para alguns dos casos dessa tabela,

dada a observação anterior de que as regras que mais obtiveram sucesso na previsão de formas em *-mento* podem ter tido seu escore de confiabilidade subestimado pelo MGL. Nesse caso, a aparência de que a competição não é decidida por esse escore pode ser um artefato do modo como compomos nossa lista de treinamento – a partir de dados de um dicionário geral.

Especificamente, vemos que as bases *aparar*, *destemperar*, *empoderar*, *enamorar*, *regrar* e *tratorar* são abrangidas pela terceira ilha de confiabilidade da Tabela 6, que gera formas em *-mento*. Da mesma forma, *fatiar* e *prestigiar* são abrangidas pela segunda ilha listada na nessa tabela. Restam os outros casos de palavras em *-ção* não atestadas, com escores medianos e que não se diferenciam por mais de 0,1 do escore de uma forma alternativa: *abrigação*, *outorgação*, *valetação*, *emparedação*, *capotação* e *agendação*. As regras que geram as formas alternativas, em *-mento*, nestes casos são diversas e não se mostraram especialmente confiáveis. Assim, elas oferecem uma fraca sugestão de que *-mento* pode estar servindo como a escolha default nos casos em que a competição não é decidida pelos escores de confiabilidade. Entretanto, fica em aberto a possibilidade de que outros fatores estejam em jogo, e coloca-se também a necessidade de coleta e investigação de mais dados da faixa de escores intermediários em trabalhos futuros.

É importante notar que, na visão de morfologia discutida na Seção 3.1, a preferência por *-mento* em contextos não marcados seria esperada, uma vez aceita a premissa de que *-mento* e *-ção* são geralmente sinônimos na expressão da nominalização canônica, juntamente com a observação de que *-ção* tem se associado com bastante frequência à formação de nomes de sentido especial, com um componente frequentativo/iterativo (Rocha, 1999), como *bateção*, *beijação*, *ferveção*, *pegação*, *xingação*, *zoação*, etc. Nesses contextos, *-mento* pode garantir a possibilidade de expressão da nominalização canônica, quando já não há outro nome consagrado para esta função (cf. *batida*/*batimento*, *beijo*, *fervimento*, *pegada*/*pegamento*, *xingamento*, *zoeira*).

5.1.4 Associação entre confiabilidade e probabilidade de atestação

Testamos até agora o grau de concordância entre as predições do MGL e os dados encontrados no corpus de forma mais ou menos discreta. Isto é, sempre que o MGL atribuiu um escore mais alto para uma forma em *-mento*, por exemplo, interpretamos isso como uma escolha do modelo por esta forma, em oposição a uma forma com *-ção*. Verificamos, então, para cada uma dessas escolhas se ela é atestada no corpus. Entretanto, o output desse aprendiz é muito mais informativo do que uma escolha discreta entre duas formas; para cada uma delas, o MGL atribui um escore

de boa-formação, de 0 a 1, correspondente ao valor de confiabilidade da melhor regra capaz de gerá-la. Esse escore gera predições gradientes sobre a boa-formação de formas linguísticas, que podem ser testadas em diversos tipos de análise. Vimos, por exemplo, que Albright e Hayes (2003) testaram a correlação desse escore com a probabilidade de produção de formas do passado simples do inglês em um experimento, e com escores de avaliação do passado de pseudopalavras por informantes.

Em nosso estudo, testamos o grau de associação entre a robustez das ilhas de confiabilidade de *-ção* e *-mento* e a probabilidade de que as formas descritas por elas sejam atestadas em nosso corpus. Para isso, realizamos uma análise de regressão logística, com o objetivo de testar a hipótese de que há uma associação significativa entre os escores de boa-formação, ou a diferença entre o escore de uma forma e o de sua rival, e a atestação de uma forma (a variável dependente).[21] Encontramos um modelo com dois preditores, escore e sufixo, capaz de prever com sucesso razoável se um output do MGL é atestado ou não (χ^2 = 74.59, df = 2, p < 0.0001). A probabilidade de atestação de uma forma é dada pela seguinte equação.

$$\text{Prob}\{\text{atestação}\} = \frac{1}{1 + \exp(-X\hat{\beta})}, \text{ em que}$$

$$X\hat{\beta} = -7,919433 + (11,7619 \times escore) + (2,126 \times [mento])$$

$$\text{e } [mento] = 1 \text{ se o sufixo for } mento,\ 0 \text{ se não for.}$$

Ambos os preditores tem papel significativo no modelo estatístico (escore de confiabilidade: coeficiente = 11,7619, Wald z = 4,81, p < 0,0001; sufixo *-mento*: coeficiente = 2,126, Wald z = 5,11, p < 0,0001). Esse resultado corrobora a hipótese de que há uma associação positiva entre o escore de confiabilidade previsto pelo MGL e a probabilidade de uma forma de nominalização ser atestada no corpus. A existência desta associação está de acordo com a ideia de que a geração de novas palavras com esses sufixos é sensível às ilhas de confiabilidade detectadas pelo MGL. A informação nova que nos é dada pela regressão logística é de que essa sensibilidade é dependente da robustez dessas ilhas, medida pelo escore de confiabilidade. No Gráfico 8 (página 70), podemos ver que a probabilidade de que uma forma seja utilizada aumenta em função desse escore.

[21] A regressão logística é bastante utilizada na linguística para o estudo de variáveis dependentes categóricas, graças à implementação disponível no pacote VARBRUL. Esta implementação, no entanto, é limitada a variáveis independentes categóricas, o que diminui sua utilidade para este trabalho, já que o escore de confiabilidade atribuído pelo MGL é uma variável contínua. As análises estatísticas empreendidas neste trabalho foram realizadas no ambiente R (R Core Team, 2014).

O coeficiente positivo para *-mento* no modelo estatístico sugere, ainda, que há uma preferência por formas com esse sufixo, pois, dado um mesmo escore de confiabilidade, há uma expectativa maior de que essas formas sejam atestadas do que as alternativas em *-ção*. Esse resultado pode ser visto com mais clareza no Gráfico 8, em que se representa a probabilidade de atestação em função do escore de confiabilidade para cada um dos sufixos. Cada output do MGL é representado no gráfico por um círculo vermelho (no caso do sufixo *-ção*) ou por um triângulo azul (no caso do sufixo *-mento*). Outputs que foram atestados no corpus aparecem no topo, e os que não foram, na parte inferior do gráfico.

A preferência por *-mento* nesse teste faz-nos perguntar se esse afixo não se mostraria mais produtivo do que *-ção* caso desconsiderássemos bases contendo os sufixos *-izar* e *-ificar*, como fizemos para a análise estatística. Na Tabela 10, abaixo, trazemos esse resultado na segunda linha, *-ção**.

Sufixo	Palavras (V)	*Hapax Legomena* (n_1)	Índice de produtividade $\mathcal{P}$
-ção	1.359	162	0,00187
*-ção**	980	97	0,00112
-mento	586	77	0,00089

Tabela 10 – Produtividade dos sufixos *-ção* e *-mento*; no caso de *-ção**, desconsideram-se bases em *-izar* ou *-ificar* (N = 86.653 *tokens*, em cada caso).

Neste caso, a diferença entre *-ção** e *-mento*, em termos do índice $\mathcal{P}$, diminui consideravelmente, mas o primeiro ainda se mostra mais produtivo, mesmo após os sufixos que potenciam a adição de *-ção* serem desconsiderados. Deixamos em aberto, para trabalhos futuros, a investigação dessa disparidade entre o resultado do teste do MGL, e da regressão logística, e o resultado obtido no cálculo do índice $\mathcal{P}$. É possível que essa disparidade desapareça com um exame cuidadoso de outros fatores, como a frequência de *-ção* em usos frequentativos/iterativos. Pode ser, ainda, que essa vantagem de *-mento* desapareça em uma rodada do MGL que considere uma lista de treinamento mais de acordo com a língua atual. É importante lembrar que, como discutido na subseção anterior, o aprendiz utilizado neste trabalho foi treinado com uma lista de palavras extraídas de um dicionário geral, que contém diversos itens em desuso. Como consequência disso, é possível que a confiabilidade de algumas generalizações envolvendo *-mento* tenha sido subestimada pelo modelo, devido à influência de contraexemplos espúrios. Sem estes, é possível que mais atestações de *-mento* fossem explicadas pelo escore de confiabilidade e não pela mera escolha

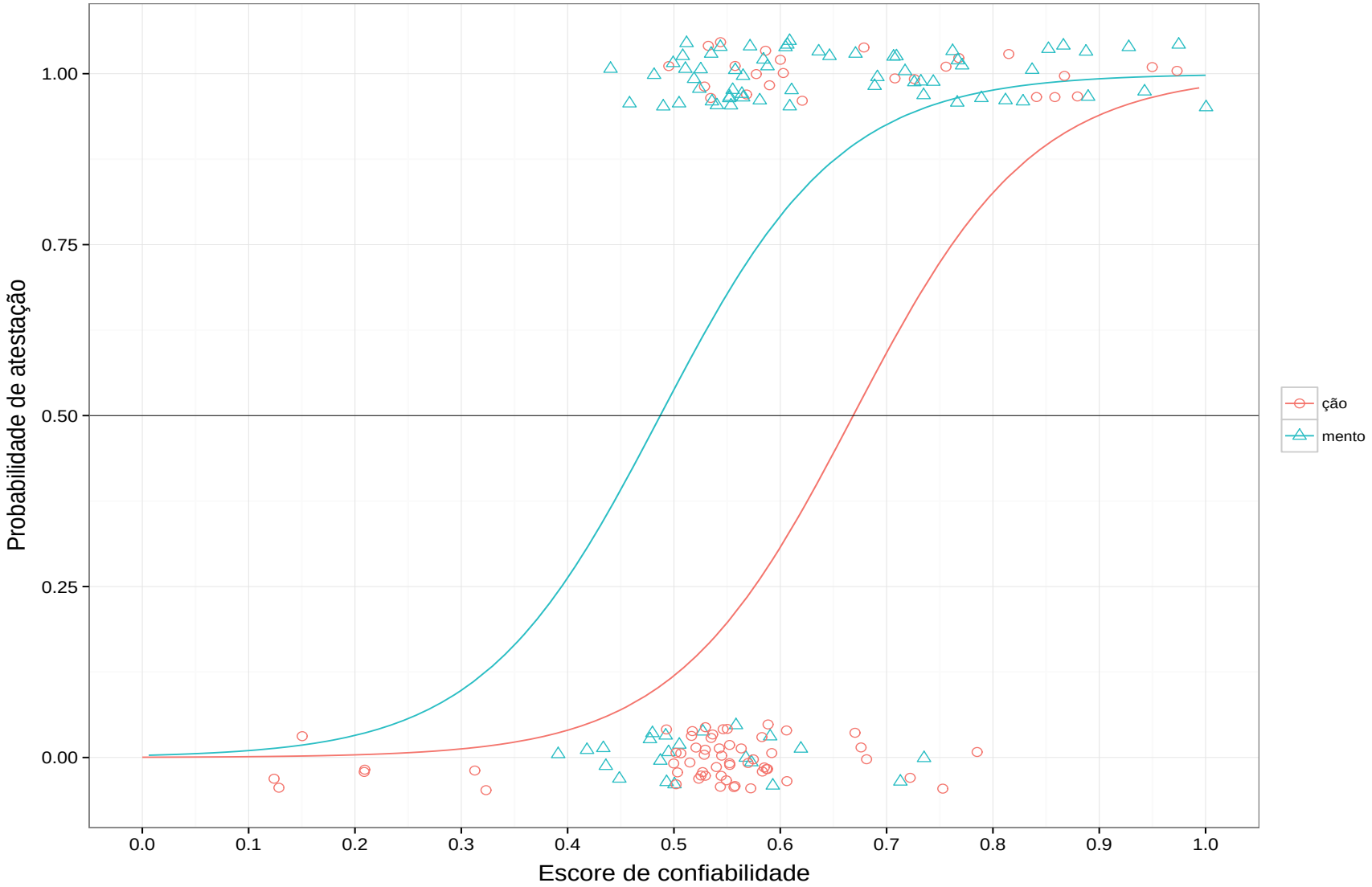

Gráfico 8: Probabilidade de atestação de outputs do MGL, em função do sufixo utilizado e do escore de confiabilidade.

do sufixo.

