Estudando o português tal como é usado: o serviço AC/DC

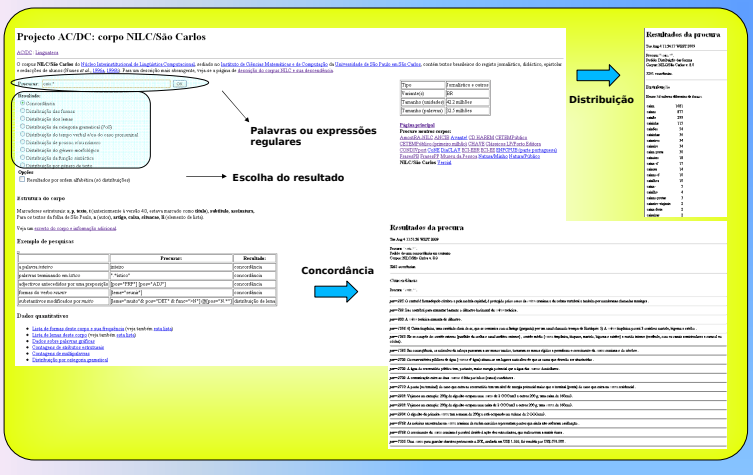
Luís Fernando Costa, Diana Santos e Paulo Rocha
{luis.costa, diana.santos}@sintef.no; paulo.rocha@di.uminho.pt
Linguateca, SINTEF ICT
PB 124, Blindern NO-0314 Oslo, Norway
http://www.linguateca.pt

Acesso a Corpos / Disponibilização de Corpos: http://www.linguateca.pt/ACDC/

Panorâmica dos corpos

Tipo de texto/Variante	Portuguesa	Brasileira	Mista	Nº de unidades
Jornalístico	Avante!, CETEMPúblico, DiaCLAV, Natura/Minho, Natura/Público		CHAVE, CONDIVport	384.593.915
Literário	Clássicos LP/Porto Editora, Vercial		ENPCPUB (parte portuguesa)	10.933.121
Correio electrónico		ANCIB	CoNE	2.531.587
Institucional	ECI-EE			31.863
Transcrição de texto oral			Museu da Pessoa	513.468
Misto	FrasesPP	NILC/São Carlos, AmostRA-NILC, ECI-EBR, FrasesPB	CD HAREM	43.459.055
Nº de unidades	264.904.779	44.825.938	132.332.292	442.063.009

Interface



Construção dos corpos

Corpo normalizado e marcado com divisões estruturais (no mínimo parágrafos e frases)

```
<p>
<s>
Dentre as FEIRAS DE SERVIÇOS (entretenimento puro e simples ), já contamos com as: BADALAÇÃO, CARTÕES
VIRTUAIS, CHAT, CINEMA, GIFS, HUMOR, JOGOS, SHOWS e TEATROS.
</s>
<s>
Neste momento estamos dando início a novas feiras de serviço, entretenimento e informação.
</s>
</p>
```

```
<s>
Em      [em] <sam-> PRP @ADVL>
este    [este] <dem> <-sam> DET M S @>N
momento  [momento] N M S @P<
estamos  [estar] <fmc> V PR 1P IND VFIN @FAUX
dando    [dar] V GER @IMV @#ICL-AUX<
início   [início] N M S @<ADVL
a        [a] PRP @<ADVL @<PIV
novas    [novo] ADJ F P @>N
feiras   [feira] N F P @P<
de       [de] PRP @N<
serviço  [serviço] N M S @P<
$,
entretenimento   [entretenimento] N M S @P<
e        [e] <co-prparg> KC @CO
informação       [informação] N F S @P<
$.
</s>
```

Corpo anotado pelo PALAVRAS (http://beta.visl.sdu.dk/visl/pt/)

http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

http://cwb.sourceforge.net/

CWB

Formato AC/DC simples (forma, lema, PoS, tempo/caso, número/pessoa, género, inf. sintáctica)

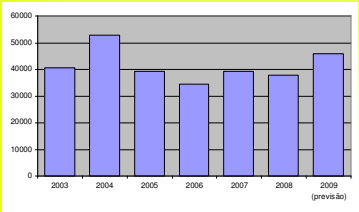
```
<s>
Neste       em+este     PRP+DET_dem  0   S   M   ADVL>+>N
momento     momento     N            0   S   M   P<
estamos     estar       V_fmc        PR_IND  1P  0   FAUX
dando       dar         V            GER 3   0   IMV_#ICL-AUX<
início      início      N            0   S   M   <ADVL
a           a           PRP          0   0   0   <ADVL<PIV
novas       novo        ADJ          0   P   F   >N
feiras      feira       N            0   P   F   P<
de          de          PRP          0   0   0   N<
serviço     serviço     N            0   S   M   P<
,           ,           PU           0   0   0   PONT
entretenimento  entretenimento  N        0   S   M   P<
e           e           KC_co-prparg  0   0   0   CO
informação  informação  N            0   S   F   P<
.           .           PU           0   0   0   PONT
</s>
```

Formato AC/DC enriquecido com mais informação estrutural e semântica:

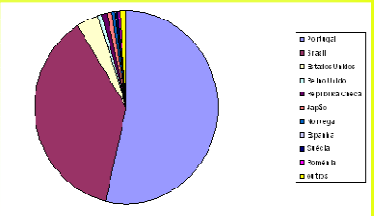
- variante
- campo semântico
- grupo da cor
- secção do jornal
- tema
- data de publicação
- autor
- obra
- etc.

Quem utiliza?

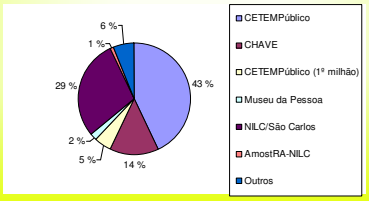
Evolução do número de visitas



Origem das consultas



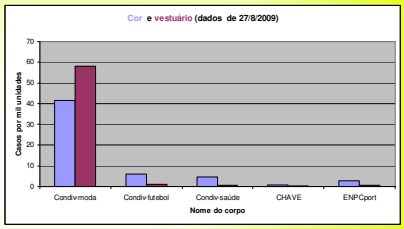
Pesquisas por corpo



Trabalho em progresso

Nova anotação de todos os corpos
Com uma versao mais recente do PALAVRAS, todos os corpos foram reanotados em 2008 ou 2009

Marcação da cor e de outros campos semânticos
1519 cores (lemas) detectadas até agora nos corpos do AC/DC
247 elementos de vestuário (lemas)
Entidades mencionadas



Melhoria da interface
Uso de procuras anteriores para sugerir reformulação
Distribuição cruzada com mais de um campo
Repositório colaborativo de procuras com descrição

Melhoria da anotação de forma cooperativa
Primeiros passos na melhoria de anotação de forma cooperativa
Escrita de regras condicionais
Sistema de revisão através da rede

UMIC Agência para a Sociedade do Conhecimento
FCCN Fundação para a Computação Científica Nacional
POS_Conhecimento Programa Operacional Sociedade do Conhecimento
União Europeia FEDER/FSE