Proceedings of Symposium in Information and Human Language Technology. Uberlândia, MG, Brazil, October 2–5, 2017.

Formação de gentílicos a partir de topônimos: descrição linguística e aprendizado automático

Roger A. de M. R. Antunes[1], Thiago A. S. Pardo[2], Gladis M. B. Almeida[1]

Núcleo Interinstitucional de Linguística Computacional (NILC)

[1] Programa de Pós-Graduação em Linguística, Universidade Federal de São Carlos

[2] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

rograntunes@gmail.com, taspardo@icmc.usp.br, gladis.mba@gmail.com

Resumo. *O presente artigo tem como objetivo descrever as regras envolvidas na transformação de topônimos em gentílicos, de modo a identificar regularidades. A partir dessas regularidades, desenvolve-se um algoritmo capaz de gerar gentílicos de forma automática. Como base teórica, são considerados conceitos da Morfologia Derivacional e, do ponto de vista metodológico, toma-se como fonte topônimos e gentílicos do Instituto Brasileiro de Geografia e Estatística (IBGE), bem como se criam procedimentos para tornarem os dados manipuláveis. Realiza-se também um processo complementar de aprendizado automático. Como resultado, obtém-se boa acurácia na predição de gentílicos, revelando regras e atributos novos e relevantes para a tarefa.*

Abstract. *This paper aims to describe the rules required in the transformation of toponyms into demonyms in order to identify regularities. From these regularities, we developed an algorithm that automatically generates demonyms for toponyms of interest. As a theoretical basis, the concepts of Derivational Morphology are considered, and, concerning the methodology, we used data about cities and demonyms provided by the Brazilian Institute of Geography and Statistics (IBGE) website, for which we produced procedures to make this data tractable. A complementary process of automatic learning was also carried out. As a result, a good accuracy was obtained in the prediction of demonyms, revealing new and relevant rules and features for the task.*

1. Introdução

Este trabalho surgiu no contexto de elaboração do Vocabulário Ortográfico Comum da Língua Portuguesa[1] - VOC (Almeida et al, 2013), especificamente durante a inserção, nessa base lexical, dos gentílicos de todos os países[2] que têm como língua oficial o

[1] VOC é uma grande base lexical, organizada numa plataforma *web*, que hospeda os Vocabulários Nacionais dos países de língua oficial portuguesa, integrantes da Comunidade dos Países de Língua Portuguesa (CPLP). É um instrumento previsto no Acordo Ortográfico de 1990. Disponível em: <voc.cplp.org>. Acesso em 13/mai/2017.

[2] São os seguintes os países que integram a CPLP: Angola, Brasil, Cabo Verde, Guiné-Bissau, Moçambique, Portugal, São Tomé e Príncipe, e Timor-Leste.

português. Pelo fato de constituir-se numa tarefa humana bastante morosa, observou-se a necessidade de estudar as regularidades morfológicas na formação dos gentílicos, de maneira que fosse possível gerar, de forma automática, os gentílicos a partir dos topônimos inseridos na base do VOC.

Gentílico é a palavra que designa o local (país, região, estado, município, povoação) onde alguém nasceu. A formação do gentílico ocorre a partir do topônimo (nome próprio dos lugares). Para ilustrar, a partir do topônimo *Ibaté*, temos o gentílico *ibateense*. Constitui nosso objeto de descrição e análise o conjunto dos gentílicos referentes aos 5.570 municípios brasileiros. A lista contendo todos os municípios bem como sua forma gentílica correspondente foi obtida a partir do *site*[3] oficial do IBGE em 2014. Esse conjunto foi analisado exaustivamente, com o intuito de observar as regras de combinação morfológicas. Assim, pudemos chegar à proposta de algoritmos capazes de gerar automaticamente gentílicos a partir de topônimos. Amparados em conceitos da Morfologia Derivacional, levantamos a produtividade dos morfemas formadores de gentílicos e encontramos nove sufixos acoplados aos topônimos que são responsáveis pela geração dos gentílicos de todos os 5.570 municípios: *-ense, -ano, -ino, -ista, -eiro, -eno, -enho, -ito* e *-ado*.