Há, ainda, outros fatores não discutidos neste trabalho que podem estar implicados na competição entre esses afixos. Uma consideração de tais fatores poderia explicar por que *-ção* ainda encontra mais contextos de formação de palavras novas do que *-mento* nos dados do *corpus*. Entre esses fatores podem estar o número de sílabas da base, cuja relevância encontra plausibilidade inicial por esses sufixos se diferenciarem nesta dimensão, e a semântica da nominalização resultante (por exemplo, estado x ação/processo); ambos estes fatores mostraram-se significativos no estudo de Silveira (2015). De toda sorte, é notável que a gramática resultante da aplicação do MGL, em nosso estudo, tenha obtido um grande sucesso de predição das formas nominalizadas, baseando-se apenas em parâmetros fonotáticos. É provável que a consideração de outras variáveis possa aumentar ainda mais esse potencial preditivo. Isso coloca a necessidade de confrontar as variáveis consideradas nos estudos da área, empreendendo análises estatísticas mais refinadas em trabalhos futuros.

6 Considerações finais

Neste trabalho, concentramos nossa atenção na competição dentro da morfologia derivacional, entendendo-a como um dos fatores que determinam a produtividade de padrões morfológicos. É muito comum, neste domínio, que haja construções alternativas para dar conta de necessidades comunicativas, e a escolha entre uma delas tende a bloquear a opção pela outra, limitando a produtividade de padrões rivais. Buscamos, assim, avançar na compreensão tanto de como se dá esse bloqueio (e de porque ele existe) quanto de como se dá a escolha entre construções alternativas.

Para a primeira dessas questões, exploramos brevemente uma abordagem pragmática do bloqueio, que não sofre dos problemas aos quais as abordagens gramaticais, dentro da morfologia lexical, estão sujeitas. Nessa perspectiva, derivam-se efeitos de bloqueio de princípios conversacionais gerais que têm como efeito a preferência pelo uso de meios já existentes (e não marcados) para exercer funções não marcadas na língua. Sugerimos que uma preferência semelhante existe na competição entre expressões novas, na medida em que usuários de uma língua parecem ser sensíveis ao grau de confiabilidade demonstrado por padrões rivais em contextos de formação de palavras. Em cada um desses contextos, a confiabilidade de um padrão seria inversamente proporcional ao número de exceções a sua aplicação encontradas no léxico, fornecendo uma medida do quão usual, esperado, ou não marcado, ele é no contexto relevante.

Utilizamos como domínio empírico para o estudo dessa proposta a competição entre *-mento* e *-ção* no português. Observamos que esses afixos são ambos produtivos no português, ainda que nosso levantamento mostre uma maior probabilidade de formação de palavras novas em *-ção*. Crucialmente, a vantagem deste sufixo, já estabelecida no século XVII, não fez com que *-mento* perdesse sua produtividade, como aconteceu com seu cognato na língua inglesa. Nossa hipótese de que a estabilidade de *-mento* no português foi amparada pela existência de contextos em que esse afixo tinha um grau de confiabilidade particularmente alto, no sentido discutido anteriormente, foi corroborada por um teste do *Minimal Generalization Learner* de Albright e Hayes (1999), aplicado a dados do Dicionário Houaiss e de um levantamento de *corpus*.

O modelo mencionado postula um procedimento de aprendizagem que explora justamente essa noção de confiabilidade na formulação de regras morfofonológicas. Com a sua aplicação, foi possível descobrir algumas ilhas de confiabilidade que se mostraram bastante robustas nos dados de treinamento (compostos por palavras

dicionarizadas) e na geração de novas palavras (a partir de bases extraídas de um *corpus* de textos escritos). Interessantemente, essas ilhas não se restringiram a contextos tradicionalmente descritos como facilitadores da aplicação dos afixos estudados (como diante da terminação *ecer* no caso de *-mento* (Rocha, 1999)); assim, descobrimos outros contextos fonotáticos que influenciam a resolução da competição entre os nominalizadores do português, detalhados na Tabela 5.

As regras descobertas pelo modelo tiveram um sucesso considerável na geração dos dados da lista de testes, sobretudo no caso do sufixo *-mento*. O sucesso de predição mostrou-se proporcional ao escore de confiabilidade atribuído pelo MGL a cada uma dessas regras, o que corrobora a predição de Albright e Hayes (1999) de que, na produção de novas formas linguísticas, falantes não se guiam apenas pela existência das regras em sua gramática, mas também pelo grau de suporte que estas encontram no léxico da língua.

Outro resultado interessante de nosso teste é o de que, na faixa de escores de confiabilidade medianos, em que o modelo não decide claramente entre formas em *-mento* ou *-ção* por meio desses escores, a forma atestada no *corpus* foi, na maioria dos casos, a que continha o sufixo *-mento*. Por um lado, isso nos sugere um caráter de *default* para este sufixo, na medida em que seria escolhido como a forma de nominalização sempre que não houvesse nenhuma pressão por outra escolha. Por outro lado, essa conclusão não parece estar de acordo com os resultados de nosso levantamento da produtividade sincrônica de *-ção* e *-mento*, pois *-ção* se mostra mais produtivo mesmo quando desconsideramos bases terminadas em *-izar* e *-ificar*. No final do capítulo anterior, discutimos a necessidade de se explorar outras variáveis linguísticas no conjunto de dados, que possam explicar essa discrepância.

Cabe mencionarmos aqui a necessidade de se testar o conjunto de regras obtidas pelo MGL por meio de outros métodos. Por exemplo, um estudo experimental de produção de pseudopalavras, a exemplo dos de Prasada e Pinker (1993) e Albright e Hayes (2002), pode nos dar mais informações sobre como usuários da língua resolvem a competição entre padrões de nominalização nesses contextos em que as regras não se diferenciam em termos de confiabilidade. Isso é possível porque esse tipo de estudo nos permite controlar mais livremente o formato das bases testadas, sem depender da casualidade das ocorrências de um corpus. Outra possibilidade é testarmos a correlação entre os escores de confiabilidade atribuídos pelo MGL e dados escalares de avaliação de pseudopalavras; isso nos permitiria fazer um uso completo do fato de o modelo ter escores contínuos como output, como em Albright (2002), Albright e Hayes (2002) e Albright e Hayes (2003).

Abordagens experimentais também são úteis para o teste de formas que têm

poucas chances de aparecer em um *corpus* de tamanho limitado. Por exemplo, a ilha de confiabilidade que se mostrou mais preditiva diante da lista de treinamento (na primeira linha da Tabela 6) prevê a adição de mento a bases terminadas em [ear] e [oar]. Entretanto, foram encontradas no corpus, entre as palavras não dicionarizadas, apenas nominalizações formadas a partir de bases com a primeira dessas terminações. Por isso, não nos foi possível testar se a produção de novas palavras diante da terminação [oar] também seguiria a predição do modelo, ou se a generalização correta envolve apenas [ear].

De modo geral, este trabalho se acrescenta ao corpo de evidências que apontam para a conclusão de que o uso de padrões morfológicos é fortemente influenciado pelas frequências de atestação prévia desses padrões no léxico. A hipótese específica deste estudo, seguindo Albright e Hayes (2002) e trabalhos posteriores, é de que esses efeitos de frequência se estabelecem com referência a contextos de diferentes níveis de generalidade; de que esses contextos são definidos estruturalmente; e de que, ao adquirirem um sistema morfológico, falantes distinguem generalizações estatisticamente robustas daquelas que contêm mais contra exemplos e, assim, são relativamente fracas.

Bibliografia

Albright, Adam (2002). “Islands of reliability for regular morphology: Evidence from Italian”. Em: *Language* 78, pp. 684–709.

Albright, Adam e Bruce Hayes (1999). “An automated learner for phonology and morphology”. Em: *Ms. http://www. humnet. ucla. edu/humnet/linguistics/people/hayes/learning/learner. pdf*.

— (2002). “Modeling English past tense intuitions with minimal generalization”. Em: *Morphological and Phonological Learning: Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON)*. Association for Computational Linguistics. Philadelphia, pp. 58–69.

— (2003). “Rules vs. analogy in English past tenses: a computational/experimental study”. Em: *Cognition* 90.2, pp. 119–161.

Aronoff, Mark (1976). *Word formation in generative grammar*. Cambridge, Massachusetts: MIT Press, p. 134.

Baayen, Harald (1992). “Quantitative aspects of morphological productivity”. Em: *Yearbook of morphology 1991*. Springer, pp. 109–149.

Baayen, R Harald (2002). *Word Frequency Distributions*. MIT Press.

Basilio, Margarida (1996). “Gramática do Português Falado – Volume IV: Estudos Descritivos”. Em: ed. por Ataliba Castilho e Margarida Basílio. Campinas: Editora da Unicamp/FAPESP. Cap. Formação e uso da nominalização deverbal sufixal no português falado, pp. 223–33.

— (2008). *Formação e classes de palavras no português Brasil*. Editora Contexto.

Bauer, Laurie (2001). *Morphological productivity*. Vol. 95. Cambridge University Press.

Blutner, Reinhard (1998). “Lexical pragmatics”. Em: *Journal of Semantics* 15.2, pp. 115–162.

Bolinger, Dwight L. (1948). *Forms of English, Accent, Morpheme, Order*. Cambridge, Mass.: Harvard University Press. Cap. On Defining the Morpheme, pp. 183–189.

Bybee, Joan e Carol Lynn Moder (1983). “Morphological classes as natural categories”. Em: *Language* 59.2, pp. 251–270.

Charikar, Moses S (2002). “Similarity estimation techniques from rounding algorithms”. Em: *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. ACM, pp. 380–388.

Chomsky, Noam e Morris Halle (1968). *The sound pattern of English*. New York, NY: Harper & Row.

Corbin, Danielle (1987). *Morphologie dérivationnelle et structuration du lexique*. Vol. 193. Walter de Gruyter.

Di Sciullo, Anna-Maria e Edwin Williams (1987). "On the definition of word". Em: *Linguistic Inquiry Monographs* 14.

Dowty, David R (1979). *Word meaning and Montague grammar: The semantics of verbs and times in generative semantics and in Montague's PTQ*. Vol. 7. Springer.

Evert, Stefan e Anke Lüdeling (2001). "Measuring morphological productivity: Is automatic preprocessing sufficient". Em: *Proceedings of the Corpus Linguistics 2001 conference*, pp. 167–175.

