Linguística com corpos na era da internetização

Diana Santos

Linguateca
www.linguateca.pt

Linguística/PLN com corpos: Algumas reflexões, salteadas de exemplos em português

Diana Santos

Linguateca
www.linguateca.pt

Índice

- Linguística com corpos
  - Rede como corpo
  - Abordagens computacionais e linguísticas: dividir para perder o controlo
- Casos concretos
  - A cor
  - Emoções
  - Processamento temporal
- Os corpos na Linguateca
  - Alguma apresentação
  - Balanço

Linguateca

Nota sobre terminologia

- Durante muitos anos usei *corpus* e *corpora* em itálico visto que era o padrão usado em Portugal pelas autoridades linguísticas
  - Mas cansei-me de corrigir alunos a dizer ou escrever *a corpora*, *os corpus* e *o corpora*
  - Depois assustei-me ao ver que outros começavam a escrever *córpora*, *córpus*
  - Decidi aportuguesar de vez, para *corpo* e *corpos*
- *Corpus linguistics*: os conhecidos "noun compounds" do inglês e os problemas que a sua tradução comporta, e a vagueza do *de*
  - **Linguística com corpos**
  - Linguística de corpo na mão
  - Linguística feita sobre corpos
  - Linguística tendo como objecto de estudo o(s) corpo(s)

O corpo não é o objecto de estudo

- There is a void at the heart of corpus linguistics. The name puts 'corpus' at the centre of the discipline.**1** In any science, one expects to find a useful account of how its central constructs are taxonomised and measured, and how the subspecies compare. But to date, corpus linguistics has measured corpora only in the most rudimentary ways, ways which provide no leverage on the different kinds of corpora there are. (Kilgarriff, 2001)

1 Alternative names for the field (or a closely related one) are "empirical linguistics" and "data-intensive linguistics". By using an adjective rather than a noun, these seem not to assert that the corpus is an object of study. Perhaps it is equivocation about what we can say about corpora that has led to the coining of the alternatives.

A equivocação é de K.! A linguística com corpos fala sobre a língua, não sobre o corpo!

Um corpo linguístico é...

EdV 2006

- Uma colecção **classificada** de **objectos linguísticos** para **uso** em PLN/LC/L
- Uso: estudo, medição, teste, avaliação
- Objectos linguísticos: textos, frases, palavras, entrevistas, erros ortográficos, entradas de dicionário, citações, pareceres jurídicos, filmes, imagens com legendas, traduções, correcções, telefonemas, WOZ, programas ...
- Classificada:
  - A nível dos parâmetros da recolha (que categorias considerar)
  - A nível da escolha (todos, alguns, amostra,...)
  - A nível dos fenómenos (tipo de erro, tipo de tradução, tipo de texto, ...)
  - A nível dos constituintes (análise sintáctica, semântica, fonológica, discursiva, etc.)
  - Avaliação

Quatro tipos de usos principais

EdV 2006

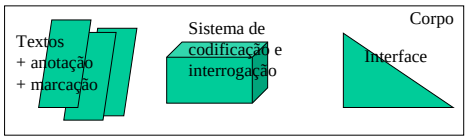
1. Ter uma ideia do problema/conhecer
   - consultor, familiarizador, treinador
2. Medir
3. Avaliar
   - uma hipótese, um sistema, um método
4. Criar outras coisas
   - dicionários ou estruturas de conhecimento, materiais de teste
   - sistemas de RAP, sistemas de ensino e jogos
   - terminologias, almanaques e catálogos, sistemas de detecção (de plágio, de spam, ...)

A internetização...

- Não faz da rede um corpo (não é replicável, não permite anotação, nem distribuição)
- "Web as corpus": Mais um slogan com falta de rigor ☺
- Amostras da rede (útil para obter textos e dados)
- Fragmentos da rede (útil para estudar a linguagem de certas comunidades)
- Uso da rede como
  - confirmadora, estimadora, consultora
- A rede como fonte de todo o conhecimento ☹, ou como liberadora de muito conhecimento (e desconhecimento) difíceis de alcançar antes

A rede não substitui os corpos...

