Sobre a construção de um recurso léxico de elementos nominais agentivos e de ação para o processamento computacional do português brasileiro

Building a lexicon of agentives and action names for computational processing of Brazilian Portuguese

Andréa Feitosa dos Santos[*]
Carlos Eduardo Atencio Torres[**]
Hélio Leonam Barroso Silva[***]

RESUMO: O principal objetivo desse trabalho é descrever o processo de elaboração de um léxico computacional de elementos nominais agentivos e de ação do Português Brasileiro, construído principalmente para contribuir com a expansão dos recursos léxicos existentes para esta língua. Busca-se, além disso, mostrar como a descrição das línguas encontra espaço dentro do Processamento de Linguagem Natural fazendo uso das Tecnologias de Estados Finitos. O léxico abrange os seguintes aspectos da derivação dos chamados falsos aumentativos: (i) formação sufixal com *–ão* de nomes agentivos; (ii) formação sufixal com *–ão* de adjetivos agentivos; (iii) formação sufixal com *–ão* de nomes de ação. Para derivar os diferentes sentidos presentes nesses três contextos, constatados no português padrão, argumentamos a favor de uma derivação para o sufixo *–ão* em que esse associa-se a uma base verbal. Assim, o léxico foi criado com um duplo objetivo: fornecer subsídios linguísticos para uma aplicação do Processamento de Linguagem Natural e contribuir com a caracterização de elementos nominais agentivos e de ação da língua portuguesa.

PALAVRAS-CHAVE: Léxico computacional. Transdutores de estados finitos. Análise morfológica. Análise derivacional.

ABSTRACT: The main objective of this paper is to describe the process of developing a computational lexicon of agentive nominal elements and action of Brazilian Portuguese built mainly to expand existing lexical resources for this language. We aim, in addition, to show how the description of the languages fits inside the Natural Language Processing making use of Finite State Technologies. The lexicon covers the following aspects of the derivation of the so-called false augmentative: (i) suffixal formation with *–ão* of agentive names; (ii) suffixal formation with *–ão* of agentive adjectives; (iii) suffixal formation with *–ão* of action names. To derive different meanings present in these three contexts, observable in standard Portuguese, we argue in favor of a derivation for the suffix *-ão* that is associated with a verbal base. Thus, the lexicon was created with a dual purpose: to provide language support to an application of natural language processing and contribute to the characterization of agentive nominal expressions and to the action of the Portuguese language.

KEYWORDS: Computational lexicon. Finite state transducers. Morphological analysis. Derivational analysis.

[*] Doutora em Linguística. Pesquisadora do Grupo de Redes de Computadores, Engenharia de Softwares e Sistemas (GREat) da Universidade Federal do Ceará (UFC) – andreafeitosasantos@gmail.com
[**] Mestre em Ciência da Computação. Analista de sistemas do Grupo de Redes de Computadores, Engenharia de Softwares e Sistemas (GREat) da Universidade Federal do Ceará (UFC) – carlostorres@great.ufc.br
[***] Acadêmico do curso de Licenciatura em Letras – Universidade Federal do Ceará – heliobarroso@great.ufc.br

1. Introdução

Muitas ferramentas para o Processamento de Linguagem Natural (PLN) requerem ou se beneficiam de recursos linguísticos, tais como: gramáticas e léxicos. Para tarefas como a análise sintática automática, um léxico morfológico é uma fonte altamente valiosa de informação (SAGOT, 2014).

Um analisador morfológico normalmente constitui uma das primeiras etapas de uma *pipeline* de processamento computacional da linguagem de alto nível. Portanto quanto maior e mais qualitativa for a cobertura da base de dados de um recurso léxico, melhores serão as condições de atuação do analisador, pois o não reconhecimento de uma palavra pode comprometer toda a análise de uma sentença nas etapas de processamento sintático e semântico posteriores ao processamento morfológico (ALENCAR, 2014; MAHLOW, 2011; SMARSARO, 2007, MOTA, 2000).

Para o português brasileiro (PB), até onde sabemos, o dicionário computacional com maior cobertura e livremente disponível é o Dicionário de Palavras Simples Flexionadas para o Português Brasileiro, DELAF_PB (MUNIZ, 2004). Esse recurso léxico possui aproximadamente 880.000 formas simples flexionadas.

No entanto, tal número é ainda bastante pequeno quando comparado a recursos léxicos de outras línguas. O The Leffe (MOLINERO, SAGOT, NICOLAS, 2009), léxico morfológico de larga escala e livremente disponível para o espanhol, possui 1.590,000 entradas. Já o DeLex (SAGOT, 2014), léxico computacional para o alemão, possui 2.3 milhões de entradas.

Como novas palavras são constantemente formadas em línguas naturais, pela adição de afixos a palavras já presentes no léxico, torna-se necessária uma frequente atualização dos recursos. Devido a isso, a expansão de um léxico computacional é uma tarefa contínua.

Observando-se o grande potencial de produtividade do sufixo *–ão* na formação de deverbais agentivos, i. e. formas que trazem em seu conteúdo a ideia de agente, e de deverbais de ação, i.e. formas que trazem em seu conteúdo a ideia de resultado decorrente de uma ação, a principal motivação para se desenvolver um recurso léxico de elementos nominais para o processamento computacional do português do Brasil foi que o DELAF_PB, mesmo já sendo muito útil no processamento do português, sob a perspectiva do PLN, pode ser ainda mais abrangente, pois possui em sua base apenas 130 palavras variando entre as noções de agentivos e nomes de ação.

