Para que serve um corpo
O que é, o que significa a sua anotação, exemplos de uso

Diana Santos

ILOS

d.s.m.santos@ilos.uio.no

Setembro de 2013

O que é um corpo?

Há sempre várias formas de definir uma palavra ou um conceito...

Formal	Através da sua forma
Funcional	Através da sua função
Relacional	Através da relação com outros conceitos conhecidos
Histórica	Através da sua origem e do seu desenvolvimento

Como engenheira, estou mais preocupada com, e vou privilegiar, a sua função.
Reparando que a maior parte das palavras não se aprendem através de definição, mas sim da observação do seu uso.

A minha definição

Um corpo é

- uma colecção
- classificada
- de objectos linguísticos
- para uso em Processamento de Linguagem Natural/Linguística Computacional/Linguística

"Objetos linguísticos" podem ser textos, frases, palavras, entrevistas, erros ortográficos, entradas de dicionário, citações, pareceres jurídicos, filmes, imagens com legendas, traduções, correcções (de textos de alunos de língua ou de tradução), telefonemas, simulações do tipo Wizard of Oz, programas...

Amostra da língua

O que é uma amostra: um pouco de algo maior para saber se se quer compara, se se quer reconhecer, se se quer conhecer, e que se leva consigo, ou que pelo menos é mais portável do que o todo.
Teoria da amostragem: amostra representativa, amostra aleatória, ... Qual é o problema? É que nem todas as amostras são igualmente fiáveis (para representar o todo).
O tamanho da amostra também é relevante: amostra de urina, amostra de sangue, amostra de tecido estampado, amostra de limões, amostra do próximo livro...

Duas operações básicas que todo o corpo permite

De notar que estamos a falar de um corpo eletrónico, ou seja, a conjunção de três coisas relacionadas:

1. um conjunto de textos
2. um conjunto de informação a marcar/classificar esses textos,
3. e uma interface que permita consultar os dois primeiros

Um corpo eletrónico digno do seu nome oferece duas operações:

1. concordância (ver algo em contexto)
2. distribuição (ver algo por parte) (frequência)

Isto porque um corpo tem sempre mais alguma informação do que o próprio texto.

Problemas de um corpo

Não quero dar a ideia de que um corpo é panaceia para todos os estudos!

- É língua sem contexto. A maior parte das vezes não sabemos quando, quem, ou onde.
- Não é fácil escolher o que se analisa e como se analisa: há sempre imensas simplificações envolvidas:
  - a forma pode ter mudado: remover itálicos, corrigir gralhas
  - variante ortográfica
  - normas de transcrição

  Há sempre muitas análises discutíveis, seja qual for o nível linguístico em que se está: estrangeirismo? nome de marca? cor? substantivo? erro ou dialeto? ditransitivo ou transitivo? dativo possessivo ou objeto indireto? uma expressão com várias palavras ou não? uma variante criativa ou um engano? um erro de tradução ou uma escolha consciente?

Exemplos de dificuldades de análise

where the pearl was buried → onde a pérola estava guardada (Santos 95)

> *At first sight, the word* enterrar *(the standard translation of* bury*) would be the right choice, and thus the preference of the translator in using a less specific verb would illustrate a strong preference of the target language, going against exact translation in that particular translation pair.*
> *However, Lauri Carlson (p.c.) has pointed to me that* bury *in English is also used with the figurative meaning "hide". If this is not the case with* enterrar *(which I believe it is true, i.e.,* enterrar *is not standardly used to express "esconder debaixo de alguma coisa"), then we could have the English sentence using* bury *to convey "hide under something", while* enterrar *would mean the more specific "hide underground". So, there would not be a (lexical) term in Portuguese with the meaning of figurative* bury*, and the translator would have been obliged to use a subsumption strategy.*

Exemplos de dificuldades de análise

Podia ter dado mais...: transitivo ou ditransitivo? (Bacelar do Nascimento et al. 93)

> *Existe uma ocorrência que foi contabilizada como bitransitiva, mas que é ambígua na sua interpretação:* eu acho bem que as raparigas hoje não queiram esta vida, não é. Não é que eu tenha sido infeliz com ela, mas reconheço que, podia ter dado mais. *As duas interpretações em confronto correspondem a "a vida ter-me dado mais" e "eu dar mais" (V+OD).*

