AELIUS FALADO 1.0

Mônica Rigo Ayres - Bolsista UFRGS, PIBIC-CNPq
monicarigoayres@hotmail.com

Gabriel de Ávila Othero – Orientador UFRGS, CNPq
gabriel.othero@ufrgs.br

OBJETIVOS

Com o avanço da tecnologia temos melhorias em vários âmbitos da ciência, e com a linguística não poderia ser diferente. Portanto, nossa pesquisa pretende contribuir com a melhoria de um programa de anotação morfossintática automática, o etiquetador Aelius, e além disso, proporcionar à equipe do VARSUL uma ferramenta automática de *POS tagging* de qualidade, robusta e gratuita de anotação de corpora de língua falada.

LINGUÍSTICA DE CORPUS

A Linguística de Corpus trata de coletar, compilar e explorar conjuntos de textos para pesquisa linguística de uma determinada língua. O foco de nosso trabalho é a parte da *Anotação, que* é apenas uma das muitas tarefas que podem ser executadas na área do *Processamento Natural da Linguagem.*

CORPUS

O corpus utilizado é do Banco de Dados do projeto VARSUL, de fala espontânea, e foi coletado em Porto Alegre nos anos 90. Foram anotados e revisados 20 trechos totalizando 410 páginas e 154.530 palavras.

ANOTADOR AUTOMÁTICO

O anotador automático morfossintático utilizado em nossa pesquisa é o Aelius, que foi desenvolvido pelo prof. Leonel Alencar, da Universidade Federal do Ceará, que coordena o grupo CompLin – Computação e Linguagem Natural. Esse projeto surgiu da necessidade de tornar acessível a estudantes e pesquisadores de linguística a análise automática de textos.

PROCESSO

Os trechos do VARSUL foram anotados automaticamente pelo programa, após, corrigimos manualmente a etiquetagem feita pelo anotador automático, analisando cuidadosamente as etiquetas que apareceram e seus contextos. Posteriormente, analisamos os erros que ocorreram na anotação e propomos melhorias para o Aelius, com a finalidade de que acerte um número ainda maior de etiquetas em sua anotação.

PRINCIPAIS IMPASSES

O Aelius é um anotador programado para analisar corpus de língua escrita, sendo sua etiquetagem baseada em documentos históricos e de língua escrita. Dessa maneira, possui limitações com marcadores discursivos, nomes próprios compostos, hesitações, derivação imprópria, interjeições, etc., além de não reconhecer gírias e expressões típicas da região Sul do Brasil.

RESULTADOS/CONCLUSÕES

O anotador utilizado possui muitos méritos, tendo grande número de acertos, mesmo tratando de língua falada. Seus erros são de tipos variados, alguns deles abrangem também língua escrita, mas em nossa pesquisa propomos soluções para os erros típicos de língua falada, como interjeições, truncamentos, onomatopeias, casos de aférese e marcadores conversacionais. Com isso esperamos que sejam feitas as melhorias necessárias no software do Aelius para que sua taxa de acerto seja ainda mais satisfatória. Sua acurácia é de 95,4% para textos de língua falada, e esperamos que com nossas sugestões implementadas a acurácia se torne ainda maior.