E, para aproveitar essa potencialidade do DELAF_PB, esse artigo vem exatamente

colaborar com um recurso léxico de elementos nominais agentivos e de ação para o processamento computacional do português do Brasil. Para a construção do recurso, implementou-se em Foma (HULDEN, 2009) - ambiente para o processamento de línguas naturais com base na Tecnologia de Estados Finitos (TEF) - um transdutor de estados finitos arquitetado para gerar e analisar formas deverbais com sufixação em *–ão*. Desse modo, o trabalho vem contribuir efetivamente com os recursos já existentes, através dos 141.170 novos vocábulos obtidos a partir da implementação do transdutor.

Portanto, diante da necessidade de expansão dos recursos léxicos para português brasileiro, esse trabalho tem duplo objetivo: i. fornecer subsídios linguísticos fundamentados para aplicações do PLN, utilizando a TEF e ii. contribuir com a caracterização dos falsos aumentativos, sobretudo as expressões nominais agentivas e os nomes de ação, de modo a contribuir com a formalização do português para o propósito do PLN, pois esta língua, apesar de ser a quinta língua mais falada no mundo e a terceira no mundo ocidental, ainda tem muito a ser tratado para fins computacionais.

Esse trabalho está organizado da seguinte maneira. Na segunda seção, apresentamos a importância do léxico para o PLN. Na seção 3, descrevemos o processo de formação de elementos nominais agentivos e de ação e argumentamos a favor da análise deverbal da sufixação em *–ão* nesse processo derivacional. A seção 4 apresenta uma análise dos aspectos morfológicos envolvidos nos deverbais encontrados no DELAF_PB. Na seção 5, demonstramos o processo de elaboração do léxico e apresentamos os resultados alcançados. Finalmente, a última seção sumariza as principais características do léxico e aponta possíveis direções para pesquisas futuras.

2. PLN e o Léxico

No âmbito do PLN, segundo Freitas (2013, p. 1032),

> léxicos se referem ao componente de um sistema que contém informação (semântica e/ou gramatical) sobre palavras ou expressões, enquanto o termo dicionário, normalmente remete a objetos (impressos ou eletrônicos) destinados a leitores humanos.

Assim, uma das tarefas de PLN realizadas por sistemas, tais como os de *Extração de Informação* (EI), *Perguntas e Respostas* e *Sumarização de Textos*, tomam por base léxicos gerais ou específicos de um determinado domínio (MAHLOW, 2011). No entanto, Duran

(2013) chama atenção para o problema da falta de um pré-conhecimento no computador. A autora coloca o exemplo de uma pessoa estrangeira que tenta identificar o gênero de uma palavra com base na sua língua nativa. Pensando dessa forma, o computador também deve possuir uma língua nativa e, por isso, um recurso léxico precisa ser descrito exaustivamente.

Em computação, para se resolver alguns problemas, primeiramente deve-se verificar a existência de um algoritmo fechado para se chegar a uma solução (CORMEN, 2002). Caso não exista, como acontece na maioria dos problemas computacionais cotidianos, dependendo do problema, deve-se avaliar um tipo de solução com outros tipos de recursos: heurísticas, algoritmos probabilísticos, estatísticos ou métodos de aprendizado computacional (RUSSELL, 2004). Estas soluções são sempre usadas no PLN e empregam, juntamente com recursos computacionais, o conhecimento de especialistas em linguística, como na construção de recursos léxicos.

Portanto, o sucesso de um sistema com PLN dependerá tanto dos algoritmos como dos subsídios lexicais fornecidos por linguistas. Tais subsídios devem abranger idealmente toda riqueza lexical dos sistemas linguísticos e, devido à característica versátil da língua, os recursos léxicos devem se expandir continuamente.

A importância dos recursos léxicos se deve, principalmente, à existência de palavras que merecem tratamento especial. Torres (2012), em seu trabalho sobre identificação de estruturas sujeito – verbo – objeto (SVO) em textos não formatados, utilizou diferentes regras para identificar esse padrão, mas os resultados mostraram que, às vezes, um erro de classificação morfossintática fazia com que outros erros fossem desencadeados em processos posteriores.

Torres (2012), em seu trabalho, apresenta o exemplo da classificação errônea para a palavra *aí*. Em sua implementação esse elemento de natureza adverbial estava sendo identificado como substantivo e isto, consequentemente, causava um grande impacto negativo no sistema. A solução adotada foi a adição dessa entrada lexical na lista de *stopwords*[1], porém, isto fez o sistema perder robustez, ou seja, a capacidade de encontrar uma solução diante de um determinado problema.

Em outro caso, a palavra *que* era ora identificada como determinante, ora como substantivo, ora como pronome. No entanto, esta palavra não pôde ser colocada na lista de *stopwords* por ser uma palavra importante para o reconhecimento das orações subordinadas.

[1] Lista de palavras que o administrador de um sistema define, para que sejam ignoradas durante o processamento de um determinado texto.

Nota-se, portanto, que a solução para isto seria uma base léxica suficientemente grande e não um conjunto de regras para tentar abranger a maioria dos casos.

Desse modo, uma base léxica suficientemente grande e bem fundamentada é extremamente útil para tarefas nada triviais, como a EI, que, atualmente, utiliza técnicas robustas como os métodos de estados finitos para solucionar problemas decorrentes do fato de as informações estarem de certo modo desestruturadas nos textos (JURAFSKY & MARTIN, 2009, p. 725).

Sendo o léxico um repositório de palavras, esse deve consistir de uma lista de todas as palavras de uma língua, incluindo nomes próprios. Uma vez que seria inconveniente ou mesmo impossível, por várias razões, listar todas estas palavras, léxicos computacionais são normalmente estruturados com uma lista de cada raiz e afixo das línguas, juntamente com uma representação morfotática que nos diz como eles podem se encaixar (JURAFSKY & MARTIN, 2009, p. 54). Mais sobre essa forma de se estruturar um léxico computacional, falaremos na seção 5 do presente artigo.