Neste artigo, em especial, nosso interesse concentra-se nos gentílicos formados a partir de topônimos unigramas, cerca de 52% da lista do IBGE, que são aqueles constituídos apenas por uma palavra, tais como Palmas, Uberlândia e Valinhos. Sendo assim, casos como São Paulo (bigrama), Rio de Janeiro (trigrama) e Barão de Monte Alto (tetragrama) não são objeto deste artigo. Em uma primeira linha de trabalho, mostramos que é possível descrever e sistematizar os processos morfológicos de produção da maioria dos gentílicos. De forma complementar à análise humana, utilizando aprendizado automático, demonstramos que é possível prever o sufixo de formação do gentílico com boa acurácia. O trabalho relatado, com sua abrangência e resultados, é inédito para o português.

Na Seção 2, introduzimos os conceitos básicos desta pesquisa e os principais trabalhos relacionados. Na Seção 3, o conjunto de dados trabalhado é descrito. Nas Seções 4 e 5, respectivamente, apresentamos as análises linguísticas e automáticas realizadas. Algumas considerações finais são feitas na Seção 6.

2. Conceitos básicos e trabalhos relacionados

O principal processo na formação de gentílicos é a derivação sufixal. A derivação, processo muito produtivo de formação de palavras no português, é a adição de um afixo a uma base ou radical (Alves, 1990; Correia e Almeida, 2012). A derivação é sufixal quando o afixo vem à direita da base (marialv<u>ense</u>) e é prefixal quando o afixo vem à esquerda (<u>des</u>abastecer) (Sandmann, 1992). Um aspecto relevante a considerar na derivação sufixal é o fato de o sufixo determinar a categoria gramatical da palavra resultante (Kedhi, 1992; Basílio, 2004). Podemos exemplificar por meio da base *feliz*, na qual adicionamos o sufixo *-mente*, que vai transformar o adjetivo *feliz* no advérbio *felizmente*.

[3] <www.cidades.ibge.gov.br>.

Entre os autores que já trataram dos gentílicos, podemos citar Melo e Gomes (2000), que tiveram como objeto de seu estudo a análise morfológica da formação de gentílicos a partir de topônimos referentes aos 26 estados brasileiros e suas respectivas capitais. Esse trabalho também dá relevo para questões referentes à toponímia e ao percurso histórico (pautado em dicionários) responsável pelas nomeações.

Dignos de nota também são os trabalhos de Areán-García (2009, 2012), situados na morfologia histórica, que descrevem a genealogia semântica dos gentílicos e dos agentivos formados por *-ist*a ao longo de todo o período de criação do Estado Brasileiro em comparação com a língua portuguesa europeia e outras línguas também europeias.

Ressalte-se que, nesses trabalhos citados, não havia a abrangência e a preocupação de automatizar o processo de formação de gentílico a partir de um topônimo, o que dá o caráter de originalidade à pesquisa relatada neste artigo.

3. Conjunto de dados

O *site* do IBGE foi escolhido como fonte de dados por se tratar de um instituto oficial e conter diversos tipos de informações referentes a todos os municípios brasileiros, incluindo aí os gentílicos. Ressalte-se que foi utilizada nesta pesquisa uma versão anterior do *site* denominada "Cidades"[4].

Embora o *site* seja extremamente rico e traga um conjunto muito diverso e detalhado de informações, ele está num formato que impede o processamento computacional, razão pela qual foram desenvolvidos *scripts* que possibilitaram a transformação daqueles dados em tabelas, o que nos permitiu total manipulação. Assim, foi construída uma lista em formato de tabela, de topônimos associados aos seus respectivos gentílicos, contendo 5.570 linhas, correspondendo ao total de municípios brasileiros. Dos 5.570 municípios, foram considerados neste trabalho apenas as formas unigramas, como mencionamos na Seção 1.

4. Descrição linguística

Esta seção está dividida em duas partes: a primeira descreve o método de trabalho e os recursos utilizados, assim como os primeiros apontamentos; e a segunda apresenta os resultados da análise e os processos resultantes.

