Linguateca

Um centro de recursos distribuído para o processamento computacional da língua portuguesa

Diana Santos, Alberto Simões, Ana Frankenberg-Garcia, Ana Pinto, Anabela Barreiro, Belinda Maia, Cristina Mota, Débora Oliveira, Eckhard Bick, Elisabete Ranchhod, José João Dias de Almeida, Luís Cabral, Luís Costa, Luís Sarmento, Marcirio Chaves, Nuno Cardoso, Paulo Rocha, Rachel Aires, Rosário Silva, Rui Vilela, Susana Afonso

Taller de Herramientas y Recursos Lingüisticos
para el Español y el Portugués @ IBERAMIA

Linguateca – Porquê?

- Reconhecimento do PLN como uma área estratégica pelas autoridades portuguesas para a Ciência e Tecnologia
  - Livro Verde para a Sociedade de Informação
- Necessidade de existência de recursos públicos e gratuitos
  - Criação de infra-estrutura de disponibilização
- Estimulo à colaboração entre actores

Linguateca – Porquê?

- Processamento Computacional do Português (vs. PLN Geral)
- As agências de distribuição internacionais (LDC ou a ELRA) não poderão competir, para o português, com uma organização dos próprios falantes e investigadores
  - *controlo* de qualidade
  - *gestão de prioridades* dos recursos humanos

Linguateca – Quem somos?

- Linguateca possui vários Pólos
  - Pólo de Oslo
  - Pólo de Odense no VISL
  - Pólo de Lisboa do COMPARA
  - Pólo de Braga
  - Pólo de Lisboa no LabEL
  - Pólo do Porto
  - Pólo de Lisboa no XLDB

- Diferentes competências, sinergias e interesses mas...

1 Único Objectivo!

Linguateca – Quem somos?

- 5 Colaboradores a tempo inteiro
- 5 Colaboradores a tempo parcial
- 4 Bolseiros de doutoramento
- 5 Bolseiros
- Mais
  - 6 Responsáveis Científicos (Pólos)
  - Vários parceiros habituais
  - Vários colaboradores pontuais

Objecto de trabalho

- Lingua português, em todas as suas variantes
- Trabalhamos em colaboração com:
  - NILC
  - UFRGS
- Por motivos burocráticos não é possível ter Pólos da Linguateca no Brasil...

Modelo IRA

- I: Informação
  - www.linguateca.pt
  - ...
- R: Recursos
  - Disponibilização de corpora
  - Disponibilização de ferramentas
  - ...
- A: Avaliação
  - Organização de Avaliações Conjuntas
  - Morfolimpiadas, CLEF, HAREM...

I: Informação

- Sítio na rede www.linguateca.pt
  - facilitar o acesso aos recursos **já** existentes
  - 1.7M+ visitas desde Julho de 1998
  - Catálogo:
    - publicações sobre PLN do português
    - recursos existentes
      - Corpora, Léxicos e dicionários, Enciclopédias e Tesauros...
    - projectos e actores
      - Grupos, centros e institutos, Projectos, Projectos europeus, Projectos internacionais, Assoc. e instituições, Empresas...
    - Ferramentas
      - Ferramentas para o português, Ambientes de Desenvolvimento e Ferramentas para outras línguas

I: Informação

- Sítio na rede www.linguateca.pt
  - Informação interessante
    - Gestão da área, Cursos na rede, Revistas electrónicas, Listas electrónicas, Grupos de discussão
  - Forum
    - bolsa de emprego, notícias e conferências
  - Sistema de procura dedicado:
    - Busca
  - Serviço de Perguntas / Respostas de auxílio a todos os interessados
    - 90 perguntas respondidas desde 01/04 sobre vários temas

R: Recursos

- Desenvolvimento colaborativo de recursos e ferramentas para o português
- Disponibilização gratuita de toda a produção
- Desenvolvimento contínuo e melhoria dos recursos actuais: vários projectos

R: Recursos

- AC/DC
- COMPARA
- CETEMPúblico & CETENFolha
- Floresta Sintá(c)tica
- AnELL
- Corpógrafo
- NATools
- CHAVE
- WPT 03
- Esfinge
- M. Trad. Distribuídas
- Museu da Pessoa
- GREASE
- Linguarudo
- TrAva e Corta

R: Recursos

- AC/DC (Acesso a Corpora/Disponib. Corpora)
  - 250M+ de palavras em português, nos registos jornalístico, literário, didáctico e correio electrónico
- COMPARA
  - O maior corpus paralelo revisto, com textos em português e inglês e as suas traduções. Interface de pesquisa DISPARA
- CETEMPúblico e CETENFolha
  - Dois corpora de texto jornalístico de grandes dimensões separados em extractos, e integralmente disponíveis.
- WPT 03
  - A maior recolha de documentos da web portuguesa (14Gb texto). Cerca de 3,5 milhões de documentos + log de registos das pesquisas no tumba!