Na próxima seção, com o intuito de fundamentar a construção do nosso recurso léxico, descrevemos o processo de formação dos elementos nominais agentivos e de ação e argumentamos a favor da análise deverbal da sufixação em *–ão* nesse processo derivacional.

3. Elementos nominais agentivos e de ação

A possibilidade de criação de novos vocábulos em português se dá de muitas formas. A compreensão dos processos de formação de palavras de uma língua, tais como a sufixação, é uma parte crítica da morfologia derivacional. Entender o comportamento morfológico e semântico do processo de sufixação é inestimável para a compreensão da produtividade dos processos morfológicos derivacionais.

A derivação morfológica é uma das maneiras de se perceber a enorme riqueza das línguas naturais. Esse assunto é sempre retomado e mencionado por linguistas e gramáticos, dentre os quais pode-se citar Câmara Júnior (2009), Rocha (2008), Bechara (2009) e Cunha e Cintra (2013).

Em geral, em suas análises, todos se posicionam claramente em favor do aspecto não obrigatório e assistemático que caracteriza o processo derivacional de criação de novas palavras. Essa é a principal diferença dos sufixos flexionais, ou desinências, os quais indicam que um vocábulo se dobra a novos empregos do mesmo vocábulo (derivação natural) e por isso

não se devem confundir com os sufixos derivacionais, destinados a criar novos vocábulos – derivação voluntária (CÂMARA JÚNIOR, 2009).

Contrariamente à acepção assistemática do processo derivacional, é possível defender a sistematicidade dos processos morfológicos de flexão e derivação no português, como se pode observar em diversos trabalhos, dentre os quais pode-se citar Monteiro (2002), Basílio (2006) e Silva (2009). No tocante ao sufixo *–ão*, o autor se posiciona a favor da formação dos agentivos derivados a partir desse sufixo, no entanto deixa em aberto se as novas palavras são derivadas de uma base verbal ou nominal. A partir disto, esse trabalho apresenta um argumento a favor da análise deverbal da sufixação em *–ão* e da sua sistematização na formação de agentivos e também de nomes de ação, mostrando que ela é preferível a uma análise que considere um nome como origem da derivação.

Com base na observação do Dicionário de Palavras Simples para o Português Brasileiro, O DELAS_PB[2] (MUNIZ, 2004), verificou-se que o léxico nominal representa o maior grupo de palavras desse dicionário. Na perspectiva da morfologia derivacional, a possibilidade de criação de novos vocábulos nominais do PB se dá, especialmente, pela derivação sufixal.

Uma análise do morfema *–ão* no Aulete Digital revelou que esse dicionário eletrônico, apesar de seu inestimável valor, atribui ao sufixo em questão, além da interpretação original de grau aumentativo, apenas a noção de nomes de ação. No entanto, em seus estudos sobre a polissemia desse morfema, Pezatti (1989) defende a presença, nessa forma sufixal, do valor semântico agentivo, mesmo que carregue um traço aumentativo. Esta ideia é corroborada por Santos (2009) e Rodrigues e Vale (2013), para os quais a sufixação é o fenômeno responsável pela formação dos nomes agentivos e dos nomes de ação derivados através do morfema *–ão*.

A derivação sufixal em *-ão* constitui processo gramatical e semântico de considerada riqueza e flexibilidade de uso no léxico. Um primeiro ponto a ser ressaltado é que há uma considerável quantidade de formas derivadas desse sufixo. A forma mais produtiva com esse sufixo em PB são os casos de grau aumentativo, como *narigão* e *meninão*. Outra forma bastante produtiva em Português é a formação de alguns gentílicos, como em *gascão*, que não designa aumentativo e sim denota valor semântico de *aquele que nasceu na Gasconha* (SANTOS, 2009; ARMELIN, 2011).

[2] O DELAS_PB possui aproximadamente 67.500 entradas canônicas associadas a suas regras de flexão. Dentre estas, os adjetivos e os nomes somam 50.145 entradas. Esse número corresponde a, aproximadamente, 75% dos vocábulos.

Outro ponto interessante que queremos ressaltar é que, se compararmos a derivação sufixal em *–ão* com outros formadores de aumentativo em PB, como *–aço* e *–aréu*, verifica-se que esses morfemas, ao serem alternados em um mesmo contexto morfológico com a forma *–ão*, não recebem os sentidos agentivos e nem designam nomes de ação, como se vê no morfema em questão (SANTOS, 2009). Observe os exemplos:

(1) Mulher – mulherão – mulheraço
(2) Mundo – mundão – mundaréu

Em (1) e (2) o grau se dá pela adjunção de sufixos, mas sem que haja a mudança da classe gramatical. Os novos vocábulos permanecem com a mesma classe gramatical das bases de derivação (ROCHA, 1986). Agora, vejamos o que acontece se tentarmos fazer o mesmo teste com bases verbais.

(3) Pedir – pidão - *pidaço - *pidaréu
(4) Comichar – comichão - *comichaço – *comicharéu

Dado que *pedir* e *comichar* pertencem à classe dos verbos, esses não podem formar vocábulos nominais com grau de aumentativo. Se os vocábulos nominais *pidão* e *comichão* podem ser derivados de verbo, então temos um indício de que não se trata de um processo derivacional de grau, mas sim de um processo derivacional produtivo na formação de agentivos e resultado de ação.

