Departamento de Letras

Elaboração de Dicionário Eletrônico

Aluno: Mauro Ricardo Rebello de Paiva
Orientadora: Prof. Dra.Violeta de San Tiago Dantas Barbosa Quental

Introdução

O grupo de pesquisa da área de Processamento de Linguagem Natural do Português, da PUC-Rio, atuou 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 a 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 foi responsável por um subprojeto. O grupo da PUC-Rio, sob a coordenação da Prof[a]. Violeta Quental, desenvolveu o subprojeto Aprendizagem Automática de Informações Lexicais, do qual o projeto Prepoly faz parte.

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 University of Southern Denmark 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, ao invés de ser analisada item a item. Nesse caso, a expressão é fixa, com significado metafórico, de adjetivo atributivo. Veja-se o exemplo, retirado de consulta ao parser Palavras:

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 de consulta ao parser abaixo, que a presença no léxico do sintagma preposicional marcado como expressão multivocabular acarreta uma análise incorreta:

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.

A etapa atual do projeto tem como objetivo a análise de expressões multivocabulares coletadas nos corpora de língua portuguesa disponíveis na internet, bem como o estudo estatístico e descritivo dessas expressões de acordo com os princípios da composicionalidade, substitucionalidade, e modificabilidade - que são os critérios citados na literatura da área para a definição de multivocábulos - para a produção posterior de identificações individuais relativas a marcas contextuais e a possíveis restrições quanto ao seu uso como multivocábulos.

Metodologia

A partir da lista de 1400 sintagmas preposicionais fornecida pelo Prof. Eckhard Bick, da Universidade de Aarhus, Dinamarca, estão sendo estudadas as características de formação e atuação desses sintagmas em seus contextos de ocorrência, de modo a identificar as situações em que devem ser analisados como expressões fixas, e as situações em que devem ser analisados como multivocábulos. Para definir contextos de ocorrência e frequência de ocorrência foram realizadas buscas por concordance destas expressões em corpora da língua portuguesa.

Os vocábulos são analisados segundo os critérios da composicionalidade – principio através do qual se obtêm o significado do todo a partir do significado das partes; da substitucionalidade, que julga a possibilidade de se realizarem substituições de morfemas de uma expressão sem alterar seu reconhecimento e significado; e da modificabilidade, que verifica se é possível a inserção de palavras no corpo da expressão e se essa modificação altera seu significado.

Assim, por exemplo, seguindo os princípios da composicionalidade, é realizada a análise da expressão "de efeito" na sentença "Foi observada uma alteração de efeito inverso ao registrado anteriormente (...)". Neste exemplo, a expressão "de efeito" foi composta a partir da preposição "de" somada ao substantivo "efeito", que exprimem, separadamente, um conteúdo semântico que se mantém na sentença. No entanto, quando a expressão é analisada na frase "Como eu não sabia mais o que falar, comecei a soltar frases de efeito para ver se o convencia (...)", semanticamente ela não mais exprime o significado composicional da sentença anterior, mas representa um significado distinto que abrange a expressão como um todo, e não somente suas partes separadamente. Nesses casos, "de efeito" seria classificada como uma expressão sem composicionalidade, ficando sujeita a uma análise etimológica apurada e à cogitação de seu uso como multivocábulo.

Ao analisarmos a expressão "de bolso", por exemplo, na frase "acabo de comprar uma edição de bolso do diário de um mago (...)", constatamos que o mesmo significado obtido na sentença não poderia ser substituído, por exemplo, pelos das expressões "para bolso", ou "de miniatura", mesmo que estas representassem também o tamanho pequeno da edição de um livro. A expressão "de bolso" seria classificada pois como uma expressão sem substitucionalidade, já que seu significado não seria reconhecido se houvesse substituições em sua estrutura gramatical.

De acordo com o princípio da modificabilidade, ao analisarmos a expressão "com bons olhos", por exemplo, na sentença " (...) eu vejo com bons olhos esse casamento (...)", entendemos o significado da expressão no contexto ao qual se refere, de aprovação, de gosto, etc... Se fosse acrescentada à sentença a palavra "castanhos", resultando em "vejo com bons olhos castanhos esse casamento", mesmo que a compreensão do significado da expressão se mantivesse, esta nova expressão seria identificada como muito pouco usual, causando estranheza e dificuldade de entendimento.

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. Ficaríamos, 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 retirados dos diversos corpora do português os resultados das buscas pelas expressões de caráter adjetival em seus contextos de aparição. Na fase atual da pesquisa, estão sendo analisados os sintagmas preposicionais em seus contextos de ocorrência de acordo com as premissas da composicionalidade, substitucionalidade e modificabilidade, para sua posterior classificação como multivocábulo ou não, e para a formalização de suas propriedades.

Para automatizar a busca por expressões preposicionais multivocabulares, estaremos a partir de agora usando a ferramenta LinguisticTools, desenvolvida por CAMINADA (2008), que utiliza medidas estatísticas para apresentar resultados relevantes em corpora.

Referências

1- Bick, E. 2000. *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework* , Århus 2000.

2- Caminada, Nuno. *Identificação automática de expressões cristalizadas preposicionais em corpora da língua portuguesa*. Dissertação de Mestrado, Instituto Militar de Engenharia, Rio de Janeiro, 2008.

3- Davies e Ferreira. Corpus do Português. http://www.corpusdoportugues.org/

4- Linguateca - http://www.linguateca.pt/

5- Projeto PLN-BR - http://www.nilc.icmc.usp.br/plnbr/