4.1. Método de trabalho e recursos utilizados

Para realizar a descrição, iniciamos pelos topônimos que recebem os sufixos *-ano, -ino, -ista, -eiro, -eno, -enho, -ito* e *-ado*, considerados os menos convencionais, já que, na maioria dos casos, o que se têm são as formações em *-ense*, considerado o sufixo prototípico para a formação de gentílicos. Sendo assim, na primeira etapa, que foi responsável pela identificação das extremidades dos topônimos e sua associação aos sufixos gentílicos, o sufixo *-ense* não foi contemplado, por ser considerado o padrão, com 91,5% de ocorrências.

Para que fosse possível diferenciar os topônimos que recebem um ou outro sufixo formador de gentílico, foram delimitadas suas terminações (extremidades),

[4] <http://cidades.ibge.gov.br/xtras/home.php>.

utilizando-se apenas critérios grafemáticos de sequências finais de letras. Na prática, foram separadas as últimas três, quatro ou cinco letras dos topônimos, dependendo do caso, separação esta que visou à distinção das unidades finais. Por exemplo, para o sufixo *-ano* (que se apresentou em 4,83% das ocorrências), agrupamos todas as extremidades de topônimos cujos gentílicos se formavam com este sufixo, sendo elas: <aba> (Piracicaba (SP) - *piracicab**ano***), <aça> (Mombaça (CE) - *mombaç**ano***), <aia> (Atibaia (SP) - *atibai**ano***), etc.

Com esse processo, percebeu-se que, embora as extremidades dos topônimos fossem responsáveis pela escolha de um ou outro sufixo, muitas delas permitiam a geração de mais de uma forma de gentílico, cada forma realizada com sufixos distintos, inclusive com o sufixo *-ense*. Por exemplo, a partir do topônimo Colômbia (SP), podem-se formar os gentílicos *colombiano* e *colombiense*.

Além de observar os grafemas que compunham as terminações dos topônimos, foi preciso também levar em conta algumas adequações morfofonológicas[5] responsáveis pela elisão ou crase da vogal temática final, quando as bases recebem o sufixo de gentílico, como nos casos de: Penedo (AL) - *penedense* (elisão: <ed**o**> + **-e**nse), Morungaba (SP) - *morumgab**ano*** (crase: <ab**a**> + **-a**no). Esses padrões foram chamados de regras de ligação grafemática e são identificadas em todas as construções de gentílicos.

4.2. Descrição e análise linguística: dados e processos resultantes

Por meio do método apresentado na subseção anterior, pudemos observar, que:

1. a derivação está presente em praticamente todos os gentílicos, excetuando-se apenas casos arbitrários como Salvador - *soteropolitano*, não tratados aqui;
2. a escolha do sufixo varia conforme a extremidade do topônimo;
3. *-ense* é o sufixo padrão, e os outros oito sufixos ocorrem em minoria, ainda que aplicável em diversos topônimos, o que amplia as possibilidades combinatórias para a geração dos gentílicos;
4. algumas extremidades de topônimos podem receber mais de um sufixo para formar gentílicos.

Procedemos, então, à estruturação, na forma de um banco de dados morfológicos, de todos esses elementos descritos, tais como: as extremidades dos topônimos associadas aos sufixos gentílicos com que ocorrem, a lista de extremidades que não se formam com *-ense* e a lista de ligação grafemática, responsável pela concatenação dos morfemas. Todos esses dados organizados permitiu-nos gerar os gentílicos a partir das unidades toponímicas unigramas, seguindo o esquema algorítmico explicitado no Quadro 1.

O algoritmo funciona da forma descrita a seguir. Inicialmente, apresenta-se o topônimo para o qual se deseja obter os gentílicos possíveis, juntamente com o banco de dados morfológicos, no qual se baseia todo o processo. No passo 1, é consultado o

[5] "É objeto da morfonologia, forma haplológica de morfofonologia, o estudo das mudanças que se operam no corpo fônico dos elementos, bases ou radicais e afixos ou flexões, que se unem para formar vocábulos ou unidades lexicais novas, compostos ou derivados, ou variantes flexionais de um mesmo vocábulo ou unidade lexical." (Sandmann, 1997, p. 50)