Grice, H Paul (1975). "Logic and conversation". Em: *Syntax and Semantics 3: Speech Acts*. Ed. por Peter Cole e Jerry L. Morgan. Vol. 3. New York: Academic Pres, pp. 41–58.

Grodt, Aline (2009). "Um estudo sobre produtividade derivacional no português falado no sul do Brasil". Em:

Hickmann, Maya (1997). "The Acquisition of French as a Native Language: Structural and Functional Determinants in a Crosslinguistic Perspective". Em: *Journal of Speech-Language Pathology and Audiology* 21.4.

Horn, Laurence (1984). "Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature". Em: *Meaning, Form, and Use in Context: Linguistic Applications*. Ed. por Deborah Schiffrin. Washington, DC: Georgetown University Press, pp. 11–42.

Horn, Laurence R (1978). "Lexical incorporation, implicature, and the least effort hypothesis". Em: *Papers from the Parasession on the Lexicon. Chicago Linguistics Society. Chicago*, pp. 196–209.

Kastovsky, Dieter (1986). "The problem of productivity in word formation". Em: *Linguistics* 24.3, pp. 585–600.

Kiparsky, Paul (1982). "Lexical phonology and morphology". Em: *Linguistics in the morning calm*.

— (1983). "Word-formation and the lexicon". Em: *Proceedings of the 1982 mid-America linguistics conference*. Vol. 3. Department of Linguistics, University of Kansas Lawrence, Kansas, p. 22.

Lindsay, Mark e Mark Aronoff (2013). "Natural selection in self-organizing morphological systems". Em: *Morphology in Toulouse: Selected Proceedings of Décembrettes*. Vol. 7.

Lodge, Anthony (2008). "Sociolinguistic stratification in 19th-century Paris". Em: *Sprachen und Sprechen im städtischen Raum* 2.1400, p. 103.

Lüdeling, Anke e Stefan Evert (2003). “Linguistic experience and productivity: corpus evidence for fine-grained distinctions”. Em: *Proceedings of the 2003 Corpus Linguistics Conference, Lancaster*.

Mayerthaler, Willi (1977). *Studien zur theoretischen und zur französischen Morphologie: Reduplikation, Echowörter, morphologische Natürlichkeit, Haplologie, Produktivität, Regeltelescoping, paradigmatischer Ausgleich*. Vol. 40. Walter de Gruyter.

McCawley, James D (1978). “Conversational implicature and the lexicon”. Em: *Syntax and semantics* 9, pp. 245–259.

Mikheev, Andrei (1997). “Automatic rule induction for unknown-word guessing”. Em: *Computational Linguistics* 23.3, pp. 405–423.

Miyagawa, Shigeru (1981). *Complex verbs and the lexicon*. Coyote Papers, Vol 1. University of Arizona Linguistics Circle.

Orengo, Viviane Moreira e Christian R Huyck (2001). “A Stemming Algorithmm for the Portuguese Language.” Em: *SPIRE*. Vol. 8, pp. 186–193.

Pinker, Steven (1998). “Words and rules”. Em: *Lingua* 106.1–4. Language Acquisition Knowledge Representation and Processing, pp. 219 –242.

Plag, Ingo (2000). “On the mechanisms of morphological rivalry: A new look at competing verb-deriving affixes in English”. Em: *Anglistentag 1999 Mainz*. Ed. por B. Reitz e S. Rieuwerts. Trier: Wissenschaftlicher Verlag Trier, pp. 63–76.

Poser, William J (1992). “Blocking of phrasal constructions by lexical items”. Em: *Lexical Matters*. Ed. por Ivan A. Sag e Anna Szabolcsi. CSLI Publications, pp. 111–130.

Prasada, Sandeep e Steven Pinker (1993). “Generalisation of regular and irregular morphological patterns”. Em: *Language and cognitive processes* 8.1, pp. 1–56.

Quadros, Emanuel Souza de (2009). *A estrutura e o uso da parassíntese no português*. Trabalho de Conclusão de Curso. Instituto de Letras, UFRGS.

— (2011). “Reflexões acerca da Produtividade Morfológica e de sua Medição: estudo de sufixos nominalizadores do português”. Em: *ReVEL* 9.ed. especial n. 5.

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Viena, Áustria.

Rocha, Luiz Carlos de Assis (1999). “A nominalização no português do Brasil”. Em: *Revista de Estudos da Linguagem* 8.1, pp. 5–51.

Sandmann, Antônio José (1988). *Formação de palavras no português contemporâneo*. Curitiba: Scientia et Labor, p. 12.

Silveira, Luciana Morales da (2015). “O emprego de -ção e de -mento no português falado no sul do Brasil”. Diss. de mestrado. Programa de Pós-Graduação em

Letras, UFRGS.

Teyssier, Paul (1982). *História da língua portuguesa*. Vol. 5. Livraria Sá da Costa.

Toynbee, Paget (1896). *A historical grammar of the french language*.

Van Marle, Jaap (1992). "The relationship between morphological productivity and frequency: a comment on Baayen's performance-oriented conception of morphological productivity". Em: *Yearbook of Morphology 1991*. Springer, pp. 151–163.

Wunderlich, Dieter (1996). "Minimalist morphology: the role of paradigms". Em: *Yearbook of morphology 1995*. Springer, pp. 93–114.

Yang, Charles (2005). "On productivity". Em: *Linguistic variation yearbook* 5.1, pp. 265–302.

APÊNDICE A – Scripts

Versões mais recentes de todos os scripts apresentados aqui podem ser encontradas no repositório virtual `https://github.com/shoeki/ling`. Neste endereço, também se encontram todos os arquivos suplementares ao uso desses scripts no contexto deste trabalho, incluindo listas de exclusão, de substituição e as listas de dados analisados nesta dissertação.

A.1 populate.py

```
1  #!/usr/bin/python
2  # -*- coding: utf-8 -*-
3
4  import sqlite3 as db
5  import argparse
6  import json
7  import re
8  import nltk
9  import stemmer
10 from progress.bar import Bar
11
12 cmdline = argparse.ArgumentParser(description='Alimenta o banco de dados a partir de
       um arquivo JSON.')
13 cmdline.add_argument('data')
14 cmdline.add_argument('corpus')
15 args = cmdline.parse_args()
16 ##############################
17
18 with open(args.data, 'r') as source:
19     data = json.load(source)
20
21 with open('subs.json', 'r') as lista_grafia:
22     grafia_subs = json.load(lista_grafia)
23
24 con = db.connect('data/corpus.db')
25 st = stemmer.RSLPStemmer()
26
27
28 with con:
29
30     cur = con.cursor()
31
32     def getstem(token):
33         '''
34         Encontra a stem apropriada para um token, corrigindo erros ágrficos.
35         '''
36
```

```
37          stem = st.stem(token)
38
39          if stem in grafia_subs:
40              return grafia_subs[stem]
41          else:
42              return stem
43
44
45      bar = Bar('Textos', max = len(data))
46
47      for text in filter(lambda t: t['corpo'], data):
48
49          # Prepare data
50          if text['titulo']:
51              tit = text['titulo'][0]
52          else:
53              tit = ' '
54          # text should be stripped of excessive newline characters
55          corpo = re.sub('\n\s+', '\n', ''.join(text['corpo']))
56          h = str(text['hash'])
57
58          # Insert data into the appropriate fields
59          try:
60              cur.execute('INSERT INTO Texto (textoid,autor,titulo,corpo,data,corpus)
                    VALUES (?, ?, ?, ?, ?, ?)', (h, '', tit, corpo, text['data'][0],
                    args.corpus))
61          except db.IntegrityError:
62              ''' Como o campo hash áest marcado para ser único em init.py,
63              caso se tente inserir um texto duplicado no banco de dados,
64              uma çãexceo IntegrityError é imediatamente gerada. '''
65              continue
66
67          # Get tokens and stems
68          pattern = r'\w+(-\w+)*' # definition of 'word'
69          tokens = [(tk, getstem(tk)) for tk in
70              map(lambda w: w.lower(), nltk.regexp_tokenize(corpo, pattern))]
71
72          stems = list(set([(tk[1], ) for tk in tokens if tk[1]]))
73
74          pos = 0
75          cur.executemany('INSERT OR IGNORE INTO Palavra(palavra) VALUES(?)', stems)
76          for tk in tokens:
77              cur.execute('INSERT INTO Token VALUES (?, ?, ?, ?, ?)',
78                      (tk[0].lower(), pos, tk[1], h, args.corpus))
79              pos = pos + 1
80
81          bar.next()
82      bar.finish()
```

A.2 stemmer.py

```
1 #!/usr/bin/env python
```

```
2  # -*- coding: utf-8 -*-
3
4  '''
5  Stemmer apenas para a ãflexo.
6  '''
7
8  from __future__ import (absolute_import, division,
9                          print_function, unicode_literals)
10 from future.builtins import *
11 import codecs
12 import string
13 import json
14 from nltk.data import load
15 from nltk.stem.api import StemmerI
16 import argparse
17
18 with open('subs.json','r') as lista_grafia:
19     grafia_subs = json.load(lista_grafia)
20
21 class RSLPStemmer(StemmerI):
22     """çã
23     Adaptao da classe original ídistribuda com o NLTK.
24     """
25
26     def __init__ (self):
27         self._model = []
28
29         self._model.append(self.read_rule("step0.pt"))
30         self._model.append(self.read_rule("step1.pt"))
31         self._model.append(self.read_rule("step5.pt"))
32
33     def read_rule (self, filename):
34         rules = load('nltk:stemmers/rslp/' + filename, format='raw').decode("utf8")
35         lines = rules.split("\n")
36
37         lines = [line for line in lines if line != ""]     # remove blank lines
38         lines = [line for line in lines if line[0] != "#"]  # remove comments
39
40         # NOTE: a simple but ugly hack to make this parser happy with double '\t's
41         lines = [line.replace("\t\t", "\t") for line in lines]
42
43         # parse rules
44         rules = []
45         for line in lines:
46             rule = []
47             tokens = line.split("\t")
48
49             # text to be searched for at the end of the string
50             rule.append(tokens[0][1:-1]) # remove quotes
51
52             # minimum stem size to perform the replacement
53             rule.append(int(tokens[1]))
54
55             # text to be replaced into
```

```
56              rule.append(tokens[2][1:-1]) # remove quotes
57
58              # exceptions to this rule
59              rule.append([token[1:-1] for token in tokens[3].split(",")])
60
61              # append to the results
62              rules.append(rule)
63
64          return rules
65
66      def stem(self, word):
67          word = word.lower()
68
69          # the word ends in 's'? apply rule for plural reduction
70          if word[-1] == "s":
71              word = self.apply_rule(word, 0)
72
73          # the word ends in 'a'? apply rule for feminine reduction
74          if word[-1] == "a":
75              word = self.apply_rule(word, 1)
76
77          # noun reduction
78          prev_word = word
79          if word == prev_word:
80              # verb reduction
81              prev_word = word
82              word = self.apply_rule(word, 2)
83
84          return word
85
86      def apply_rule(self, word, rule_index):
87          rules = self._model[rule_index]
88          for rule in rules:
89              suffix_length = len(rule[0])
90              if word[-suffix_length:] == rule[0]:       # if suffix matches
91                  if len(word) >= suffix_length + rule[1]:  # if we have minimum size
92                      if word not in rule[3]:               # if not an exception
93                          word = word[:-suffix_length] + rule[2]
94                          break
95
96          return word
97
98  def getstem(token):
99      '''
100     Encontra a stem apropriada para o token, corrigindo erros de grafia.
101     '''
102
103     st = stemmer.RSLPStemmer()
104     stem = st.stem(token)
105
106     if stem in grafia_subs:
107         return grafia_subs[stem]
108     else:
109         return stem
```

```
110
111 if __name__ == "main":
112
113     comando = argparse.ArgumentParser(description='Stemmer para retirar as õflexes')
114     comando.add_argument('input')
115     comando.add_argument('output')
116     comando.add_argument('exceptions')
117     args = comando.parse_args()
118
119     stemmer = RSLPStemmer()
120
121     words_by_stem = {}
122
123     # carrega a lista de çõsubstituies de ortografia
124     with codecs.open(args.exceptions, 'r', encoding='utf-8') as grafia_ex:
125         grafia_subs = json.load(grafia_ex)
126
127     with codecs.open(args.input, 'r', encoding='utf-8') as source:
128         for word in source:
129             word = word.split('\t')[0]
130             word = string.rstrip(word)
131             stem = stemmer.stem(word)
132             if stem in grafia_subs:
133                 stem = grafia_subs[stem]
134             words_by_stem.setdefault(stem, []).append(word)
135
136     with codecs.open(args.output, 'w', encoding='utf-8') as output:
137         json.dump(words_by_stem, output, indent = 4)
```

