Linguateca 10 anos: festejo ou luto?

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*)

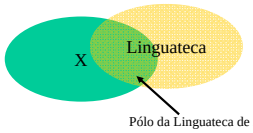
Modelo IRA

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

www.linguateca.pt

Pormenores de organização

- Núcleo no SINTEF ICT
- Todos os outros pólos são incluídos numa organização que faz I&D na área do processamento computacional da língua portuguesa



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 Web
- 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 estivemos a competir com a comunidade
- Não criámos algo para ficar
- Há algumas iniciativas que talvez mereçam permanecer
  - Avaliações conjuntas
  - Novos corpos e mais informação / anotação dos recursos existentes
  - ???
- Muitos recursos podem ser considerados como suficientemente bons para poderem ser usados e melhorados pela comunidade

No fim (2008)

- Provavelmente o sítio com mais informações sobre o processamento de uma língua (de todas as línguas do mundo): continuará na FCCN
- 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

Financiamento

Período	Total	SINTEF	
Maio 1998 – Maio 2000:	270 k€	270 k€	
Maio 2000 – Maio 2003:	1.4 M€	970 k€	
Maio 2003 – Dez. 2006:	1.8 M€	1 039 k€	
Dez. 2006 – Dez. 2008:	1.7 M€	900 k€	
10 anos e 7 meses	4.8 M€	3.2 M€	(66%)

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

Objectivo da presente reunião

- O que é/foi a Linguateca?
- O que correu bem e mal durante a sua existência?
- O que poderia ter sido feito melhor?
- Há coisas a fazer e/ou maneiras de nos organizarmos que devam ser tentadas?
- Há futuro para o tipo de actividades que iniciámos ou para outras?

- Como ficará patente, somos todos muito diferentes...