

Corpos linguísticos da Linguateca: apresentação

Diana Santos

Linguateca
www.linguateca.pt

Breve história dos corpos na Linguateca

- 1. tornar acessível (na rede) o que já existia
- 2. criar e/ou melhorar (analisando) esses corpos
- 3. adicionar a dimensão da tradução → COMPARA
- 4. adicionar a dimensão da revisão humana → Floresta Sintá(c)tica
- 5. adicionar a possibilidade de criar corpos próprios
- 6. adicionar a dimensão de corpos comparáveis

Projecto AC/DC

Corpógrafo

Linguateca, um centro de recursos distribuído

- Projecto gerido pela FCCN, financiado pelo POSI

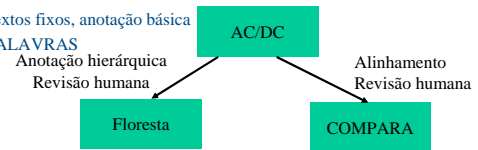
Modelo IRA

- Informação
 - Recursos
 - Avaliação
- www.linguateca.pt



Semelhanças e diferenças

- Textos fixos, anotação básica do PALAVRAS
- Textos da responsabilidade e escolha dos participantes



Breve história (cronologia e liderança)

- 1998 início do projecto AC/DC (Diana Santos)
- 1999 início do projecto COMPARA/DISPARA (Diana Santos e Ana Frankenberg-Garcia)
- 1999 início do projecto Floresta Sintá(c)tica (Diana Santos e Eckhard Bick)
- 2002 início do projecto Corpógrafo (Belinda Maia)
- 2004 início da anotação sintáctica do COMPARA (português)
- 2007 início da anotação sintáctica do COMPARA (inglês)

Uso de corpos no ensino da língua portuguesa

- Para motivar/inspirar o professor
o professor como perito, sempre a aprender
- Para criar materiais de ensino
- Para criar materiais de treino
- Para criar materiais de teste
o professor como profissional, com ferramentas melhores
- Para ajudar o próprio aluno a aprender
- Para motivar/inspirar os alunos
o professor como pedagogo, inovador e facilitador

Projecto AC/DC

- Início em 1998-1999
- Uso do PALAVRAS (Bick, 2000) a partir de 1999
- Criação de corpos também puramente distribuíveis
 - CETEMPúblico
 - CETENFolha
 - CHAVE
- Listas de frequências
- Uso para a criação de recursos de avaliação
- Disponibilização de recursos criados no âmbito da avaliação

Galeria dos corpora

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



Rocha (2007)

Galeria dos corpora

- Entrevistas (texto oral transcrito)
 - Museu da Pessoa
- Mensagens de correio electrónico
 - Listas: ANCIB
 - SPAM: CoNE
- Recursos de avaliação
 - CDHAREM
- CETEMPúblico (primeiro milhão)



Rocha (2007)

The screenshot shows a web interface for the Projecto AC/DC corpus. It includes a search bar, a list of filters (e.g., Ano, Tipo, Destino), and a table of search results. The interface is in Portuguese and provides detailed options for filtering the corpus data.

Breve descrição do projecto AC/DC

- Acesso a Corpora / Disponibilização de Corpora
- 20 corpora em português
- Cerca de 346 milhões de palavras
- Aprox. 15 milhões de frases
- Variantes portuguesa e brasileira
- Jornalístico, literário, texto didáctico, entrevistas, listas electrónicas ...
- Interface em Perl ao ambiente de corpos IMS-CWB
- Uso do analisador sintáctico automático PALAVRAS para a anotação gramatical

Anotando os corpos

Cada corpus é anotado sintacticamente.