A.3 init.py

```
1  #!/usr/bin/python
2  # -*- coding: utf-8 -*-
3
4  ''' çãInicializao do banco de dados
5  '''
6
7  import sqlite3 as db
8
9  con = db.connect('data/corpus.db')
10
11 with con:
12
13     cur = con.cursor()
14
15     ''' A unidade ábsica do banco de dados é o token (entendido como cadaê
16     ocorrncia de uma palavra). Cada token faz êreferncia à palavra que ele
17     instancia e ao texto em que aparece.
18     '''
19
20     cur.executescript('''
21         CREATE TABLE Corpus(
```

```
22              nome    TEXT PRIMARY KEY
23              );
24          CREATE TABLE Texto(
25              textoid TEXT PRIMARY KEY,
26              hash    TEXT UNIQUE,
27              autor   TEXT,
28              titulo  TEXT,
29              corpo   TEXT,
30              data    TEXT,
31              corpus  TEXT REFERENCES Corpus(nome)
32              );
33          CREATE TABLE Palavra(
34              palavraid INTEGER PRIMARY KEY,
35              palavra TEXT UNIQUE
36              );
37          CREATE TABLE Token(
38              token   TEXT,
39              pos     INTEGER,
40              stem    INTEGER REFERENCES Palavra(palavraid),
41              texto   INTEGER REFERENCES Texto(textoid),
42              corpus  TEXT REFERENCES Corpus(nome)
43              );
44          ''')
```

A.4 freqlist.py

```
1  #!/usr/bin/env python3
2  # -*- coding: utf-8 -*-
3
4  import re
5  import argparse
6  import tools
7  import random
8  import statistics
9  import sqlite3 as db
10 from operator import itemgetter
11
12
13 class Sample:
14
15     def __init__(self, freqs, size, iterations = 1000):
16         self.freqs = freqs
17         self.size = size
18         self.stems = list(map(lambda i: i[0], freqs))
19         self.samples = [dict(s) for s in self.sample(iterations)]
20         self.average = self.averageSample()
21         self.gfreq = self.averageRanks(self.average)
22
23
24     def randomSample(self):
25         ''' Given a list of pairs (type, freq), this should output a random sample
               of N
```

```
26          these pairs.
27          '''
28
29          outfreqs = {}
30
31          infreqs = dict(self.freqs)
32          instems = self.stems.copy()
33
34          for c in range(0, self.size):
35              stem = random.choice(instems)
36
37              try:
38                  outfreqs[stem] = outfreqs[stem] + 1
39              except KeyError:
40                  outfreqs[stem] = 1
41
42              # Remove stem if this was its last token.
43              infreqs[stem] = infreqs[stem] - 1
44              if infreqs[stem] == 0:
45                  instems.remove(stem)
46
47          return outfreqs.items()
48
49
50      def sample(self, iterations):
51          ''' Returns a certain number of random samples.'''
52
53          iters = []
54
55          for c in range(0, iterations):
56              iters.append(self.randomSample())
57
58          return iters
59
60
61      def averageSample(self, iterations = 1000):
62          ''' Returns the average of a number of iterations.'''
63
64          stats = {}
65
66          for stem in self.stems:
67              stats[stem] = {}
68              freqlist = [s[stem] for s in self.samples if stem in s.keys()]
69              stats[stem]['mean'] = sum(freqlist) / iterations
70              stats[stem]['median'] = statistics.median(freqlist)
71
72          return stats
73
74
75      def averageRanks(self, stats):
76
77          ranks = {}
78
79          for item in stats.items():
```

```
80              try:
81                  ranks[int(item[1]['median'])] = ranks[int(item[1]['median'])] + 1
82              except KeyError:
83                  ranks[int(item[1]['median'])] = 1
84
85          return sorted(ranks.items(), key=itemgetter(0))
86
87
88
89  def getList(aff, corpus, N=0, db_file = 'data/corpus.db'):
90
91      con = db.connect(db_file)
92
93      exceptions = tools.getExceptions(aff)
94      regex = re.compile(tools.afixo[aff], re.UNICODE)
95
96      def match(word):
97          if type(word) == str and regex.search(word) and word not in exceptions:
98              return 1
99          else:
100             return 0
101
102     with con:
103
104         cur = con.cursor()
105
106         con.create_function('match', 1, match)
107
108         if corpus == 'all':
109             cur.execute('SELECT stem, count(*) AS tk FROM Token WHERE match(stem) =
                    1 GROUP BY stem ORDER BY tk DESC')
110         else:
111             cur.execute('SELECT stem, count(*) AS tk FROM Token WHERE corpus = ? AND
                     match(stem) = 1 GROUP BY stem ORDER BY tk DESC', (corpus,))
112
113     output = cur.fetchall()
114
115     if N > 0:
116         output = output[:N]
117
118     return output
119
120
121 if __name__ == '__main__':
122
123     cmd = argparse.ArgumentParser(description='Por exemplo: ./freqlist.py all'
124                                       ' mento data/out_mento.tfl')
125     cmd.add_argument('corpus', help='Nome do corpus dentro do banco de dados.'
126                                     ' "all" seleciona todo o banco.')
127     cmd.add_argument('afixo', help='êReferncia do afixo desejado, que precisa estar'
128                                     ' definida em tools.py')
129     cmd.add_argument('output', help='Arquivo de output que áreceber a lista de'
130                                     ' êfrequncias.')
131     args = cmd.parse_args()
```

```
132
133     afixo = args.afixo
134     corpus = args.corpus
135
136     freqs = getList(afixo, corpus)
137
138     with open(args.output, 'w') as tfl:
139         tfl.write('f' + '\n')
140         for step in freqs:
141             tfl.write(str(step[1]) + '\n')
142
143     with open(args.output + '.list', 'w') as tfl:
144         for step in freqs:
145             tfl.write(step[0] + '\t' + str(step[1]) + '\n')
146
147     print('Tokens: ', str(sum(map(lambda f: f[1], freqs))))
```

A.5 tools.py

```
1  # -*- coding: utf-8 -*-
2
3  import re
4  from datetime import datetime as dt
5  import json
6  import sqlite3
7
8  db = sqlite3.connect('data/corpus.db')
9  db.text_factory = lambda x: x.decode('utf-8')
10
11 exceptions_file = 'exceptions.json'
12
13 afixo = {
14         'mento': 'mento$|mnto$|mentu$',
15         'ção': '(?<!.iza|fica)çã(o$ç|ao$)',
16         'ura': '(t|d)ura$',
17     }
18
19 def match(word):
20     if type(word) == str and regex.search(word) and word not in exceptions:
21         return 1
22     else:
23         return 0
24
25 def getExceptions(aff):
26     '''
27     Dado um afixo, retorna a lista de çõexcees correspondente, carregada a
28     partir de um arquivo JSON.
29     '''
30
31     try:
32         with open(exceptions_file, 'r') as exs:
33             return json.load(exs, object_hook=lambda dic: dic[aff])
```

```
34      except TypeError:
35          return {}
36
37
38  def getContexts(query, n=20):
39      '''
40      Busca contextos de êocorrncia da <query> no corpus, retornando, no ámximo,
41      <n> contextos.
42      '''
43
44      with db:
45          cur = db.cursor()
46          tokens = cur.execute('SELECT corpo FROM Texto WHERE textoid IN (SELECT texto
                FROM Token WHERE token = ? LIMIT ?)',
47                               (query, n)).fetchall()
48
49      return map(lambda t: t[0], tokens)
50
51  def getTokens(affix, corpus, N=0):
52      '''
53      Dado um afixo, retorna todos os tokens que o instanciam no corpus.
54      '''
55
56      global regex, exceptions
57      regex = re.compile(afixo[affix], re.UNICODE)
58      exceptions = getExceptions(affix)
59
60      with db:
61
62          cur = db.cursor()
63
64          db.create_function('match', 1, match)
65
66          if corpus == 'all':
67              cur.execute('SELECT token, stem FROM Token WHERE match(stem) = 1')
68          else:
69              cur.execute('SELECT token, stem FROM Token WHERE corpus = ? AND match(
                    stem) = 1', (corpus,))
70
71      tokens = cur.fetchall()
72
73      print(len(tokens))
74
75      if N > 0:
76          tokens = tokens[:N]
77
78      return tokens
```