- Embora os alimente, os distribua e os proselitize
- Um corpo valioso precisa de ter
  - documentação associada
  - controlo de versões
  - revisão humana
- Um corpo na rede é na realidade um trio



Metodologia da linguística com corpos

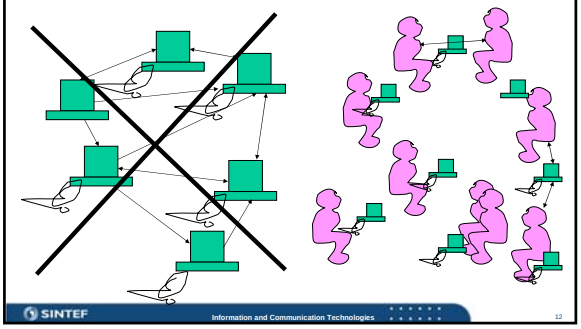
- O que é um corpo?
  - Um instrumento, um objecto de estudo, ou uma metodologia de estudo?
- O que é que se faz com corpos?
  - Estuda-se a língua, avaliam-se sistemas, prepara-se materiais de ensino, teste, ..., treinam-se sistemas
- Perguntas típicas a um corpo
  - Há correlação entre duas propriedades observáveis?
  - Há correlação entre uma categoria de classificação e uma propriedade observável?
  - Diferenças e semelhanças entre dois objectos
- O que é que não se pode fazer com corpos?
  - Obter dados negativos, distinguir entre errado ou em falta, analisar o contexto

Ver Fillmore (1992), Sampson (2003), Sorace & Keller (2005), Gries (2006), Pustejowsky & Rumshisky (2008), Santos (2008)

A importância da compreensão

- *Should it ever come about that linguistics can be carried out without the intervention and suffering of a native-speaker analyst, I will probably lose interest in the enterprise* (Fillmore, 1992: 59)
- Se vier a ser possível fazer linguística sem o esforço e a intuição dos falantes dessa língua (como língua materna), tal não é para mim!

Sociedade da informação, do conhecimento



Métodos automáticos de avaliação?

- O engano dos métodos automáticos de avaliação, por exemplo em tradução automática:
- BLEU usa um conjunto de traduções de referência (humanas) para avaliar/comparar traduções (automáticas)
  - Ou seja: medição automática de semelhança com a tradução humana, o que é aceitável em 90% dos casos
  - Mas: produzir essas traduções custa!
  - Então: as pessoas começaram a usar muitas traduções de frases diferentes (por exemplo no EuroParl) para aproximar o BLEU...
  - Neste momento muitas avaliações/medidas relatadas na literatura de TA apenas significam semelhança literal com "traduções" feitas no parlamento europeu, num corpo estafado/estafadíssimo...

Frequência, e unidades

ESSLLI 2007

- O que é mais importante: o que é mais ou o que é menos frequente?
  - stopwords em RI
  - palavras de frequência média para indexação
  - palavras raras em estudos de atribuição de autor ou de detecção de plagiarismo
- O que é uma palavra?
  - Avaliação de correcção ortográfica: `correcçãoortográfica`
  - *Morfolimpíadas* e o pesadelo da atomização (15,9% das formas e 9,5% dos tipos não eram consensuais)
  - Citando Sinclair sobre expressões com várias palavras ...
  - A pontuação conta para comparar análises sintácticas?

Para diferentes níveis de análise, diferentes unidades...

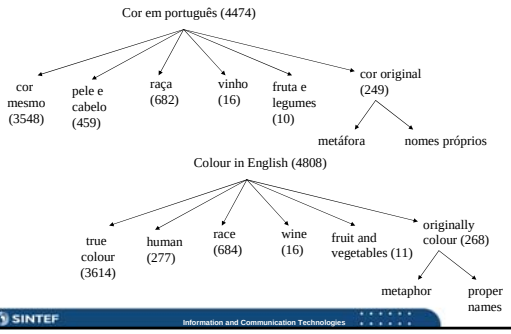
- If *set in train* always occurs together in this sequence when it has the obvious meaning, then the three words constitute **one** choice. As soon as learners have appreciated that each phrase operates as a whole, more or less as a single word, (...) they have a new word *set in train*. Not many learners 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 uso da cor: Um estudo exploratório num corpo paralelo

