Uma Metodologia Semi-Automática para Detecção de Neologismos do Português Brasileiro[+]

Ieda Maria Alves[1], Bruno Oliveira Maroneze[1], Mariangela de Araujo[1], Thiago Alexandre Salgueiro Pardo[2], Sandra Maria Aluísio[2]

[1] Faculdade de Filosofia, Letras e Ciências Humanas (FFLCH)
Rua do Lago, 717. Cidade Universitária, São Paulo-SP, Brasil
www.fflch.usp.br
[2] Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
CP 668 – ICMC-USP, 13.560-970 São Carlos-SP, Brasil
http://www.nilc.icmc.usp.br

iemalves@usp.br, maronezebruno@yahoo.com.br, armariangela@ig.com.br,
{taspardo, sandra}@icmc.usp.br

Resumo. Apresenta-se, neste artigo, uma metodologia semi-automática para detecção de neologismos da língua portuguesa do Brasil. Descrevem-se os trabalhos realizados até então, de forma manual, em geral, e as ferramentas computacionais desenvolvidas, juntamente com os recursos de língua, para auxiliar nesta tarefa.
Palavras-chave: neologismos, lingüística de córpus, léxicos

1. Introdução

Neologismos são unidades lexicais formalmente novas ou formas já existentes empregadas com novo sentido que ocorrem em uma língua, muitas das quais são posteriormente incorporadas a ela.

O estudo de neologismos é importante tanto sob a ótica da Lingüística quanto da Lingüística Computacional: na primeira, estudam-se a formação dos neologismos e o impacto destes na língua; na segunda, os neologismos são essenciais para a manutenção e atualização dos repositórios de dados lingüístico-computacionais (os léxicos, por exemplo) que servem de base para o desenvolvimento de pesquisas nessa área. Na Lingüística Computacional, em particular, esse tipo de pesquisa se enquadra na subárea denominada Lingüística de Córpus.

Este trabalho tem o objetivo de apresentar os resultados iniciais da investigação do uso de um ferramental computacional que permite a detecção semi-automática de neologismos da língua portuguesa. Apresentam-se, inicialmente, uma breve história das pesquisas nessa área e o Projeto TermNeo (histórico, metodologia, resultados obtidos, limitações). Em seguida, na Seção 3, é descrita a metodologia desenvolvida para a detecção semi-automática de neologismos. Na Seção 4, é relatado um estudo de caso. Considerações finais são apresentadas na Seção 5.

[+] Este trabalho contou com o apoio do CNPq.

2. A busca por neologismos

O estudo dos neologismos teve início na década de 50, e sua observação sistemática foi iniciada no início da década de 60, pelo Prof. Bernard Quemada, no *Laboratoire d´Analyse Lexicologique du Centre d´Etude du Vocabulaire Français*, Besançon, França. A partir dessa iniciativa pioneira, vários projetos de observação da neologia surgiram em outros países, inclusive no Brasil, onde, em 1988, foi implementado o *Observatório de Neologismos Científicos e Técnicos do Português Contemporâneo do Brasil* (Projeto TermNeo), com a finalidade de coletar, analisar e difundir aspectos da neologia geral e da neologia científica e técnica do português contemporâneo. Desenvolve também, desde 1993, a observação sistemática da neologia do que se costuma chamar de língua geral, em oposição às línguas de especialidade (*Base de Neologismos do Português Brasileiro Contemporâneo*). Para tanto, foi coletado um córpus constituído por jornais e revistas (*Folha de S. Paulo*, *O Globo*, *IstoÉ* e *Veja*), a partir de janeiro de 1993, observados por amostragem (um veículo por semana). Nesses veículos, são coletados neologismos de caráter vernáculo e estrangeiro. Como princípio metodológico, consideram-se neológicas as unidades lexicais que não estão incluídas em um córpus de exclusão, ou seja, um conjunto de dicionários da língua.

A coleta tem sido efetuada por meio da leitura tradicional, com o registro das unidades lexicais neológicas em fichas lexicais exemplificadas no sítio do projeto. Buscando tornar a coleta de neologismos mais ágil e precisa, e observando-se a metodologia já empregada em outros observatórios (como o ferramental SEXTAN), estabeleceu-se uma parceria com lingüistas computacionais para a elaboração de uma ferramenta que permita a extração automática de candidatos a neologismos, os quais são avaliados por humanos em uma etapa posterior.

