ANOTAÇÃO AUTOMÁTICA DO CORPUS DO VARSUL

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

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

OBJETIVOS

Anotar automática e manualmente novos trechos do corpus do VARSUL.

Melhorar o funcionamento do Aelius, uma ferramenta gratuita de anotação morfossintática do português.

Proporcionar à equipe do VARSUL uma ferramenta automática de *POS tagging* de qualidade, robusta e gratuita de anotação de corpora falados.

CORPUS

Corpus de fala espontânea, coletado na região Sul do Brasil.

Foram anotados e revisados:

7 trechos com cerca de 20 minutos de fala transcrita

76 páginas de texto
21.337 palavras

ANOTAÇÃO

A anotação é apenas uma das muitas tarefas que podem ser executadas na área do *Processamento Natural da Linguagem.*

Utilizamos o anotador automático Aelius, desenvolvido pelo prof. Dr. Leonel Alencar, da UFC.

PROCESSO

1. Foram selecionados excertos do corpus do VARSUL para a anotação
2. Os trechos do VARSUL foram anotados automaticamente
3. Corrigimos manualmente a anotação feita pelo anotador automático
4. Confrontamos a anotação automática com a manual e analisamos as principais deficiências do Aelius

PRINCIPAIS IMPASSES

O Aelius é um anotador de corpus escrito.

Sua etiquetagem é baseada em documentos históricos.

Por isso, possui limitações com marcadores discursivos, nomes próprios compostos, hesitações, derivação imprópria, interjeições, etc.

RESULTADOS

O anotador utilizado possui muitos méritos, tendo grande número de acertos.

Seus erros são de tipos variados e ainda serão investigadas suas causas e como fazer as melhorias necessárias em seu software para que aumentem ainda mais as taxas de acerto, que até agora são de 95%.