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

Projecto AC/DC

- 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

Corpógrafo

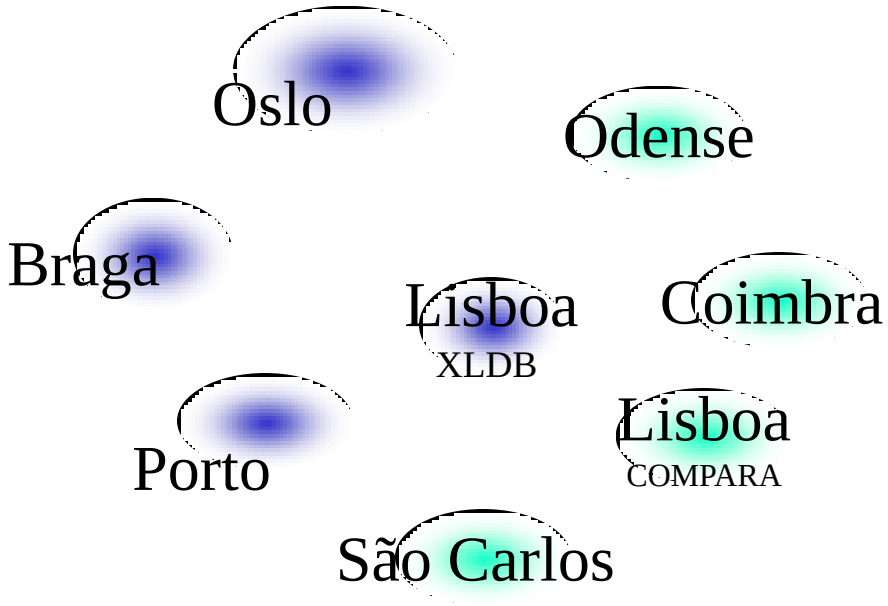
Linguateca, um centro de recursos distribuído

- Projecto gerido pela FCCN, financiado pelo POSI

Modelo IRA

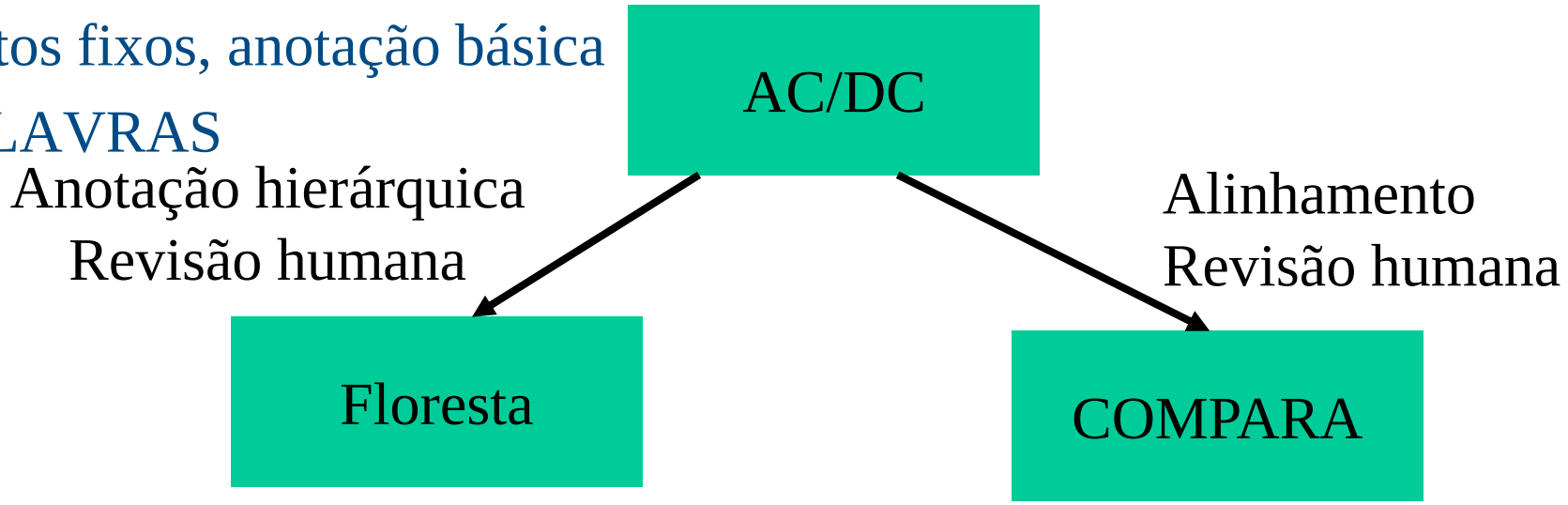
- **I**nformação
- **R**ecursos
- **A**valiação

