www.linguateca.pt

Linguateca: presente e futuro

Diana Santos & Marcirio Chaves

Estrutura da apresentação

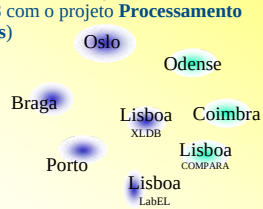
- ✓ Linguateca – Panorâmica
  - ✓ Origem, Objetivos, Resultados
  - ✓ Equipe, Investigação, Estrutura
- ✓ HAREM

Linguateca, um projeto para o português

- ✓ Centro de recursos distribuído para o processamento computacional da língua portuguesa
- ✓ Projecto financiado pela FCT através do POSI (2000-2006)
- ✓ Primeiro pólo no SINTEF ICT, Oslo, começou em 2000 (actividade no SINTEF começou em 1998 com o projeto **Processamento Computacional do Português**)

Modelo IRA

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



Linguateca num relance

- ✓ > 1000 links Mais de 1.500.000 visitas ao site

- ✓ Recursos públicos
- ✓ Incentivar a investigação e colaboração
- ✓ Medida e comparação formal

- ✓ Uma língua, muitas culturas
- ✓ Cooperação usando a Web
- ✓ Não à adaptação direta das aplicações para o inglês

A origem da Linguateca

- ✓ Resultado da participação no Livro Branco, que identificou
- ✓ Problemas: falta de ...
  - ✓ recursos públicos
  - ✓ cooperação entre os grupos, Brasil e Portugal
  - ✓ avaliação
  - ✓ esforço na manutenção e disponibilização de recursos
- ✓ Soluções: Projeto piloto dedicado à
  - ✓ Criação de recursos públicos (desenvolvimento, questões legais, etc.)
  - ✓ Organização de avaliações conjuntas
  - ✓ Criação de um portal dedicado à área
- ✓ Em rede (juntando mão-de-obra a grupos de investigação de acordo com os pressupostos da Linguateca)

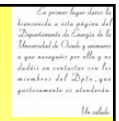
Alguns objetivos da Linguateca

- ✓ Fazer com que o PLN do português seja tão qualificado como o das outras línguas
- ✓ Impedir que as pessoas continuassem a trabalhar em PLN do inglês com a desculpa de que não havia recursos para o português
- ✓ Evitar que os grupos jogassem fora (ou guardassem secretamente) os seus recursos em vez de os disponibilizar, ajudando-os e contribuindo para essa tarefa
- ✓ Conseguir colaboração entre os vários países de língua portuguesa para tratarem todas as variantes e não só a "sua"
- ✓ Medir o progresso em várias áreas cimentando e incrementando a colaboração entre os vários atores (avaliações conjuntas)

Alguns resultados: Informação

- ✓ Portal constantemente actualizado, www.linguateca.pt
- ✓ Catálogo de recursos, atores e publicações
- ✓ Resposta a todos os usuários
- ✓ Manutenção de um repositório
- ✓ Documentação sobre os recursos criados pela Linguateca
- ✓ Informação sobre as avaliações conjuntas
- ✓ Publicações no âmbito da Linguateca

Alguns resultados: Recursos



- ✓ Serviços na Web para dar acesso a corpora e ferramentas
  - ✓ AC/DC (Acesso a corpora/Disponibilização de corpora)
  - ✓ COMPARA
  - ✓ Esfinge
  - ✓ SIEMÊS
- ✓ Várias ferramentas
  - ✓ Atomizadores e separadores de frases
  - ✓ Sistemas de REM
  - ✓ Alinhadores à palavra
- ✓ Criação de corpora, colecções, ou dados para distribuição
  - ✓ CETEMPúblico, CETENFolha, WPT03
  - ✓ GKB (*Geographic Knowledge Base*) e Geo-Net-PT01
  - ✓ REPENTINO (REPositório para reconhecimento de ENTidades com NOme)
  - ✓ Colecção douradas: CHAVE, Morfolimpíadas e HAREM

