Comparador: forma de auscultar corpos no AC/DC

Alberto Simões

19 de Julho de 2014

Resumo

Com a quantidade de diferentes corpos e a grande diversidade entre o que está anotado e marcado em cada um, é necessário desenvolver ferramentas que possam permitir ao linguista comparar fenómenos de várias perspetivas diferentes.

O Comparador, já mencionado em vários textos sobre o AC/DC ou a Linguateca em geral [2, 3], mas nunca ainda documentado independentemente, é uma destas ferramentas, que agora com a Gramateca urge tornar disponível para uma comunidade maior de interessados.

Os requisitos do Comparador foram discutidos e acordados entre Diana Santos e Alberto Simões, e a documentação foi feita em conjunto, mas o desenvolvimento é apenas da responsabilidade de Alberto Simões.

1 Objetivo

Embora os dados e os resultados existam acessíveis para todos os que usam o AC/DC na sua pesquisa, a quantidade de informação associada aos corpos e a diversidade do material podem tornar as pesquisas complicadas para um linguistia desprevenido.

Poder comparar diferentes pesquisas com uma única interface, e obter os resultados diretamente num único lugar, torna muito mais fácil o reaproveitamento do material e a própria interação com o AC/DC, da mesma forma que o Ensinador[1] [4] tornou possível usar o mesmo material num contexto didático de sala de aula.

Outras ferramentas existem ou foram desenvolvidas para reuso do material do AC/DC para outros fins ou por outros tipos de utilizadores, como é o caso do Ordenador, do Distribuidor, do Castor, do Ensinador e do VARRA.

A inspiração original do Comparador, para que conste, foi um programa que comparava a frequência na rede (Internet) no final dos anos 90, usado para ensinar inglês para estrangeiros. (NOTA: estamos à procura da referência exata!)

[1] `http://www.linguateca.pt/Ensinador/`

2 Como utilizar

O Comparador, como todas as ferramentas associadas ao AC/DC, resume-se, do ponto de vista do utilizador, a uma interface na Rede em que se escolhe um conjunto de parâmetros e se obtém o resultado em HTML.

Do ponto de vista de interface, as duas pesquisas a comparar são especificadas em paralelo, e o resultado aparece em paralelo no écran, ou numa tabela única.

Seguem-se alguns exemplos com a sua motivação linguística. Para exemplos práticos, ver a página de Ajuda do próprio Comparador[2].

2.1 A mesma pesquisa em dois corpos diferentes

Uma das questões mais básicas da linguística com corpos é a da importância do corpo para uma determinada conclusão: será que ter feito uma pesquisa num dado corpo é suficiente, ou se tivesse usado outro não poderia ter chegado à mesma conclusão?

Sendo o AC/DC um serviço que disponibiliza o aceso a todos os corpos cujos donos nos dêem autorização (sem impedir ou prejudicar que os mesmos corpos sejam interrogados e aprimorado noutros locais), o AC/DC é o local privilegiado para a comparação de corpos (ou melhor de pesquisas em corpos).

2.2 A mesma pesquisa em duas variantes diferentes

Outra das áreas que tem importância fulcral para a política linguística da língua portuguesa, é conhecer a diversidade da nbossa língua, e e em particular saber se as diferenças entre variantes são absolutas (categóricas) ou uma questão de grau e de variabilidade intrínsica a todas as línguas.

Para isso nada como poder comparar todas as análises em variantes diferentes.

2.3 A mesma pesquisa em dois géneros diferentes

Outra questão essencial na linguística com corpos é a de saber quais as generalizações a fazer para a língua toda, a língua geral, ou apenas para géneros específicos. Para isso é preciso que se possa recuperar o género dos dados a que recorremos, e que se possa pedir distribuição por género e tipo de texto, pese embora a grande dificuldade de comcordância na delimitação dos mesmos.

2.4 A mesma pesquisa com dois analisadores diferentes

É preciso também reconhecer que a análise subjacente à maior parte do material presente no AC/DC tem dois vieses: a teoria linguística implícita no PALAVRAS[1], e o desempenho (prático) do mesmo analisador.