- Um campo semântico relativamente simples
- Poucas (?) palavras, poucos (?) conceitos, pouca (?) necessidade de interpretação

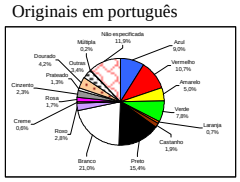
1. Pegar nos textos e marcá-los (Silva et al. 2007-hoje)
2. Escolha das categorias de anotação
   1. O que fazer com cores descoradas? Ou sem cor?
   2. O que fazer com cores usadas noutros campos semânticos?
3. Estudar as relações entre cor e o resto (sintaxe, léxico, autores...)
4. Estudar a tradução da cor
5. Estudar as diferenças entre as línguas

Classificação das cores no COMPARA

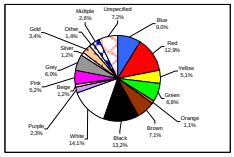


Distribuição da cor nos textos originais

Originais em português



Originais em inglês



- Mais branco, mais preto
- Muito mais cor não especificada
- Mais rosa, mais cinzento, mais marrom

A cor na tradução, no COMPARA

- Diferentes metáforas: *sorriso amarelo -> wan smile; romance cor de rosa -> ?; blue movies -> filmes tristes; nódoas negras -> bruises; armas brancas -> knives; fazer a vida negra -> give a hard time; red herring; paint the town red -> ?; brown off -> maçar-se*
- Culturas diferentes: *black coffee -> café, red meat -> carne mal passada; correio azul -> first class stamp; red demands -> intimações; red tape; Black Maria -> ramona; red-light district,*
- Vagueza: *golden -> dourado*, ou *de ouro, de prata*
- Criatividade do tradutor: *putty-coloured -> cor de massa de vidraceiro; civic redbrick -> novas e sem tradição; azuleleca -> tricklight*
- Convenções diferentes: *brown paper -> papel pardo, goldfish -> peixe vermelho; claras -> egg whites; página em branco -> blank page*; *dark purposes -> negros propósitos;*

O género, ou assunto, é importante...

- No EuroParl
  - Cores dos partidos
  - Cores dos nomes dos deputados
  - Cores de locais/edifícios usados metonímicamente para referir o governo de um país, a Casa Branca
- Sobre futebol
  - cartões
  - equipas
- Gastronomia
  - cerveja loira, carnes brancas, açúcar amarelo, manjar branco
  - alourar a cebola, dourar o assado, corar o frango,

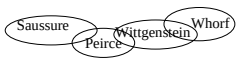
Mais uma digressão colorida

- Há um contínuo entre o léxico e a gramática, mas ambas são categorização (simplificação da experiência)
- Gramática: as coisas a que a língua dá mais importância, e para as quais faz regras especiais
- Cores: *Colorless green ideas sleep furiously*
  - Red green houses: agramatical (duas palavras de cor não fazem parte de um SN)
  - *Green* pode ter o sentido de não maduro...
- Número, género, roupa, responsabilidade, afecto, regra geral ou caso particular
  - Calçar os sapatos, calçar as meias, pôr as meias, vestir o vestido, pôr o chapéu, apertar o cinto
  - Diminutivos: livrinho, casinha, priminha, arrozinho, prendinha

A análise de emoções

- Durante muito (demasiado) tempo, a análise da informação/factos na língua teve primazia na linguística
- Profundamente errado de todos os pontos de vista!
  - Não há história sem interpretação
  - Os primeiros adjectivos são os emotivos, tais como *bom* ou *mau*
  - A função expressiva da língua é tão (ou mais) importante que a informativa
- Ellis (1993) refere três enganos fundamentais
  - Que a função primordial da língua é a comunicação
  - Que há simples / intelectuais (redondo, quadrado) e complexos (bom)
  - Que uma classificação ou categorização relaciona/junta coisas diferentes

Ellis (1993)



- Os casos simples não interessam (podem ser explicados por qualquer teoria!)
- São os difíceis que são relevantes, que constituem um desafio às teorias
- Seja como for, é mais importante comunicar:
  - Estou feliz. ou O céu é azul.
- Categorização pode ser feita pela forma, pela função, ou pela atitude/emoção, e/ou por várias destas questões
  - Árvores, ervas daninhas, venenos, livros, poemas, guloseimas, ...