Alguns resultados: Avaliações conjuntas

- ✓ Selecionar uma área
- ✓ Criar recursos para a avaliar, em consenso com os participantes
- ✓ Criar programas de avaliação
- ✓ Organizar um evento
- ✓ Publicar os resultados

- ✓ Morfolimpíadas (análise morfológica sem contexto)
- ✓ CLEF (RI cruzada e Resposta Automática a Perguntas - RAP)
- ✓ HAREM (Reconhecimento de Entidades Mencionadas - REM)

Estrutura da Linguateca

- ✓ Pólo XLDB de Lisboa: Web portuguesa, RI e RI geográfica
- ✓ Pólo do Porto: terminologia, corpora especializados, avaliação de tradução automática
- ✓ Pólo de Braga: ferramentas, tradução automática, gestão e validação de recursos
- ✓ Pólo de Oslo: organização, portal, avaliações conjuntas, RAP
- ✓ Pólo COMPARA: corpora paralelos
- ✓ Pólo Odense: floresta sintática
- ✓ Pólo Coimbra: ontologias lexicais

Quem é o público/usuários da Linguateca?

- ✓ Pessoas envolvidas no desenvolvimento de aplicações de PLN
- ✓ Consumidores de dados (linguistas)
- ✓ Utilizadores de programas que envolvem PLN

Quem é a equipe?

- ✓ Sêniores: *Diana Santos, José João Dias de Almeida, Elisabete Ranchhod, Eckhard Bick, Belinda Maia, Ana Frankenberg Garcia, Mário J. Silva, Paulo Gomes*
- ✓ Contratados para as tarefas "básicas" da Linguateca (um por pólo): *Luís Costa, Nuno Cardoso, Rui Vilela, Luís Miguel Cabral, António Silva*
- ✓ Contratados à tarefa: *Paulo Rocha, Susana Afonso, Raquel Marchi, Rosário Silva*
- ✓ Doutorandos: *Marcirio Chaves, Alberto Simões, Nuno Seco, Luís Sarmento*
- ✓ Bolsistas para tarefas mais curtas: *Susana Inácio, Ana Sofia Pinto*
- ✓ Amigos/associados: *Rachel Aires, Cristina Mota, Anabela Barreiro*

Investigação (para tese de doutorado)

- ✓ Rachel Aires: É possível categorizar os textos da Web segundo as necessidades de informação dos usuários? **(concluída Agosto 2005)**
- ✓ Marcirio Chaves: É possível gerar ontologias geográficas úteis a partir da análise de textos em português?
- ✓ Alberto Simões: É possível aumentar significativamente os exemplos usados na Tradução Automática baseada em exemplos (TABE) com o processamento inteligente de corpora comparáveis?
- ✓ Nuno Seco: Métodos de criação e de avaliação de uma ontologia lexical para o português
- ✓ Luís Sarmento: Análise semântica robusta

Investigação (para tese de mestrado)

- ✓ Alberto Simões: Alinhador à palavra **(concluída Setembro 2004)**
- ✓ Luís Miguel Cabral: Extrator de informação na rede para o catálogo de publicações
- ✓ Rui Vilela: Extração de informação

Outras investigações feitas na Linguateca

- ✓ Métodos de criação de uma floresta sintática (treebank)
- ✓ Métodos de RAP
- ✓ Anotação sintática do português
- ✓ Usabilidade com base nos diários (logs) de serviços na Web
- ✓ Métodos de avaliação
- ✓ Detecção de entidades mencionadas (EM)
- ✓ Criação de serviços na rede
- ✓ Avaliação de recursos

