

ELABORAÇÃO DE DICIONÁRIO ELETRÔNICO

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

Introdução

Esta pesquisa faz parte do Projeto Prepoly, desenvolvido em colaboração com o Prof. Eckhard Bick, da Aarhus University, e se situa dentro do escopo do Projeto Processamento de Linguagem Natural do Português do Brasil, PLN-BR, financiado pelo CNPq e desenvolvido em parceria com as universidades USP/São Carlos, Unisinos, Federal de São Carlos, PUCRS e Mackenzie. O objetivo geral do projeto Prepoly é a análise de expressões prepositivas do português, especialmente quando constituem expressões multivocabulares.

Quando submetidas à análise de um parsing, expressões multivocabulares são etiquetadas preferencialmente como um vocábulo único, por questões como economia de processamento e por corresponderem mais adequadamente ao fenômeno da composição lexical. A expressão “em pêlo”, por exemplo, se analisada como multivocábulo, seria classificada como sintagma preposicional com função adjetival, como em “nu ‘em pêlo’”, e não como sintagma adverbial formado de preposição e substantivo, como em “carrapato ‘em pêlo’ de cachorro”.

Para a decisão de análise de sintagmas preposicionais como expressões multivocabulares, tem de se realizar a coleta dessas expressões em seus contextos de ocorrência, e analisá-las lingüística e estatisticamente quanto à sua relevância como multivocábulos, para posterior inserção no léxico computacional – no caso dessa pesquisa o léxico utilizado pelo parser Palavras (Bick 2000) -, de modo a constarem nos resultados de análise morfossintática automática da língua portuguesa.

Objetivos

Esta etapa do projeto tem como objetivo a coleta de expressões multivocabulares disponíveis nos corpora de língua portuguesa disponíveis na internet, bem como a análise estatística e descritiva dessas expressões a partir de seus contextos de ocorrência, e a produção de identificações individuais relativas a marcas contextuais e a possíveis restrições quanto ao seu uso como multivocábulos.

Metodologia

A partir de uma lista prévia de 1400 sintagmas preposicionais fornecida pelo Prof. Eckhard Bick, da Universidade de Aarhus, Dinamarca, está sendo realizada a captação dessas expressões em seus contextos de ocorrência, de modo a identificar as situações em que devem ser analisadas como multivocábulos.

Assim, por exemplo, é realizada a busca pela expressão “sem jeito” em seus contextos frasais em diversos corpus de língua portuguesa. Estão sendo considerados para isso corpora compilados para pesquisa acadêmica, como os da LINGUATECA (<http://www.linguateca.pt>) e o Corpus do Português de Davies & Ferreira, disponível no endereço <http://www.corpusdportugues.org/>. Além desses, buscamos também as expressões na

ferramenta Google, para dar conta de usos mais informais da língua, não contemplados nos corpora citados.

Em seguida, são estabelecidos critérios de análise quanto às aparições dessa expressão como *multiword*. Supondo, por exemplo, que uma das sentenças encontradas tenha sido a seguinte: “Respondi, meio **sem jeito** (...)”. Submetida à análise do parser Palavras, a expressão em negrito será analisada morfológicamente como sintagma preposicional (**sem=jeito** [sem=jeito] PP) com função adjetival (ADJ) predicativa em relação ao sujeito (@SC), sendo ignorado seu possível caráter composicional. Veja-se a análise a seguir:

respondi [responder] <fmc> V PS 1S IND VFIN @FMV
, [,] PU
meio [meio] <det> <quant> ADV @<ADVL
sem=jeito [sem=jeito] ADJ @<SC

Da mesma forma, uma expressão como “de fininho”, em “Saiu **de fininho**”, deveria ser analisada como sintagma preposicional com função adverbial, mas não consta ainda do léxico do Palavras. Assim, essa sentença será analisada como

saiu [sair] V PS 3S IND VFIN
de [de] PRP
fininho [fininho] N M S

perdendo-se a informação de que se trata de uma expressão fixa com significado cristalizado.

A partir dos resultados das buscas, nos casos em que a ocorrência da expressão como multivocábulo concorre com a ocorrência da expressão como sintagma composicional, há de se decidir sobre sua permanência com etiquetas que indiquem quando o parser deverá analisá-la como multivocábulo, sua eliminação da lista, por não ser estatisticamente relevante à análise, ou seu acréscimo à lista, quando não faz parte do léxico já compilado.

Por último, tendo permanecido na lista, são realizadas indicações em linguagem formal quanto ao contexto de aparição da expressão, e quanto ao seu uso como multivocábulo. “De fininho”, por exemplo, seria analisado da mesma forma como “de repente” (**de=repente** [de=repente] PP) como Sprep, e acrescido da informação <+sair>, por sua frequência de aparecimento com esse verbo.

Conclusões

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 contextos de ocorrência nos quais estas palavras estão inseridas, para sua classificação como multivocábulo ou não, e para a formalização de suas propriedades. A pesquisa por expressões preposicionadas adverbiais está sendo conduzida paralelamente no âmbito do projeto Prepoly.

Referências

- [1] Bick, E. 2000. *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, Århus 2000.
- [2] Davies e Ferreira. Corpus do Português. <http://www.corpusdoportugues.org/>
- [3] Projeto PLN-BR - <http://www.ronaldomartins.pro.br/plnбр/>
- [4] Linguateca - <http://www.linguateca.pt/>