Padrões de subjectividade

- "Sentiment analysis" – "opinion mining" Pang & Lee (2008)
- Revisão da literatura focada em "acesso à informação" ☹ de uma das áreas de aplicação mais mexidas na presente década
- Detecção da polaridade
  - Positivo ou negativo
- Sistema de busca de opiniões
  - Quer opiniões ou factos?
  - Que parte dos documentos/frases apresenta opiniões?
  - Qual a opinião/sentimento dominante?
  - Como resumir/agregar as modas?

Alguns problemas na área das emoções

- Ironia, sarcasmo
- Razões para a atitude, em vez da atitude:
- Importância do domínio:
  - rápido – bom para carros, mau para comida
  - grande – bom para casas, mau para portáteis
  - pesado – bom para carros, mau para malas, ? para filmes ou livros
  - leve – bom para mobiliário (?), comida (?), ? para filmes ou livros
- Subjectivo: varia com o receptor/anotador
- Convenções culturais "c'est pas mal"
  - tu estás óptima
  - nível da expectativa: "maravilhosa"

Atitudes e avaliação

- boas notícias e más notícias
  - não só dependem de para quem
  - como não são necessariamente subjectivas (nem objectivas)
- *conservador* ou *socialista* (idem para *maluco* ou *terrível*)
  - bons ou maus conforme quem os profere
  - mesmo *honesto* ou *consequente* dependem de a quem se referem e caracterizam quem os profere
- contraste é muito eficiente... *Isto é horrível mas... adorei*
- seis emoções (em inglês?) / sete pecados mortais / dez mandamentos
  - anger, disgust, fear, happiness, sadness, and surprise

Exprimir opiniões ou estados de espírito

- Presença ou frequência?
  - Dizer dez vezes que é bom torna algo melhor do que dizer é bom uma vez?
- Número de hapax legomena são indicadores de subjectividade
- Criatividade também
- O poder subtil da negação
  - Nem tudo é o que parece
  - No melhor pano cai a nódoa
  - Não desgosto de ... Nada me agrada mais/menos do que ...
  - Evita a facilidade das tramas de Hollywood
- Mudanças na língua: *gay*, *rapariga*, *rabino*, *marechal*
- Avaliação de avaliações (recensões): foram úteis?

Interlúdio eleições americanas

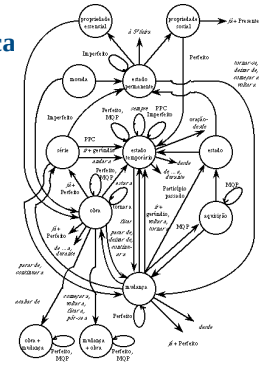
- A linguística com corpos (?) é usada activamente pelos meios de comunicação social e pelos analistas políticos actuais
- O discurso de Obama vs. Mc Cain
  - Quem diz mais "não"?
  - Quem usa mais "I" ou "You"?
  - Quem usa mais verbos ou mais substantivos?
  - Quem tem frases mais compridas?
- Quais os tópicos mais discutidos nos discursos de José Sócrates, ou de Lula? Quais as polémicas/notícias mais discutidas, sobre estes políticos, ao longo do tempo?
- Convenções de cores opostas às europeias: vermelho/azul

O processamento temporal

- Um hiato enorme entre os meios relacionados com o tempo (verbal) na língua e a concepção de tempo na nossa sociedade
  - Linguagem de "ler as horas"
  - Visão do passado e do futuro
  - Visão da história
- Tempo e aspecto
  - A categorização temporal está intimamente ligada à forma como categorizamos acontecimentos e processos
- Tradução do tempo
- Ontologias geográficas

Uma visão gráfica

O sistema aspectual do português



Línguas diferentes: causas

- contextos diferentes; falantes diferentes
- informação implícita diferente; constante uso criativo

Admira que as línguas divirjam???