Com base nisto vemos que, além de grau de aumentativo, as formas de sufixação em *–ão* realmente apresentam duas significações a mais para esse morfema, a saber, *nomina agentis* (agente) e *nomina actionis* (ação ou resultado). A questão aqui colocada é saber se um substantivo ou adjetivo terminado em *–ão* deriva de fato de um verbo.

Nesse trabalho partimos da defesa dos processos deverbais envolvidos na derivação sufixal em –ão, pois, como será visto um pouco mais adiante, fica claro, que por uma questão de economia, o processo deverbal é mais econômico que uma derivação verbal somada a uma derivação nominal.

Observe os exemplos abaixo:

(5) **Agente (nomina agentis)**
brigão

chorão
fujão
pidão

Nos exemplos em (5), vemos que o sufixo *-ão* confere à base o valor semântico agentivo "aquele que *briga/chora/foge/pede*", associada a um traço de intensidade. Os vocábulos *brigão, chorão* e *fujão* em (5) podem ter sido derivados tanto das bases nominais *briga*, *choro* e *fuga*, como das bases verbais *brigar*, *chorar* e *fugir.* Somente o vocábulo *pidão* nos permite afirmar que se trata de uma derivação deverbal, uma vez que não há uma base nominal da qual pudesse ter sido derivado.

Agora vejamos os seguintes exemplos:

(6) **Ação ou resultado de um verbo (nomina actionis)**
beliscão
apertão
esfregão
cutucão

Nos exemplos em (6) o sufixo atribui à base o valor semântico de "ação ou resultado de uma ação". Note que tanto *beliscão*, como *esfregão* e *cutucão* possuem apenas as bases verbais *beliscar*, *esfregar* e *cutucar*. Somente o vocábulo *apertão* pode ter sido derivado ou da base nominal *aperto* ou da base verbal *apertar*.

Observe ainda que tanto em (5) como em (6), o sufixo transmite a ideia de agente e de ação ligada ao significado existente nos verbos, mas também transforma o verbo em um nome. Nesses casos, o sufixo não só empresta significado acessório ao semantema como muda a palavra de uma classe ou função para outra. Desse modo, com base nos contrastes dos exemplos em (5) e em (6), nos posicionamos em favor da derivação verbal desses vocábulos.

Nesse trabalho apresentamos um argumento a favor da análise deverbal da sufixação em *–ão* para a formação de agentivos e nomes de ação, mostrando que ela é preferível a uma análise que considere um nome como a origem da derivação. Na próxima seção apresentamos uma análise dos aspectos morfológicos envolvidos nos deverbais encontrados no DELAF_PB e fazemos uma classificação tipológica desses deverbais.

4. Análise do DELAF_PB

Nesse trabalho tratamos, especialmente, de derivação sufixal. Em português, a sufixação é um fenômeno que constitui processo gramatical e semântico de considerada riqueza e

flexibilidade de uso no léxico. Esse processo derivacional é responsável pela formação dos nomes agentivos e dos nomes de ação, derivados através do morfema *–ão*.

Devido a necessidade de descrever o léxico do português para fins de processamento automático, um vasto número de vocábulos em *-ão* foi observado, a fim de fazer uma descrição do conteúdo desta forma sufixal. Com o intuito de sistematizar os aspectos morfológicos envolvidos nesta derivação sufixal, foi feito um levantamento das entradas lexicais do DELAF_PB.

O DELAF_PB, até onde sabemos, é o recurso léxico do PB com maior cobertura e que está livremente disponível para download no site do Núcleo Interinstitucional de Linguística Computacional (NILC). Mesmo possuindo, aproximadamente, 880.000 palavras flexionadas, esse recurso léxico possui apenas 130 palavras variando entre as noções de agentivos e nomes de ação.

Para chegarmos a esse número, inicialmente foram selecionados do DELAF_PB todos os vocábulos terminados em *–ão*, incluindo nomes, verbos, adjetivos, advérbios, etc. Em seguida, por estarmos tratando somente do caso dos falsos aumentativos[3], foram consideradas apenas as palavras da classe dos adjetivos e dos substantivos. Por fim, como não estamos considerando os casos de aumentativos, foram retiradas do DELAF_PB, de modo automático, as palavras que estavam definidas por esse dicionário como possuidoras de grau.

Diante da nossa escolha em fazer uma distinção entre o sufixo *-ão* em contexto de uso aumentativo e esse mesmo sufixo em outros contextos, como os deverbais agentivos e de ação, foram desconsideradas as palavras terminadas em *–ção*, embora saibamos que esse sufixo é claramente um formador de um deverbal que resulta de um verbo de ação, como em *chateação* e *aporrinhação*. Foram retirados ainda os gentílicos, como *afegão* e os nomes próprios, como *Abraão*.

Outros verbetes foram descartados da análise, quando se constatava que suas terminações não se configuravam de fato em sufixo. Para exemplificar tal procedimento, pode-se citar *bastão***,** cujo sentido não está atrelado ao sentido do lexema (seja **basto* ou **bastar*), ou ainda *boião*, cujo sentido de *vaso* ou *fogão* não está de maneira alguma ligado ao verbo *boiar*. Outros exemplos de vocábulos retirados são: *cafetão, cordão, corrimão, furacão, garanhão, ocasião*. Ao final, dos 130 vocábulos obtidos, observamos que 37 são adjetivos e 93

[3] Adotamos a noção de falso aumentativo presente em Rodrigues e Vale (2013) para nos referirmos aos casos de sufixação em –ão que não recebem interpretação de aumentativo.

são substantivos.

