Linguateca: sete anos devotados ao processamento computacional do português

Diana Santos
www.linguateca.pt

O que é a Linguateca?

- Um projecto de infraestrutura para o processamento computacional da língua portuguesa
- proposto no seguimento da reflexão em torno da área feita para o Livro Branco (1999)
- é preciso desenvolver recursos para a nossa língua
- é preciso organizar avaliações conjuntas para aumentar a qualidade dos sistemas
- é preciso estabelecer um portal que informa sobre a área

Um soneto...

Sete anos de pastor Jacob servia
Labão, pai de Raquel, serrana bela;
mas não servia o pai, servia a ela,
e a ela só por prémio pretendia.

Os dias, na esperança de um só dia,
passava, contentando-se com vê-la;
porém o pai, usando de cautela,
em lugar de Raquel lhe dava Lia.

Vendo o triste pastor que com enganos
lhe fora assi negada a sua pastora,
como se não a tivera merecida,

Começa de servir outros sete anos,
dizendo: Mais servira, se não fora
para tão longo amor tão curta a vida!

Linguateca, um centro de recursos distribuído

- Projecto gerido pela FCCN, financiado pelo POSI

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

www.linguateca.pt

Participantes/capital humano

- séniores: *Diana Santos, José João Almeida, Eckhard Bick, Belinda Maia, Ana Frankenberg Garcia, Mário J. Silva, Paulo Gomes, Luís Costa*
- colaboradores: *Luís Miguel Cabral, Susana Inácio, Rosário Silva, Ana Sofia Pinto, Paulo Rocha, Cláudia de Freitas, Sérgio Matos, Hugo Oliveira, Pedro Martins*
- bolseiros de doutoramento: *Rachel Aires* (2005), *Marcirio Chaves, Alberto Simões, Nuno Seco, Anabela Barreiro, Nuno Cardoso*
- passado: *Signe Oksefjell, Susana Afonso, Raquel Marchi, Renato Haber, Alex Soares, Pedro Moura, Débora Oliveira, Isabel Marcelino, Cristina Mota, Luís Sarmento, António Silva, Rui Vilela*

34 pessoas ao todo relacionadas com a Linguateca

Uma organização virtual

A *Linguateca* em poucas palavras

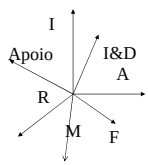
- > 1000 entradas 3 milhões de visitas ao nosso sítio
- AC/DC, CETEMPúblico, COMPARA … Recursos consideráveis
- *Morfolimpíadas* A primeira avaliação conjunta para o português, seguida pelo CLEF e pelo HAREM

- Recursos públicos
- Investigação e colaboração
- Comparação e medição formal

- Uma língua, muitas culturas
- Cooperação usando a Web
- Não adaptar aplicações do inglês

A evolução do modelo IRA

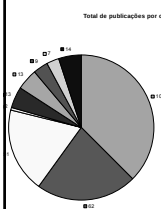
- Inicialmente: Informação, Recursos e Avaliação
- Mais eixos foram surgindo
  - Manutenção (dos recursos)
  - Apoio ao utilizador
  - Investigação (teses)
  - Formação

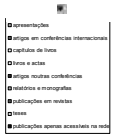


Teses no âmbito da Linguateca

- Rachel Aires: Uso de marcadores estilísticos para a busca na Web em português **(Agosto 2005)** [Sandra Aluísio & Diana Santos]
- Alberto Simões: Parallel Corpora word alignment and applications **(Setembro 2004)** [José João D Almeida]
- Nuno Cardoso: Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas **(Outubro 2006)** [Mário Silva & Eugénio Oliveira]
- Luís Miguel Cabral: SUPeRB - **S**istema **U**niformizado de **Pe**squisa de **R**eferências **B**ibliográficas. (Março 2007) [Diana Santos & Eugénio Oliveira]

Publicações





cerca de 350 publicações e apresentações desde 1998

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 Web
  - através de serviços
  - desenvolvendo ambientes
  - permitindo a invocação remota de ferramentas
- Tornamos disponível para ser "levantado"
  - corpora e listas de frequências
  - ambientes
  - ferramentas computacionais

- Nota: até 2004 ainda distribuímos 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 separados de frases) para português
- **NATools**: alinhadores de textos paralelos (frasese e palavras)
- **WebJspell**: correcção ortográfica interactiva na Web

- CETEMPúblico, CETENFolha
- **Floresta Sintá(c)tica**: a primeira floresta sintáctica para o português
- **COMPARA**: o maior corpus 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
  - 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.
- HAREM II (2007-)

REM, *reconhecimento de entidades mencionadas*

- Identificação e classificação de nomes próprios (e expressões numéricas) em texto -- em português

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***.

PUC-Rio 2006

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.

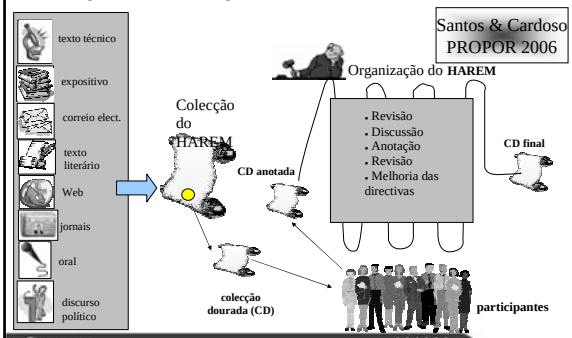
Eça de Queirós nasceu na ***Póvoa de Varzim*** em *1845*, e faleceu *1900*, em ***Paris***. Estudou na ***Universidade de Coimbra***.

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

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

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

Criação da colecção dourada do HAREM



A Linguateca em termos de projectos

- AC/DC
- Floresta Sintá(c)tica
- Morfolimpíadas
- COMPARA
- CLEF/CHAVE/ Esfinge
- HAREM
- Corpógrafo
- NATools
- WPT03
- PAPEL

Objectivos, passados e futuros

- 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
  - CLEF: QolA, GeoCLEF, ...
  - **Floresta Sintá(c)tica**
  - COMPARA