# A arte dos corpora em português: O projecto AC/DC

*Paulo Rocha*
*Linguateca, Pólo de Coimbra*
http://www.linguateca.pt/Coimbra/



## Breve descrição do projecto AC/DC

- **A**cesso a **C**orpora / **D**isponibilização de **C**orpora
- Utiliza o IMS Corpus Query Processor (CQP)
- Cerca de 20 corpora
- Aprox. 346 milhões de palavras
- Aprox. 15 milhões de frases
- Variantes portuguesa e brasileira
- Jornalístico, literário, texto didáctico, entrevistas, listas electrónicas …



## Breve resenha histórica

- 1998: iniciado o projecto
- 1999: primeiros corpora acessíveis na Rede
- 2000: primeiros corpora anotados na Rede
- 2001: disponibilizado o CETEMPúblico na Rede
- 2002: idem para o CETENFolha
- 2003: todos os corpora anotados
- 2007: nova interface



## Acedendo ao projecto

http://www.linguateca.pt/ACDC/




## Algumas simples pesquisas

- Pesquisar uma palavra
  - DiaCLAV: "palavra"
- Pesquisar formas
  - DiaCLAV: "palavr.+"
- Pesquisar por período de tempo
  - Condivport: "Eusébio"
- Pesquisar por outros campos
  - variante PT/BR: Condivport: "Académica"
  - fonte DiaCLAV: "Académica"



## Galeria dos corpora (I)

- Jornalístico generalista
  - CETEMPúblico
  - CETENFolha (→ São Carlos)
  - NatPúblico
- Jornais regionais
  - NatMinho
  - DiaCLAV
- Jornalístico especifico
  - Desportivo : CONDIVport
  - Político: Avante!
- Literário
  - Vercial
  - ClassLPPE
  - ENPCPub



# Galeria dos corpora (II)

- Entrevistas (texto oral transcrito)
  - Museu da Pessoa

- Mensagens electrónicas
  - Listas: ANCIB
  - SPAM: CoNE

- Recursos de avaliação
  - CDHAREM
  - CHAVE
- Outros
  - ECI-EBR
  - ECI-EE
  - FrasesPP / FrasesPB

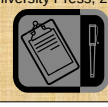


# Anotando os corpora

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

```
Cada corpus é anotado sintacticamente.
```

PALAVRAS



```
$START
Cada        [cada] <quant> DET M S @>N
corpus      [corpus] N M S @SUBJ>
é           [ser] <fmc> V PR 3S IND VFIN @FAUX
anotado     [anotar] V PCP M S @IMV @#ICL-AUX<
sintacticamente ALT sintaticamente      [sintático] ADV @<ADVL
$.
```

Formato AC/DC

```
Cada            cada        DET_quant   0       S   M       >N              0
corpus          corpus      N           0       S   M       SUBJ>           0
é               ser         V_fmc       PR_IND  3S  0       FAUX            0
anotado         anotar      V           PCP     S   M       IMV_#ICL-AUX<   0
sintacticamente sintático   ADV         0       0   0       0               0
.               .           PONT        0       0   0       0               0
```



# Pesquisando os corpora anotados

- Pesquisar substantivos
  - `[pos="N"]`
- Pesquisar nomes próprios
  - `[pos="PROP"]`
- Pesquisar um lema
  - `[lema="procurar"]`

- Verificar uso de preposições com verbos
  - (distr. por lema) `[lema="procurar"] @[pos="PRP"]`
- Ver a função de uma palavra
  - (distribuição por função) `"eu"`



# Lista de frequências

- Formas (palavras / unidades)
- Lemas
- Palavras x PoS
- Lemas x PoS

```
424835 ,
299804 de
212139 a
189739 e
178499 .
165586 que
138619 o
124526 do
121939 da
 67088 em
 62090 os
 60718 para
…
```

```
487095 ,
413800 o
264130 de
260359 de+o
193380 .
163301 que
152173 e
115547 ser
110276 em+o
102561 um
 82407 a+o
…
```

```
19630 ano
10700 dia
 9372 presidente
 7240 pessoa
 7119 equipa
 6515 hora
 6044 jogo
 5943 parte
 5867 vez
 5527 cidade
 5513 projecto
…
```



# Limitações do AC/DC

- Não existe desambiguação a nível de palavra
  - manga vs. manga
  - banda vs. banda vs. banda vs …

- Não permite identificar sintagmas
- Não permite pesquisar árvores



# Limitações do AC/DC

Para superar algumas dessas faltas,…

… existe a FLORESTA SINTÁ(C)TICA

# Publicações


Diana Santos. "Disponibilização de *corpora* de texto através da WWW". In Palmira Marrafa & Maria Antónia Mota (eds.), *Linguística Computacional: Investigação Fundamental e Aplicações. Actas do I Workshop sobre Linguística Computacional da Associação Portuguesa de Linguística* (Lisboa, 25-27 de Maio de 1998), Lisboa: Colibri, pp. 323-335.

Diana Santos & Elisabete Ranchhod. "Ambientes de processamento de corpora em português: Comparação entre dois sistemas". In Irene Rodrigues & Paulo Quaresma (eds.), *Actas do IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR'99)* (Evora, 20-21 de Setembro de 1999), pp. 257-268.

Diana Santos & Eckhard Bick. "Providing Internet access to Portuguese corpora: the AC/DC project". In Maria Gavrilidou, George Carayannis, Stella Markantonatou, Stelios Piperidis & Gregory Stainhauer (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)* (Athens, 31 May-2 June 2000), pp. 205-210.

Susana Afonso. "Clara e sucintamente: um estudo em corpus sobre a coordenação de advérbios em -mente". In Amália Mendes & Tiago Freitas (orgs.), *Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística (APL 2002)* (Porto, 2-4 de Outubro de 2002), Lisboa: APL, pp. 27-36.

Diana Santos & Luís Sarmento. "O projecto AC/DC: acesso a corpora/disponibilização de corpora". In Amália Mendes & Tiago Freitas (orgs.), *Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística (APL 2002)* (Porto, 2-4 de Outubro de 2002), Lisboa: APL, pp. 705-717.

Diana Santos. "Aonde vamos em relação a *aonde*". *the ESPecialist* **25.1** (2004), pp. 85-103. São Paulo.

Paulo Rocha & Diana Santos. "Disponibilizando a <OBRA>Colecção Douradа</OBRA> do <ACONTECIMENTO>HAREM</ACONTECIMENTO> através do projecto <LOCAL|ORGANIZACAO|ABSTRACCAO>AC/DC</LOCAL|ORGANIZACAO|ABSTRACCAO>". In Diana Santos & Nuno Cardoso (eds.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca:, 2007.

Santos, Diana. "Breves explorações num mar de língua", *Ilha do Desterro* **52** (2007). No prelo.




# PERGUNTAS