Das 37 bases adjetivais somente 31 foram analisadas, pois as outras 6 possuíam classificação duvidosa. Dos 31 vocábulos classificados como adjetivos, após consulta no Aulete Digital das entradas lexicais encontradas, observamos que 14 possuíam tanto nomes como verbos com a mesma significação, sugerindo que podem ter sido derivados tanto de uma base nominal como de uma base verbal. Os outros 17 vocábulos possuem registros no Aulete Digital apenas para verbos com a mesma significação, sugerindo que podem ter sido derivados exclusivamente de uma base verbal[4].

Em termos percentuais podemos dizer que 43,24% podem ter ambas as bases como ponto de derivação. Os outros 56,76% possuem somente a base verbal como ponto derivacional. Confira a tabela 1 abaixo:

Tabela 1: Classificação tipológica dos deverbais adjetivais em –ão.

ADJETIVOS					
Base Nominal	Base Verbal	Agentivo	Resultado de ação	Agentivo ou Resultado de ação	Instrumento
Briga	brigar	brigão	-	-	-
...	...	...	...	...	...
Subtotal		14	-	-	-
Base Verbal		Agentivo	Resultado de ação	Agentivo ou Resultado de ação	Instrumento
falastrar		falastrão	-	-	-
...		...	...	...	...
Subotal		17	-	-	-
Total:		31	-	-	-

Já em relação aos substantivos, 28 podem ter sido derivados tanto de uma base nominal como de uma base verbal e 65 podem ter apenas uma base verbal. Confira alguns exemplos na tabela 2.

[4] Foi utilizado como critério para esta afirmação a inexistência de uma entrada lexical nominal no Aulete Digital.

Tabela 2: Classificação tipológica dos deverbais nominais em –ão.

NOMES					
Base Nominal	Base Verbal	Agentivo	Resultado de ação	Agentivo ou Resultado de ação	Instrumento
brida	bridar	bridão	-	-	-
acordo	acordar	-	acordão	-	-
abano	abanar	-	-	abanão	-
poda	podar	-	-	-	podão
...	...	...	...	...	...
Subtotal		9	16	2	2
Base Verbal		Agentivo	Resultado de ação	Agentivo ou Resultado de ação	Instrumento
lambuzar		lambuzão	-	-	-
arranhar		-	arranhão	-	-
chupar		-	-	chupão	-
...		...	...	...	...
Subtotal		28	34	2	-
Total		37	50	4	2

Em termos percentuais podemos dizer que 30,11% podem ter ambas as bases como ponto de derivação e que somente a base verbal pode ser a base derivacional em 69,9% dos casos.

Outro aspecto considerado relevante em nossa análise do DELAF_PB é a noção presente dos novos vocábulos resultantes da derivação. Observamos que, dos 31 vocábulos classificados como adjetivos, todos possuem somente a noção de agentividade.

Em relação aos substantivos, observamos que, quando consultados no Aulete Digital, dos 93 vocábulos nominais, 37 são considerados por esse dicionário puramente como agentivos, 50 podem ser interpretados como resultado de ação, 4 possuem ambas as noções e somente dois

trazem a noção de instrumento.

Em termos percentuais a noção de agentividade presente nos substantivos em *–ão* corresponde a 39,79 % dos vocábulos. Enquanto as noções de ação, ambas as noções e de instrumento correspondem a 53,77 %, 4,30 % e 2,16 %, respectivamente. Note que, se os novos vocábulos formados são nominais, a produtividade desse sufixo é maior para a noção de resultado de uma ação.

Esses números sugerem que a derivação em *–ão*, na acepção aqui proposta, é um processo deverbal, pois nem sempre há uma base nominal correspondente ao significado presente no novo vocábulo. Supomos, então, que a derivação pode ocorrer de uma base nominal desde que esta base já seja uma forma deverbal. Sem uma base verbal que origine um nome base para uma posterior sufixação em *–ão*, não há como se obter um vocábulo em *–ão*, que extrapole a noção de grau aumentativo e represente as noções agentiva e ação.

Como a construção de um léxico dever ser linguisticamente bem fundamentada, nesta seção embasamos o presente trabalho no quadro da Morfologia Derivacional para descrever o fenômeno da sufixação em *–ão* na formação de agentivos e de nomes de ação. A próxima seção discuti como representar uma versão do léxico para a derivação dos agentivos e dos nomes de ação, incluindo como usar TEF para modelar combinações mórficas de derivação.

5. Elaboração do léxico

Esta seção fala sobre a construção de um recurso léxico de elementos nominais agentivos e de ação para o processamento computacional do Português do Brasil. A fim de criar o recurso léxico para os deverbais agentivos e deverbais de ação aqui propostos, o primeiro passo foi, é claro, reutilizar recursos léxicos disponíveis para o Português. Reutilizar esses recursos é uma maneira prática para iniciar o desenvolvimento de um novo. Portanto, partindo da hipótese da derivação verbal, tomamos os radicais das formas verbais infinitivas do DELAF_PB e construímos a arquitetura do nosso transdutor de estados finitos.

Uma maneira de gerar, de modo automático, recursos léxicos para o PLN se dá pela compilação de transdutores de estados finitos. Transdutores de estados finitos permitem modelar os mais diferentes fenômenos, desde o funcionamento de uma máquina até processos de formação de palavras. Entende-se por *estado* um modo transitório atribuído ao objeto que está sendo modelado. A depender do objeto, estímulos permitem a *transição* de um estado para outro. No caso de uma máquina, por exemplo, os estados ON - OFF são claramente definidos

e motivados. No caso das línguas naturais o importante são as transições. Um conjunto de estados interligados por transições é chamado de *rede*. O termo *finito* é utilizado para se referir ao "número de estados da rede que, para o presente propósito, pode ser satisfatoriamente entendido como não-infinito (KARTTUNEN, 2003, p.2-5)".