3. Ferramentas computacionais para busca de neologismos

Diversas ferramentas para a busca automática de candidatos a neologismos foram desenvolvidas. Na Figura 2, mostra-se o ferramental desenvolvido.

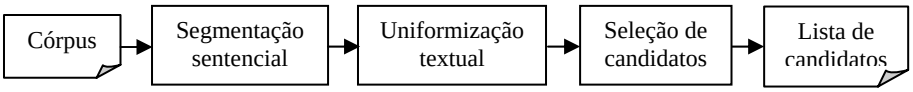

Figura 2. Metodologia de busca automática por candidatos a neologismos

A segmentação sentencial delimita as sentenças dos textos que compõem o córpus. Isso é necessário porque, após a identificação dos termos candidatos a neologismos, estes serão exibidos na sentença em que ocorrem para os especialistas humanos. Na metodologia exibida, o processo de segmentação é realizado pelo SENTER (Pardo, 2006), um segmentador sentencial automático para o português do Brasil.

O processo de uniformização textual consiste em substituir nos textos do córpus as ocorrências de números e endereços eletrônicos por termos genéricos. Por exemplo, para a sentença “No dia 2, João não foi à aula”, o resultado desse processo seria “No dia NÚMERO, João não foi à aula”. A utilização de termos genéricos simplifica a verificação humana posterior que é realizada para identificação dos neologismos. Além disso, garante-se que os elementos substituídos não sejam considerados candidatos a neologismos.

Por fim, o processo de seleção de candidatos a neologismos realiza a busca de todas as palavras do córpus em dicionários de exclusão. Se uma palavra não ocorre em nenhum dicionário consultado, a palavra é considerada candidata a neologismo. Como resultado desse processo, produz-se como saída um arquivo contendo os candidatos a neologismos, todas as sentenças em que ocorrem e uma indicação se os candidatos são nomes próprios ou não, a qual é feita em função da capitalização da primeira letra dos candidatos: se maiúscula, então o candidato em questão pode ser um nome próprio. A exibição de todas as sentenças em que um candidato a neologismo ocorre permite que o especialista humano faça uma análise completa dos sentidos do candidato no córpus.

4. Estudo de caso

Foram analisados dois números de cada ano da revista *Veja On-line*, de 2001 até 2006. Para os anos de 2001 a 2005, foram escolhidos os últimos números dos meses de Abril e Outubro. Para o ano de 2006, os números de 15 de Fevereiro e 22 de Fevereiro. No total, o córpus coletado contém mais de 445 mil palavras.

Após a coleta do córpus, procedeu-se à extração automática dos candidatos a neologismos e à sua análise. Utilizaram-se, para isso, três léxicos computacionais: o léxico do ReGra (Nunes et al., 1996), com mais de 1.5 milhões de entradas; o léxico do Unitex-PB (Muniz, 2004), com aproximadamente 950 mil entradas; o REPENTINO (Sarmento et al., 2006), uma base com cerca de 450 mil nomes.

Apenas dois neologismos se repetiram em todos os anos: *ex-governador* e *ex-ministro*, por se referirem a cargos políticos muito comumente mencionados em textos jornalísticos. Nessa mesma situação estão *ex-deputado* (que só não foi atestado em 2005) e *ex-prefeito* (que só não foi atestado em 2003).

Também se pôde observar como certos neologismos deixaram de ser usados. O neologismo *talibã*, por exemplo, foi bem freqüente no ano de 2001 (9 ocorrências), o ano em que o regime talibã foi derrubado com a guerra do Afeganistão. Em 2002, já são encontradas apenas 3 ocorrências, e em 2003, apenas uma. Em 2004, novamente são encontradas 3 ocorrências. Em 2005 e 2006, já não se encontram mais ocorrências de *talibã* na amostra estudada.

Notam-se casos de neologismos que se referem a fatos surgidos mais recentemente, como *bioplastia* (a partir de 2005), *ipod* (a partir de 2004), *itunes* (a partir de 2005) e *mensalão* (a partir de 2005). Curiosamente, o ferramental detectou como candidato a neologismo uma unidade lexical já relativamente antiga, *camelódromo*, registrada na amostra apenas a partir de 2005. Isso aconteceu pelo fato de tal palavra não estar contida em nenhum dos léxicos consultados.

