Nos bastidores da Gramateca: uma série de serviços

Alberto Simões e Diana Santos

http://www.linguateca.pt

Motivação

Demasiado texto anotado para podermos conhecê-lo sem ferramentas informáticas

- potenciar a procura e apresentar a repartição
- permitir comparações
- permitir análises alternativas
- visualizar sob diferentes ângulos

Temas de investigação

- Usabilidade
- Comparação de anotadores sintáticos
- Gramática baseada em corpos
- Estudo de diferenças entre teorias linguísticas

AC/DC

Acesso a corpos, disponibilização de corpos

- servindo 27 corpos diferentes, 1,5 mil milhões de palavras
- anotada com muita informação associada

Há 15 anos servindo as comunidades do processamento computacional da língua portuguesa e da linguística, graças ao PALAVRAS e ao OpenCWB (`http://cwb.sourceforge.net/`).

Serviços

Procura	`http://www.linguateca.pt/ACDC/`
Distribuidor	`http://www.linguateca.pt/Distribuidor/`
Comparador	`http://www.linguateca.pt/Comparador/`
Ordenador	`http://www.linguateca.pt/Ordenador/`
VARRA	`http://www.linguateca.pt/VARRA/`
Ensinador	`http://www.linguateca.pt/Ensinador/`
Ensinador paralelo	`http://dinis2.linguateca.pt/paraensinador/`
Rêve	`http://www.linguateca.pt/Reve/`
Repositório GitHub	`https://github.com/linguateca`

Exemplos da sua invocação

Distribuição mais fina

Linguateca
Ajuda

Distribuidor

Expressão de pesquisa: ?pos=/N/ gen pessnum

Corpo: ECI-EE **Tipo de resultado:** html

procurar

gen	Frequência Total	pessnum	Frequência Parcial	%
M	3354	P	1344	40,07%
		S	2010	59,93%
F	3246	S	2347	72,30%
		P	899	27,70%
M/F	24	S	10	41,67%
		P	14	58,33%

Linguateca
Ajuda

Distribuidor

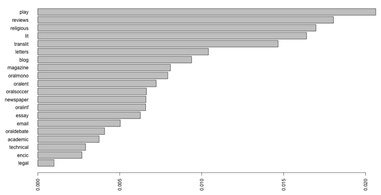
Expressão de pesquisa: ?sema=/cor:raça/ tema lema

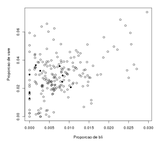
Corpo: CONDIV **Tipo de resultado:** html

procurar

tema	Frequência Total	lema	Frequência Parcial	%
saúde	59	cor	1	1,69%
		branco	34	57,63%
		preto	1	1,69%
		negro	22	37,29%
		branca	1	1,69%
futebol	9	branco	4	44,44%
		negro	4	44,44%
		preto	1	11,11%
moda	6	branco	3	50,00%
		preto	3	50,00%

Capacidades de visualização





Comparação de procuras

Linguateca
AC/DC
Ajuda

Comparador

Procurar:	[lema="verão" & variante="PT"]	Procurar:	[lema="verão" & variante="BR"]
Corpo:	CHAVE	Corpo:	CHAVE
Distribuir por:	semana	Distribuir por:	semana

☑ fundir numa única tabela

Limite mínimo de frequência: 0

procurar

30	260	41
33	238	23
32	231	24
31	213	19
35	185	45
27	197	26
34	194	28
28	191	26
29	192	22
26	149	20
38	106	62

Linguateca
AC/DC
Ajuda

Comparador

Procurar:	[sema=".*coragem.*"]	Procurar:	[sema=".*coragem.*"]
Corpo:	CHAVE	Corpo:	CONDIV
Distribuir por:	lema	Distribuir por:	lema

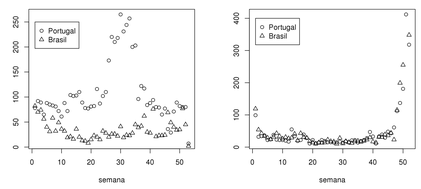
☑ fundir numa única tabela

Limite mínimo de frequência: 30

procurar

apoiar	12623	462
herói	3901	119
determinação	2928	232
coragem	2983	160
incentivo	2803	45
incentivar	1992	90
heroína	1997	6
ousar	1449	59
estímulo	1156	164
corajoso	843	33
encorajar	840	21
bravo	703	125
ousadia	707	53
incitar	610	34

Inspeção fina de correlações



Criação de material didático

Linguateca
Reiniciar Ensinador

Ensinador

A procurar "[word="quandoIse"]~ [pos!="V.º"]~ [temcagr=".ºSUBJ"].lema" no corpus C-Oral-Brasil v. 3.2 [121 entradas.]