A.6 g2pbr.py

```
1  #!/usr/bin/env python3
2
```

```
 3 import re
 4 import sys
 5
 6 class Phon:
 7     '''çã
 8     Aplicao de regras ófonolgicas para o êportugus.
 9
10     A ideia é ler um arquivo com regras que podem ser aplicadas a strings.
11     A çãmotivao inicial para este script foi criar um sistema de çãtranscrioé
12     fontica/ófonolgica a partir de palavras escritas. Como isso pode ser feito
13     de diversas formas, dependendo da variedade do êportugus, as regras ãso
14     dadas por um arquivo de texto carregado pelo script.
15     '''
16
17     vowels = 'aeiou'
18     nasal_v = 'ãẽĩõũ'
19     nasals = 'mn'
20     consonants = 'bcdfghjklmnpqrstvxywzZS'
21
22     def __init__(self, rules='rules.pt'):
23
24         self.model = self.read(rules)
25
26
27     def read(self, source):
28         '''
29         Cada linha do arquivo de regras tem o formato:
30         input   output çõ exceesçõ
31
32         excees é uma lista de palavras separadas por ívrgulas.
33
34         Linhas de ácomentrio çcomeam com '#'.
35         '''
36
37         rules = []
38
39         with open(source, 'r') as s:
40             for line in s:
41                 if line[0] != '#':
42                     rule = list(map(lambda s: s.strip(), (line.split('\t'))))
43                     try:
44                         rule[2] = rule[2].split(',')
45                     except IndexError:
46                         rule.append('')
47                     rules.extend(self.expand(rule))
48
49         return rules
50
51
52     def expand(self, rule):
53         '''
54         As regras podem ser escritas com ávariveis; por exemplo, VC denota
55         qualquer vogal seguida de uma consoante. Esta çãfuno expande essa
56         regra abreviada e tem como valor de retorno uma lista de regras
```

```
57          que abrange todas as possibilidades de çãinstanciao das ávariveis.
58
59          A çãinstanciao de uma ávarivel deve ser a mesma no input/contexto
60          da regra e no output. Assim, se VN -> V pode ser instanciada por an -> a,
61          mas ãno por an -> e.
62          '''
63
64          i = 0
65          cxs = [[list(rule[0]), rule[1]]]
66
67          while i < len(rule[0]):
68
69              if rule[0][i] == 'V':
70                  l = []
71                  for cx in cxs:
72                      for v in self.vowels + self.nasal_v:
73                          cx[0][i] = v
74                          out = cx[1].replace('V', v, 1)
75                          # é preciso dar conta das çõnasalizaes
76                          if rule[1][i] == 'Ṽ':
77                              out = cx[1].replace('Ṽ',
78                                      self.nasal_v[self.vowels.find(v)])
79                          l.append([cx[0][:], out])
80
81              if rule[0][i] == 'Ṽ':
82                  l = []
83                  for cx in cxs:
84                      for v in self.nasal_v:
85                          cx[0][i] = v
86                          out = cx[1].replace('Ṽ', v)
87                          l.append([cx[0][:], out])
88
89              if rule[0][i] == 'C':
90                  l = []
91                  for cx in cxs:
92                      for c in self.consonants:
93                          cx[0][i] = c
94                          out = cx[1].replace('C',c)
95                          l.append([cx[0][:], out])
96
97              if rule[0][i] == 'N':
98                  l = []
99                  for cx in cxs:
100                     for n in self.nasals:
101                         cx[0][i] = n
102                         out = cx[1].replace('N', n)
103                         l.append([cx[0][:], out])
104
105             try:
106                 cxs = l[:]
107             except UnboundLocalError:
108                 pass
109             i += 1
110
```

```
111         return list(map(lambda l: (''.join(l[0]).replace('0', ''),
112             l[1], rule[2]), cxs))
113
114
115     def apply_rule(self, word, rule):
116
117         return re.sub(rule[0], rule[1], word).replace('0', '')
118
119
120     def run(self, words):
121
122         for rule in self.model:
123             words = list(map(lambda w: self.apply_rule(w, rule), words))
124
125         return words
126
127
128 if __name__ == '__main__':
129
130     args = sys.argv[1:]
131     ph = Phon()
132
133     try:
134         with open(args[0], 'r') as words:
135             with open(args[1], 'w') as output:
136                 for line in words:
137                     output.write('\t'.join(ph.run(line.split())) + '\n')
138     except IndexError:
139         print('Indique o arquivo de input e o arquivo de output')
```

APÊNDICE B – Listas

B.1 Lista de palavras com o sufixo -*mento* no corpus geral

Cada palavra é listada com sua respectiva frequência de atestação no corpus.

julgamento	4255	empreendimento	781	isolamento	363
pagamento	3424	monitoramento	759	impedimento	349
investimento	3316	rendimento	730	saneamento	348
atendimento	3151	entendimento	714	alojamento	333
desenvolvimento	2861	fechamento	699	gerenciamento	326
tratamento	2441	andamento	690	assentamento	324
crescimento	2258	reconhecimento	689	requerimento	312
equipamento	2222	policiamento	641	atropelamento	305
treinamento	1695	abastecimento	606	regimento	296
procedimento	1607	medicamento	604	aquecimento	286
estabelecimento	1517	desaparecimento	603	deslizamento	279
conhecimento	1506	alagamento	579	vencimento	275
comportamento	1491	vazamento	561	adiamento	273
orçamento	1447	deslocamento	529	rompimento	265
cruzamento	1349	posicionamento	529	licenciamento	259
levantamento	1288	esclarecimento	514	revezamento	256
funcionamento	1274	questionamento	479	faturamento	244
lançamento	1273	afastamento	477	constrangimento	242
relacionamento	1253	desabamento	456	recebimento	218
financiamento	1243	aproveitamento	445	indiciamento	210
ferimento	1197	acampamento	439	enfrentamento	208
rebaixamento	1187	fornecimento	439	recolhimento	204
acontecimento	1160	desmatamento	423	surgimento	203
estacionamento	1127	sofrimento	411	comprometimento	188
congestionamento	1107	cancelamento	408	armamento	178
planejamento	1094	encerramento	407	desentendimento	177
envolvimento	1074	descumprimento	401	ressarcimento	169
pensamento	1044	regulamento	396	acolhimento	166
sentimento	959	acompanhamento	392	patrulhamento	164
cumprimento	879	pronunciamento	376	enriquecimento	163

fortalecimento	161	endividamento	82	ensinamento	52
descontentamento	160	linchamento	80	prolongamento	52
acostamento	158	envenenamento	78	ordenamento	51
consentimento	158	afogamento	77	recrutamento	51
compartilhamento	149	abatimento	76	condicionamento	50
racionamento	146	parcelamento	76	confinamento	49
encaminhamento	145	detalhamento	73	convencimento	49
processamento	139	estiramento	70	derramamento	49
ajustamento	135	juramento	70	aprofundamento	48
agradecimento	131	arrependimento	69	afundamento	47
desabastecimento	128	carregamento	69	alargamento	47
desligamento	125	enfraquecimento	67	remanejamento	47
superfaturamento	122	envelhecimento	67	esquecimento	46
desdobramento	121	ressecamento	67	ressentimento	46
engarrafamento	121	falecimento	66	engajamento	45
tombamento	119	preenchimento	66	comparecimento	44
entrosamento	114	prosseguimento	66	desmoronamento	44
xingamento	111	mapeamento	65	mandamento	44
sepultamento	109	amadurecimento	64	arrendamento	43
agravamento	106	bombeamento	64	endurecimento	43
acabamento	103	rastreamento	62	fretamento	43
sangramento	103	resfriamento	62	renascimento	43
armazenamento	101	adiantamento	60	entroncamento	42
aperfeiçoamento	100	aliciamento	60	engavetamento	41
congelamento	98	cadastramento	60	seguimento	41
escoamento	98	destelhamento	60	suprimento	41
favorecimento	92	esgotamento	60	chamamento	40
aprimoramento	90	loteamento	60	mantimento	40
cerceamento	90	credenciamento	59	zoneamento	40
espancamento	90	discernimento	59	adestramento	39
salvamento	90	aparecimento	57	direcionamento	39
arquivamento	84	alinhamento	56	acirramento	38
agendamento	83	relaxamento	55	enquadramento	38
arrombamento	83	revestimento	54	rolamento	38
desconhecimento	83	distanciamento	53	transbordamento	38
ligamento	83	desarmamento	52	apontamento	37
reestabelecimento	83	enforcamento	52	descarrilamento	37

cabimento	36	assessoramento	18	indeferimento	12
esvaziamento	36	refinanciamento	18	nivelamento	12
estranhamento	34	cabeamento	17	regramento	12
contingenciamento	33	desflorestamento	17	acasalamento	11
destacamento	33	reabastecimento	17	ajuizamento	11
enxugamento	33	abafamento	16	cercamento	11
assoreamento	32	aleitamento	16	descobrimento	11
batimento	32	barateamento	16	descolamento	11
provimento	32	descredenciamento	16	encanamento	11
apedrejamento	31	calçamento	15	fardamento	11
aparelhamento	30	cozimento	15	pertencimento	11
sucateamento	30	dimensionamento	15	soterramento	11
embasamento	29	formigamento	15	acondicionamento	10
estrangulamento	29	melhoramento	15	aditamento	10
estreitamento	29	apoiamento	14	apodrecimento	10
encarceramento	28	deferimento	14	arrefecimento	10
acionamento	27	emplacamento	14	avivamento	10
desapontamento	27	enterramento	14	contentamento	10
desmembramento	27	passamento	14	desaquecimento	10
recapeamento	27	ranqueamento	14	desmantelamento	10
merecimento	26	afrouxamento	13	emagrecimento	10
adensamento	25	amortecimento	13	empobrecimento	10
empoderamento	25	aterramento	13	esquartejamento	10
desprendimento	24	açodamento	13	fingimento	10
agrupamento	23	beneficiamento	13	reassentamento	10
represamento	22	prejulgamento	13	recrudescimento	10
oferecimento	21	relançamento	13	tabelamento	10
alongamento	20	subfinanciamento	13	tensionamento	10
derretimento	20	aborrecimento	12	aconselhamento	9
recadastramento	20	alisamento	12	chaveamento	9
banimento	19	arredondamento	12	cometimento	9
capotamento	19	desvirtuamento	12	descarregamento	9
encolhimento	19	enchimento	12	empacotamento	9
reflorestamento	19	escapamento	12	encantamento	9
trancamento	19	esmagamento	12	encarecimento	9
alistamento	18	fatiamento	12	espaçamento	9
asfaltamento	18	fuzilamento	12	içamento	9

reposicionamento	9	escoramento	6	esfolamento	4
abortamento	8	exaurimento	6	espalhamento	4
adoecimento	8	florescimento	6	estofamento	4
afunilamento	8	redirecionamento	6	extravasamento	4
agenciamento	8	refinamento	6	fichamento	4
apagamento	8	retardamento	6	grampeamento	4
aprisionamento	8	solapamento	6	justiciamento	4
comedimento	8	sufocamento	6	letramento	4
contigenciamento	8	acobertamento	5	perdimento	4
desassoreamento	8	acoplamento	5	referenciamento	4
descerramento	8	alastramento	5	reordenamento	4
divertimento	8	assentimento	5	replanejamento	4
encastelamento	8	balizamento	5	repovoamento	4
engessamento	8	bronzeamento	5	travamento	4
escalonamento	8	comissionamento	5	abalroamento	3
reaproveitamento	8	desarquivamento	5	acendimento	3
repartimento	8	descongelamento	5	adernamento	3
ressurgimento	8	esfacelamento	5	amesquinhamento	3
abrandamento	7	esfriamento	5	apaziguamento	3
apadrinhamento	7	fraturamento	5	arrebatamento	3
deslumbramento	7	incitamento	5	atordoamento	3
destombamento	7	reaparelhamento	5	avacalhamento	3
entupimento	7	reaquecimento	5	açulamento	3
internamento	7	tratoramento	5	balanceamento	3
justiçamento	7	afinamento	4	compadecimento	3
madeiramento	7	aquartelamento	4	congraçamento	3
ornamento	7	baleamento	4	desalojamento	3
realinhamento	7	carvoejamento	4	descaramento	3
redimensionamento	7	clareamento	4	desencadeamento	3
revigoramento	7	coroamento	4	desentrosamento	3
silenciamento	7	corrimento	4	desmascaramento	3
achatamento	6	defloramento	4	distensionamento	3
aforamento	6	descasamento	4	emparedamento	3
arrastamento	6	desgarramento	4	enamoramento	3
açoitamento	6	desordenamento	4	encoleiramento	3
desinvestimento	6	empilhamento	4	entrelaçamento	3
despojamento	6	envelopamento	4	equacionamento	3