Como esperado, as unidades lexicais neológicas *pré-candidato* e *pré-candidatura* foram atestadas apenas nas proximidades de anos eleitorais (2001/2002 e 2005/2006), não tendo sido registradas em 2003 e 2004, anos distantes das eleições para presidente e governador.

Certos empregos de diminutivos e aumentativos foram recorrentes, mas relativamente esporádicos: *coleguinhas* (2001 e 2006), *comecinho* (2004 e 2005) e *corpão* (2001 e 2005).

Os neologismos mais recorrentes na amostra estudada pertencem, em geral, às seguintes áreas:

- política: *lulista*, (2002, 2005 e 2006), *político-partidária* (2004 e 2006), tucanato (2002, 2003, 2005 e 2006) e nomes de cargos, como *primeira-ministra* (2001 a 2003) e *vice-prefeito* (2004 e 2006);
- economia e negócios: *custo-benefício* (2003 e 2004), *concorrencial* (2003 e 2004) e nomes de cargos, como *diretor-geral* (2001, 2002, 2004 e 2006) e *economista-chefe* (2002-2004 e 2006);
- tecnologia: (celular) *pré-pago* (2003 e 2005), *pósitrons* (2003 e 2004), *walkman* (2002 e 2006);
- medicina, beleza e saúde: *anabolizante* (2004 e 2006), *botox* (2003-2005), *clonado* (2002, 2003 e 2005), *intragástrico* (2001 e 2002), *vasectomizado* (2003 e 2004).

Também se nota a presença de neologismos relacionados aos atentados terroristas e suas decorrências, como *antiamericanismo* (2003 e 2004) e *antiterror* (2004 e 2006).

Por fim, nota-se uma série de unidades lexicais não-neológicas, porém de baixa freqüência, que, por isso, não constam dos léxicos de exclusão. É o caso, por exemplo, de *café-da-manhã* (raramente grafado com hífen – 2003 e 2004), *mamute* (2002 e 2005) e *séquito* (2001 e 2004), entre outras.

Com o uso das ferramentas computacionais, pode se verificar: o trabalho de 2 a 3 dias para busca manual de neologismos se reduziu a menos de 2 horas; em geral, aproximadamente 98% dos candidatos a neologismos indicados pelo ferramental como não sendo nome próprio eram neologismos, de fato; como esperado, apenas 13% dos candidatos indicados como possíveis nomes próprios eram neologismos.

5 Considerações finais

Os dados já obtidos têm sido utilizados para atualizar os repositórios de dados lexicais existentes utilizados pelos grupos de pesquisa envolvidos neste trabalho. O ferramental computacional desenvolvido encontra-se disponível *on-line* para uso pela comunidade de pesquisa.

Entre as limitações do ferramental desenvolvido, vale citar a impossibilidade de se detectar neologismos semânticos, isto é, unidades lexicais antigas com novos sentidos. A investigação de metodologias (semi-)automáticas para identificar estes neologismos constitui um possível trabalho futuro.

Referências

Muniz, M.C.M (2004). *A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB*. Dissertação de Mestrado. Instituto de Ciências Matemáticas e de Computação, USP/São Carlos, 72p.

Nunes, M.G.V. et al. (1996). The design of a Lexicon for Brazilian Portuguese: Lessons learned and Perspectives. In the *Proceedings of the II Workshop on Computational Processing of Written and Speak Portuguese*, pp. 61-70. CEFET-PR, Curitiba.

Observatori de Neologia (2004). *Metodología del trabajo en neología: criterios, materiales y procesos*. Sèrie Monografies, 9. Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada.

Pardo, T.A.S. (2006). *SENTER: Um Segmentador Sentencial Automático para o Português do Brasil*. Série de Relatórios do NILC. NILC-TR-06-01. São Carlos-SP, Janeiro, 6p.

Sarmento, L.; Pinto, A.S.; Cabral, L. (2006). REPENTINO: A Wide-scope Gazetteer for Entity Recognition in Portuguese. In the *Proceedings of the 7th Workshop on Computational Processing of Written and Spoken Portuguese*. Itatiaia-RJ, Brazil.