Para se obter o transdutor, compilamos no Foma a codificação proposta na nossa arquitetura derivacional. O Foma é um compilador, uma linguagem e uma biblioteca. Trata-se de um ambiente para o processamento de línguas naturais, baseado na TEF, capaz de construir e operar transdutores de estados finitos (HULDEN, 2009). Esse compilador permite a descrição de palavras formadas por processos derivacionais e flexionais. A análise das derivações e das flexões é baseada na descrição de regras morfológicas, representadas na arquitetura do transdutor de estados finitos.

Todo transdutor tem duas faces, uma face de superfície, onde estão as formas de superfície (*lower words*) e uma face subjacente, onde estão as formas subjacentes (*upper words*). Diz-se que um transdutor analisa quando o *input* corresponde a uma forma subjacente. E inversamente, quando o *input* é uma forma de superfície, diz-se que o transdutor gera.

Para codificar a arquitetura de um transdutor e mapear os pares *forma subjacente* e *forma de superfície* é necessário utilizar o formalismo *lexc*. Nesse formalismo, por um lado, o código é formado por blocos LEXICON. Vejamos um exemplo:

```
LEXICON                 Radicais
...                     ...
tripudi                 PoS;
tropic                  PoS;
tropeç                  PoS;
...                     ...
LEXICON                 PoS
...                     ...
ão:^ão+Nag+MASC+SG      #;
...                     ...
```

Note que no nosso transdutor estão codificados o LEXICON dos radicais verbais e o LEXICON das características morfossintáticas em questão, nomeados como *Radicais* e *PoS*, respectivamente.

Por outro lado, o desenho da arquitetura depende das decisões linguísticas adotadas. Por exemplo, codificamos as etiquetas morfossintáticas e as classes gramaticais, tais quais: V(erbo), Nag (nomes agentivos), Nac (nomes de ação), A(adjetivos agentivos), MASC(ulino),

SG(singular), FEM(inino) e PL(ural). Note que as etiquetas Nag e Nac expressam a distinção entre os deverbais agentivos e os de ação. No desenho da arquitetura também estão estabelecidos o bloco dos radicais verbais e o bloco das informações derivacionais e flexionais.

Na arquitetura do nosso transdutor está codificado que somente os nomes e os adjetivos agentivos recebem traços de gênero e número. Por outro lado, os nomes que designam resultado de ação recebem apenas os traços de número, pois diferentemente dos agentivos, estes nomes não alternam em gênero, como demonstram os exemplos abaixo.

(7) **agentivos**
o babão – a babona
b. resultado de ação
o remendo - *a remenda

Como já foi dito, o nosso transdutor foi construído com base nas formas verbais infinitivas do DELAF_PB, mas como a arquitetura do transdutor foi construída com base apenas em radicais verbais, para evitar que o transdutor hipergerasse, eliminamos as marcas de infinitivo e modelamos nossa arquitetura com base apenas nos seus radicais.

Para tanto, foram excluídos, de modo automático, os casos de verbos monossilábicos, pois ao terem suas marcas de infinitivo eliminadas, perderiam também a informação semântica presente no radical, como em (8):

(8) ter – t-
dar – d-

Ao serem retiradas as marcas de infinito desses verbos, o que se obtém é *t-* e *d-*, que certamente não são morfemas do português.

Regras de alternância ortográfica que ocorrem em determinados verbos também não foram modeladas nesse trabalho, ou seja, não foram modeladas as alternâncias morfofonológicas que ocorrem nas vogais dos radicais verbais, como em (9):

(9) pedir – pidão

Informados os radicais e modeladas as regras de derivação e flexão morfológicas, o próximo passo foi compilar no Foma a codificação referente ao processo de formação dos deverbais para a geração do transdutor.

Daí então, o arquivo é lido e compilado no Foma do seguinte modo:

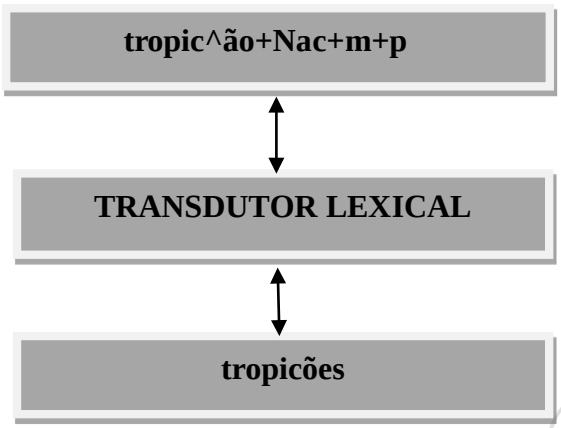

Figura 1: Representação da relação de mão dupla existente entre a forma subjacente *tropic^ão+Nac+m+p* e a forma de superfície *tropicões*.

Note que, no transdutor, a forma de superfície corresponde a *tropicões* e na forma subjacente encontram-se o lema e os traços e classes gramaticais[5].

Os conjuntos de palavras morfologicamente relatados pela derivação podem ser representados usando um grafo, tal como na Figura 1, onde o *input* representa o radical da forma derivada e o *output* seu código flexional. Assim, o grafo gera as seguintes entradas para o recurso léxico:

tropicões: tropic^ão+Nac+MASC+PL
tropicões: tropic^ão+Aag+MASC+PL
tropicões: tropic^ão+Nag+MASC+PL
tropiconas: tropic^ão+Aag+FEM+PL
tropiconas: tropic^ão+Nag+FEM+PL
tropicona: tropic^ão+Aag+FEM+SG
tropicona: tropic^ão+Nag+FEM+SG
tropicão: tropic^ão+Nac+MASC +SG
tropicão: tropic^ão+Aag+MASC +SG
tropicão: tropic^ão+Nag+MASC +SG

[5] Chamamos atenção para o símbolo "^" que representa o limite do morfema.