www.linguateca.pt



Semelhanças e diferenças

- Textos fixos, anotação básica do PALAVRAS



- Textos da responsabilidade e escolha dos participantes

Corpógrafo

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

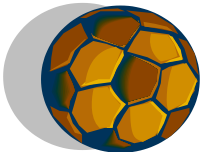
- Início em 1998-1999
- Uso do PALAVRAS (Bick, 2000) a partir de 1999
- Criação de corpos também puramente distribuíveis
  - CETEMPúblico
  - CETENFolha
  - CHAVE
- Listas de frequências
- Uso para a criação de recursos de avaliação
- Disponibilização de recursos criados no âmbito da avaliação

Galeria dos corpora

- Jornalístico generalista
  - CETEMPúblico
  - CETENFolha (→ São Carlos)
  - NatPúblico

- Jornais regionais
  - NatMinho
  - DiaCLAV

- Jornalístico especifico
  - Desportivo : CONDIVport
  - Político: Avante!



- Literário
  - Vercial
  - ClassLPPE
  - ENPCPub

Rocha (2007)

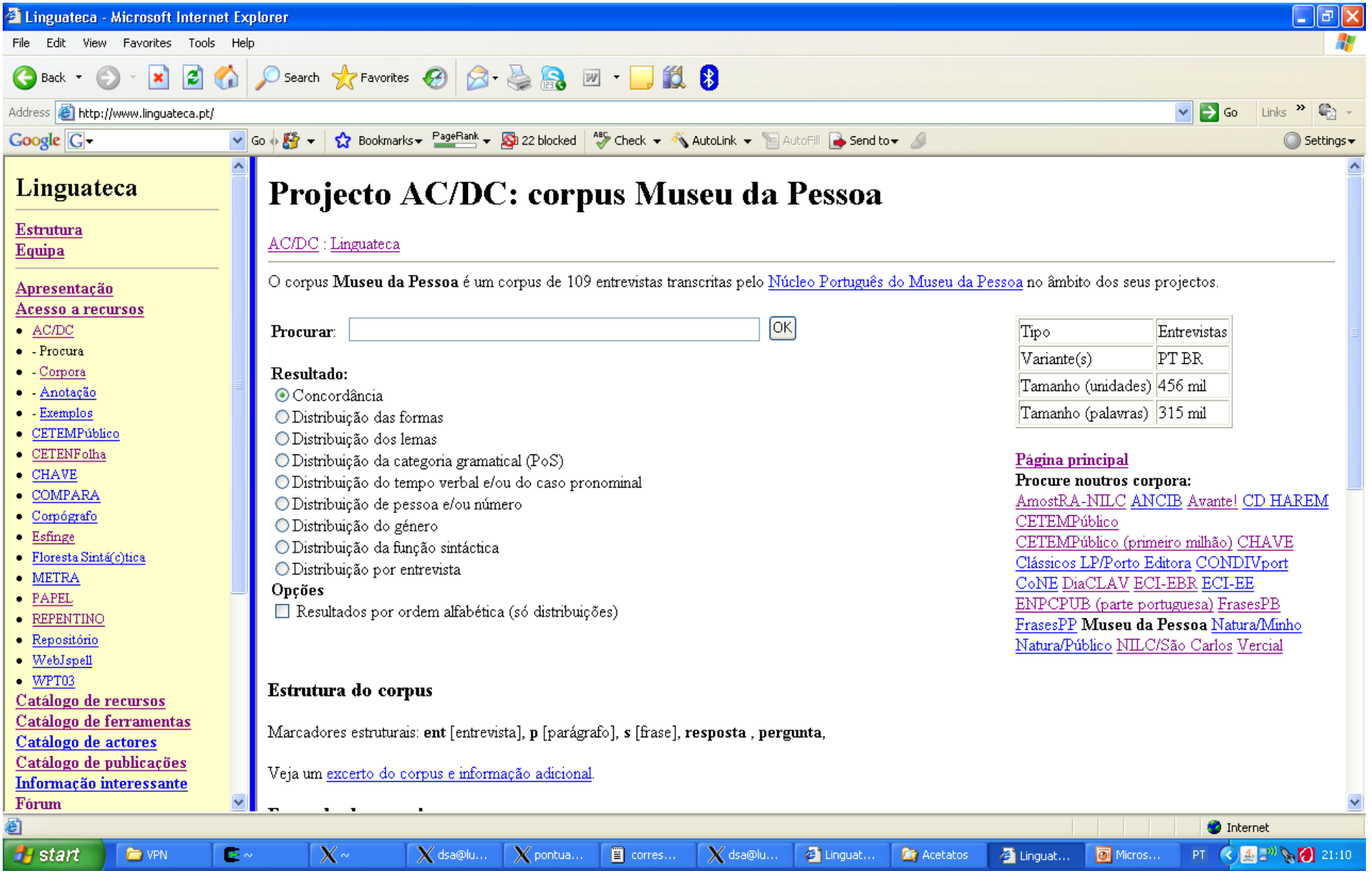
Galeria dos corpora

- Entrevistas
  (texto oral transcrito)
  - Museu da Pessoa

- Mensagens de correio electrónico
  - Listas: ANCIB
  - SPAM: CoNE

- Recursos de avaliação
  - CDHAREM
- CETEMPúblico
  (primeiro milhão)

Rocha (2007)


Linguateca - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.linguateca.pt/
Go Links
Google Go Bookmarks PageRank 22 blocked Check AutoLink AutoFill Send to Settings
Linguateca
Estrutura
Equipa
Apresentação
Acesso a recursos
AC/DC
- Procura
- Corpora
- Anotação
- Exemplos
CETEMPúblico
CETENFolha
