Sandra Maria Aluísio (ICMC-USP)

Ieda Maria Alves (FFLCH-USP)

Mariangela de Araujo (FFLCH-USP)

Bruno Oliveira Maroneze (PG-FFLCH-USP)

Thiago A. S. Pardo (ICMC-USP)

Esta exposição tem a finalidade de apresentar os resultados parciais obtidos por uma parceria estabelecida entre pesquisadores do Projeto TermNeo (Projeto Observatório de Neologismos Científicos e Técnicos do Português Brasileiro Contemporâneo) e pesquisadores do NILC (Núcleo Interinstitucional de Lingüística Computacional), da Universidade de São Paulo, com a finalidade de se construir uma ferramenta computacional de extração automática de neologismos. Apresenta-se o Projeto TemNeo e os córpora construídos pelo NILC, a metodologia empregada para a elaboração da ferramenta e os resultados parciais obtidos.

1. Projeto "Observatório de Neologismos Científicos e Técnicos do Português Brasileiro Contemporâneo" (TermNeo)

O Projeto Observatório de Neologismos Científicos e Técnicos do Português Brasileiro Contemporâneo (Projeto TermNeo) tem o objetivo geral de coletar, analisar e difundir aspectos da neologia geral e da neologia científica e técnica do português contemporâneo do Brasil e de elaborar glossários e dicionários terminológicos em algumas das áreas estudadas.

No âmbito desse projeto, a *Base de Neologismos do Português Brasileiro Contemporâneo*, dedicada especialmente ao estudo da neologia geral, cumpre o objetivo de coletar e analisar a neologia do português contemporâneo do Brasil, observada em um córpus jornalístico, fornecendo subsídios para o estudo da evolução do léxico português (variante brasileira) e para a elaboração de

repertórios de unidades lexicais neológicas. Como objetivos específicos, tem também a finalidade de constituir um córpus para a elaboração de dissertações e teses e, ainda, de fornecer subsídios para a elaboração de dicionários de língua e dicionários neológicos do português brasileiro.

Para cumprir esses objetivos, observa-se a neologia do português brasileiro desde janeiro de 1993 por meio de uma coleta sistemática e por amostragem nos veículos:Jornal *O Globo* (primeiro domingo de cada mês)Revista *IstoÉ* (segunda semana de cada mês)Jornal *Folha de S. Paulo* (terceiro domingo de cada mês)Revista *Veja* (quarta semana de cada mês)

A metodologia seguida para o desenvolvimento do Projeto é a da leitura "manual" para a coleta dos candidatos a neologismos. O caráter neológico dos mesmos tem levado em conta o princípio metodológico do *córpus de exclusão*, segundo o qual não é neológica a unidade lexical que não está inserida em um conjunto de obras lexicográficas brasileiras, apresentado a seguir:

ACADEMIA BRASILEIRA DE LETRAS. *Vocabulário ortográfico da língua portuguesa*. Rio de Janeiro: Bloch, 1981 (para o córpus correspondente ao período de 1993 a 1998)