banco de extremidades toponímicas associadas aos oito sufixos gentílicos (excetuando-se *-ense*), buscando-se pelos possíveis sufixos associados às extremidades existentes (de 3, 4 ou 5 letras) do topônimo de interesse. Caso essa consulta retorne resultados (conforme checagem no passo 2), segue-se para o passo 3. No passo 3, consulta-se se alguma das extremidades toponímicas pode se associar ao sufixo *-ense*. Este passo ajuda a impedir a sobregeração dos gentílicos. Se o passo 3 sinalizar negativamente, deve-se gerar o gentílico somente com o sufixo encontrado na consulta realizada no passo 1, o que é feito no passo 4. Quando a etapa 3 sinaliza positivamente, pula-se para o passo 5, que prevê a geração de variações do gentílico com o sufixo identificado no passo 1 e também com o sufixo *-ense*. No caso da consulta do passo 2 não retornar resultado, pula-se diretamente para o passo 6, que prevê somente a utilização do sufixo *-ense* para a formação do gentílico (o que representa o caso padrão, ou *default*). No passo 7, depois que já está(ão) selecionado(s) o(s) sufixo(s) que formará(rão) o(s) gentílico(s), há necessidade de consulta ao banco das regras de adequação grafemática, pois, ao concatenar os elementos morfológicos envolvidos na derivação, às vezes, é preciso adicionar ou suprimir letras. No passo 8, as regras identificadas (se houver alguma) são aplicadas, produzindo-se os gentílicos. No passo 9, removem-se eventuais diacríticos residuais dos gentílicos produzidos. Os gentílicos resultantes são retornados/exibidos pelo passo 10.

Quadro 1. Algoritmo de produção de gentílicos

Dados de entrada: topônimo de interesse, banco de dados morfológicos
Dados de saída: gentílicos possíveis para o topônimo de interesse
Início do algoritmo
1. Buscar no banco de dados as extremidades possíveis do topônimo (com 3, 4 ou 5 letras) e, para cada extremidade encontrada (se houver alguma), o sufixo associado a ela
2. Se a busca acima encontrar extremidades e sufixos no banco de dados, então
    3. Checar, no banco de dados, se cada extremidade encontrada pode se associar ao sufixo *-ense*
        4. Se a extremidade não se associar a *-ense*, então utilizar o sufixo encontrado no passo 1 para produzir o gentílico
        5. Senão, se a extremidade puder se associar a *-ense*, então utilizar o sufixo encontrado no passo 1 e também o sufixo *-ense* para construir os possíveis gentílicos
6. Senão, se a busca não encontrar extremidade alguma no banco de dados, então utilizar somente o sufixo *-ense* para construir o gentílico
7. Buscar no banco de dados por regras de adequações grafemáticas que devem ser aplicadas na construção de cada possível gentílico
8. Aplicar as regras de adequações grafemáticas identificadas (caso haja alguma) para a adequação dos gentílicos
9. Se houver diacríticos nos gentílicos produzidos, remover os diacríticos
10. Retornar os gentílicos produzidos
Fim do algoritmo

Como se nota, a lógica do trabalho foi de descrever as exceções para chegar à regularidade, portanto, tudo que não fez parte das particularidades dos oito sufixos é caracterizado como padrão, com sua construção realizada pelo sufixo *-ense*.

Para exemplificar as etapas do algoritmo, assume-se que o topônimo *Poconé (MT)* é fornecido como entrada. No passo 1, busca-se e encontra-se sua extremidade associada ao sufixo *-ano*. No passo 2, confirmando que a busca no passo 1 teve sucesso,

ativa-se o passo 3, que verifica que a extremidade <oné> não se liga a *-ense*, passando, portanto, para o passo 4, que é a etapa responsável pela associação da base *Poconé* a *-ano*. Passa-se, então, ao passo 7, em que se percebe que não há necessidade de aplicação de regras específicas para supressão ou adição de elementos. Então, no passo 8, simplesmente concatenam-se os elementos (*Poconé* + (vazio) + *ano*). No passo 9, remove-se o diacrítico referente ao acento (´). No passo 10, por fim, é retornado o resultado, que é o gentílico *poconeano*.