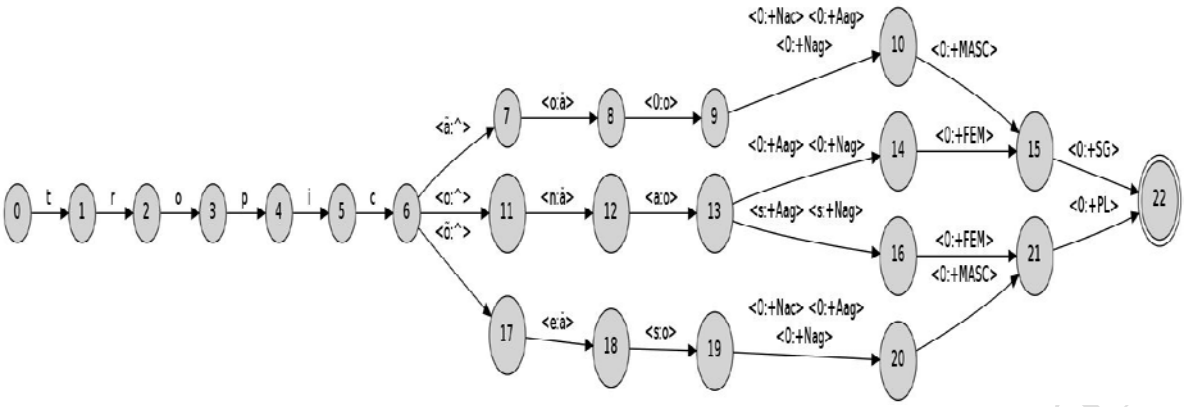

Figura 2: transdutor modelando uma alternância morfológica na derivação de agentivos: inserção de *–ão* após o radical verbal.

Cada caminho percorrido pelo transdutor no momento de uma análise ou de uma geração corresponde a um par *forma subjacente-forma de superfície*. Desse modo, nosso transdutor possui 14.241 radicais, que, ao serem compilados no Foma, causaram uma explosão de novas formas.

Com a implementação, foram obtidos 141.170 caminhos, cada um desses correspondendo a uma nova entrada do recurso léxico. Esse considerável enriquecimento do recurso léxico se deve a duas razões, quais sejam, a quantidade de radicais verbais obtidos do DELAF_PB e ao fato de o sufixo *–ão* ser realmente fértil em suas acepções. Ademais, o transdutor possui 6.948 estados, 22.232 transições e tem um tamanho ínfimo de 348,1 KB.

6. Considerações finais

Nesse artigo um recurso léxico de elementos nominais agentivos e de ação para o processamento computacional do Português do Brasil foi apresentado. O léxico foi implementado em Foma e, com os seus 141.170 vocábulos, contribui consideravelmente com os recursos já existentes. Até onde sabemos, esse recurso léxico é o primeiro a trazer um número significativamente grande de expressões nominais agentivas e de ação geradas pela sufixação em *-ão*. A existência de um léxico com deverbais agentivos e deverbais de ação pode ser de grande valor, já que, como se observou no DELAS_PB, as expressões nominais representam um grande grupo de palavras do PB.

O trabalho trata da importância dos recursos léxicos para o PLN, descreve o processo de formação dos nominais agentivos e de ação e argumenta a favor da análise deverbal da

sufixação em *–ão*, com base na análise dos aspectos morfológicos envolvidos no processo de elaboração do léxico.

Com o objetivo de fornecer subsídios linguísticos fundamentados para as aplicações do PLN utilizando a TEF e de contribuir com a caracterização dos falsos aumentativos, o léxico apresentado é mais um recurso para o processamento do PB, que contribui, com a sua cobertura, para a expansão dos léxicos computacionais já existentes.

No entanto, o processo de elaboração de um recurso léxico não consistiu apenas na análise e geração de palavras, mas incluiu, também, uma fundamentação linguística baseada na morfologia derivacional, com o objetivo de caracterizar o comportamento das palavras no contexto da sufixação em *–ão*.

O processo de elaboração do léxico se justifica, principalmente, pela preocupação em incorporar nos recursos léxicos já existentes palavras que, embora em um contexto morfológico apareçam como derivações sufixais em *–ão*, não correspondem a vocábulos que expressam grau de aumentativo.

É necessário considerar que, para além da sufixação em *-ão*, na elaboração da arquitetura do nosso transdutor, nós também tratamos o correspondente plural *-ões*, assim como as formas femininas *–ona* e *–onas*. Para trabalhos futuros, as palavras geradas e analisadas pelo nosso transdutor podem, ainda, contribuir para a investigação, na língua, das restrições semântico-pragmáticas e morfofonológicas dos verbos envolvidos na derivação em questão.

Desse modo, pode-se afirmar que os resultados obtidos foram satisfatórios, pois a discussão gerada conseguiu aprofundar o problema da definição da origem dos agentivos e dos nomes de ação no português do Brasil e, consequentemente, colaborar para a expansão dos recursos léxicos atualmente existentes para o Português.

Referências Bibliográficas

ALENCAR, L. F. *et. al*. JMorpher: a Finite-State Morphological Parser in Java for Android. In: BAPTISTA, J.; MAMEDE, N. (Eds.). **Proceedings of the 11th International Conference on Computational Processing of Portuguese**, **PROPOR 2014**. São Carlos: USP-São Carlos, 2014. p. 59-69.