HAREM é Avaliação de Reconhecimento de Entidades Mencionadas

- ✓ Problema: Identificar e classificar nomes próprios em contexto em texto em português, dada uma tabela inicial e quanto à morfologia
- ✓ A forma mais básica de semântica
- ✓ Três tarefas:
  - ✓ Identificar uma EM
  - ✓ Classificá-la morfologicamente
  - ✓ Classificá-la pelo tipo de entidade a que se refere
- ✓ Organização de uma avaliação conjunta
  - ✓ Criar uma colecção dourada anotada com as soluções
  - ✓ Fornecer aos sistemas participantes grandes quantidades de texto (coleção HAREM)
  - ✓ Avaliar (através da comparação automática com as soluções)

O HAREM

- ✓ Calendário:
  - ✓ Iniciado em setembro 2004
  - ✓ Registo dos participantes finalizado em outubro 2004
  - ✓ Coleção HAREM distribuída: 14 fev. 2005
  - ✓ Prazo final para a recepção dos resultados dos sistemas: 16 fev. 2005
  - ✓ Resultados finais: outubro de 2005
  - ✓ Encontro em 2006
- ✓ Participantes:
  - ✓ 10 sistemas (2 Brasil, 1 Dinamarca, 1 Espanha, 1 México, 5 Portugal)
- ✓ Organizadores: Diana Santos, Nuno Cardoso, mais
  - ✓ Coleção dourada: Paulo Rocha, Susana Afonso, Anabela Barreiro
  - ✓ Avaliação: Nuno Seco, Rui Vilela

Originalidade do HAREM: organização

- ✓ Separação da tarefa em três
  - ✓ identificação
  - ✓ classificação semântica (categoria e tipo)
  - ✓ classificação morfológica (género e número)
- ✓ Uso de vários gêneros textuais
  - ✓ jornalístico,Web, entrevistas, texto técnico, literário, email, ...
- ✓ Distribuição de uma colecção com meta-dados
  - ✓ variante, gênero textual, ...

Originalidade do HAREM: colecção dourada

- ✓ Classificação das EMs em contexto
  O Brasil venceu a Copa (PESSOA $_{GRUPO}$), O Brasil assinou o tratado (ORGANIZACAO $_{ADMINISTRACAO}$), O Brasil tem muitos rios (LOCAL $_{ADMINISTRATIVO}$), Por amor ao Brasil (ABSTRACCAO $_{IDEIA}$), ...
- ✓ Novas categorias motivadas para o português
  - ✓ PESSOA, ORGANIZAÇÃO, LOCAL, VALOR
  - ✓ ABSTRACÇÃO, COISA, OBRA, ACONTECIMENTO
- ✓ Tratamento de vagueza
  - ✓ categorias |
  - ✓ uso do marcador <ALT> para delimitar diferentes alternativas de identificação
- ✓ Diretivas precisas
  - ✓ quanto à interpretação
  - ✓ quanto à delimitação

Originalidade do HAREM: Método de avaliação

- ✓ Cenário global e selectivo
  - ✓ possível escolher um subconjunto de categorias e/ou tipos
- ✓ Avaliação absoluta ou relativa (na classif. semântica e morfológica)
  - ✓ contando com as correctamente identificadas, ou com todas as EMs na CD
- ✓ Considerar parcialmente identificadas
- ✓ Várias medidas originais
- ✓ Resultados
  - ✓ por categoria
  - ✓ por género
  - ✓ por variante

Resultados do HAREM

- ✓ Colecção dourada pública
- ✓ Arquitectura pública (programas em Perl e Java)
- ✓ Dez sistemas prontos a atacar o problema de REM em português (quantos haveria sem o HAREM?)
- ✓ Uma primeira medida do estado da técnica em português
- ✓ Objectivos científicos
  - ✓ Medir a dificuldade do problema para o português
  - ✓ Pôr em relevo as especificidades do português
  - ✓ Verificar se as EMs podiam ser discriminadoras de gênero textual

Considerações Finais

- ✓ Panorâmica Linguateca
- ✓ Modelo IRA
- ✓ HAREM
- ✓ Escola de Verão (Junho)
- ✓ Geo-Net-PT01 - http://xldb.fc.ul.pt/geonetpt