PALAVRAS

```

SETAAT
Cada [cada] <quant> DET M S @>N
corpus [corpus] M M S @>N
d [ser] <func> V PR 3S IND VFIN @PAUX
anotado [anotar] V PCP M S @IMV @ICL-AUX<
sintacticamente ALT sintaticamente [sintático] ADV @>ADV
S
    
```

Formato AC/DC

```

cada      cada      DET_quant      S      M      S      >N      0
corpus    corpus    M              S      M      S      @>N      0
d         ser        V_func         PR_3ND 3S 0    IND VFIN @PAUX 0
anotado   anotar    V              PCP    M      S      @IMV @ICL-AUX< 0
sintacticamente  sintaticamente  ADV          S              0
    
```

Rocha (2007)

Assuntos

- O sentido de uma palavra
 - ambiguidade
 - frequência
 - confusabilidade
- Ordem numa frase
 - “colocações”
 - sufixos
 - peso
- Conotações
- Formas de tratamento
- Material didáctico (escolha múltipla, reescrita, correcção)

Perceber o sentido de uma palavra

- Em contexto
- Através do registo em que é usada
- Através das palavras com que co-ocorre
- Através das palavras semelhantes com que pode ser contrastada
- Através da tradução

Simple selecção de exemplos

- Palavras difíceis que é preciso explicar aos alunos
preterir, premonição, intervindo
- Diferenças subtis no sentido
grade vs. gradeamento; argumento vs. guião
- Colocações: como é que se usam estes adjectivos
claro, engraçado
- Comparações
mais ADJ do que importante, forte, rápido, grave, baixo

Temas mais avançados de gramática

- O uso de *cujo* e que relações são mais frequentes
- *seus* vs. *deles*
- Material metafórico: o que é conceptualizado como uma luta?
(renhido)
- *Cedo e tarde* (725/2366 Vercial; 10925/58080 CETEMPúblico)
- Organização textual e lexical
 - Descrições de acidentes (de automóveis)
 - Recensões de concertos
 - Caracterização de pessoas

Qual a tradução correcta de *skuffelse*?

- *skuffelse*: decepção, desapontamento, desencantamento, desencanto, desgosto, desilusão, desengano, engano, frustração

Procura:

"decepção|desapontamento|desencantamento|desencanto|desgosto|desilusão|desengano|engano|frustração".

Pedido: Distribuição das formas

desilusão 1783 frustração 1513 engano 1288 decepção 1017
desencanto 743 desgosto 715 desapontamento 354 desencantamento
30 desengano 14

Santos (2004)

Dupla negação e polaridade negativa

- *Não desgosto*
- *Não me importo*
- *Não acho* (no sentido “acho que não”)

Procura: [lema="desgostar"]; Pedido de uma concordância em contexto; Corpus: CETEMPúblico v1.7 172 ocorrências.

Já esta semana, no começo da sua digressão asiática, Christopher travou com as autoridades chinesas uma guerra de palavras à distância, ao afirmar-se «profundamente **desgostado**» com a perseguição a dissidentes, ao que Pequim respondeu que ele se estava a «intrrometer irresponsavelmente» numa questão doméstica .

De resto, com fundas certezas, ambos os realizadores não **desgostariam** de se ver associados . No estado actual das investigações do mistério Rossini, o diagnóstico parece ser um complexo de «dolce farniente» e **desgosto** das diabruras («diavolerie») do romantismo rompante .

Esta é a coisa que mais me **desgosta**, em termos globais .

Eu não **desgosto** delas, o Taveira é um homem com grande talento, vê-se que há um impulso de arquitecto nele. ”

Aliás, como bom liberal, não **desgosto** de ficar relativamente sozinho .

Confesso que não **desgostei** .

Procura: [lema="desgostar"]; Pedido de uma concordância em contexto; Corpus: CETEMPúblico v1.7 172 ocorrências. **EDITADO**

De resto, com fundas certezas, ambos os realizadores não desgostariam de se ver associados .

Eu não desgosto delas, o Taveira é um homem com grande talento, vê-se que há um impulso de arquitecto nele. ”

Aliás, como bom liberal, não desgosto de ficar relativamente sozinho .

Confesso que não desgostei .

Não acho ...

- Mas eu **não acho** que Senna, hoje, seja espectacular .
Tratam-me bem e **não acho** que haja nisto algum cinismo .
«O facto de ter sido escolhido na Taça Davis não quer dizer nada; **não acho** que haja uma grande diferença entre nós», afirmou o campeão nacional .
R. -- Eu **não acho** que haja nenhuma relação directa .
Eu **não acho** que o budismo seja uma religião na mesma acepção que as outras .
R. -- Não, **não acho** que Cavaco Silva seja mais ou menos democrático do que qualquer outro primeiro ministro que estivesse na posição dele .
A Confederação dispõe agora de três batalhões, mas eu, como Presidente, **não acho** necessário enviar para a zona a sua totalidade .
Portanto, **não acho** exagerado esperar fazer a ratificação rapidamente .