Seleccione as concordâncias que deseja usar. OK

- ☐ Ah, é, uai / / 24 se n **morrer** antes deles. (morrer)
- ☐ Lá em casa, aviso se cê **for** lá. (ser)
- ☐ Se eu fizesse / / 280 se eu **fizesse** . (fazer)
- ☐ Eu, se eu **fosse** cê fazia isso não, Carlão / / 204 se eu fosse cê matava aqui o'S/ / 205 eu te avisei, Zé / / 206 p ' cê + CAR (ser)

Linguateca

Ensinador Paralelo

Eu não sei o que se passa , mas há coisa misteriosa que eu não ____ edvinhar .	☑	er . Far har holdt seg innestengt sammen med fetter Baltasar hele
Faleu-me em criados mortos ; mas eu não ____ entender ...	☐	jeg kunne ikke here helt hva hun sa til meg .
Prometia partir para Coimbra logo que o _____ fazer sem receio de Teresa sofrer na sua ausência .	☐	behandling . Han lovet å dra til Coimbra så snart han var viss på at Teresa ikke ville bli utsatt for noe ubehag i hans fravær .
Um apenas adiantava coisa que não ____ alumiar a justiça , e vinha a ser que o mato , nas vizinhanças do local , fora chapotado .	☑	et til , ikke etter . En av dem la til , uten at det hjalp de som underskrev saken . stærre , at småskogen i nærheten av funnstedet var hogget ned .
Nesta escuridade a justiça não ____ dar passo algum .	☐	Uten flere opplysninger kunne ikke myndighetene utrette noe som helst .
Tanto ao velho como ao morgado convinha apagar algum indício que _____ envolvê-los no mistério daquelas duas mortes .	☐	Det passet både Tadeu og Baltasar godt at enhver opplysning som kunne blande dem inn i drapsmysteriet ble undertrykt .

Reanotação de dados anotados

Linguateca

Rêve

Partes do corpo

Classifique a palavra segundo o seu sentido:

- referindo-se ao corpo humano (corpo);
- usando uma palavra do corpo para outros fins:
  - corpo:faculdade
  - corpo:outros
  - corpo:animal
  - corpo:partedeobjeto
  - corpo:sentimento
  - corpo:vegetal
  - corpo:grupo
  - corpo:doença
  - corpo:posição
  - corpo:emoção
  - corpo:lugar
  - corpo:movimento
- ou não se referindo de todo ao corpo humano: 0

Se a palavra fizer parte de uma expressão maior também relacionada com o corpo, indique essa conexão no campo de comentário, com o sufixo EVP, por exemplo `corpoEVP` ou `corpo:lugarEVP`.

Quando considerar que mais de uma classificação é possível, indique ambas no comentário ligadas por _.

E001-PT-332: E na aula recebem-me com uma salva de **palmas** .	corpo:outros
E001-PT-905: Eu andei com caixotes às **costas** , todos nós.	corpo
E003-PT-24: A ferramenta é esta que tenho na **mão** : a colmadeira, o lareiro e a estaca.	corpo
E103-PT-39: Eu enjoo muito e como fui de carreira, ia muito ' amorrinhada ', uma mulher que ia atrás de mim, num sítio qualquer bateu-me nas **costas** e disse-me: -- Saía aqui.	corpo
E105-PT-91: Os guardas passaram por mim e eu levava o alforges às **costas** e eles não me viram.	corpo
E133-BR-1060: Meu irmão já até levantou a **mão** para minha mãe, ela tem medo que ele faça alguma besteira.	corpo:outros *corpo:outrosEVP_co*

Referências

Luís Costa, Diana Santos e Paulo Alexandre Rocha. 2014. Estudando o português tal como é usado: o serviço AC/DC. In *The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)* (São Carlos, Brasil, 8-11 de Setembro 2009).

Eckhard Bick. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press.

Cláudia Freitas, Diana Santos, Hugo Gonçalo Oliveira & Violeta Quental. 2014. VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC. In Simone Sarmento, Tony Berber Sardinha, Livia Pretto Mottin & Ana Maria T. Ibaños. *Pesquisas e perspetivas em lingüística de corpus*, Mercado de Letras, pp. 199-232.

Diana Santos. 2014. Corpora at Linguateca: Vision and roads taken. In Tony Berber Sardinha & Telma de Lurdes São Bento Ferreira (eds.), *Working with Portuguese Corpora*, Bloomsbury, 2014, pp. 219-236.

Alberto Simões & Diana Santos. 2011. Ensinador: corpus-based Portuguese grammar exercises. In *Procesamiento del Lenguaje Natural*, 47, pp. 301–309.

Agradecimentoss

A Linguateca foi financiada pelo governo português e pela União Europeia através do MCTES, a UMIC, e a FCT de 1998 até 2010, e gerida e apoiada pela infraestrutura informática da FCCN desde 2000. O trabalho mais recente foi financiado pela Universidade de Oslo. Estamos também gratos ao Research Computing Group desta universidade.

FCCN Fundação para a Computação Científica Nacional · MCTES Ministério da Ciência, Tecnologia e Ensino Superior · FCT Fundação para a Ciência e a Tecnologia · UMIC