It would surely be surprising, and a very strong empirical claim, that different languages using different means to express 'meanings' always arrived at exactly the same end (Keenan, 1978)

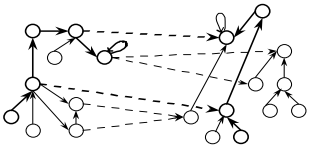
Línguas diferentes: resultados / factos

- gramática diferente
- itens lexicais diferentes
- convenções diferentes sobre implícito/explícito
- regras diferentes para obrigatoriedade e opcionalidade
- diferentes estratégias discursivas
- coisas diferentes que fazemos com as palavras
- diferentes metáforas convencionalizadas
- diferentes realidades descritas/facilmente invocadas
- diferentes provérbios/cultura

O modelo da rede de tradução

- Descrição independente das duas línguas (categorias diferentes em línguas diferentes)
- Pontes entre diferentes categorias
- Vagueza (vários pedaços de significado associados a um mesmo "operador")

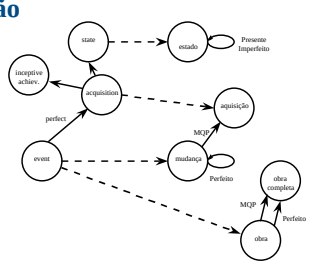
Ver Santos (2004)



A rede de tradução

- "pluperfect" traduzido de muitos modos
- Arcos de tradução nunca preservam o sentido

Ver Santos (2004)



Exemplo concreto: falha aspectual

- Àquele "sentido" não é possível associar o mesmo tipo de padrão aspectual nas duas línguas
  - ***Ficou*** *na estrada a vê-los afastarem-se*
  - *he stood in the road watching them leave*
  - *I saw him* ***sit*** *there all afternoon*
  - *No prédio fronteiro, viu o calafate* ***sentado*** *à mesa*
  - *In his house Kino* ***squatted*** *on his sleeping mat,* ***brooding****.*
  - *E Kino em casa,* ***acocorado*** *na esteira,* ***meditava*** *longamente.*

Tradução do tempo verbal

- *Kino's people had sung of everything that happened or existed.*
- *A gente de Kino cantara tudo o que acontecera ou existira*
- *A gente de Kino cantara sobre tudo o que* ***acontecia*** *ou* ***existia***

passado | presente

- *He was trapped as his people were always trapped*
- *Estava peado, como todos os da sua raça sempre tinham estado*
- *Estava peado, como todos os da sua raça sempre* ***estavam***

A tradução...

- Depende dos predicados
  - no original I loved her, I met her
  - na tradução gathered juntaram-se amontoavam-se
- Depende dos argumentos dos predicados
  - No original the road, the girl climbed
  - enganou-se enganou-o was mistaken, cheated
- Depende da situação de enunciação
  - Valha-me Deus!
- Depende da língua de chegada
  - Nossa senhora Madonna, Our lady

Ontologias geográficas

- Outro campo completamente diferente?
- Não é preciso corpos nem linguística?

Interesting issues (1)

- Names change, roles change!
- Topic: *African capitals...*

Santos & Cardoso (2008)

Típica pergunta do GikiCLEF

```
<title> Que guerras foram travadas na Grécia?
  </title>
<description>Documentos relevantes descrevem
  guerras passadas em solo grego, quer na Grécia
  antiga, quer na moderna.</title>
```

- Faz sentido marcar/computar Grécia como um lugar imutável?
- Faz sentido desenvolver ontologias geográficas separadas do tempo, ou do ponto de vista?
  - Malvinas/Shetland
  - Palestina/Israel
  - Europa de Leste, América do Norte, "Europa" (UE, futebol, festival da canção)

Voltando aos corpos na era da internetização

- Uma breve história (pessoal)
- Competências para fazer linguística com corpos
- Problemas com que a nossa comunidade se debate

História dos corpos (a minha versão)

Os tempos mudam...