O algoritmo foi validado manualmente. Todos os topônimos utilizados tiveram seus gentílicos produzidos pelo algoritmo. Inicialmente, verificou-se na listagem do IBGE se o gentílico produzido era o esperado. Caso não fosse, buscou-se o gentílico produzido no dicionário eletrônico Houaiss (2009) e/ou na *web* (na Wikipedia e no Google), para atestar sua viabilidade. Com essa forma de validação, foi possível verificar que: (i) as tarefas realizadas pelo algoritmo são capazes de gerar gentílicos a partir de topônimos, tomando como base a lista extraída do IBGE; e (ii) nenhuma lista de gentílicos é tão completa, no sentido de possuir uma abrangência em relação à existência de uma forma de gentílico para cada cidade, quanto a do IBGE, pois nem todos os outros meios de validação registram os gentílicos gerados pelo algoritmo e/ou possuem gentílicos para os topônimos processados.

5. Aprendizado automático

O aprendizado automático foi realizado sobre a base de municípios brasileiros com dois principais objetivos: (i) verificar, de forma automática, a regularidade do processo morfológico para produção dos gentílicos e (ii) identificar e extrair eventuais padrões interessantes de produção de gentílicos.

No primeiro caso, se o aprendizado realizado pela máquina demonstra uma boa acurácia, pode-se corroborar a análise humana anterior de que, pelo menos parcialmente, há processos regulares, baseados nas características dos topônimos, que podem ser reutilizados. Casos arbitrários (por exemplo, o topônimo *Niterói* e seu gentílico *fluminense*) pouco contribuiriam para o aprendizado.

No segundo caso, a máquina, com seus métodos de aprendizado, pode identificar padrões que a análise e introspecção humana podem não reconhecer. A análise humana normalmente é um processo caro e lento, sujeito às falhas e inconsistências humanas. A análise automática, por outro lado, pode ser executada em larga escala, de maneira relativamente eficiente, em grandes bases de dados, podendo identificar conhecimento relevante adicional não detectado pelo humano, mesmo que seu método de busca de padrões seja significativamente mais limitado do que o nível que a cognição humana atinge. É essa perspectiva que se destaca, em que os padrões identificados pela máquina podem revelar processos interessantes que podem enriquecer a análise humana e indicar, inclusive, novas possibilidades de análise.

Neste trabalho, como um dos objetivos é explicitar novos conhecimentos, foram utilizados métodos simbólicos de aprendizado, que podem produzir árvores de decisão e regras de classificação. Relatamos, principalmente, o resultado produzido pelo método PART (Frank e Witten, 1998), que produz regras de classificação. Ele foi adotado porque foi capaz de produzir um conjunto mais compacto de regras (em relação aos demais métodos testados) e com boa acurácia.

Para modelar a tarefa em questão como um problema de aprendizado de máquina, cada topônimo foi considerado uma instância de aprendizado e caracterizado por um conjunto de oito atributos, a saber:

1. atributo de tamanho relativo do topônimo, em que topônimos com 5 letras ou menos são considerados "pequenos", topônimos com 6 a 10 letras são considerados de tamanho "médio" e topônimos com mais de 10 letras são considerados "grandes" (definiu-se empiricamente o intervalo de número de letras para cada tamanho especificado);
2. atributo que armazena os três últimos caracteres do topônimo (concatenados), em linha com parte da descrição linguística relatada anteriormente neste artigo;
3. cinco atributos que consistem, isoladamente, nos cinco últimos caracteres dos topônimos (que nomeamos, de forma abreviada, de c a c-4, sendo o último caractere representado pelo atributo 'c', o penúltimo pelo atributo 'c-1', e assim por diante);
4. atributo que armazena o estado onde se localiza a cidade relativa ao topônimo.

A cada instância de aprendizado foi associada sua classe, que, neste caso, é o sufixo que deve ser utilizado para constituir o gentílico correspondente. Como ilustração, o Quadro 2 mostra como seria a representação do topônimo *Piracicaba* (cujo gentílico é *piracicabano*) como uma instância para aprendizado de máquina. Todos os topônimos unigramas foram representados dessa forma. No total, foram utilizados 245 topônimos para o aprendizado, nomeadamente aqueles que não se formam com o sufixo *-ense*.

Quadro 2. Exemplo de representação do topônimo *Piracicaba* como uma instância para aprendizado de máquina

Topônimo	Atributos								Classe
	tamanho	3 últimos caracteres	c-4	c-3	c-2	c-1	c	estado	
Piracicaba	médio	aba	i	c	a	b	a	SP	ano

O aprendizado de máquina foi realizado na ambiente WEKA (Witten et al., 2011). Em um primeiro momento, o conjunto completo de instâncias foi utilizado para aprendizado e também avaliação da acurácia, buscando-se confirmar a regularidade do processo de formação de gentílicos.