erguimento	3	desenraizamento	2	rejulgamento	2
escorregamento	3	desfavorecimento	2	reparcelamento	2
esfarelamento	3	desfolhamento	2	reprocessamento	2
povoamento	3	desmerecimento	2	secamento	2
propagandeamento	3	desnivelamento	2	sensoriamento	2
reasfaltamento	3	despejamento	2	sentenciamento	2
rebatimento	3	desprovimento	2	sequenciamento	2
recondicionamento	3	destalhamento	2	sobrestamento	2
reinvestimento	3	disciplinamento	2	televisionamento	2
religamento	3	embelezamento	2	tingimento	2
repatriamento	3	emparelhamento	2	trucidamento	2
abrigamento	2	emperramento	2	acanhamento	1
acatamento	2	encurtamento	2	acertamento	1
achincalhamento	2	enlouquecimento	2	acorrentamento	1
aclaramento	2	enrijecimento	2	acumpliciamento	1
afrontamento	2	escalpelamento	2	agarramento	1
ajuntamento	2	esquentamento	2	agigantamento	1
amarelamento	2	estancamento	2	arrebentamento	1
aparamento	2	flagelamento	2	assanhamento	1
apenamento	2	gotejamento	2	atravessamento	1
apensamento	2	guinchamento	2	atrevimento	1
arbitramento	2	inchamento	2	aviltamento	1
arejamento	2	inundamento	2	avistamento	1
arrecadamento	2	locupletamento	2	bloqueamento	1
arruamento	2	outorgamento	2	borramento	1
atingimento	2	perecimento	2	caimento	1
barramento	2	pisoteamento	2	colhimento	1
branqueamento	2	polimento	2	compreendimento	1
curtimento	2	reagendamento	2	coqueamento	1
custeamento	2	reaparecimento	2	desaceleramento	1
desalinhamento	2	reatamento	2	desbaratamento	1
desbordamento	2	rebuscamento	2	descabimento	1
descadastramento	2	recarregamento	2	desentupimento	1
descarnamento	2	recobrimento	2	desfalecimento	1
descomprometimento	2	reequipamento	2	desfazimento	1
descongestionamento	2	reerguimento	2	despedimento	1
desemperramento	2	refreamento	2	desregramento	1

desrepresamento	1	engolimento	1	patenteamento	1
destemperamento	1	enraizamento	1	patrolamento	1
destravamento	1	ensaibramento	1	pinçamento	1
destreinamento	1	envasamento	1	pressentimento	1
desvelamento	1	enxovalhamento	1	prestigiamento	1
desvinculamento	1	esbanjamento	1	protelamento	1
diligenciamento	1	escamoteamento	1	rabaixamento	1
dobramento	1	escaneamento	1	reagrupamento	1
embaralhamento	1	esfaqueamento	1	recenseamento	1
empalamento	1	espelhamento	1	recredenciamento	1
encadeamento	1	fracionamento	1	reencantamento	1
enclausuramento	1	hasteamento	1	referimento	1
encobrimento	1	intrometimento	1	relembramento	1
encorajamento	1	jateamento	1	retreinamento	1
endeusamento	1	mascaramento	1	taxiamento	1
enfaixamento	1	padecimento	1	travestimento	1
enferrujamento	1	pareamento	1	valetamento	1

B.2 Lista de palavras com o sufixo -*ção* no corpus geral

Cada palavra é listada com sua respectiva frequência de atestação no corpus.

informação	8407	criação	2966	constituição	2077
manifestação	7811	opção	2961	contratação	1982
operação	6234	administração	2795	inflação	1980
investigação	6010	edição	2780	divulgação	1968
educação	5082	formação	2766	fiscalização	1933
participação	4965	negociação	2740	apresentação	1898
competição	4540	declaração	2588	colocação	1857
confederação	4399	avaliação	2571	reação	1786
eleição	4291	atuação	2454	proteção	1732
organização	3820	oposição	2378	realização	1725
construção	3545	aprovação	2331	preocupação	1716
redução	3454	acusação	2139	fundação	1685
votação	3447	classificação	2137	punição	1649
instituição	3388	corrupção	2119	circulação	1561
produção	3364	licitação	2107	alteração	1545
associação	3027	federação	2101	reivindicação	1487

instalação	1458	apuração	801	cassação	562
publicação	1417	reprodução	776	coordenação	556
ligação	1391	obrigação	774	prorrogação	555
legislação	1374	reintegração	770	escalação	553
intervenção	1372	consolação	752	prestação	539
geração	1367	graduação	746	projeção	530
recuperação	1343	recomendação	736	destruição	521
liberação	1335	interdição	730	articulação	515
comparação	1322	convocação	728	lotação	510
marcação	1313	representação	722	substituição	510
inscrição	1311	composição	720	gravação	508
orientação	1283	determinação	718	confirmação	505
ocupação	1280	prevenção	703	correção	502
autorização	1240	identificação	701	libertação	490
exposição	1222	tradução	701	diminuição	484
exploração	1197	implantação	695	habitação	484
paralisação	1184	renovação	678	utilização	484
definição	1081	tramitação	676	duração	481
programação	1074	detenção	667	variação	481
concentração	1072	demarcação	663	motivação	479
execução	1063	depredação	646	adaptação	477
indicação	1045	contribuição	642	discriminação	476
condenação	1037	afirmação	641	sinalização	475
mobilização	1012	indenização	634	finalização	474
reclamação	978	interpretação	633	interrupção	474
explicação	977	doação	628	perseguição	470
distribuição	965	evolução	628	importação	448
reeleição	929	proibição	621	solicitação	447
restrição	928	arrecadação	620	isenção	437
resolução	880	violação	616	observação	436
atração	868	comemoração	611	habilitação	430
alimentação	862	inauguração	594	embarcação	416
aplicação	848	documentação	592	atualização	409
ampliação	845	integração	589	valorização	408
movimentação	821	internação	585	remoção	405
delegação	809	eliminação	568	transformação	405
preparação	805	remuneração	564	revelação	400

exportação	396	atribuição	294	inserção	217
vacinação	393	inspiração	293	adequação	216
nomeação	392	combinação	290	poluição	216
exibição	391	elevação	290	complicação	215
consideração	386	reposição	288	qualificação	214
assunção	383	inovação	286	captação	213
desoneração	383	oração	286	comprovação	211
regulamentação	383	localização	284	delação	211
preservação	377	revalidação	284	coligação	210
especulação	374	fabricação	281	manipulação	208
recepção	372	comercialização	279	concepção	206
iluminação	365	aceitação	275	citação	205
indignação	362	condução	273	humilhação	205
contradição	359	elaboração	273	superação	200
percepção	357	erradicação	272	exumação	199
desocupação	353	provocação	267	imposição	198
pontuação	351	conservação	266	irritação	198
alegação	345	extinção	264	precipitação	197
animação	345	frustração	262	gestação	193
infecção	345	aceleração	259	plantação	193
modificação	339	regularização	256	civilização	190
compensação	338	retaliação	250	distorção	188
rejeição	334	demolição	247	reconstrução	188
retenção	334	medição	246	narração	187
separação	334	anulação	245	revogação	187
legalização	332	natação	245	sustentação	186
colaboração	330	simulação	244	devolução	185
aproximação	327	implementação	238	invenção	185
celebração	320	decoração	230	agremiação	184
notificação	318	absolvição	228	reestruturação	183
instrução	316	dedicação	228	abstenção	182
demonstração	307	destinação	228	reputação	182
cooperação	303	desapropriação	227	terceirização	182
limitação	303	descrição	226	modernização	181
prostituição	299	privatização	226	cotação	180
antecipação	296	contaminação	221	exoneração	179
satisfação	296	pichação	218	extradição	179

mineração	177	conciliação	137	computação	104
capacitação	175	oscilação	137	degradação	104
contenção	174	especialização	136	democratização	103
regulação	174	ilustração	136	tributação	103
fixação	172	ocultação	136	transcrição	102
amamentação	171	dominação	135	liquidação	101
aparição	171	descriminalização	134	renegociação	101
traição	171	prescrição	133	escavação	100
obtenção	169	reconstituição	133	aspiração	99
reformulação	169	salvação	133	medicação	99
adulteração	168	imaginação	132	quitação	99
interação	165	reprovação	132	deliberação	98
constatação	162	revitalização	132	expedição	98
verificação	161	infiltração	129	congregação	97
desvalorização	158	injeção	127	interceptação	97
duplicação	158	inundação	127	intimidação	96
imigração	158	certificação	126	validação	96
apreciação	157	conscientização	123	navegação	95
anotação	156	consolidação	122	promulgação	95
edificação	155	reabilitação	120	visitação	94
extração	155	intoxicação	118	colonização	93
admiração	154	pavimentação	117	negação	92
averiguação	153	introdução	116	peregrinação	92
filiação	152	distinção	115	perfuração	92
argumentação	151	vinculação	115	tentação	92
autuação	148	contestação	113	alfabetização	91
continuação	148	responsabilização	112	sonegação	91
armação	147	homologação	111	forração	90
reparação	146	numeração	110	gratificação	90
restauração	146	apelação	108	obstrução	89
desaceleração	144	denominação	108	inflamação	88
migração	144	intimação	107	ventilação	88
criminalização	143	aglomeração	106	menstruação	87
repetição	142	conspiração	106	proposição	87
locação	140	escoriação	106	segregação	86
tripulação	140	transposição	106	mutilação	85
mediação	139	acomodação	105	degustação	84

incorporação	84	proclamação	64	recreação	50
repartição	84	respiração	64	mecanização	48
formulação	83	reinauguração	63	patinação	48
normalização	83	conotação	62	empolgação	47
pacificação	83	emancipação	61	intensificação	47
restituição	81	perturbação	61	intuição	47
flexibilização	80	decapitação	59	padronização	47
implicação	80	inclinação	59	privação	47
imunização	79	mutação	59	resignação	47
hospitalização	78	globalização	57	suspeição	47
difamação	77	insinuação	57	cogitação	46
incitação	77	castração	56	desinformação	46
encenação	75	confraternização	56	desmilitarização	46
urbanização	75	decretação	56	distração	46
falsificação	74	desnutrição	56	industrialização	46
instauração	74	moderação	56	nutrição	46
receptação	73	pregação	56	simplificação	46
ostentação	72	reconciliação	56	agitação	45
proliferação	72	evacuação	55	desorganização	45
recordação	72	reorganização	54	devoção	45
irrigação	71	detecção	53	retração	45
visualização	71	propagação	53	alienação	44
dissertação	70	sedução	53	coação	44
penetração	70	vibração	53	delimitação	44
configuração	69	destituição	52	depilação	44
formalização	69	encarnação	52	diferenciação	44
abolição	67	unificação	52	desidratação	43
redemocratização	67	apropriação	51	estagnação	43
especificação	66	experimentação	51	luxação	43
iniciação	66	inalação	51	objeção	43
reapresentação	66	inquietação	51	retratação	43
decomposição	65	intermediação	51	ponderação	42
deterioração	65	procuração	51	presunção	42
efetivação	65	acumulação	50	equiparação	41
recomposição	65	caracterização	50	polarização	41
refrigeração	65	complementação	50	reedição	41
disseminação	64	reavaliação	50	reinserção	41