R: Recursos

- Floresta Sintá(c)tica
  - Primeiro “treebank” para a língua portuguesa. Árvores analisadas pelo analisador sintáctico PALAVRAS. Texto proveniente dos corpora CETEMPúblico e CETENFolha).
- AnELL: Anotador Electrónico LabEL Linguateca
  - Serviço público de anotação automática de textos, via web, que utiliza o INTEX. Possibilidade de revisão manual.
- Corpógrafo
  - Plataforma web que permite coleccionar textos em vários formatos, formar e analisar corpora, extrair terminologia e criar bases de dados terminológicas e ontologias.

R: Recursos

- NATools
  - pacote que inclui um alinhador à frase e um outro à palavra, um gerador de dicionários probabilísticos, um módulo de classificação/avaliação da probabilidade de tradução de dois textos, um extractor de terminologia bilingue multi-palavra
- TrAva e CorTA
  - O TrAva é uma ferramenta construída essencialmente para a criação de material de teste para a tradução automática. O CorTA é Corpus de Traduções automáticas Avaliadas
- CHAVE
  - colecção de documentos do Público de 1994 e 1995 + conjunto de perguntas e suas respostas para fazer avaliação conjunta de sistemas de resposta automática a perguntas.

R: Recursos

- Esfinge
  - Sistema de resposta a perguntas de domínio geral em português, que implementa a arquitectura descrita por Brill* e que explora a redundância existente na rede. Participação na tarefa Q&A do CLEF2004
- Memórias de Tradução Distribuídas
  - Serviço em desenvolvimento rede destinado a permitir empresas de tradução, comunidades de tradutores ou mesmo tradutores independentes consultar as memórias de tradução de outros tradutores
- Museu da Pessoa
  - A Linguateca associou-se ao Museu da Pessoa português para tirar partido das histórias recolhidas: fonte de corpora orais.

• Eric Brill. “Processing Natural Language without Natural Language Processing”,
•in A. Gelbukh (ed.), CICLing 2003, LNCS 2588, Springer-Verlag Berlin Heidelberg, 2003, pp. 360-369

R: Recursos... E Projectos

- GREASE (Geographic Reasoning for Search Engines)
  - Métodos, algoritmos e arquitecturas informáticas para que um sistema auxilie o utilizador a encontrar páginas na rede, escritas em língua portuguesa, *com um âmbito geográfico próximo à sua localização*.
- Linguarudo
  - Estudo da arquitectura de RI para a Web usando PLN em português
  - apresentação dos resultados conforme as necessidades do usuário.
  - Disponibilização de um corpus de páginas da Web brasileira categorizadas por tipo de necessidade de informação.

A: Avaliação

- Motivar e Organizar Avaliações Conjuntas
  - Reunião dos investigadores
  - Partilha de experiências
  - Motivar os investigadores para trabalhar em conjunto
  - Discussão dos critérios de avaliação entre os participantes
  - Troca da diferentes pontos de vista sobre a área
  - Esforços iniciados em 2002

A: Avaliação

- Muito trabalho envolvido...
  - Calendarização
  - Organização Geral
  - Preparação dos recursos
  - Desenvolvimento dos programas de medição
  - Determinação e publicação de resultados
  - Organização de Encontro
  - Discussão de resultados
  - Publicação de documentação
  - Distribuição de recursos construídos
- ... mas compensador!

A: Avaliação

- Morfolimpiadas
  - Avaliação de analisadores morfológicos, verificadores ortográficos e radicalizadores
  - 7 participantes
  - Março a Junho de 2003
- AVALON' 2003
  - Encontro de Avaliação Conjunta de Sistemas de Processamento Computacional do Português (livro no prelo)
- CLEF 04
  - participação na organização do CLEF incluindo o português nas pistas de RI e RP.
  - participação no exercício de avaliação com os sistemas Esfinge e tumba!

A: Avaliação

- Em preparação: HAREM
  - reconhecimento de entidades mencionadas
  - 17 participantes
  - Avaliação até 19 de Janeiro de 2005
  - Produção cooperativa de um corpus de referência com marcação manual de EM

Perspectivas

- Divulgação de recursos e investigação continua a ser prioritária
  - fomenta a discussão
  - junta os investigadores
  - acelera a investigação
- Trabalhar na produção e disponibilização de recursos continua a ser prioritário
  - para quê partir sempre do zero?
- Continuar a organizar Avaliações Conjuntas:
  - boa forma de fomentar a discussão sobre uma área específica
  - Possibilidade de aferir o estado de desenvolvimento
  - Perceber quais a prioridades/necessidades de investigação

Conclusão

- Projecto com mais de 6 anos
  - Disponibilização gratuita de informação e recursos a muitos utilizadores e investigadores
  - Interacção com centenas de investigadores das mais variadas áreas
  - Organização de Avaliações Conjuntas (participação)
  - Desenvolvimento de massa-crítica e apoio a grupos de investigação em várias insitituições
  - Motivação de investigadores para o estudo do português
  - ...
- Contributo geral para o desenvolvimento da Processamento Computacional da língua portuguesa
- Continuaremos a trabalhar...