Utilizando-se somente o atributo de três últimos caracteres para o aprendizado (simulando-se parte da análise linguística já relatada anteriormente), a acurácia atingida pelo aprendizado foi de 86,5%, ou seja, foi possível, com base nas regras aprendidas, prever corretamente o sufixo (a classe) do gentílico correspondente em 86,5% dos casos (ou seja, 212 casos, no total). Nesse cenário, foram aprendidas 114 regras, sendo a mais produtiva (a primeira do conjunto) a seguinte regra: SE os 3 últimos caracteres do topônimo forem *lis*, ENTÃO a classe é *-ano*.

Utilizando-se mais atributos da representação (o tamanho relativo e os 5 últimos caracteres isolados, além do atributo anterior), produziu-se uma acurácia de 88,9% (uma melhoria de 2,7% em relação ao resultado anterior), com 107 regras aprendidas.

Incorporando-se o atributo de estado (que, em um primeiro momento, poderia parecer irrelevante), atingiram-se impressionantes 99,1% de acurácia (uma melhoria de 11,4% em relação ao resultado acima), com 100 regras aprendidas, sendo que a maior diferença para a versão anterior foi a previsão correta dos gentílicos terminados em *-ino*. Basicamente, o aprendizado identificou que a maior parte dos gentílicos terminados em *-ino* provém dos estados de Goiás e Tocantins (das regiões Centro-Oeste e Norte do Brasil, respectivamente). Algumas regras do conjunto aprendido são exemplificadas abaixo:

> SE os 3 últimos caracteres do topônimo forem *aba*, ENTÃO a classe é *-ano*
>
> ...
>
> SENÃO SE o estado for *Paraná* E o último caractere for *s*, ENTÃO a classe é *-ano*
>
> SENÃO SE o estado for *Minas Gerais* E os 3 últimos caracteres do topônimo forem *lis* E o tamanho for *grande*, ENTÃO a classe é *-ano*
>
> ...
>
> SENÃO SE os 3 últimos caracteres do topônimo forem *lis* E o estado for *Goiás*, ENTÃO a classe é *-ino*

Corroborando a descrição linguística realizada e a relevância do aprendizado automático, ao se realizar a seleção de atributos no ambiente WEKA, os atributos relativos aos três últimos caracteres do topônimo e ao estado mostram-se como os mais relevantes.

Em outra perspectiva, ao se realizar a avaliação sobre conjuntos de dados diferentes dos utilizados para o aprendizado (que consiste em uma boa prática em aprendizado de máquina), utilizando-se o esquema de validação cruzada de 10 pastas (com todos os atributos), atingem-se 69,7% de acurácia, sendo que a maioria dos erros ocorre para as classes *-ano* e *-ino*.

Por fim, vale citar mais um resultado de aprendizado automático. Avaliando-se conjuntos de regras aprendidos automaticamente, obteve-se um conjunto de regras enxuto e elegante com o método JRip (Cohen, 1995), com uma boa acurácia de 88,1% (utilizando-se todo o conjunto de dados para aprendizado e avaliação). O conjunto é exibido a seguir. Como se pode ver, os três últimos caracteres e o estado mantêm-se como atributos muito relevantes na classificação.

> SE os 3 últimos caracteres do topônimo forem *aré*, ENTÃO a classe é *-eno*
>
> SENÃO SE o estado for *Goiás*, ENTÃO a classe é *-ino*
>
> SENÃO SE o antepenúltimo caractere do topônimo for '*n*' e o último caractere for '*e*', ENTÃO a classe é *-ino*
>
> SENÃO SE os 3 últimos caracteres do topônimo forem *tes*, ENTÃO a classe é *-ino*
>
> SENÃO SE o estado for *Tocantins* e os 3 últimos caracteres do topônimo forem *lis*, ENTÃO a classe é *-ino*

SENÃO a classe é *-ano*

É relevante adicionar que todos os métodos de aprendizado de máquina investigados foram executados no WEKA com suas configurações padrões. Além disso, ressalta-se que muitos outros métodos foram avaliados, mas relatamos aqui somente os mais promissores.