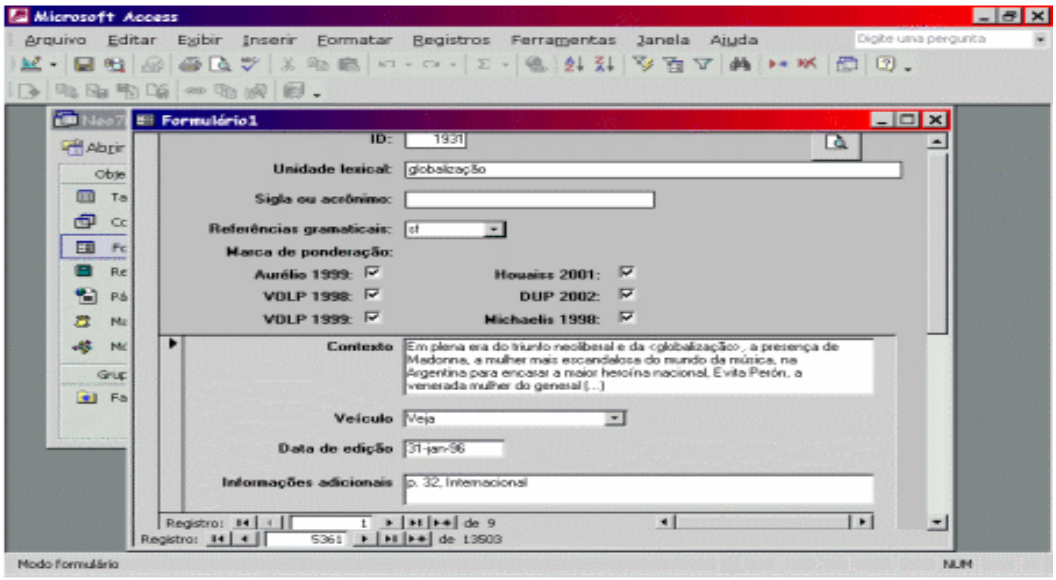
- *Vocabulário ortográfico da língua portuguesa*. 2 ed. Rio de Janeiro: Academia, 1998. 1 ed. 1981 (para o córpus coletado em 1999)
- *Vocabulário ortográfico da língua portuguesa*. 3 ed. Rio de Janeiro: Imprensa Nacional, 1999. 1 ed. 1981 (para o córpus coletado de 2000 a 2004)
- *Vocabulário ortográfico da língua portuguesa*. 4 ed. Rio de Janeiro: Imprensa Nacional, 2004. 1 ed. 1981 (para o córpus coletado a partir de 2005)

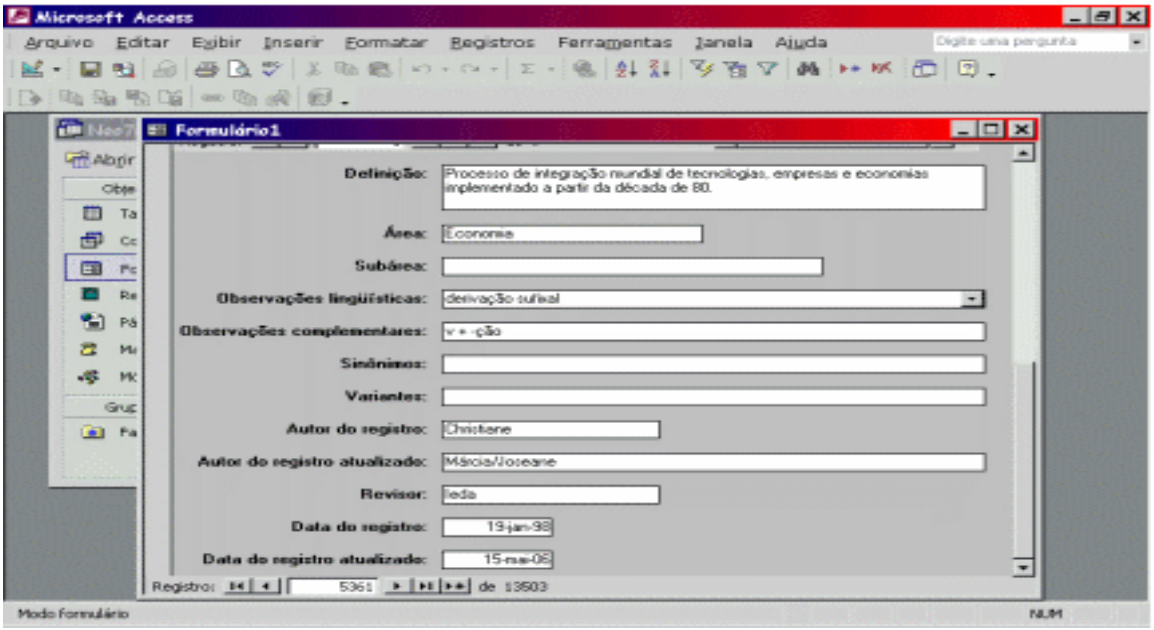
  FERREIRA, Aurélio Buarque de Holanda. *Novo dicionário da língua portuguesa.* 2 ed. Rio de Janeiro: Nova Fronteira, 1986. 1 ed. 1975 (para o córpus correspondente ao período de 1993 a 1999)
- *Novo Aurélio Século XXI: o dicionário da língua portuguesa.* 3 ed. Rio de Janeiro:

Nova Fronteira, 1999. 1 ed. 1975 (para o córpus coletado a partir de 2000)

- MICHAELIS: moderno dicionário da língua portuguesa. São Paulo: Melhoramentos, 1998. (para o córpus coletado a partir de 1999)
- HOUAISS, Antonio e VILLAR. Mauro. *Dicionário Houaiss da língua portuguesa.* Rio de Janeiro: Objetiva, 2001 (para o córpus coletado a partir de 2002)

Os neologismos coletados filtrados pelo córpus de exclusão são em seguida inseridos em uma base de dados Access, que apresenta os seguintes campos (ou a seguinte configuração):





Os dados da *Base*, relativos ao período entre janeiro de 1993 e dezembro de 2000 (cerca de 13 500 unidades lexicais neológicas), têm possibilitado diversas análises quantitativas e qualitativas. Os dados quantitativos revelam que os estrangeirismos correspondem a 17% dos neologismos coletados e os restantes 83% são representados por processos vernáculos: fonológicos (2%), derivados prefixais (30%), derivados sufixais (10%), compostos (22%), compostos sintagmáticos (13%), semânticos (4%), truncações (2%), outros processos como formações com siglas e reduções (2%).

Apesar do auxílio de recursos como textos jornalísticos informatizados disponibilizados pela Internet, dicionários eletrônicos e programas (Folio Builder, Wordsmith Tools) que ajudam na busca de prefixos, sufixos, formações recorrentes, essa metodologia tradicional não permite a análise de um córpus de grandes dimensões. Para os alunos-bolsistas que trabalham no Projeto, dificuldades adicionais são representadas no que concerne ao reconhecimento de neologismos semânticos, de compostos e de derivados prefixais grafados sem hífen (observa-se, por exemplo, nos veículos do córpus, a concorrência entre formas compostas como *auxílio moradia / auxílio-moradia* e formas prefixadas como não *reconhecimento / não-reconhecimento*) e de formações sintagmáticas, muitas vezes confundidas com sintagmas livres.

2. Núcleo Interinstitucional de Lingüística Computacional (NILC-USP)

Desde 1993, o NILC constrói córpus para contribuir em pesquisas de Processamento de Língua Natural, especialmente do Português Brasileiro. Alguns desses córpus, descritos em www.nilc.icmc.usp.br: Córpus NILC, CórpusDT, córpus paralelos, CórpusGIS, RHETALHO, TeMário, Brazilian Portuguese Treebank, Lácio-Web.O Córpus NILC, que é utilizado para a parceria estabelecida com o Projeto TermNeo, é constituído por 40 milhões de palavras repartidas em 40 textos em prosa divididos em textos corrigidos, textos não-corrigidos e textos

semicorrigidos. Esse córpus apóia, especialmente, o projeto ReGra (Revisor Gramatical do Português Brasileiro), em constante atualização desde 1993 e constituído por 68.530 canônicas e 1.563.136 entradas (formas flexionadas, palavras compostas e locuções (5.763 das entradas são locuções nominais, prepositivas, adjetivas, adverbiais, conjuntivas)).

Os pesquisadores do NILC consideram também, para a parceria com o Projeto TermNeo, os dados do Projeto Unitex-PB (http://nilc.icmc.usp.br:8180/unitex-pb/index.html), que integra o Projeto Unitex (França), em parceria com Eric Laporte. Esses dados são representados por um.