CHAVE
COMPARA
Corpógrafo
Esfinge
Floresta Sintá(c)tica
METRA
PAPEL
REPENTINO
Repositório
WebJspell
WPT03
Catálogo de recursos
Catálogo de ferramentas
Catálogo de actores
Catálogo de publicações
Informação interessante
Fórum
Projecto AC/DC: corpus Museu da Pessoa
AC/DC : Linguateca
O corpus Museu da Pessoa é um corpus de 109 entrevistas transcritas pelo Núcleo Português do Museu da Pessoa no âmbito dos seus projectos.
Procurar: OK
Resultado:
Concordância
Distribuição das formas
Distribuição dos lemas
Distribuição da categoria gramatical (PoS)
Distribuição do tempo verbal e/ou do caso pronominal
Distribuição de pessoa e/ou número
Distribuição do género
Distribuição da função sintáctica
Distribuição por entrevista
Opções
Resultados por ordem alfabética (só distribuições)
Tipo Entrevistas
Variante(s) PT BR
Tamanho (unidades) 456 mil
Tamanho (palavras) 315 mil
Página principal
Procure noutros corpora:
AmostRA-NILC ANCIB Avante! CD HAREM
CETEMPúblico
CETEMPúblico (primeiro milhão) CHAVE
Clássicos LP/Porto Editora CONDIVport
CoNE DiaCLAV ECI-EBR ECI-EE
ENPCPUB (parte portuguesa) FrasesPB
FrasesPP Museu da Pessoa Natura/Minho
Natura/Público NILC/São Carlos Vercial
Estrutura do corpus
Marcadores estruturais: ent [entrevista], p [parágrafo], s [frase], resposta , pergunta,
Veja um excerto do corpus e informação adicional.
Internet
start VPN dsa@lu... pontua... corres... dsa@lu... Linguat... Acetatos Linguat... Micros... PT 21:10

