

Introdução

Diana Santos

Abril 1992

Esta publicação tem como objectivo descrever o corpus coligido no INESC em 1990, como ferramenta para o estudo do português, tanto para o teste de hipóteses linguísticas como para a avaliação de ferramentas computacionais que lidem com a nossa língua.

1 Porquê este corpus

Sendo de dimensão ínfima comparado com os corpora dignos desse nome (cf. por exemplo a dimensão do Corpus do Português Fundamental, ou do Lancaster-Oslo-Bergen Corpus para o inglês), é suficientemente grande, parece-nos, para experimentar metodologias, ferramentas, e para transmitir uma primeira impressão àqueles linguistas, que, formados numa perspectiva essencialmente teórica, não estão habituados a observar a sua língua em quantidade.

É também útil, em nossa opinião, para determinar o desempenho de gramáticas ou outros analisadores da língua portuguesa que não tenham sido desenhados especificamente para ele, visto que a sua dimensão exclui à partida o desenho de uma gramática que “apenas” cubra este ou aquele tipo de construções exemplificado no corpus.

De facto, a sua criação foi precisamente determinada pela necessidade de avaliar o desempenho de um analisador sintáctico do português então em desenvolvimento no INESC, associada à necessidade de orientação em relação aos problemas a descrever em seguida. Isto porque, sendo o nosso objectivo o de construir ferramentas de processamento de linguagem natural de cobertura vasta (“broad-coverage”), a ordem pela qual se vão “atacando” os problemas tem de ser dirigida pelos dados, e não arbitrária.

2 Esforço comum

Os investigadores envolvidos na selecção (manual) do corpus foram (por ordem alfabética e em partes iguais) António Colaço, Carla Fernandes, Dalila Rosales e Diana Santos. Como revisores, trabalharam no corpus Carla Fernandes, Diana Santos e Rui Marques.

O preenchimento do dicionário de forma a obter o corpus anotado foi feito quase exclusivamente por Rui Marques.

3 Descrição do conteúdo

As contribuições aqui coligadas são de variados autores e olham para o corpus de maneiras diversas. Passo a descrevê-las de forma resumida.

O primeiro texto, da minha autoria, é uma pequena descrição do corpus, estilística e lexical, cujo objectivo é apenas fornecer alguns ponteiros para a sua apreciação global.

O segundo artigo, da autoria de Rui Marques, constitui um esforço de categorização sintáctica do corpus sem ter acesso a ferramentas computacionais de análise sintáctica. Por

isso, o estudo incide principalmente sobre palavras gramaticais ou advérbios, e sobre noções de ordem em vez de análise de constituintes, ainda que coloque problemas tão interessantes como a identificação da função gramatical do primeiro sintagma de uma frase, ou da ligação adverbial a sintagmas verbais, nominais, ou à frase como um todo.

O terceiro artigo, da autoria de Regina Reis, versa a interessante questão das palavras que aparecem no corpus mas não em dicionários de língua corrente.

Segue-se uma descrição das ferramentas computacionais desenvolvidas no INESC para o processamento de corpora de texto em geral, da autoria de José Carlos Medeiros.

O quinto documento, a que chamámos “Português Quantitativo”, descreve uma experiência e algumas medidas efectuadas sobre o corpus, usando as ferramentas descritas no texto anterior. neste artigo descreve-se além disso a forma de obter o corpus anotado que também incluímos nesta colectânea, e qual o sentido das anotações.

Segue-se o corpus simples e a sua versão anotada, finalizando com a descrição das fontes usadas na sua recolha.