Pesquisando corpos:
Foi você que pediu um Corpo Ferreira?

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

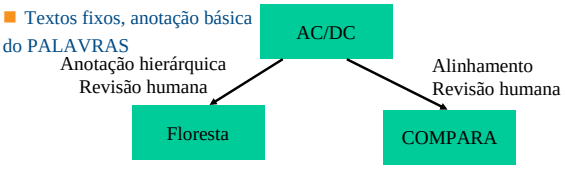
Projecto AC/DC

- 3. adicionar a dimensão da tradução → COMPARA, CorTrad
- 4. adicionar a dimensão da revisão humana → Floresta Sintá(c)tica, COMPARA
- 5. adicionar a possibilidade de criar corpos próprios
- 6. adicionar a dimensão de corpos comparáveis

Corpógrafo

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)

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)

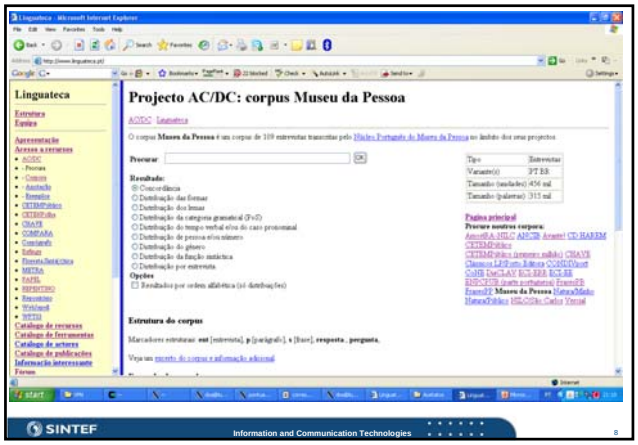
Galeria dos corpora

- Entrevistas (texto oral transcrito)
  - Museu da Pessoa

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

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

Rocha (2007)



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)

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 encontrada no CETEMPúblico 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

Santos (2008)

Problemas típicos dos 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

Santos (2008)

Uso de corpos no ensino da língua portuguesa

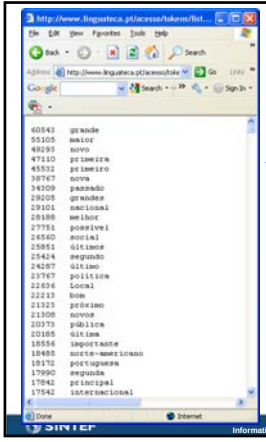
- Para motivar/inspirar o professor

  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

Santos (2008)



Adjectivos mais frequentes no CHAVE:

Formas e lemas

Audiência do AC/DC

1. A idade da inocência
2. A hora da verdade
3. A tempo?

Pressuposto deste curso:
As pessoas interessadas em estudos linguísticos, e em corpos anotados, estão mais habilitadas do que um vulgar engenheiro a compreender uma sintaxe de procura não trivial, visto que são especialistas em línguas e linguagens☺

EBraLC (2008)

Exemplos práticos de exploração

- Que verbos são mais usados com o advérbio *depressa*?
- Que sujeito é mais comum para *despedir* ?
- O que é mais azul em português?
- E cor-de-rosa?

- Qual a construção mais comum com o verbo *dizer*?
- Que verbos são mais usados em orações relativas?
- Com que verbo *não* é mais frequente?

- Quando é que *sempre* é usado?

A interacção (primeira volta)

- Escolha do que se quer procurar
- Escolha do que se quer receber
  - Concordância
  - Distribuição
    - Por palavra
    - Por lema
    - Por categoria gramatical
    - Por autor
    - etc

EBraLC (2008)

Interlúdio sobre categorização

- A que categoria semântica pertence *Salamandra Roxa* e *Globo Azul* ?

PPLJ1(1071): *Helena passou a outro envelope*, **Salamandra Roxa** , spoilt

PBPC2(123): *Petrus me disse que eu sabia despertar Ágape, através do* ***Globo Azul***.

- A que categoria gramatical pertence *A Grã-Bretanha Precisa das Suas Universidades.* ?

EBDL2(724): *O seu carro está estacionado lá fora, um Renault vermelho com seis anos, com um autocolante amarelo no vidro traseiro: «****A Grã-Bretanha Precisa das Suas Universidades.****»*

- A que grupo de cor pertence *cor de osso* ?

O que significa *verde*? Como interpretá-lo?

- *PBOL1(248): Desejaria ser, em parte, como essa adolescente, e sustentar com doçura, ano após ano, também emoldurada, meu ramo sempre* ***verde****, sua corola imortal.)* [sema="cor_naomaduro"]
- *PBCB2(202): Um cacho de bananas* ***verdes*** *no chão da cozinha lembra-me que passei o dia a chá e bolacha.* [sema="cor_naomaduro"]
- *PBJS1(1126): Não entendo por que é que, pelo menos, não alugaram um cupê para andar em volta do parque -- disse o capitão, referindo-se às elegantes carruagens de aluguel decoradas com espelhos, seda* ***adamascada*** *e contornos de prata, verdadeiras camas ambulantes, que eram anunciadas diariamente nos jornais.* [sema="cor_0"]
- *EBDL1T2(867): A área tem muito* ***verde****, é arborizada e quieta.* [sema="cor_0"]

EBraLC (2008)

O PALAVRAS

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

- Desenvolvido por Eckhard Bick desde 1994
- O primeiro analisador sintáctico para o português disponível na rede e usável por terceiros
- Um sistema dependencial baseado em restrições (constraint grammar dependency parsing)
- Uma teoria subjacente sobre o português
- Atribui sempre um par **forma:função**
- Usado pela Linguateca desde 1999-2000
  - Nova atomização
  - Muitas melhorias sugeridas através da colaboração
  - Como todos os sistemas automáticos, não é perfeito...

EBraLC (2008)

Um exemplo pouco animador

CF185-7 Ele sequestrou e violentou três meninos com a intenção de lhes transmitir o vírus da Aids de que se sabia portador.

com	[com] PRP @<ADVL
a	[o] <artd> DET F S @>N
intenção	[intenção] N F S @P<
de	[de] PRP @N<
lhes	[eles] PERS M/F 3P DAT @DAT>
transmitir	[transmitir] V INF @IMV @#ICL-P<
o	[o] <artd> DET M S @>N
vírus	[vírus] N M S @<ACC
de	[de] <sam-> PRP @N<
a	[o] <artd> <-sam> DET F S @>N
Aids	[aids] PROP F S @P<
de	[de] PRP @ADVL>
que	[que] KS @SUB @#FS-<ACC @#FS-<ACC
se	[se] PERS M/F 3S/P ACC @SUBJ>
sabia	[saber] V IMPF 3S IND VFIN @FMV
portador	[portador] N M S @<ACC

Formato AC/DC e COMPARA e CorTrad

word	**lema**	**pos**	**T**	**P**	**G**	**func**	**...**
com	com	PRP	0	0	0	<ADVL	0
a	o	DET_artd	0	S	F	>N	0
intenção	intenção	N	0	S	F	P<	0
de	de	PRP	0	0	0	N<	0
lhes	eles	PERS	DAT	3P	M/F	DAT>	0
transmitir	transmitir	V	INF	3	0	IMV_#ICL-P<	0
o	o	DET_artd	0	S	M	>N	0
vírus	vírus	N	0	S	M	<ACC	0
da	de+o	PRP+DET_artd	0	S	F	N<+>N	0
Aids	aids	PROP	0	S	F	P<	0
de	de	PRP	0	0	0	ADVL>	0
que	que	KS	0	0	0	SUB_#FS-<ACC #FS-<ACC	0
se	se	PERS	ACC	3S/P	M/F	SUBJ>	0

T temcagr: tempo, caso, grau **P pessnum**: pessoa, número **G gen**: género

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

Santos (2008)

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")

TaLC (2008)

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** .

TaLC (2008)

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** .

TaLC (2008)

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

TaLC (2008)

Exemplo de criação de material didáctico em português para estrangeiros

- *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)

TaLC (2008)

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

TaLC (2008)

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

TaLC (2008)

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 ...

TaLC (2008)

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

Forma, maneira, modo e *caminho*

No DiaCLAV (um corpo do AC/DC)

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.*"]**
- **[word="modo" & pos="N.*"] @[pos="PRP.*"]** etc.

Duas maneiras diferentes de obter uma tabela grande

[word="forma|modo|caminho|maneira" & pos="N.*"] [pos="de"] etc.

Mas a classificação não é óbvia!

Santos et al. (2007)

- Em última análise, a forma como a distinção está codificada em qualquer floresta é arbitrária! (**mas pode ser resolvida com um sistema de procura adequado**)
- `icl`, `acl`, `fcl` existem, mas `rcl` ou `subcl` não existem
- O tempo está marcado no verbo, ou só/também na oração?
- O género está marcado no SN, ou só no seu núcleo, ou em cada palavra passível de ter género?
  - *um índio pele vermelha* : que género deve ser marcado em *vermelha*? e em *pele*? e em *pele vermelha*?
- 3/4 razões para ter um adjectivo como núcleo
  - elipse, propriedade, indeterminação: *jovens alemães*

E as necessidades do utilizador não são precisas!

Santos et al. (2007)

- sintagma nominal *é quando* sintagma nominal
- SN complexos, mas sem oração no meio
- verbos que aparecem após uma citação
- frases em que a ordem sujeito-verbo-objecto é quebrada
  - frases que têm os 3, e verbo sem auxiliares?
  - VSO, SOV, OSV, OVS, VOS
- encontre um SN com a maior quantidade possível de dependentes
  - pai [de família [de emigrantes [dos subúrbios [de Moscovo [de 1900]]]]]
  - cão [de caça] [de loiça] [da Bélgica] [do meu pai] [do tempo da Grande Guerra]
- orações em que o particípio não exerce uma função verbal

Sintaxe da procura no IMS-CWB

- [atributo="valor"]
- [atributo!="valor"] } ou uma sequência destes testes de atributos ligada por &
- Uma sequência, que pode ser modificada por {min, max}, ou por * ou +, significando qualquer número incluindo zero ou não

```
[lema="comer"][pos="DET.*"]* [pos="N.*" & func="<ACC"]
```

- Os valores são descritos por expressões regulares
  - `. qualquer caracter`
  - `[a-d] qualquer dos caracteres a, b, c, ou d`
  - `[,:?.] qualquer dos caracteres ,, :, ?, or .` } caracteres
  - `[^axl9] todos os caracteres excepto a, x, l ou 9`
  - `+ um ou mais`
  - `* zero ou mais`
  - `{2,7} pelo menos 2 e no máximo sete` } modificadores

EBraLC (2008)

Mais informações sobre a sintaxe do IMS-CWB

- Listas de valores: separadas por |
  - (passagem|estrada|via|rua)
  - [lema="forma|maneira|modo"]
- Caracteres especiais: precedidos de \
  - [lema="casar\+.*"]
- Indicação de sobre que unidade deve ser calculada a distribuição: @
  - [lema="dar"] []* @[func="<ACC"]
  - "caixa" @[pos="ADJ.*"]
- Atalhos: [word="pena"] pode ser escrito "pena" (e, se sozinho, pena no COMPARA e no AC/DC)
- Restrição a um atributo estrutural ou a um número de unidades:
  - within s, within 5

EBraLC (2008)

Exploração inicial: Algumas questões talvez surpreendentes

- Adjectivos que modificam a palavra *pedra*
- Lema de substantivos femininos "regulares" ou de advérbios
- Função de gerúndio
- Categoria gramatical de palavras em *–ado*
- Categoria gramatical de *amigo*

- Categoria gramatical das palavras em *-endo*
- Categoria gramatical de *alto*

- Qual o género morfológico do que é comparado em *mais X do que*?

O estudo de *embora* (inspirado por *home*)

- Quais as categorias gramaticais relevantes?
- Quais os verbos associados?
- Problemas na anotação...

- E *fora*?

- *Inspirado em/por*: qual a diferença?

Outras questões sem auxílio automático

- Quantas *correntes* há no DiaCLAV? Distribuição de sentidos ...
- Oposição privativa: palavras que significam algo e o seu oposto
  - gosto: bom gosto, mau gosto
  - sorte: boa sorte, má sorte
  - ficar: mudar ou permanecer
- Vagos quanto a de propósito ou não
  - envergonhado
  - enganado
- Vagos quanto a uma característica ou a um sentimento
  - apaixonado
  - engraçado

EBraLC (2008)

Mensagem final

- O mais importante é a informação que se pode obter, e desenvolver a metodologia de como obtê-la
- É importante perceber as potencialidades (e as limitações) de cada corpo, e idealmente usar vários
- **Mas**: não fiquem ofuscados!
- É importante compreender que um corpo ajuda-nos a fazer linguística
- Mas o mais importante é
  - Explorar
  - Descobrir
  - Validar

EBraLC (2008)

Nos bastidores do AC/DC

- Qual o género do ECI-EBR?
  - Está dividido em parágrafos, sem qualquer separação de textos e portanto de géneros
- No DiaCLAV, N1223 e N1228 são iguais :-(
  - Um sistema que encontre / identifique artigos iguais...
- Vale a pena anotar e rever a anotação de texto cheio de erros? Saúde do ConDiv
  - transmissível ao macaco, encontra-se no sistema nervoso central e dá origem **ima** inflamação aguda dos cornos anteriores, ou neurónios motores, da substância cinzenta <**Li** medula-**espinhal**; a existência deste vírus foi admitida, desde 1908, após os trabalhos dos **i** investigadores Landsteiner e Popper; em 1913 outros cientistas - Flexner, Levaditi e Noguels conseguiram cultivá-lo em tecidos, extraindo da substância nervosa humana .

Novas funcionalidades do AC/DC

- Anotação da cor e de outros campos semânticos
  - Roupa
  - Lugares, pessoas, organizações, etc.
- Procura apoiada em relações semânticas
- Comparação contrastiva entre duas procuras alternativas
  - Duas caixas de procura independentes
- Reuso de procuras mais complexas, a partir da colaboração dos utilizadores
  - Primeiro passo: lista de perguntas já respondidas
  - Segundo passo: criação de macros

Referências

- Bick, Eckhard. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, 2000.
- Esta apresentação reutilizou muitas anteriores, identificadas por rectângulos circundando a referência, falando de corpos e sua utilização. Cito apenas as que não são exclusivamente minhas:
  - Paulo Rocha. "Arte dos corpora em português: o projecto AC/DC". *Encontro Um passeio pela Floresta Sintá (c)tica* (Coimbra, Portugal, 28 de Setembro de 2007).
  - Diana Santos, Eckhard Bick & Susana Afonso. "Floresta Sintá (c)tica: apresentação e história do projecto". *Encontro Um passeio pela Floresta Sintá (c)tica* (Coimbra, Portugal, 28 de Setembro de 2007).