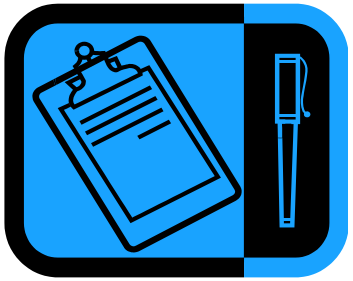
Breve descrição do projecto AC/DC

- **A**cesso a **C**orpora / **D**isponibilização de **C**orpora
- 20 corpos 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

```
$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

Rocha (2007)

Trabalho com corpos linguísticos: Vantagens e desvantagens

- Imersão controlada
- Fonte incansável de diálogo
- Texto real em vez de exemplos artificiais

- Interferência de muitos outros fenómenos
- Nada de ”texto limpo”
- Demasiado e no entanto insuficiente

Santos (2004)

Problemas típicos com utilizadores avançados do AC/DC

- As categorias com que trabalham não são as que estão marcadas (ou que são subjacentes à anotação)

  não há (e se calhar não pode haver) uma terminologia gramatical 100% consensual -- cada linguista / gramático faz as distinções que lhe parecem pertinentes e de certa forma divide a língua de forma diferente.

- Há erros na anotação ou no próprio texto do corpo

  É humanamente impossível rever 350 milhões de palavras ☹

- O género textual do corpo com que estão a trabalhar não é apropriado para tirar conclusões

Exemplos de pedidos/comentários

- *preciso de dados em que as gerundivas ocorrem como orações pequenas do tipo:*

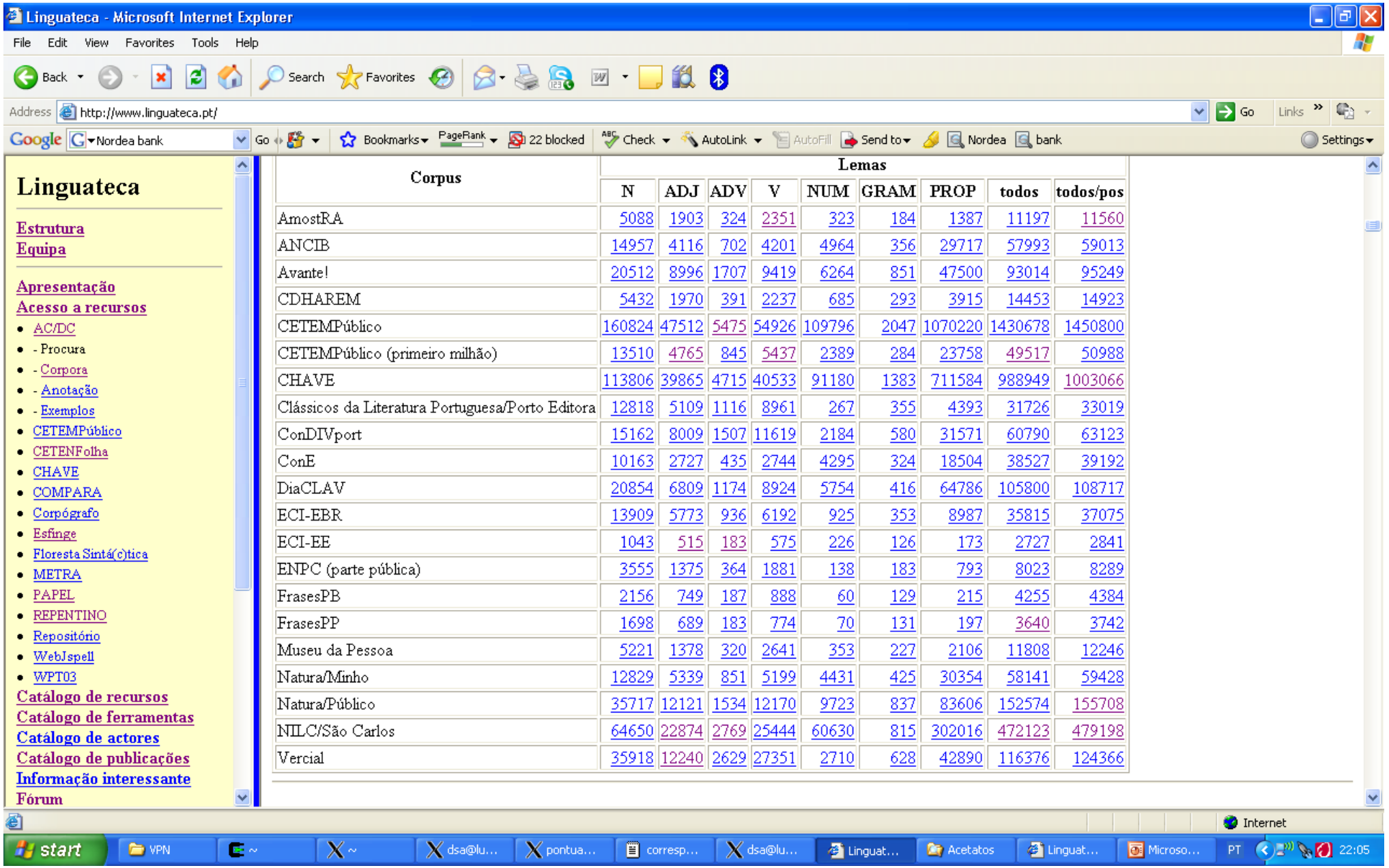
Ela saiu cantando

O navio afundou matando 100 pessoas

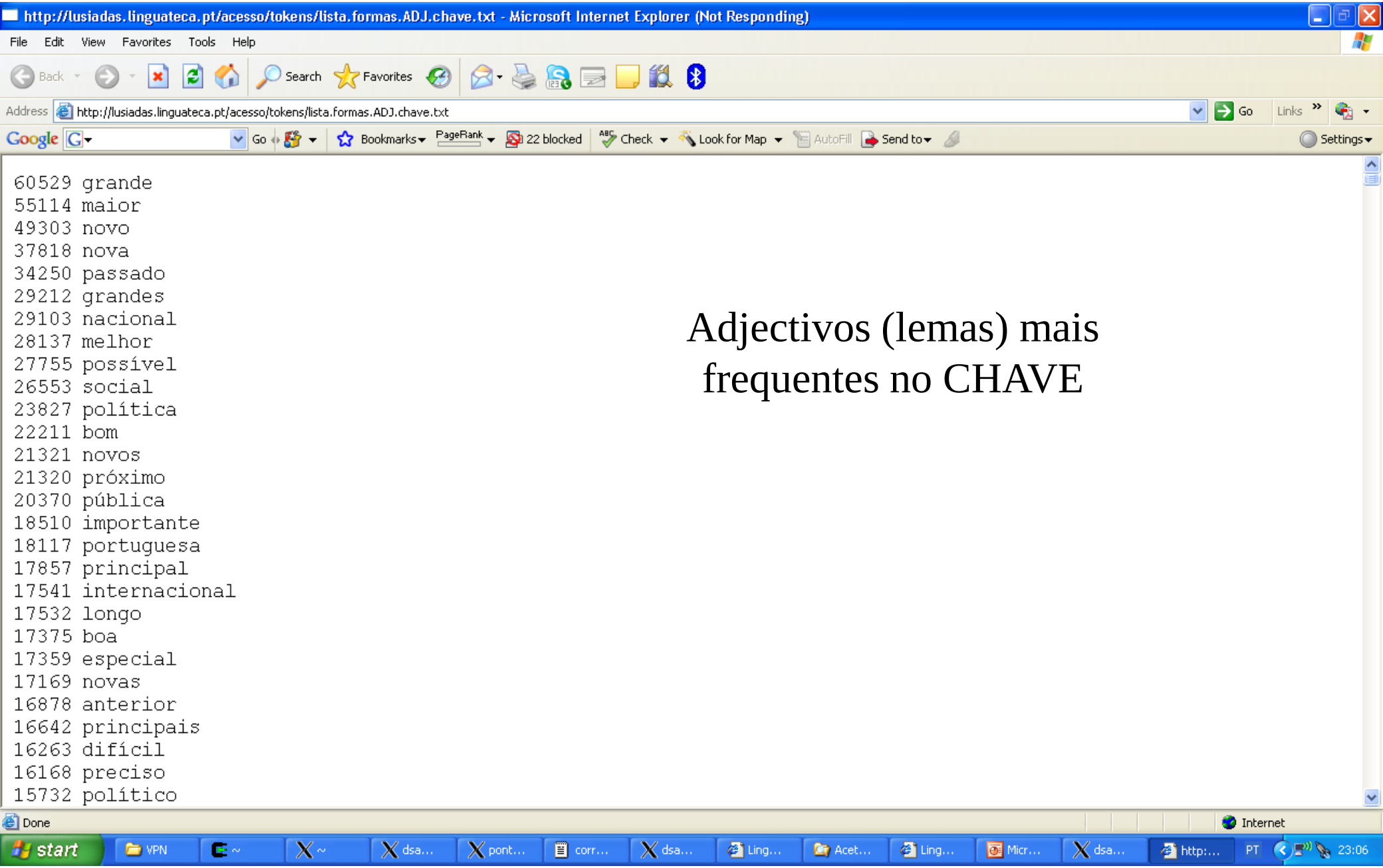
- *Uma pequena gralha na seguinte frase encotrada no CETEMPública v. 1.7.*

"O anónimo Sr. J. Manuel Almeida nunca será notícia, anda que lhe saia a sorte grande, a menos que morra de comoção ao receber o prémio ."

anda que => ainda que



Corpus	Lemas								
	N	ADJ	ADV	V	NUM	GRAM	PROP	todos	todos/pos
AmostRA	5088	1903	324	2351	323	184	1387	11197	11560
ANCIB	14957	4116	702	4201	4964	356	29717	57993	59013
Avante!	20512	8996	1707	9419	6264	851	47500	93014	95249
CDHAREM	5432	1970	391	2237	685	293	3915	14453	14923
CETEMPúblico	160824	47512	5475	54926	109796	2047	1070220	1430678	1450800
CETEMPúblico (primeiro milhão)	13510	4765	845	5437	2389	284	23758	49517	50988
CHAVE	113806	39865	4715	40533	91180	1383	711584	988949	1003066
Clássicos da Literatura Portuguesa/Porto Editora	12818	5109	1116	8961	267	355	4393	31726	33019
ConDIVport	15162	8009	1507	11619	2184	580	31571	60790	63123
ConE	10163	2727	435	2744	4295	324	18504	38527	39192
DiaCLAV	20854	6809	1174	8924	5754	416	64786	105800	108717
ECI-EBR	13909	5773	936	6192	925	353	8987	35815	37075
ECI-EE	1043	515	183	575	226	126	173	2727	2841
ENPC (parte pública)	3555	1375	364	1881	138	183	793	8023	8289
FrasesPB	2156	749	187	888	60	129	215	4255	4384
FrasesPP	1698	689	183	774	70	131	197	3640	3742
Museu da Pessoa	5221	1378	320	2641	353	227	2106	11808	12246
Natura/Minho	12829	5339	851	5199	4431	425	30354	58141	59428
Natura/Público	35717	12121	1534	12170	9723	837	83606	152574	155708
NILC/São Carlos	64650	22874	2769	25444	60630	815	302016	472123	479198
Vercial	35918	12240	2629	27351	2710	628	42890	116376	124366



Adjectivos (lemas) mais frequentes no CHAVE

Exercícios com o AC/DC: questões e resolução

Diana Santos

Linguateca
www.linguateca.pt

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

Simples 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 «intrometer 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 **pessnum**: 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 caráter 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.*"]**