Exemplos de dificuldades de análise

CP46-1 Na mesma ocasião iniciaram-se investigações que incidiram sobre o árbitro madeirense Marques da Silva, **também ele suspeito de se ter deixado corromper**. (Bick et al. 2007)

> *Em CP46-1, existem duas configurações sintácticas para o trecho a negrito sem que se verifiquem diferentes interpretações a nível da semântica. Numa das configurações o trecho é considerado uma oração sem verbo, na outra um sintagma nominal. Estas diferentes representações sintácticas têm, além disso, consequências ao nível das funções dos seus constituintes imediatos.*

Exemplos de dificuldades de análise

Santos e Gasperin 2002

> *One example is the choice between encoding a particular syntactic difference as PoS or as constituent function. In* Três quartos do hotel foram ocupados pela polícia *one can represent the difference by assigning the PoS noun to* quartos *in one interpretation and the PoS numeral in the other. Alternatively, one can have both parses tagging* quartos *as noun at the PoS level, but individuated by their function inside the NP* Três quartos do hotel *(having either* quartos *or* hotel *as NP head). (...) The same liberty at making distinctions can be seen in the three sentences* Ele está de volta, De volta da mãe, ele apressava-se *or* Comprou o bilhete de volta, *where a parser can give the same PoS, viz. preposition noun, to the three instances of* de volta, *but separate them by function (e.g. by AJP, AVP and PP), or actually perform three different tokenizations as well: "de volta", "de volta de", and "de" "volta".*

Exemplos de dificuldades de análise

Onde marcar o quê? Santos et al. (2007)

- *ex-capitão das FAPLA*: quantas palavras, e quais?
- O género deve ser marcado no SN, ou só no seu núcleo, ou em cada palavra passível de ter género? Ou seja: em *um índio pele vermelha*: que género deve ser marcado em *vermelha*? e em *pele*? e em *pele vermelha*?

Exemplos de dificuldades de análise

Semântica (Pontes)

- *ele precisa de apanhar sol* (ele é que o diz)
- *ele precisa de apanhar uma boa tareia* (ele não concorda)
- *ele precisa de dançar para se sentir feliz* (ele talvez não saiba)

Qual o sujeito de *precisa*? E o papel semântico de *ele*?

Exemplos de dificuldades de análise

Qual o tempo que devemos marcar?

- *ele tinha o casaco vestido*
- *ele tinha vestido o casaco*
- *ele tinha vestido o casaco, o carapuço e as luvas*

O que é que ele tinha vestido? : mais que perfeito ou *ter* com particípio passado?

Exemplos de dificuldades de análise

presidente da Câmara de Oslo: quantas EM, e qual a sua classificação? Defendendo a marcação das várias alternativas. <ALT> <ORGANIZACAO>Governo</ORGANIZACAO> de <PESSOA>Mário Soares</PESSOA> | <ORGANIZACAO>Governo de Mário Soares<ORGANIZACAO> </ALT>. Santos e Cardoso 2006

> *In addition, sometimes human annotators do not have enough information to decide, even consulting other knowledge sources, such as the Web: How many (Portuguese speaking) readers know that* Miss Brown *is a Brazilian band or that* os pastéis de Belém *is often used as a place in Lisbon?*

Exemplos de dificuldades de análise

ReRelEM: relações entre entidades mencionadas. Santos e Cardoso 2009 (EdV)
Qual a relação entre *Coimbra* e *Portugal* no seguinte texto?

> **Portugal** *perdeu para a* **Alemanha** *nas quartas de final da* **Eurocopa**. *Vi o jogo na* **Praça da República**, *e mesmo com a derrota os bares de* **Coimbra** *estavam cheios.*

E entre *Lobos* e *Vasco Uva*?

> **Lobos** *recebidos em apoteose*
> *(...), o capitão* **Vasco Uva** *explicou por que houve uma empatia tão grande entre (...)*

Exemplos de dificuldades de análise

- A que grupo de cor pertence *cor de osso*? Branco, creme ou amarelo?
- que verde é este? E adamascado?
  - Desejaria ser, em parte, como essa adolescente, e sustentar com doçura, ano após ano, também emoldurada, meu ramo sempre **verde**, sua corola imortal. Cor ou não maduro?
  - Um cacho de bananas **verdes** no chão da cozinha lembra-me que passei o dia a chá e bolacha. Cor ou não maduro?
  - A área tem muito **verde**, é arborizada e quieta. Cor ou nao cor?
  - 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. Cor ou não cor?
- A cor designa raça aqui? *A Mulata Sedutora (Uma espécie de mulher-objeto* **cor de chocolate**, *desejada por todas as raças).*
- Quantas cores em *Rímel* **colorido azul marinho** *ou castanho na ponta dos cílios também dão cor*?
- Cor política ou de raça? *Sem a ajuda soviética e vítima do bloqueio americano, o sonho do socialismo* **moreno**, *que tanto encantou a esquerda brasileira, parecia estar fadado a um final infeliz*

Primeira conclusão

- Quanto mais analisado um corpo, mais se constitui numa obra de criação.
- Quanto menos analisado um corpo, menos se pode concluir ou interrogar.

Por essa razão, o Corpus Brasileiro, embora maior, será possivelmente apenas interessante para estudos de léxico, enquanto o CHAVE pode ser usado para estudos textuais (de texto jornalístico).
Além disso, quanto mais fácil é criar/obter, mais temos esse tipo de corpo. Por isso há tantos corpos de textos jornalísticos, ou textos na Web, e agora estamos na hora dos pios (tweets).

Vantagens dos corpos

A maior vantagem é que, depois de ser criado (ou durante), pode ser interrogado vezes sem conta, e as perguntas e procuras repetidas, e melhoradas.
Presta-se também a análises quantitativas, embora pressuponha sempre uma análise qualitativa (mesmo que automática).

O que se faz com um corpo

- Miriam Schlesinger: Distinguir se o género do tradutor tem influência.
- verificar a ordem da mudança *vos-vossos* em português de Portugal
- verificar a entrada na língua da palavra *marrom* em vez de *castanho*
- como se estrutura um pedido de casamento, como se conta
- a estrutura de um conto popular (ou de uma ida ao restaurante)
- como compreendemos uma anedota? E/ou humor em geral
- a descrição física e psíquica de uma pessoa

O que se faz com um corpo II

- Caracterizar ou narrar? diferentes estratégias
- Tipos de respostas: incentivo, conflito, resposta neutra, "já chega"
- Mal-entendidos e a forma de os sanar
  - "Não era isso que eu queria dizer"
  - retractar-se
  - acusar o toque
- como se identifica que uma dada palavra mudou de sentido: *partilhar*
- a morte de uma palavra ou expressão (e os seus substitutos)

O que se faz com um corpo III

- Quais os adjetivos mais antigos? Olhando para a criação de verbos *en + ecer*
- Qual a diferença entre as diversas expressões de localização *por cima de*, *em cima de*, *acima*? É a mesma que é espelhada por *por baixo de*, *em baixo de*, *abaixo de*, *debaixo de*?
- Existe diferença entre *entre* e *por entre*?
- O que é mais frequente: verbos derivados de X que significam **pôr X em**, ou **pôr em X**? *envenenar*, *embandeirar* ou *engarrafar*, *emoldurar*?
- Como obter descrições de caráter?

O que se faz com um corpo IV

- Há prova que a ordem das cores em português segue a hipótese de Berlin e Kay?
- Existe diferença entre os processos de derivação *a-ar* ou *en-ar* ou *-ar*?
- Renata Tironi: Se compararmos um resumo com o texto de que é resumo... qual a diferença em nomes próprios (diferentes), orações relativas, conjunções, cadeias anafóricas?
- Que cor é pardo?

Alguns métodos que podem ser usados com corpos

Usando alguma artilharia estatística

1. Comparação de duas proporções
2. Correlação entre duas propriedades
3. Classificação de exemplos dado um conjunto de classes
4. Agrupamento de várias instâncias com base em características mensuráveis

Nada disto é mágica, e tudo isto pressupõe uma análise sólida das características usadas.