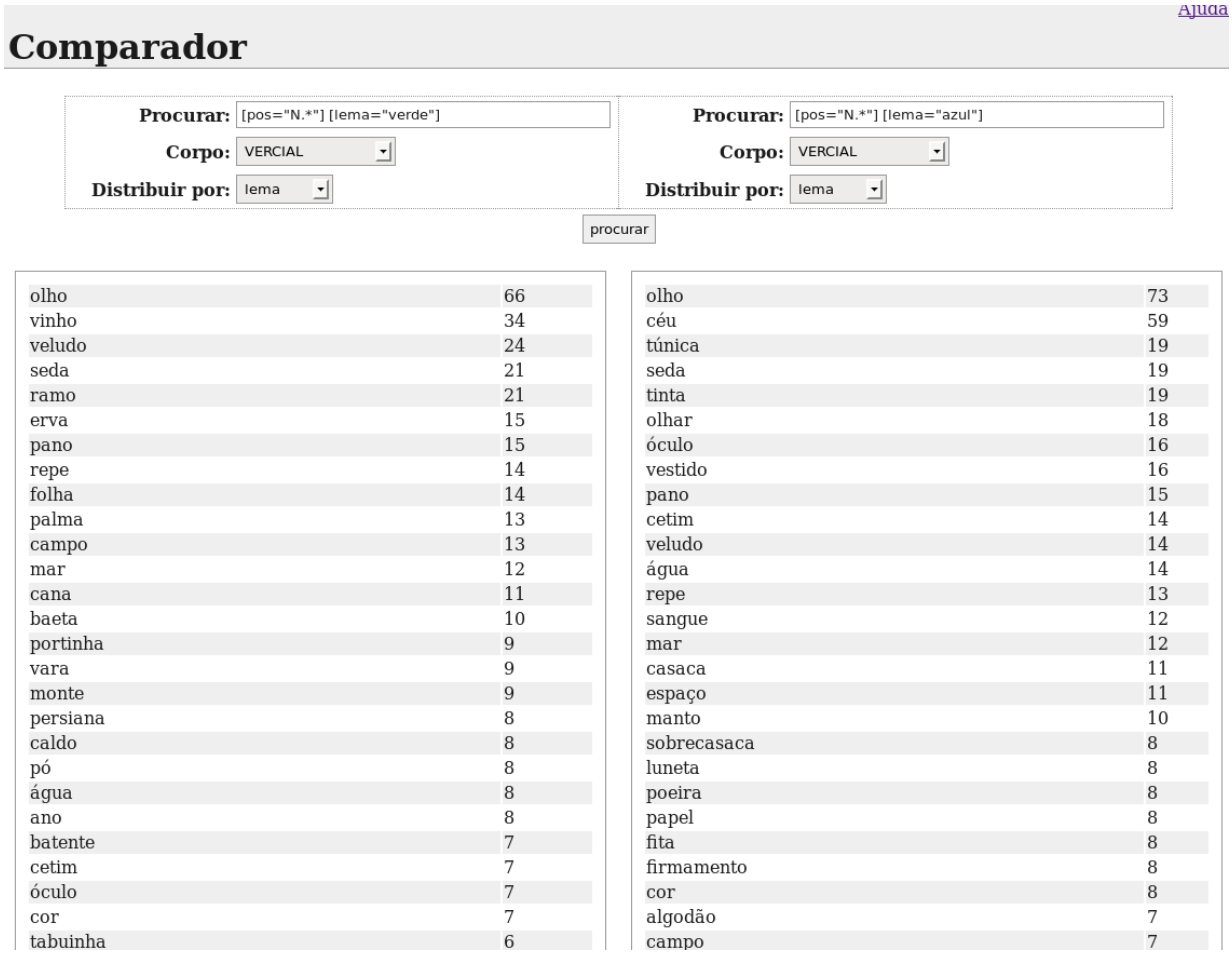
[2] `http://www.linguateca.pt/Comparador/`, escolha Ajuda

Embora estejamos gratos pela simples existência do PALAVRAS e pela possibilidade de desenvolvimento conjunto que muitos projetos com a Linguateca proporcionaram, nada obsta a que a totalidade do mesmo material seja também analisado por outros sistemas, e isso já acontece nos seguintes corpos: o FrasesPP e o AmostRA, e estamos de momento a trabalhar na mesma vertente com o ReLI e o COLONIA.

2.5 Comparando duas cores no mesmo corpo

Obviamente que a comparação não se esgota com a variação entre parâmetros externos, mas que até é provavelmente mais interessante do ponto de vista linguístico. Embora partes dessa comparação já se possam fazer através da interface padrão do AC/DC, é certamente interessante poder comparar em paralelo, por exemplo, diferentes distribuições.