Conclui-se, no geral, que o aprendizado de máquina é relevante para explicitar conhecimentos interessantes, corroborando resultados manuais e complementando a análise humana.

6. Considerações finais

Pelo que se sabe, o trabalho aqui apresentado é inédito em termos de abrangência e profundidade de anáĺise realizada para a língua portuguesa. De nosso particular interesse é a replicação desta pesquisa para outros países de língua oficial portuguesa, verificando-se em que medida os processos morfológicos identificados também ocorrem. Logicamente, nesses casos, o uso de informações geográficas específicas do Brasil (como a informação de estado) não é relevante.

Como próximo passo dessa pesquisa, vislumbra-se o desenvolvimento de uma interface web de fácil acesso e uso que, (i) além de manter um catálogo dos topônimos e seus gentílicos, (ii) permita a geração de possíveis gentílicos a partir de novos topônimos apresentados, utilizando-se o algoritmo produzido e as regras de aprendizado de máquina identificadas.

Agradecimentos

Agradecemos a valiosa colaboração de José Pedro Ferreira, pesquisador do Centro de Estudos de Linguística Geral e Aplicada (CELGA) da Universidade de Coimbra, Portugal, que desenvolveu os *scripts* que possibilitaram a captura e transformação dos dados provenientes do *site* do IBGE. Agradecemos também à CAPES e à FAPESP, pelo apoio a este trabalho.

Referências

Almeida, G. M. B.; Ferreira, J. P.; Correia, M.; Oliveira, G. M. (2013) "Vocabulário Ortográfico Comum (VOC): constituição de uma base lexical para a língua portuguesa". *Estudos Linguísticos* (São Paulo, 1978), v. 42, p. 204-215.

Alves, I. M. (1990) *Neologismo: criação lexical*. São Paulo: Ática.

Antunes, R. A. de M. R. (2017) *Formação de Gentílicos a partir de Topônimos: Proposta de geração automática.* 253 pg.. Dissertação (Mestrado) - Universidade Federal de São Carlos - UFSCar.

Areán-García, N. (2009) "A formação de nomes gentílicos com o sufixo -ista no português: algumas questões". *Estudos Linguísticos*, São Paulo, 38 (2): 31-41.

Areán-García, N. (2012) "A formação de nomes de profissionais a partir do sufixo -ista". In: Ana María Cestero Mancera, Isabel Molina Martos, Florentino Paredes García (eds) *La lengua, lugar de encuentro*. Actas del XVI Congreso Internacional

de la Alfal. Alcalá de Henares: Servicio de Publicaciones de la Universidad de Alcalá, p. 2475-2483.

Basílio, M. (2004) *Formação e classes de palavras no português do Brasil*. São Paulo: Contexto.

Cohen, W. W. (1995) "Fast Effective Rule Induction", In: *Proceedings of the Twelfth International Conference on Machine Learning,* p. 115-123.

Correia, M. e Almeida, G. (2012) *Neologia em português*. São Paulo: Parábola.

Instituto Antônio Houaiss. (2009) *Dicionário eletrônico Houaiss da língua portuguesa 1.0*. Rio de Janeiro: Objetiva.

Instituto Brasileiro de Geografia e Estatística (IBGE). Cidades@. Disponível em: http://www.cidades.ibge.gov.br/xtras/home.php.

Frank, E. and Witten, I. H. (1998) "Generating Accurate Rule Sets Without Global Optimization", In: *Proceedings of the Fifteenth International Conference on Machine Learning*, p. 144-151.

Kehdi, V. (1999) *Formação de palavras em português*. 3a edição. São Paulo: Ática.

Melo, C. R. e Gomes, J. J. (2000) "Adjetivos pátrios brasileiros". *Ao pé da Letra* (UFPE), v. 2, p. 35-40.

Sandmann, A. (1992) *Morfologia lexical*. São Paulo: Contexto.

Sandmann, A. (1997) *Morfologia Geral*. 3ª ed. São Paulo: Contexto.

Witten, I. H.; Frank, E.; Hall, M. A. (2011) *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers.