PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

DL - DEPARTAMENTO DE LETRAS

ELABORAÇÃO DE DICIONÁRIO ELETRÔNICO

Mauro Rebello[1],
Violeta de San Tiago Dantas Barbosa Quental [2].

CNPq
Conselho Nacional de Desenvolvimento
Científico e Tecnológico

[1] *Aluno do curso de Letras da PUC-Rio*
[2] *Professora e pesquisadora da área de Processamento natural da linguagem da PUC-Rio, Departamento de Letras.*

SUMÁRIO

1. Introdução

O grupo de pesquisa da área de Processamento de Linguagem Natural do Português, da PUC-Rio, atua em projeto de pesquisa integrado ao projeto "Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil - PLN-BR" (edital CTInfo/MCT/CNPq nº 011/2005), em colaboração com as Universidade de São Paulo (USP), campus de São Carlos; Universidade Federal de São Carlos (UFSCar); Universidade Estadual Paulista (UNESP), campus de Araraquara; Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS); Pontifícia Universidade Católica do Rio de Janeiro (PUC/RJ); à Universidade do Vale do Rio dos Sinos (UNISINOS); e Universidade Presbiteriana Mackenzie. Cada grupo dessas universidades é responsável por um subprojeto. O grupo da PUC-Rio, sob a coordenação da Profª. Violeta Quental, desenvolve o subprojeto Aprendizagem Automática de Informações Lexicais.

2. Objetivo

Como parte das atividades de aquisição de informações sobre o léxico, elaborou-se o projeto Prepoly, desenvolvido em colaboração com o professor Eckhard Bick, da Aarthus University e Linguateca (Pólo de Oslo). O objetivo geral do projeto é definir expressões prepositivas do português que constituem expressões multivocabulares, com grau de fixidez suficiente para serem incorporadas a um léxico - no caso o léxico do parser Palavras (Bick, 2000).

Quando analisadas por um parsing, expressões multivocabulares são etiquetadas como um vocábulo único, por questões que envolvem economia de processamento do parser e, do ponto de vista lingüístico, por apresentarem significado não-composicional e por essa análise refletir adequadamente o fenômeno de economia lexical. Dessa forma, a expressão "de gelo", por exemplo, se analisada como multivocábulo, será classificada como sintagma preposicional composto (PP), com função adjetival (**ADJ** @N<), como no extrato da saída do parser Palavras (http://visl.hum.sdu.dk/visl/pt/parsing/automatic/) em consulta realizada em 13/08/2007. Nesse caso, a expressão é fixa, com significado metafórico, de adjetivo atributivo. Veja-se o exemplo:

ela [ela] **PERS** F 3S NOM/PIV
tem [ter] **V** PR 3S IND VFIN
um [um] <quant> <arti> **DET** M S
coração [coração] **N** M S
de=gelo [de=gelo] PP

ela [ela] **PERS** F 3S NOM @SUBJ>
tem [ter] <fmc> **V** PR 3S IND VFIN @FMV
um [um] <arti> **DET** M S @>N
coração [coração] **N** M S @<ACC
de=gelo [de=gelo] **ADJ** @N<

A mesma seqüência de palavras, no entanto, não deveria ser analisada como multivocábulo na frase "traga duas pedras *de gelo* seco!", mas como sintagma preposicional composicional, formado de preposição, nome e adjetivo, com a possível interpretação de "gelo seco" como multivocábulo. Vê-se, no exemplo de saída abaixo, que a presença no léxico do sintagma preposicional marcado como expressão multivocabular acarreta uma análise incorreta pelo parser:

traga [tragar] **V** PR 3S IND VFIN

duas [dois] <card> **NUM** F P
pedras [pedra] **N** F P
de=gelo [de=gelo] PP
seco [seco] **ADJ** M S

traga [tragar] <fmc> **V** PR 3S IND VFIN @FMV
duas [dois] <card> **NUM** F P @>N
pedras [pedra] **N** F P @<ACC
de=gelo [de=gelo] **ADJ** @N<
seco [seco] **ADJ** M S @PRED>

A definição de quais expressões devem ser marcadas no léxico como compostas é, portanto, fundamental para a correção do parser, e esse é o objetivo da pesquisa atual: rever, acrescentar, modificar a lista atual de sintagmas preposicionais multivocabulares presente no léxico do Palavras. Em geral, nota-se que a atual lista contém expressões que se comportam das duas formas: composicionalmente, ou, quando em sentido metafórico, como multivocábulos. É necessário então definir em que contextos essas formas ocorrem, com qual freqüência relativa, para que se possa decidir pela validade de mantê-las no léxico como sintagmas preposicionais compostos.

Esta pesquisa usa a metodologia de busca por *concordance* em corpus, e, por isso, torna-se indispensável a consulta a corpora robustos da língua portuguesa. As buscas pelas expressões prepositivas são realizadas através do mecanismo de *concordance* presente nos corpora eletrônicos disponibilizados pela Linguateca (http://www.linguateca.pt) e Corpus do Português de Davies & Ferreira, disponível no site http://www.corpusdoportugues.org/.