veiculação	41	generalização	34	perpetuação	29
deposição	40	politização	34	postulação	29
detonação	40	titulação	34	secreção	29
imitação	40	desobstrução	33	socialização	29
impugnação	40	facilitação	33	deflagração	28
indagação	40	formatação	33	hidratação	28
devastação	39	fundamentação	33	esterilização	27
suposição	39	indexação	33	exclamação	27
abominação	38	redistribuição	33	multiplicação	27
dedução	38	acareação	32	popularização	27
desclassificação	38	contação	32	radicalização	27
descontração	38	oficialização	32	amputação	26
expropriação	38	reaproximação	32	concretização	26
fragmentação	38	rendição	32	meditação	26
internacionalização	38	ressocialização	32	penalização	26
ordenação	38	aclamação	31	remodelação	26
profissionalização	38	banalização	31	triangulação	26
racionalização	38	elucidação	31	ejaculação	25
saudação	38	enrolação	31	infestação	25
torção	38	estruturação	31	requisição	25
universalização	38	gozação	31	retificação	25
bonificação	37	lamentação	31	subordinação	25
vedação	37	precarização	31	abstração	24
compilação	36	aflição	30	aferição	24
readequação	36	consagração	30	centralização	24
reconsideração	36	coroação	30	deportação	24
depreciação	35	desaprovação	30	fertilização	24
improvisação	35	malhação	30	higienização	24
interrogação	35	perdição	30	naturalização	24
judicialização	35	realocação	30	orquestração	24
junção	35	alocação	29	priorização	24
masturbação	35	capitalização	29	sofisticação	24
oxigenação	35	contração	29	viabilização	24
deserção	34	conversação	29	badalação	23
dissolução	34	designação	29	climatização	23
estabilização	34	disponibilização	29	digitalização	23
flutuação	34	objetificação	29	diversificação	23

estatização	23	diagramação	18	contabilização	15
exaltação	23	hesitação	18	danificação	15
mitigação	23	inadequação	18	despoluição	15
reencarnação	23	instrumentalização	18	desvinculação	15
taxação	23	otimização	18	enganação	15
ativação	22	postergação	18	fidelização	15
beatificação	22	recondução	18	ideologização	15
desmoralização	22	sobreposição	18	provação	15
dotação	22	explanação	17	pulverização	15
remarcação	22	informatização	17	purificação	15
desativação	21	nacionalização	17	ramificação	15
descentralização	21	pasteurização	17	regeneração	15
excitação	21	varrição	17	saturação	15
imputação	21	verticalização	17	subcontratação	15
liberalização	21	vitimização	17	amortização	14
municipalização	21	adoração	16	aniquilação	14
absorção	20	categorização	16	desarticulação	14
aclimatação	20	cicatrização	16	desconsideração	14
agilização	20	confrontação	16	escolarização	14
conformação	20	contraposição	16	federalização	14
desincompatibilização	20	deformação	16	modulação	14
individualização	20	digitação	16	operacionalização	14
legitimação	20	elitização	16	patologização	14
procriação	20	fascinação	16	premeditação	14
baldeação	19	humanização	16	reativação	14
canonização	19	invocação	16	reeducação	14
conjugação	19	reafirmação	16	refundação	14
contemplação	19	repactuação	16	suplementação	14
culpabilização	19	repatriação	16	tipificação	14
imobilização	19	rotação	16	adivinhação	13
insubordinação	19	sedação	16	arborização	13
interiorização	19	sucção	16	camarotização	13
requalificação	19	assombração	15	cessação	13
subnotificação	19	canalização	15	coloração	13
cartelização	18	coagulação	15	cooptação	13
condecoração	18	consternação	15	desatualização	13
desconstrução	18	consumação	15	desinsetização	13

espetacularização	13	divagação	10	adulação	8
figuração	13	evangelização	10	afiliação	8
harmonização	13	inseminação	10	ascenção	8
predisposição	13	majoração	10	assimilação	8
recontratação	13	massificação	10	coisificação	8
usurpação	13	moralização	10	consecução	8
contextualização	12	ondulação	10	dedetização	8
desregulamentação	12	ratificação	10	desaparição	8
estimulação	12	reintrodução	10	descaracterização	8
fermentação	12	reprogramação	10	desconfiguração	8
idealização	12	ressignificação	10	descriminação	8
maturação	12	reunificação	10	dissecção	8
minimização	12	afetação	9	ebulição	8
prevaricação	12	alucinação	9	emigração	8
ambientação	11	anexação	9	inutilização	8
circunscrição	11	colação	9	justificação	8
desburocratização	11	cremação	9	louvação	8
desestabilização	11	degeneração	9	marginalização	8
desumanização	11	descatracalização	9	maximização	8
doutrinação	11	desinfecção	9	mercantilização	8
indução	11	desolação	9	neutralização	8
mensuração	11	entonação	9	personificação	8
miscigenação	11	evaporação	9	pulsação	8
normatização	11	floração	9	reidratação	8
readaptação	11	intepretação	9	reiteração	8
reocupação	11	internalização	9	ridicularização	8
arrumação	10	lubrificação	9	subscrição	8
cerração	10	panificação	9	subtração	8
compatibilização	10	partidarização	9	adubação	7
congratulação	10	precificação	9	amarração	7
danação	10	procrastinação	9	autenticação	7
demonização	10	reanimação	9	bolinação	7
desindustrialização	10	retribuição	9	compactação	7
desintegração	10	sujeição	9	contratualização	7
desintoxicação	10	veneração	9	conturbação	7
desmobilização	10	abdicação	8	deleção	7
desqualificação	10	acentuação	8	desagregação	7

desfiliação	7	diplomação	6	injunção	5
dilatação	7	dissipação	6	interpelação	5
diluição	7	esquerdização	6	intersecção	5
dissimulação	7	explicitação	6	islamização	5
escovação	7	fecundação	6	medicalização	5
estigmatização	7	felicitação	6	mistificação	5
execração	7	hostilização	6	motorização	5
fundição	7	irresignação	6	pactuação	5
glamurização	7	materialização	6	pegação	5
gradação	7	obstinação	6	pigmentação	5
impermeabilização	7	palpitação	6	protelação	5
incineração	7	recapitulação	6	reaparição	5
inibição	7	redefinição	6	recriação	5
institucionalização	7	reincorporação	6	reinvenção	5
militarização	7	religação	6	relativização	5
nebulização	7	reorientação	6	remição	5
recolocação	7	replicação	6	republicação	5
reinterpretação	7	ressuscitação	6	reurbanização	5
santificação	7	transpiração	6	reverberação	5
tergiversação	7	acreditação	5	sagração	5
torrefação	7	afobação	5	sensibilização	5
vacilação	7	aglutinação	5	significação	5
abreviação	6	amolação	5	sindicalização	5
aculturação	6	comiseração	5	valoração	5
apartação	6	depravação	5	abnegação	4
apelidação	6	desafetação	5	aliteração	4
argentinização	6	desaposentação	5	anunciação	4
bateção	6	descolonização	5	arguição	4
burocratização	6	desratização	5	atracação	4
cocção	6	desterritorialização	5	capitulação	4
cotização	6	deturpação	5	coabitação	4
decantação	6	dramatização	5	coalização	4
descontaminação	6	estratificação	5	coletivização	4
desertificação	6	extirpação	5	conflagração	4
desestruturação	6	forçação	5	contrafação	4
despolitização	6	fragilização	5	crucificação	4
dilapidação	6	glamourização	5	denunciação	4

derivação	4	vinificação	4	fruição	3
desarrumação	4	zoação	4	granulação	3
desconcentração	4	afixação	3	inaceitabilização	3
desestatização	4	aporrinhação	3	incriminação	3
desregulação	4	atrofização	3	incubação	3
dissociação	4	bajulação	3	instigação	3
elucubração	4	cauterização	3	integralização	3
esculhambação	4	codificação	3	intercepção	3
estadualização	4	comoditização	3	justaposição	3
evitação	4	constrição	3	lavação	3
expatriação	4	corroboração	3	memorização	3
falação	4	culminação	3	objetivação	3
favelização	4	curtição	3	personalização	3
flagelação	4	declamação	3	pressurização	3
imolação	4	decodificação	3	profanação	3
infantilização	4	deglutição	3	publicização	3
inferiorização	4	degolação	3	putrefação	3
invalidação	4	delaminação	3	reacomodação	3
involução	4	depuração	3	rearticulação	3
laceração	4	desassociação	3	redesignação	3
lactação	4	desconvocação	3	reelaboração	3
malversação	4	deslegitimação	3	regravação	3
monitoração	4	desmotivação	3	reinumação	3
mundialização	4	desnaturalização	3	relocação	3
prelibação	4	desobjetificação	3	reordenação	3
prolongação	4	desprogramação	3	ruminação	3
reclassificação	4	desproteção	3	segmentação	3
refutação	4	domesticação	3	sonorização	3
regionalização	4	empulhação	3	suavização	3
reutilização	4	encadernação	3	trepidação	3
setorização	4	encheção	3	tripartição	3
sistematização	4	erotização	3	academização	2
sociabilização	4	estereotipação	3	admoestação	2
subnutrição	4	externalização	3	afinação	2
sustação	4	extrapolação	3	amplificação	2
terminação	4	fabulação	3	arregimentação	2
unitização	4	fanatização	3	assemelhação	2

averbação	2	deseducação	2	hibridização	2
babação	2	desenergização	2	hierarquização	2
bipartição	2	desindexação	2	imbricação	2
bipolarização	2	desinfestação	2	imprecação	2
brasileirização	2	desinternação	2	incapacitação	2
brotação	2	despreocupação	2	incomodação	2
brutalização	2	despressurização	2	incrementação	2
cabralização	2	dessazonalização	2	incrustação	2
calcificação	2	dessubjetivação	2	interposição	2
calibração	2	dissecação	2	inviabilização	2
carbonização	2	editoração	2	irradiação	2
caricaturização	2	efetuação	2	judiação	2
carnavalização	2	ejeção	2	levitação	2
catalogação	2	eleitorização	2	locupletação	2
catação	2	eletrificação	2	metropolização	2
centrifugação	2	emanação	2	midiatização	2
coibição	2	emasculação	2	mitificação	2
comutação	2	encampação	2	monitorização	2
conceituação	2	equalização	2	murdoquização	2
conclamação	2	escravização	2	obliteração	2
condensação	2	esferificação	2	oneração	2
consignação	2	esfregação	2	ossificação	2
constitucionalização	2	estipulação	2	oxidação	2
contemporização	2	estrangeirização	2	partição	2
contorção	2	eternização	2	paulistização	2
cravação	2	euforização	2	piração	2
cumulação	2	evocação	2	polinização	2
datenização	2	exacerbação	2	prestidigitação	2
declinação	2	exortação	2	problematização	2
defraudação	2	expectoração	2	protagonização	2
denegação	2	expiação	2	protocolação	2
denotação	2	familiarização	2	purgação	2
desautomatização	2	feminização	2	quarteirização	2
desautorização	2	fossilização	2	recapitalização	2
desbancarização	2	fulanização	2	reclinação	2
descontratualização	2	gentrificação	2	reconfiguração	2
descoordenação	2	gourmetização	2	recriminação	2