- Corpus do português fundamental: obra monumental, mas anos e anos a digitar e a processar devido à falta de meios informáticos
- Em 1988, compilar um conjunto de mil frases independentes para testar uma gramática era obra! Obter texto electrónico era (quase) impossível, torná-lo público um feito ☺
- Em 1998, fazer um sistema que permitia a leigos consultar um corpo na rede era inovador
- Em 2000 criar o CETEMPúblico (200 milhões de unidades) era obra...
- Em 2004 criar o WPT03 (um instantâneo da Web) era obra...

Balanço da história dos corpos

- A questão do tamanho é uma falsa questão...
- A questão dos géneros também (o BNC passou do prazo? Kilgarriff et al. 2007)
- A questão da análise dos corpos será sempre o mais importante
- A questão das aplicações e dos usos é fundamental
  - O que se faz com os corpos, o que se extrai dos corpos
- Já não estamos na fase da recolha ou compilação, mas sim na fase da filtragem e da sumarização, ou da apresentação dos resultados
- Cada vez mais a linguística com corpos se vai entrelaçar, misturar, e esbater com a recolha de informação, à medida que esta inclui mais linguística, e à medida que a linguística precisa mais da RI

Dividir para reinar? O saber não é como a política

- *Contratei um engenheiro*
- *Contratei umas linguistas*

Em vez de mútuo respeito e aprendizagem

- *Eu sei de X, ele sabe de Y*

Áreas estanques com diferentes gurus e diferentes terminologias

Uns vêm de Letras, outros de informática... Mas o caminho a seguir é o mesmo, na área do processamento computacional da língua

Todos temos de saber

(Ou estar dispostos a aprender)

- Linguística
- Filosofia da linguagem
- Computação
- Métodos numéricos e estatísticos
- Métodos qualitativos

Todos temos uma responsabilidade que não podemos alienar

- Escrever bem
- Definir uma terminologia adequada
- Ensinar os mais novos

Crítica

Santos (2007)

- A maior parte das pessoas que usam as florestas não têm a noção do trabalho que lá está incluído, nem do que ainda falta ser feito
- A maior parte das pessoas que trabalham com florestas passam o tempo a criá-las ou a melhorá-las, não a usá-las
- As florestas são um investimento para o futuro, mas geralmente não acompanhado:
  - ainda não existem os utilizadores
  - os futuros utilizadores muito raramente exprimem os seus desejos e/ou necessidades (e quando os exprimem, precisam sempre de corpos muitíssimo maiores!)
  - os gramáticos (fora da equipa) estão aparentemente completamente desinteressados na existência de uma floresta sintáctica ou não para a sua língua

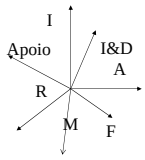
A *Linguateca* em poucas palavras

- > 2000 entradas 7 milhões de visitas ao nosso sítio
- AC/DC, CETEMPúblico, COMPARA … Recursos consideráveis
- *Morfolimpíadas* A primeira avaliação conjunta para o português, seguida pelo CLEF e pelo HAREM

- Recursos públicos
- Investigação e colaboração
- Comparação e medição formal
- Uma língua, muitas culturas
- Cooperação usando a Rede
- Não adaptar aplicações do inglês

A evolução do modelo IRA

- Inicialmente: Informação, Recursos e Avaliação
- Mais eixos foram surgindo
  - Manutenção (dos recursos)
  - Apoio ao utilizador
  - Investigação (teses)
  - Formação



Informação

- Portal
  - catálogo de projectos, actores, recursos e ferramentas
  - forum (notícias, bolsa de emprego e conferências)
  - ligações úteis (listas, informação técnica, revistas em português...)
  - um repositório de artigos ou ferramentas
- Um catálogo de publicações sobre a área
- Vários serviços ou sistemas na Web que dão acesso directo a recursos
  - Busca no AC/DC, COMPARA, Floresta
  - RAP no Esfinge
  - Corpógrafo: um ambiente para desenvolver terminologia e estudo de linguagens específicas
- Resposta a todas as perguntas que nos fazem!

Recursos (1)

- Damos acesso através da Rede
  - através de serviços
  - desenvolvendo ambientes
  - permitindo a invocação remota de ferramentas
- Tornamos disponível para ser "levantado"
  - corpos e listas de frequências
  - ambientes
  - ferramentas computacionais
- Nota: até 2004 ainda distribuíamos o CETEMPúblico em CD ☺

Recursos (2)

- **Corpógrafo**: um ambiente para fazer terminologia profissional
- **Esfinge**: um sistema de RAP (resposta automática a perguntas)
- Um atomizador robusto (e separador de frases) para português
- **NATools**: alinhadores de textos paralelos (frases e palavras)
- **WebJspell**: correcção ortográfica interactiva na Web

- CETEMPúblico, CETENFolha
- **Floresta Sintá(c)tica**: a primeira floresta sintáctica para o português
- **COMPARA**: o maior corpo paralelo editado do mundo

Avaliação conjunta

- Definir uma tarefa em conjunto
- Criar um processo de avaliação dessa tarefa
  - medidas
  - recursos
  - procedimento
- Comparar o desempenho dos vários sistemas participantes
- Tornar públicos os recursos, programas e resultados dos sistemas para
  - validação externa
  - investigação na tarefa e na metodologia de avaliação
  - organização de futuras edições
  - treino de novos participantes

Avaliação conjunta

- Morfolimpíadas (2003-2004)
  - Diana Santos (ed.). *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*, Lisboa: IST Press, 2007.
- CLEF: 2004, 2005, 2006, 2007, 2008
  - RI, RAP, RIG monolingue e cruzada
  - Actas pós-conferência publicadas pela Springer
- HAREM (2005-2007)
  - Diana Santos & Nuno Cardoso (eds.), *Reconhecimento de entidades mencionadas em português: o primeiro HAREM*. Linguateca, 2007.
- Segundo HAREM (2007-2008)
  - Cristina Mota & Diana Santos (eds.), *Desafios no reconhecimento de entidades mencionadas em português: actas do Segundo HAREM*. Linguateca, 2008.

Motivação para o HAREM

PUC-Rio 2006

- Estamos apenas a fazer o mesmo que já se fez, mas agora para português?
- Ou existem também questões científicas e de engenharia válidas a que podemos responder com esta actividade?

- É possível fazer ciência e engenharia para o português que sejam melhores do que as que foram feitas para o inglês
- embora o HAREM tenha sido feito de raiz para o português, como metodologia inovadora pode ser igualmente aplicado ao inglês ou a outra língua qualquer

É a mesma tarefa? "Só" português...

- Uma língua ser diferente é relevante?
- É só mudar os módulos (atomizador, ortografia) e os recursos (almanaques)? Adaptações menores...
- **Ou** uma língua diferente tem desafios diferentes? Assuntos diferentes sobre os quais as pessoas falam, convenções tipográficas diferentes, diferentes conceptualizações do mundo...
- Isto é uma questão que só pode ser resolvida empiricamente... experimentando ver como é para o português e depois comparando

PUC-Rio 2006

A mesma tarefa? Questões metodológicas

- Qual o conjunto de classificações que nos interessam?
- Como conseguir acordo na sua interpretação?
- É relevante a extensão a outros géneros?
- O conceito de *entidade mencionada* foi delimitado da mesma maneira? Os critérios operacionais são os mesmos?...
  - identificação parcial
  - proximidade ontológica
  - erros ortográficos, variantes diferentes
- A extensão a outros tipos de classificação é relevante?
- Como tratamos da vagueza, e da discordância (efeito de tecto)

PUC-Rio 2006

REM: categorias

ESSLLI 2007

Eça de Queirós nasceu na Póvoa de Varzim em 1845, e faleceu 1900, em Paris. Estudou na Universidade de Coimbra.

Eça de Queirós nasceu na ***Póvoa de Varzim*** em *1845*, e faleceu *1900*, em ***Paris***. Estudou na ***Universidade de Coimbra***.

Categorias semânticas I: ***Cidade***, *Ano*, ***Pessoa***, ***Universidade***

Categorias semânticas II: ***Lugar***, *Tempo*, ***Pessoa***, ***Organização***

Categorias semânticas III: ***Local administrativo***, *Data*, ***Escritor***, ***Instituição cultural***

Objectivos, passados e futuros

Santos (2007)

- 2000-2003 Estabelecimento no contexto do processamento do português (em Portugal e no Brasil)
- 2003-2005 Conhecidos em círculos internacionais como "guardiães do português"
- 2006-2008 A investigação feita sobre o português é reconhecida internacionalmente como relevante para o PLN em geral
  - HAREM, **ReRelEM**
  - CLEF: QolA, GeoCLEF, **GikiCLEF**
  - Floresta Sintá(c)tica
  - COMPARA

Mensagens principais desta palestra

- A linguística com corpos está no âmago, no centro, de todo o processamento de linguagem natural – até na recolha de informação geográfica ou nas aplicações inteligentes
- Não há "os linguistas" e os outros: redutor, e principalmente prejudicial para os linguistas!
- Saber outras línguas deve ser enriquecedor, e não empobrecedor, para a nossa língua... Por uma defesa dos direitos linguísticos, e para uma defesa do rigor e da qualidade na publicação em língua portuguesa
- A Linguateca é/foi um projecto para a língua portuguesa, onde quer que os falantes de português se encontrem, e é e foi **internacional** e **intercultural**

Bibliografia

- Ellis, John M. *Language, Thought and Logic*. Evanston IL: Northwestern University Press, 1993.
- Fillmore, Charles J. ' "Corpus linguistics" or "Computer-aided armchair linguistics" '. In Svartvik, J. (ed.), *Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82.* Berlin: Mouton de Gruyter, 1992, pp. 35–60.
- Gries, Stefan Th. "Some proposals towards more rigorous corpus linguistics". *Zeitschrift für Anglistik und Amerikanistik* **54**.2, 2006, pp. 191-202.
- Keenan, Edward L. "Some Logical Problems in Translation", in F. Guenthner & M. Guenthner-Reutter (eds.), *Meaning and Translation: Philosophical and Linguistic Approaches*, Duckworth, 1978, pp.157-89.

Bibliografia (cont.)

- Kilgarriff, Adam. “Comparing Corpora”. *International Journal of Corpus Linguistics* **6** (1), 2001, pp. 1-37.
- Kilgarriff, Adam, Sue Atkins & Michael Rundell. “BNC Design Model Past its Sell-by”. *Corpus Linguistics Conference*, Birmingham, UK, 2007.
- Nunberg, Geoffrey. “ The Non-Uniqueness of Semantic Solutions: Polysemy”. *Linguistics and Philosophy* **3** no2, 1979, pp. 143-184.
- Pang, Bo & Lillian Lee. “Opinion mining and sentiment analysis”. *Foundations and Trends in Information Retrieval* Vol. 2, No 1-2 (2008), pp. 1–135.

Bibliografia (cont.)

- Pustejovsky, James & Anna Rumshisky. “Between Chaos and Structure: Interpreting Lexical Data Through a Theoretical Lens”. *Int J Lexicography*, **21**(3), 2008, 337-355.
- Sampson, Geoffrey. “Thoughts on two decades of drawing trees”, in Anne Abeillé (ed.), *Treebanks: Building and using parsed corpora*, Kluwer Academic Publishers, 2003, pp. 23-41.
- Santos, Diana. *Translation-based corpus studies: Contrasting Portuguese and English tense and aspect systems.* Amsterdam/New York, NY: Rodopi, 2004.
- Santos, Diana. “Corporizando algumas questões”, in Stella Tagnin & Oto Araújo Vale (eds.), *Avanços da Lingüística de Corpus no Brasil*, Editora Humanitas, São Paulo, 2008, pp. 41-66.

Bibliografia (cont.)

- Silva, Rosário, Susana Inácio & Diana Santos. "Documentação da anotação relativa à cor no COMPARA". Primeira versão: 27 de Novembro de 2007. Em constante redacção.
- Sinclair, John. *Corpus, Concordance, Collocation (Describing English Language)*. Oxford University Press, Aug 1991.
- Sorace, Antonella & Frank Keller. “Gradience in linguistic data”. *Lingua* **115**, 2005, pp. 1497-1524