Realizamos, ainda, buscas através do Google, para incluir usos de linguagem mais informal, ausentes nos corpora consultados. Essas expressões prepositivas candidatas a multivocábulos têm de ser analisadas em seus contextos de ocorrência, para que, depois de avaliadas em relação à predominância de uso como compostos, se mantenham ou sejam incorporadas como tal ao léxico computacional utilizado pelo parser Palavras (Bick, 2000).

3. **Metodologia**

Na primeira etapa do projeto, a partir de uma lista prévia de 1400 sintagmas preposicionais compostos fornecida pelo Professor Eckhard Bick, houve a divisão dos sintagmas em três grupos gramaticais, a saber: locuções prepositivas com função adverbial, adjetival, ou adverbial e adjetival. Para cada nova sub-lista, foi criada uma pasta eletrônica respeitando o critério gramatical já descrito. Na lista dos sintagmas adverbiais, dispúnhamos de expressões como (“a=seu=modo”), (“ao=mesmo=passo”), (“até=debaixo=da=água”) etc...; para a dos sintagmas adjetivais, tínhamos, por exemplo, (“da=mesma=farinha”), (“das=arábias”), (“de=fachada”); e, para a lista dos sintagmas adverbiais-adjetivais, encontramos termos como (com=a=corda=no=pescoço), (com=as=mãos=na=massa), (de=araque) etc... A partir dessas listas, foram realizadas consultas para cada item nos diversos corpora eletrônicos de língua portuguesa disponíveis na internet, e a captação gradativa de cada um dos 1400 sintagmas preposicionais em seus contextos de ocorrência.

Para a análise dos sintagmas nas sentenças em que aparecem nos corpora, é necessário dispor de um grande número de ocorrências, para que possa ser realizada a análise percentual de aparição em contexto desses termos. Contudo, a consulta exclusiva nos sites de corpora mostrou-se insatisfatória, pois o número de ocorrências para cada vocábulo pesquisado era insuficiente para a realização de uma análise mais precisa. O grupo então está à procura de outros corpora eletrônicos, além de utilizar agora a ferramenta de busca Google.

A análise dos vocábulos é realizada da seguinte forma: após a captação eletrônica, são estabelecidos critérios quanto à análise desses sintagmas como *multiwords.* Supondo que a expressão "do peito" tenha sido captada neste contexto: "...falou o amigo do peito", submetemos a sentença ao analisador Palavras. O *parser* categoriza "amigo do peito" da seguinte maneira: **amigo=do=peito** [amigo=do=peito] **N** M S, onde "N" é nome, ignorando seu possível caráter composicional, como quando submetemos à análise somente a expressão "do peito", e obtemos o resultado: (**de** [de] <sam-> **PRP)** , no qual "PRP" é pronome, (**o** [o] <-sam> <artd> **DET** M S), no qual "DET" é artigo definido e determinante, e (**peito** [peito] **N**), no qual "N" é nome.

Após a análise, há de se decidir sobre a permanência ou eliminação da expressão da lista. Para este fim, primeiramente identificam-se os casos em que a ocorrência da expressão como multivocábulo concorre com a ocorrência da expressão como sintagma preposicional composicional. Se a multipalavra ainda não fizer parte do léxico já compilado, é incorporada à lista; se for estatisticamente irrelevante à análise, é excluída.

Às palavras remanescentes são agregadas, na última etapa, indicações em linguagem formalmente compreensível pelo *parser* quanto ao seu uso como multivocábulo e seu contexto de aparição. Assim, por exemplo, acrescentam-se informações sobre os verbos com os quais os sintagmas preposicionais aparecem como multivocábulos. A expressão "de quatro", por exemplo, quando composta ocorre normalmente com os verbos "ficar" ou "cair"; esse dado seria incorporado ao léxico na forma <+ficar> |<+cair>.

5. Conclusões

O maior problema encontrado na pesquisa é a inclusão de novos multivocábulos à lista, já que a busca por expressões regulares como [PREP] + [N] traria uma quantidade de respostas incalculável e, na maioria dos casos, irrelevantes para a questão da pesquisa. Fica-se, por isso, com a possibilidade de acrescentar à lista de PPs apenas as expressões que intuitivamente consideramos boas candidatas à análise como multiwords, ou aquelas que apareçam nas buscas por acaso.

Até o momento foram realizadas as buscas e análise das expressões prepositivas com função adjetival pelo bolsista. Nos próximos meses serão analisadas as expressões com função adjetival e/ou adverbial. Paralelamente, outra bolsista financiada pelo Projeto PLN-BR - Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil estuda as expressões com função adverbial. O bolsista participará da Mostra PIBIC da PUC-Rio, com apresentação em formato de pôster.

Referências

AFONSO, Susana. *Árvores deitadas: Descrição do formato e das opções de análise na Floresta Sintáctica.* Documentação principal, em constante atualização. Disponível em [http://acdc.linguateca.pt/treebank/DocumentacaoFloresta.html#Bick2000]

BICK, Eckhard. *The Parsing System Palavras, Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, Aarhus University Press, 2000.