ARMELIN, P. R. G. Sobre a interação entre as marcas de diminutivo e aumentativo no Português Brasileiro. **ReVEL,** edição especial, n.5, 2011.

BASÍLIO. M. **Formação e classes de palavras no português do Brasil.** São Paulo: Contexto, 2006.

BECHARA, E. **Moderna gramática portuguesa**. 37. ed. Rio de Janeiro: Nova Fronteira, 2009.

CÂMARA JÚNIOR, J. M. **Estrutura da Língua Portuguesa**. 42. ed. Petrópolis: vozes, 2009.

CORMEN, T. H. et al. ***Introduction to Algorithms.*** 2 ed. Cambridge: MIT Press, 2001.

CUNHA, C.; CINTRA, L. **Gramática do Português Contemporâneo**. 6. ed. Rio de Janeiro: Lexicon, 2013.

DICIONÁRIO AULETE. Dicionário online Caldas Aulete. **Aulete Digital**. Disponível em http://www.aulete.com.br/. Acesso em 09 outubro 2014.

DURAN, M. S. A importância dos recursos lexicais para o processamento automático do português. **Estudos Linguísticos**, v. 42, n.2, 2013.

FREITAS, C. Sobre a construção de um léxico da afetividade para o processamento computacional do português. **RIBLA**, Belo Horizonte, v. 13, n. 4, p. 1031-1059, 2013.

HULDEN, M. Foma: a finite-state compiler and library. **Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (Demos)**. Stroudsburg, 2009. p. 29-32.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing:** An introduction to natural language processing, computational linguistics, and speech recognition. 2. ed. New Jersey: Pearson Education, 2009.

KARTUNEN, L.; BEESLEY, K. **Finite State Morphology**. Stanford: CSLI Publications, 2003.

MAHLOW, C. Prefácio. In: MAHLOW, C.; MICHAEL, P. (Orgs). Systems and Frameworks for Computational Morphology. **Proceedings of the Second International Workshop on Systems and Frameworks for Computational Morphology (SFCM).** Zurich, 2011.

MOLINERO, M. A.; SAGOT, B.; NICOLAS, L. A morphological and syntactic wide-coverage lexicon for Spanish: The Leffe. **Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)**, Borovets, 2009. p. 264-269.

MONTEIRO. J. L. **Morfologia Portuguesa.** 4. ed. Campinas: Pontes, 2002.

MOTA, C. Analysis of Derivational Morphology by Finite State Transducers. In: DISTER, A. (Ed.). **Actes des Troisièmes Journées INTEX, *RISSH: Revue, Informatique et Statistique dans les Sciences Humaines,*** Liège: Université de Liège, 2000. p. 273-287.

MUNIZ, M.C. M. Projeto UNITEX-PB. **Núcleo Interinstitucional de Lingüística Computacional (NILC) – USP-São Carlos**, São Paulo, 2004. Disponível em http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/dicionarios.html. Acesso em 09 outubro 2014.

PEZZATTI, E. G. A gramática da derivação sufixal: três casos exemplares. **Alfa**. São Paulo, n. 33, p. 103-114, 1989.

ROCHA, L. C. de A. **Estruturas morfológicas do Português**. 2. Ed. São Paulo: WMF Martins Fontes, 2008.

RODRIGUES, R.; VALE, O. A. Análise dos falsos aumentativos no Português Brasileiro. **Anais da III Jornada de Descrição do Português**, Fortaleza: Unifor, 2013. p. 9-14.

RODRIGUES, R.; VALE, O. A. Calça, calcinha, calção: falsos diminutivos e falsos aumentativos no Português do Brasil. **Anais do II Colóquio Brasileiro de Morfologia.** Rio de Janeiro: UFRJ, 2013. p. 209-218.

RUSSEL, S. J.; NORVIG, P. **Artificial Intelligence**: a modern approach. 2. ed. New Jersey: Pearson Education, 2003.

SAGOT, B. DeLex, a freely-avaible, large-scale and linguistically grounded morphological lexicon for German. **Proceedings of the XIX International Conference on Language Resources and Evaluation (LREC),** 26-31 May, Reykjavik, Iceland, 2014. p. 2778-2784.

SANTOS, A. P. Estudo do sufixo –ão: Valores semânticos e proposta genealógica *In*: MARÇALO, M. J. *et al*. (Orgs.). **A língua portuguesa*:*** ultrapassar fronteiras, juntar culturas. Évora: Universidade de Évora, p. 1-21, 2010.

SANTOS, A. P. Para além do significado de aumentativo do sufixo –ão. **Cadernos do CNLF**, Rio de Janeiro, v. XIII, n. 4, p. 2494-2510, 2009.

SILVA, A. L. R. **Morfologia derivacional da língua portuguesa:** o sufixo –*vel* na formação dos adjetivos. 2009. 141 f. Dissertação (Mestrado em Linguística) – Programa de Pós-Graduação em Linguística. Mestrado Interinstitucional UFC/UFMA, Fortaleza, 2009.

SMARSARO, A. O léxico e o processamento de linguagem natural. **Revista (Con) Textos Linguísticos**, Vitória, n. 1, p. 49-54, 2007.

TORRES, C. E. A. **Uso de informação linguística e análise de conceitos formais no aprendizado de ontologias**. 2012. 67 f. Dissertação (Mestrado em Ciência da Computação) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2012.

Artigo recebido em: 22.02.2015
Artigo aprovado em: 14.06.2015