Dicionário de Palavras Simples para o Português Brasileiro (aproximadamente 67.500 canônicas associadas a suas regras de flexão), um Dicionário de Palavras Simples Flexionadas para o Português Brasileiro (aproximadamente 880.000 palavras flexionadas) e um Dicionário de Palavras Compostas Flexionadas para o Português Brasileiro (aproximadamente 4.000 entradas).

Consideram também os dados do REPENTINO (http://www.linguateca.pt/repentino), que consiste em um repositório de entidades nomeadas, do português europeu e do português brasileiro, construído pela Linguateca (http://www.linguateca.pt/). Esse repositório apresenta nomes próprios e locuções nominais, sendo representado por aproximadamente 450.000 entradas divididas em várias categorias conceituais.

3. Parceria FFLCH-NILC

A parceria estabelecida entre o grupo TermNeo e o NILC tem o objetivo de desenvolver uma ferramenta computacional que permita maior automatização na extração de neologismos a partir de córpora textuais. Essa ferramenta visa a permitir, além da diminuição do tempo de coleta (tornando desnecessária a leitura "a olho nu") uma maior precisão (com a eliminação do risco de se "deixar passar" um neologismo). Além disso, as contagens de freqüência a partir dos córpora

informatizados possibilitariam a realização de análises quantitativas, além das qualitativas já efetuadas. Do ponto de vista metodológico, o córpus de exclusão deixa de ser o filtro lexicográfico, evitando-se, assim, as arbitrariedades presentes nas obras lexicográficas disponíveis. O córpus de exclusão passa a ser constituído por léxicos extraídos de textos reais, o que tornaria o filtro mais confiável e baseado, efetivamente, no uso.

Para os pesquisadores do NILC, os resultados da parceria devem permitir, também, o enriquecimento e atualização dos repositórios lexicais organizados pelo grupo, ao identificar as eventuais lacunas vocabulares existentes nesses repositórios. Desse modo, espera-se conseguir uma sinergia entre os processos lingüístico (identificação de neologismos) e lingüístico-computacional (enriquecimento de repositórios lexicais).Após o encontro entre pesquisadores do NILC e do grupo TermNeo, foi proposta a realização de um teste-piloto. O córpus utilizado para o teste foi constituído pelos textos das quatro edições da revista Veja de agosto de 2001. A primeira etapa consistiu no pré-processamento do córpus: cada texto era precedido por um cabeçalho, como o que segue:29/08/2001 - Veja - Carta ao Leitor - Dicionário Houaiss - Notícias do idioma vivoPUBLICAÇÃO: Veja DATA: 29/08/2001 EDIÇÃO: 1715 PÁG.: 9 RESUMO: A saga da criação do mais completo e moderno dicionário da língua portuguesa é o tema de capa desta edição de Veja.ASSUNTO PRINCIPAL: DICIONÁRIO HOUAISSPALAVRAS-CHAVES: LINGUAGEM; PORTUGUÊS /IDIOMA/; PUBLICAÇÃO

Veja Carta ao Leitor 29/08/2001 2001

Carta ao Leitor

Dicionário HouaissNotícias do idioma vivo

Numa de suas odes mais famosas, reproduzida em parte na introdução do Aurélio, o poeta chileno Pablo Neruda diz que o dicionário não é uma tumba de...

Esse cabeçalho, que contém informações não-relevantes para a extração de neologismos, foi eliminado automaticamente. Além disso, foram resolvidos automaticamente certos problemas de formatação do córpus, como sentenças quebradas.

Em seguida, o córpus foi segmentado automaticamente pela ferramenta SENTER (http://www.nilc.icmc.usp.br/ thiago/senter.html), que, baseada na pontuação, permite identificar o início e o fim de cada sentença. Dessa forma, o resultado é uma sentença por linha. Por fim, dados como números, datas, endereços de sítios da Internet e de correio eletrônico foram substituídos por categorias genéricas, como NÚMERO, DATA etc.