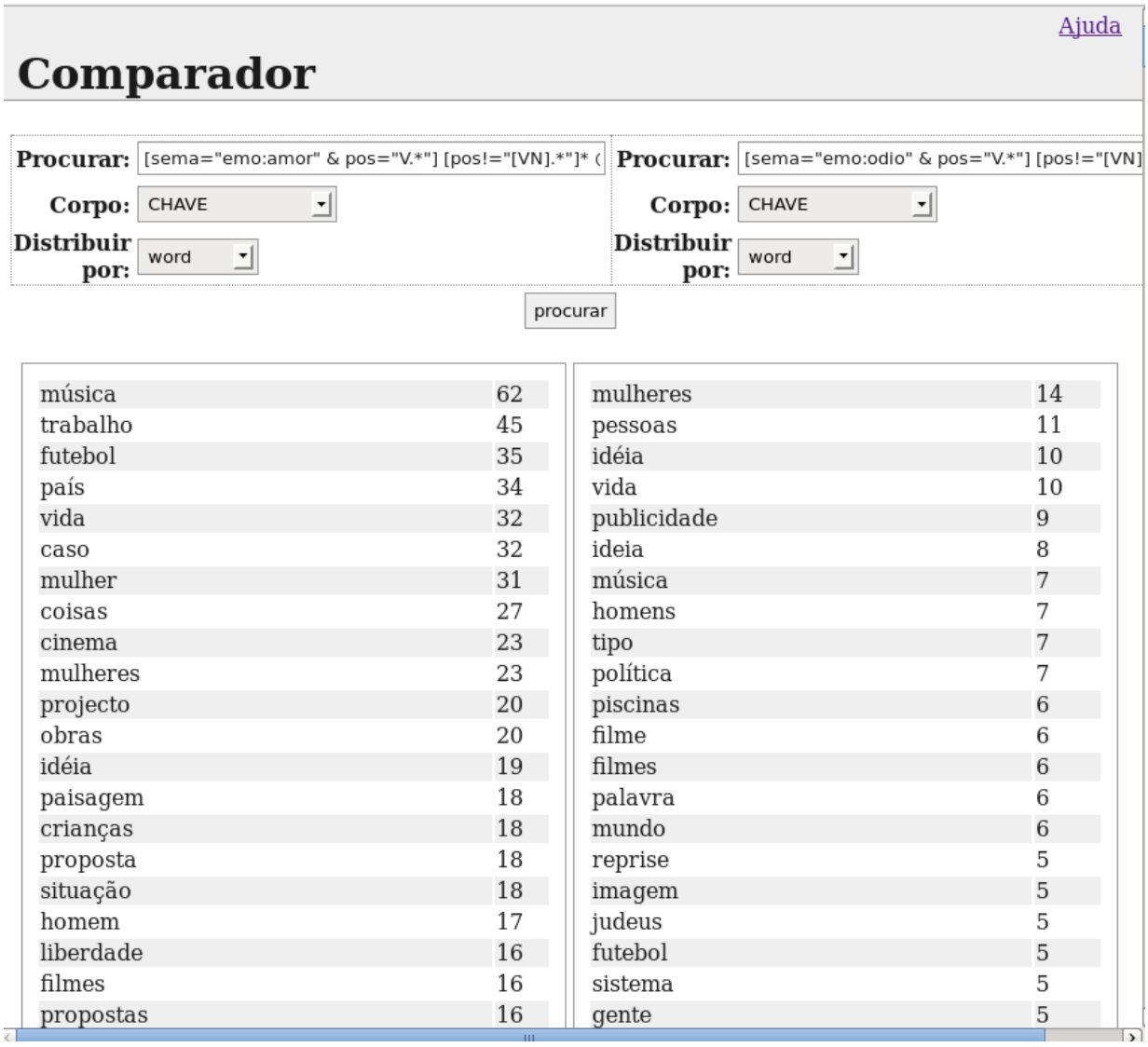
Quais as palavras mais correntes associadas com a cor verde, e com a cor azul? Uma forma muito simples (que tem de ser naturalmente refinada) é a descrita na figura seguinte:



2.6 Comparando duas emoções no mesmo género

Da mesma forma, pode ser interessante comparar quais os fenómenos que provocam diferentes emoções.

Na próxima figura temos uma simples comparação entre aquilo que inspira amor ou ódio em texto jornalístico.



música	62
trabalho	45
futebol	35
país	34
vida	32
caso	32
mulher	31
coisas	27
cinema	23
mulheres	23
projecto	20
obras	20
idéia	19
paisagem	18
crianças	18
proposta	18
situação	18
homem	17
liberdade	16
filmes	16
propostas	16

mulheres	14
pessoas	11
idéia	10
vida	10
publicidade	9
ideia	8
música	7
homens	7
tipo	7
política	7
piscinas	6
filme	6
filmes	6
palavra	6
mundo	6
reprise	5
imagem	5
judeus	5
futebol	5
sistema	5
gente	5

Agradecimentos A Linguateca agradece à Universidade de Oslo apoio financeiro para o desenvolvimento desta e outras ferramentas associadas à Gramateca.

Referências

[1] Eckhard Bick. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, Aarhus University, Aarhus, Denmark, November 2000.

[2] Diana Santos. Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties. In J.B. Johannessen, editor, *Language Variation Infrastructure: Papers on selected projects*, volume 3, pages 113–128, 2011.

[3] Diana Santos. Corpora at Linguateca: Vision and Roads Taken. In Tony Berber Sardinha and Telma São Bento Ferreira, editors, *Working with Portuguese corpora*, pages 219–236. Bloomsbury, 2014.

[4] Alberto Simões and Diana Santos. Ensinador: corpus-based Portuguese grammar exercises. *Procesamiento del Lenguaje Natural*, 47:301–309, Setembro 2011.