- 260 no CETEMPúblico; em 5609 "acho"; em 28941 "achar"
- Distribuição de **pesnum**: 1S 11238; 3S 10229; 3P 3642

Exemplo de criação de material didáctico

- *Foi ou era? Estava ou esteve?*
- Obter bons exemplos de cada caso, retirar, e pedir aos alunos para preencher
- Enquanto ____ disponível em carácter experimental, a Biblioteca Virtual Carlos Chagas chegou a receber mensagens de portadores da doença, que tiveram suas dúvidas esclarecidas por especialistas
- Este enunciado que só Jesus é a verdadeira alegria, é um enunciado que sempre ____ presente, sempre-já lá, como diz Pêcheux, na memória discursiva dos fiéis enquanto dogma religioso
- No ano em que, segundo Balzac, o poeta Dante Alighieri ____ em Paris, ele poderia ter lido todos os 1.338 volumes da biblioteca da universidade (que era, então, a maior da França)

Exemplo de mutilação: reescreva...

- [pos="PROP"] ", " "que".
- Junte de forma harmoniosa as seguintes partes de informação
- 1) A questão que envolve a população e Alfredo da Cruz tem origem na indefinição da propriedade das Capelas do Calvário.
- 2) Alfredo da Cruz é um empresário que comprou essas capelas.
- A questão que envolve a população e Alfredo da Cruz tem origem na indefinição da propriedade das Capelas do **Calvário, que** alegadamente o empresário comprou, conjuntamente com uma quinta que confina com o adro dos templos
- Ausente esteve **Dário, que** ainda não chegou de Moçambique, onde esteve ao serviço da selecção

Exemplo de correcção (?)

Pergunta: O que é que está mal nesta(s) frase(s)?

- É assim que José António Silva encara os resultados das eleições para a comissão política concelhia do PSD, que **correram** (decorreram) no último domingo, em Leiria
- É assim que José António Silva encara os resultados das eleições **contra** (para) a comissão política concelhia do PSD, que decorreram no último domingo, em Leiria
- É assim que José António Silva encara **nos** (os) resultados das eleições para a comissão política concelhia do PSD, que decorreram no último domingo, em Leiria

Procuras mais complicadas

- Procura: "de" a:[pos="N,*"] "em" @[word=a.word] within s;
Distribuição de lema
MP CHAVE
- porta 6 de boca em boca
- terra 2 de mão em mão
- colégio 1 de porta em porta
- barraca 1 de geração em geração
- porto 1 de vitória em vitória
- casa 1 de repartição em repartição ...

Expressões mais ou menos idiomáticas

- If *set in train* always occurs together in this sequence when it has the obvious meaning, then the three words constitute **one** choice. As soon as learners have appreciated that each phrase operates as a whole, more or less as a single word, (...) they have a new word *set in train*. Not many learners will confuse *set* and *say* just because they begin with s; learners are not expecting s to have meaning on its own. (Sinclair, 1991: 78)
- O problema dos espaços ou do que constitui uma palavra ou unidade lexical não é óbvio: se para a flexão é importante e necessário separar *dar uma cambalhota*, para o sentido é importante e necessário juntar

O que é que se dá e o que é que se tira?

- [lema="dar"] []* @[func="<ACC"]
- [lema="tirar"] []* @[func="<ACC"]

A dimensão moral

- envergonhado
- enganado
- apaixonado
- engraçado

Vagos quanto a de propósito ou não

Vagos quanto a uma característica ou um sentimento

Forma, maneira, modo e caminho

DiaCLAV: Houve 4 valores diferentes de **forma**.

forma	4741	4684	de	735	a	724	para	21
modo	1067	1067	de	69	entre	4	para	3
caminho	728	726	de	232	para	68	a	30
maneira	573	557	de	129	a	31	para	6

- [word="forma" & pos="N,*"] @[pos="PRP,*"]