Para esse teste-piloto, foram usados como córpus de exclusão os três léxicos do NILC já mencionados: o léxico do NILC, o Unitex-PB e o REPENTINO. Esses léxicos foram pré-processados, em primeiro lugar, para excluir informações irrelevantes para a coleta de neologismos, como informações morfossintáticas. Além disso, as locuções presentes nos léxicos foram segmentadas (por exemplo, Rio de Janeiro foi separado em três unidades), porque, nesse teste-piloto, a preocupação era com a detecção de neologismos de uma única unidade.

Finalmente, após o pré-processamento, foi efetuada a filtragem: as 253.361 palavras do córpus foram comparadas às listas dos três léxicos. Aquelas presentes no córpus, mas que não constavam dos léxicos, foram consideradas candidatas a neologismos. Os resultados obtidos (quantitativamente) foram os seguintes:

Léxico do NILC:

- 4.478 palavras candidatas (1.7%) / 4.395 palavras candidatas distintas

Unitex-PB

- 5.223 palavras candidatas (2.6%) /4.815 palavras candidatas distintas

REPENTINO

- 30.429 palavras candidatas (15.1%)/ 30.250 palavras candidatas distintas Todos

- 1.803 palavras candidatas (0.8%) / 1.633 palavras candidatas distintas

Novamente, o léxico do NILC foi o que possibilitou a melhor filtragem. A porcentagem final de candidatos foi bastante semelhante à do primeiro teste.Os 1.803 candidatos foram analisados por um lingüista do projeto TermNeo; novamente, a análise levou apenas cerca de 2 horas. Os percentuais foram muito semelhantes. Em primeiro lugar, 1.164 candidatos (64,6%) foram marcados com "talvez seja nome próprio", sobrando 639 candidatos (35,4%) com probabilidade mais alta de serem neologismos.Após a análise pelo lingüista, foram obtidos os seguintes resultados: dentre os prováveis nomes próprios, apenas 164 (14%) eram de fato neologismos, enquanto, dentre os demais, 626 (98%) eram neologismos. No total, 790 (43,8%) foram validados como neologismos.Os dados quantitativos revelaram que a ferramenta detectou os nomes próprios com alta probabilidade de acerto. Além disso, a porcentagem final de neologismos em relação ao tamanho total do córpus ficou em 0,39% em ambos os testes, o que parece revelar que essa deve ser, em média, a quantidade de neologismos presente em córpus do tipo analisado.Foram detectados alguns problemas, que a ferramenta não pôde resolver. Em primeiro lugar, alguns números, datas e endereços da Internet não puderam ser automaticamente eliminados, e foram incorretamente detectados como neologismos. Além disso, a segmentação sentencial é insuficiente, em alguns casos, e pode não permitir a correta identificação do significado e do uso do neologismo.Outros dois problemas que a ferramenta não consegue resolver referem-se a dificuldades já detectadas no processo de leitura tradicional. Em primeiro lugar, há o caso das formações sintagmáticas, que não são unidas por hífen, bem como certas composições e formações prefixais. Para detectar esses casos, foi sugerida a utilização de um analisador sintático automático (parser), que ainda precisa ser testado.Em segundo lugar, há o problema dos neologismos semânticos, para o qual ainda não foram encontradas soluções satisfatórias. Nesse caso, a provável solução será o retorno à leitura tradicional, por

amostragem.

4. Considerações finais

De maneira geral, o resultado dos testes-piloto foi bastante positivo: o tempo de coleta foi significativamente reduzido, o que permitirá a inclusão de mais veículos no córpus (possivelmente a inclusão do jornal O *Estado de S. Paulo* e da revista *Época*), bem como o aumento de edições analisadas (todas as edições das revistas semanais e jornais de dois domingos por mês). O ambiente informatizado criado para a extração de neologismos será integralmente disponibilizado para uso por outros pesquisadores que enfrentam o mesmo problema na coleta neológica. No momento, já estão disponibilizadas as ferramentas SENTER, Pre-processing program e Filtering program (www.nilc.icmc.usp.br/ thiago/neologismo.html).