reencenação	2	arbitração	1	despatologização	1
reexibição	2	arrebentação	1	despenalização	1
reificação	2	atribulação	1	despersonalização	1
reindexação	2	automatização	1	despublicação	1
reinternação	2	avacalhação	1	desrealização	1
remediação	2	azaração	1	dessalinização	1
renderização	2	barração	1	eletrocução	1
revitimização	2	bastardização	1	embolização	1
salivação	2	bifurcação	1	embromação	1
subjugação	2	camominação	1	encucação	1
subutilização	2	causação	1	epistemologização	1
superposição	2	chateação	1	esfoliação	1
tabulação	2	clarificação	1	esquematização	1
tarifação	2	completação	1	estupefação	1
territorialização	2	conceitualização	1	exasperação	1
totalização	2	conglomeração	1	exculpação	1
transmigração	2	conspurcação	1	exemplificação	1
transmutação	2	conurbação	1	expiração	1
tribulação	2	convalidação	1	extubação	1
trocação	2	cristalização	1	exultação	1
uniformização	2	defloração	1	fascistização	1
vermifugação	2	degravação	1	feminilização	1
xingação	2	demoção	1	ferveção	1
abdução	1	densificação	1	ficcionalização	1
acidificação	1	descamação	1	fornicação	1
adjetivação	1	descoloração	1	gamificação	1
adultização	1	descompatibilização	1	garotização	1
agregação	1	descompensação	1	glorificação	1
agrupação	1	desdeificação	1	gravitação	1
aleitação	1	deselitização	1	grenalização	1
angariação	1	desinfectação	1	hipertextualização	1
angulação	1	desinflação	1	homogeneização	1
animalização	1	desinibição	1	horizontalização	1
apalpação	1	desinterdição	1	idiotização	1
apolitização	1	desmistificação	1	igualização	1
aposentação	1	desopilação	1	ilibação	1
aprimoração	1	desorientação	1	importunação	1

indisponibilização	1	plastinação	1	satelitização	1
individuação	1	plutocratização	1	secação	1
instanciação	1	predação	1	securitização	1
instação	1	preterição	1	sedimentação	1
intubação	1	proscrição	1	semaforização	1
invisibilização	1	prostração	1	sexualização	1
ironização	1	puxação	1	sodomização	1
isolação	1	reabsorção	1	solidificação	1
laicização	1	realimentação	1	sufocação	1
laminação	1	recertificação	1	tonificação	1
lapidação	1	recitação	1	totemização	1
maquinação	1	recombinação	1	tradicionalização	1
masculinização	1	recompensação	1	tranquilização	1
mastigação	1	regurgitação	1	transfiguração	1
notação	1	rememoração	1	transgenitalização	1
obfuscação	1	repaginação	1	transliteração	1
obnubilação	1	repavimentação	1	transplantação	1
oportunização	1	retorção	1	trasladação	1
ovulação	1	retroação	1	trepanação	1
paginação	1	revalorização	1	triplicação	1
paparicação	1	revascularização	1	veganização	1
pejotização	1	robotização	1	vitimação	1
periclitação	1	romantização	1	vocalização	1
periferização	1	sacralização	1	vulgarização	1
permutação	1	sanitização	1		

B.3 Predições do MGL sobre as bases da lista de teste

A transcrição dos dados abaixo segue o Alfabeto Fonético Internacional (IPA), com exceção dos seguintes símbolos, com suas correspondências no IPA indicadas:

C ⇔ ʧ Z ⇔ ʒ S ⇔ ʃ
ñ ⇔ ɳ L ⇔ ʎ r ⇔ ɾ
ê ⇔ ẽ î ⇔ ĩ û ⇔ ũ

Palavra prevista	**Score**	**Concordância com o corpus**
fleksibilizasãw	0.9603511600607122	1
formalizasãw	0.9603511600607122	1

Zudisializasãw	0.9654263153541583	1
kõtasãw	0.5612950027850393	1
prekarizasãw	0.9696859383994192	1
penalizasãw	0.9654263153541583	1
iZienizasãw	0.9790311756734119	1
priorizasãw	0.9696859383994192	1
kuwpabilizasãw	0.9603511600607122	1
kartelizasãw	0.9603511600607122	1
vertikalizasãw	0.9654263153541583	1
vitimizasãw	0.9790311756734119	1
elitizasãw	0.9790311756734119	1
repaktuamêto	0.6157203369575124	0
ideoloZizasãw	0.9696859383994192	1
operasionalizasãw	0.9654263153541583	1
patoloZizasãw	0.9696859383994192	1
kamarotizasãw	0.9790311756734119	1
espetakularizasãw	0.9696859383994192	1
kõtestualizasãw	0.9603511600607122	1
ãbiêtasãw	0.856619906452439	1
normatizasãw	0.9790311756734119	1
kõpatibilizasãw	0.9603511600607122	1
demonizasãw	0.9790311756734119	1
resignifikasãw	0.9547889123314056	1
deskatrakalizasãw	0.9654263153541583	1
presifikasãw	0.9547889123314056	1
koizifikasãw	0.9647033020341809	1
merkãtilizasãw	0.9603511600607122	1
kõtratualizasãw	0.9603511600607122	1
estigmatizasãw	0.9790311756734119	1
glamurizasãw	0.9696859383994192	1
arZêtinizasãw	0.9790311756734119	1
eskerdizasãw	0.9790311756734119	1
ostilizasãw	0.9603511600607122	1
akreditasãw	0.6650424946640602	1
fraZilizasãw	0.9603511600607122	1
glamurizasãw	0.9696859383994192	1
medikalizasãw	0.9654263153541583	1

relativizasãw	0.9509088094682157	1
mûdializasãw	0.9654263153541583	1
setorizasãw	0.9696859383994192	1
zoamêto	0.872669534858661	0
atrofizasãw	0.9509088094682157	1
komoditizasãw	0.9790311756734119	1
estereotipasãw	0.5731470318949038	1
esternalizasãw	0.9654263153541583	1
fanatizasãw	0.9790311756734119	1
inaseitabilizasãw	0.9603511600607122	1
publisizasãw	0.976596958098565	1
akademizasãw	0.9790311756734119	1
babasãw	0.5731470318949038	1
brazileirizasãw	0.9696859383994192	1
brutalizasãw	0.9654263153541583	1
kabralizasãw	0.9654263153541583	1
karikaturizasãw	0.9696859383994192	1
kõstitusionalizasãw	0.9654263153541583	1
datenizasãw	0.9790311756734119	1
desbãkarizasãw	0.9696859383994192	1
desazonalizasãw	0.9654263153541583	1
eleitorizasãw	0.9696859383994192	1
esferifikasãw	0.9723642885055108	1
estrãZeirizasãw	0.9696859383994192	1
eternizasãw	0.9790311756734119	1
euforizasãw	0.9696859383994192	1
feminizasãw	0.9790311756734119	1
fulanizasãw	0.9790311756734119	1
Zêtrifikasãw	0.9723642885055108	1
gourmetizasãw	0.9790311756734119	1
îkomodasãw	0.5474734153559959	1
îviabilizasãw	0.9603511600607122	1
metropolizasãw	0.9603511600607122	1
murdokizasãw	0.9498814842639756	1
onerasãw	0.7195281071782736	1
paulistizasãw	0.9790311756734119	1
pirasãw	0.5436125234774447	1

problematizasãw	0.9790311756734119	1
protagonizasãw	0.9790311756734119	1
protokolamêto	0.7600081742324856	0
quarteirizasãw	0.9696859383994192	1
remediasãw	0.7195281071782736	1
rêderizasãw	0.9696859383994192	1
tabulasãw	0.8104605458843437	1
texitorializasãw	0.9654263153541583	1
trokasãw	0.8135853253579092	1
vermifugasãw	0.5731470318949038	1
aduwtizasãw	0.9790311756734119	1
bastardizasãw	0.9790311756734119	1
kamominasãw	0.8896825182728914	1
kõseitualizasãw	0.9603511600607122	1
epistemoloZizasãw	0.9696859383994192	1
feminilizasãw	0.9603511600607122	1
fervimêto	0.9337639899767065	0
fiksionalizasãw	0.9654263153541583	1
gamifikasãw	0.9547889123314056	1
garotizasãw	0.9790311756734119	1
grenalizasãw	0.9654263153541583	1
ipertestualizasãw	0.9603511600607122	1
orizõtalizasãw	0.9654263153541583	1
idiotizasãw	0.9790311756734119	1
îstãsiamêto	0.5551913612528647	0
îstamêto	0.7311620947761734	0
îvizibilizasãw	0.9603511600607122	1
ironizasãw	0.9790311756734119	1
maskulinizasãw	0.9790311756734119	1
obfuskasãw	0.6526419647510655	1
oportunizasãw	0.9790311756734119	1
paparikasãw	0.9382085137805416	1
peZotizasãw	0.9790311756734119	1
periklitasãw	0.7600081742324856	1
periferizasãw	0.9696859383994192	1
plastinasãw	0.8289449578245965	1
plutokratizasãw	0.9790311756734119	1

romãtizasãw	0.9790311756734119	1
semaforizasãw	0.9696859383994192	1
seksualizasãw	0.9603511600607122	1
sodomizasãw	0.9790311756734119	1
totemizasãw	0.9790311756734119	1
tradisionalizasãw	0.9654263153541583	1
trãsZenitalizasãw	0.9654263153541583	1
veganizasãw	0.9790311756734119	1
vitimasãw	0.7665514247294686	1
aZêdasãw	0.5474734153559959	0
êpoderasãw	0.7195281071782736	0
kapotasãw	0.5612950027850393	0
rãkeamêto	0.5982200948604091	1
atexamêto	0.7087546455954752	1
alizasãw	0.9603511600607122	0
fatiasãw	0.6411001972789299	0
regrasãw	0.5436125234774447	0
serkasãw	0.7384578933822437	0
deskolamêto	0.7600081742324856	1
pertêsimêto	0.9589873484855508	1
reasêtamêto	0.653144413758534	1
têsionamêto	0.8104605458843437	1
êZesamêto	0.8817511624454536	1
silêsiamêto	0.5551913612528647	1
ezaurimêto	0.6860810861413094	1
akobertamêto	0.7311620947761734	1
tratorasãw	0.573201666597758	0
baleamêto	0.5982200948604091	1
karvoeZamêto	0.9362694926773609	1
desgaxamêto	0.7087546455954752	1
Zustisiamêto	0.5551913612528647	1
referêsiamêto	0.5551913612528647	1
distêsionamêto	0.8104605458843437	1
êparedasãw	0.5665879321644713	0
enamorasãw	0.573201666597758	0
êkoleiramêto	0.7913819572047552	1
ergimêto	0.9277095587939276	1

propagãdeamêto	0.5982200948604091	1
abrigasãw	0.5731470318949038	0
amarelamêto	0.8492810135065826	1
aparasãw	0.5687634409190985	0
apenasãw	0.6652592185497399	0
atîZimêto	0.7167143077898737	1
deskarnasãw	0.6303599297956773	0
eskawpelamêto	0.7195281071782736	1
flaZelamêto	0.8492810135065826	1
gîSamêto	0.8687647160535693	1
outorgasãw	0.5731470318949038	0
pizoteamêto	0.5982200948604091	1
sekasãw	0.9237954424782447	1
akûplisiamêto	0.5551913612528647	1
avistamêto	0.7311620947761734	1
blokeamêto	0.5982200948604091	1
boxamêto	0.7087546455954752	1
kõpreêdimêto	0.9277095587939276	1
kokeamêto	0.5982200948604091	1
destêperasãw	0.7195281071782736	0
diliZêsiamêto	0.5551913612528647	1
êbaraLamêto	0.8041039423282165	1
êpalamêto	0.5609002242230826	1
êfaiSamêto	0.8687647160535693	1
êgolimêto	0.6787548754947956	1
eskamoteamêto	0.5982200948604091	1
esfakeamêto	0.5982200948604091	1
Zateamêto	0.5982200948604091	1
pareamêto	0.5982200948604091	1
patrolamêto	0.7600081742324856	1
prestiZiasãw	0.5744150009524565	0
protelamêto	0.8492810135065826	1
taksiamêto	0.5551913612528647	1
travestimêto	0.6860810861413094	1
valetasãw	0.5715024731072729	0