Linguateca: objectivos e resultados

Diana Santos

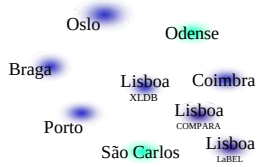
Linguateca, um centro de recursos distribuído

- Centro de recursos -- distribuído -- para o **processamento computacional da língua portuguesa**
- Projecto financiado pela FCT/POSI (2000-2006), POSC (2007-2008)
- Primeiro pólo em Oslo desde 2000 (actividade no SINTEF começou em 1998 com o projeto *Processamento Computacional do Português*)

Modelo IRA

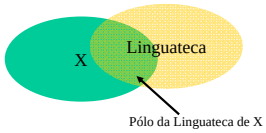
- **I**nformação
- **R**ecursos
- **A**valiação

www.linguateca.pt



Pormenores de organização

- Núcleo no SINTEF ICT
- Todos os outros pólos incluídos numa organização que faz I&D na área do processamento computacional da língua portuguesa
- Número de pessoas envolvidas ao longo dos 10 anos: 30 a 40



A evolução em termos de eixos de actuação

Linguateca num relance

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

- Recursos públicos
- Incentivar a investigação e colaboração
- Medida e comparação formal
- Uma língua, muitas culturas
- Cooperação usando a Rede
- Não à adaptação directa 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 objectivos da Linguateca: sonhos or realidade?

- 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 deitassem 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 actores (avaliações conjuntas)

Serviço à comunidade

- Não quisemos competir com a comunidade, mas sim criar condições e dados para, em conjunto, irmos mais longe
- Todos os recursos são grátis, e públicos
- Não fazemos diferença entre empresas e investigação: queremos que as empresas a façam, e que os investigadores ganhem...
- Segunda fase: avançar para projectos com mais impacto na sociedade, para demonstrar também que a área do processamento computacional da língua portuguesa pode servir para mais do que a própria comunidade que nele está envolvida

Resultados de dez anos de actuação

- Provavelmente o sítio com mais informações sobre o processamento de uma língua (de todas as línguas do mundo): www.linguateca.pt
- Bem conhecido em Portugal e no Brasil e pela comunidade internacional
- Um conjunto de recursos e ferramentas testados e documentados que podem ser usados por todos
- Estudos sobre português (RI, RIG, TA, extracção automática de terminologia, RAP, etc.)
- Materiais pedagógicos em português
- Um grupo razoável de pessoas treinadas na área e muitas outras com algum conhecimento do assunto e dos problemas

Podemos orgulhar-nos de, na Linguateca, ...

- Termos organizado a primeira avaliação conjunta para português
- Termos criado a primeira floresta (*treebank*) para o português
- O primeiro serviço de corpos linguísticos na rede para o português
- O primeiro sistema de resposta automática a perguntas na Rede para o português
- O maior corpo paralelo anotado e revisto do mundo
- O primeiro instantâneo da Rede correspondente a um país
- O primeiro ambiente público semi-automático de extracção de terminologia para o português
- Uma ontologia lexical para o português gratuita

Os fracassos da Linguateca

- As pessoas (re)usam sem citar nem dar crédito
- Alguns grupos recebem financiamento para fazer o que já há feito sem qualquer impunidade
- Muitas pessoas comparam os resultados unilateralmente com os das avaliações conjuntas sem participarem
- A maior parte das pessoas prefere participar em avaliações conjuntas/conferências "internacionais" embora sejam menos interessantes em termos científicos
- As pessoas preferem publicar na Springer (com comitês de programa falando português) e/ou em (mau) inglês

Pode dizer-se que isto é fora da nossa competência, mas é claramente contrário ao que pretendíamos

Exemplos de sucessos

Avaliações conjuntas

- HAREM: avaliação conjunta de reconhecimento de entidades mencionadas

Participaram, no total de duas edições, 17 grupos diferentes

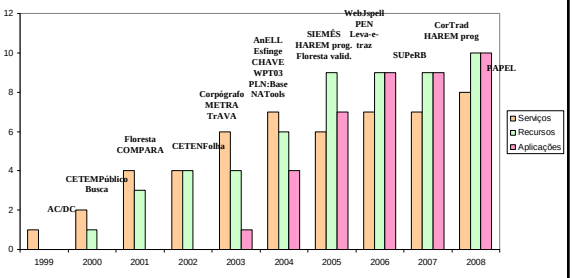
- CLEF: recolha de informação cruzada e resposta automática a perguntas

Participaram, no total de seis edições (2004 a 2009), c. 30 grupos diferentes

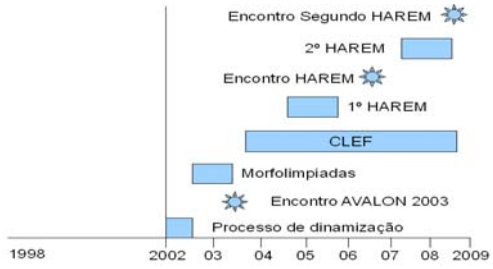
Recursos e sua disseminação

- CETEMPúblico: 578 grupos/pessoas além de a sua consulta ser pública
- AC/DC e COMPARA: 385 mil acessos/perguntas
- Corpógrafo: 1785 utilizadores

Recursos, aplicações e serviços e respectiva duração



Calendário de iniciativas de avaliação



Publicação em português

- Três livros em português e um em inglês



- Várias contribuições em revistas brasileiras e em volumes portugueses e brasileiros
- ... além da chamada "publicação internacional"
- Desenvolvimento de um sistema dedicado e motivado para tratar de referências **em português** ou referências a portugueses noutras línguas

SUPeRB

O futuro...

- Queremos fazer um portal de acesso ao material da Biblioteca Nacional com a tecnologia desenvolvida e testada que permita navegação inteligente na literatura e nos conteúdos em português
- Queremos desenvolver um sistema de relacionamento e descoberta para a produção científica em português
- Desejamos dotar o arquivo da web portuguesa de capacidades de pesquisa inéditas
- Pretendemos continuar a desenvolver sistemas de melhoria e apoio à edição, procura e manipulação da Wikipédia e de outros sistemas de disseminação de conteúdos em português
- Continuando a apoiar e estimular a publicação de recursos