

Processamento computacional do português: o que é?

Diana Santos

SINTEF Telecom and Informatics

Plano da apresentação

Razões para a identificação desta categoria

O projecto *Processamento computacional do português*

Estado da arte e da técnica

- através do catálogo
- através da área do PLN

Processamento computacional do português

- ♦ = processamento de linguagem natural; em que a linguagem natural é o português
- ♦ Não existe PLN «geral», «independente da língua» (apesar de poderem existir técnicas e metodologias que o possam ser)
- ♦ Em geral PLN não adjectivado significa «processamento do inglês»
- ♦ Para avaliar, é preciso ter um contexto / língua em que se trabalha

Trabalhar numa língua e apresentar os resultados noutra

- ♦ Criação de «textos» incoerentes
- ♦ Baseada em hipóteses não confirmadas
 - que a estrutura dos textos é idêntica
 - que a semântica dos tempos verbais correspondentes é idêntica

Last month I bought a house. It had an aquarium. Mary offered me a red fish. John gave me his frog. My fish died yesterday. It stopped breathing. It became blue. It went to the top of the aquarium.
(COLING'92, p. 335)

O projecto Proc. Comp. do Port.

- ♦ Criado pelo MCT por altura da discussão das áreas para o Livro Branco
- ♦ Três vertentes principais:
 - Catálogo
 - Criação e disponibilização de recursos
 - Avaliação
- ♦ Futuro: criação de um centro de recursos para o processamento do português

Acção política

- ♦ Identificação de alguns problemas
- ♦ Sugestão de algumas soluções
- ♦ Criação de um portal que apresenta o proc. computacional do português, bem divulgado a nível internacional
- ♦ Estabelecimento de cooperação com actores internacionais (Bick, LDC, IMS, Davies, ...)



Trabalho realizado : Recursos

- ♦ Projecto AC/DC: criação de uma interface na Web para todos os corpora livremente acessíveis
- ♦ CETEMPúblico: criação de um corpus de 180 milhões de palavras para distribuir
- ♦ COMPARA/DISPARA: criação de uma interface na Web para corpora paralelos (criação do próprio corpus: Ana Frankenberg-García)



Trabalho realizado: Catálogo

- ♦ Recursos
- ♦ Actores
- ♦ Publicações
- ♦ Informação interessante
- ♦ Sistema de busca
- ♦ Série de programas desenvolvidos para a manutenção e expansão do catálogo



Trabalho realizado : Avaliação

- ♦ Conjugadores de verbos
 - ♦ Alinhadores
- Trabalho relacionado
- ♦ Comparação de corpora
 - ♦ Motores de procura em português
 - ♦ Ambientes de processamento de corpora em português
 - ♦ Analisador sintáctico

Sem dúvida a área mais difícil !



Futuro : Centro de recursos

Além da disponibilização crescente de recursos ...

- ♦ Organização de conferências de avaliação
- ♦ Concursos para criação de recursos
- ♦ Especificação de materiais de referência
- ♦ Novos pólos: Braga, Lisboa, ...
- ♦ Conselho consultivo internacional




A situação actual (fonte: Web)

- ♦ 25 corpora
- ♦ 170 léxicos e/ou dicionários
- ♦ 80 ferramentas
- ♦ 500 publicações
- ♦ 44 páginas pessoais
- ♦ 35 empresas
- ♦ 2-12 listas electrónicas




Uma panorâmica do PLN

- Escrita e produção de um texto
- Leitura e folheamento
- Tradução
- Aprendizagem e ensino
- Sistemas de informação
- Sistemas interactivos
- Indexação
- Entrada de dados
- Segurança e identificação



Ajuda à escrita

- ▣ correctores
- ? ambientes de redacção
- «controladores» (linguagens controladas)
- ? documentação automática
- ajuda à digitação
- ▣ sistemas de ditado
- formataadores
- ▣ comunicação para deficientes



Ajuda à leitura

- extracção de informação
- sumarização
- «text mining» (garimpo de textos)
- ▣ leitura em voz alta automática
- visitas guiadas
- ? recuperação de informação multilingue




Tradução

- ▣ tradução automática
- facilitadores de tradução
- navegadores por textos noutras línguas
- estações de trabalho para tradutores
- ? tradução de fala




Ensino e aprendizagem

- ▣ sistemas tutores
- enciclopédias inteligentes
- ▣ jogos didácticos
- ▣ criação automática de exercícios



Consulta a informação

- interrogação telefónica
- perguntas a enciclopédias
- perguntas à rede
- ▣ sistemas conselheiros
- publicação personalizada



Sistemas interactivos

- ? comandos falados
- interacção com «assistentes»
- simulação animada de instruções
- reportagem automática
- jogos com interacção

Indexação

- ? criação semi-automática de tesauros
- ▣ indexação de grandes quantidades de material
- catalogação de imagens
- ? criação semi-automática de terminologias

Entrada e colecção de dados

- ? extracção automática de informação para as BDs internas de um sistema
- leitura óptica
- actualização de dados por fala
- ▣ reconhecimento de manuscritos

Segurança e identificação

- ? reconhecimento de assinaturas
- ? reconhecimento da «voz do dono»
- ▣ reconhecimento da língua / substracto linguístico
- identificação automática de um autor

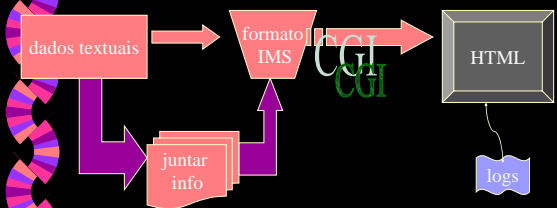
Outros

- ? estudos sociológicos
- análise de uma disciplina científica
- medição do «peso» de uma dada cultura
- criação de metáforas para uma boa interface
- ▣ planificação do ensino de uma língua

O projecto AC/DC

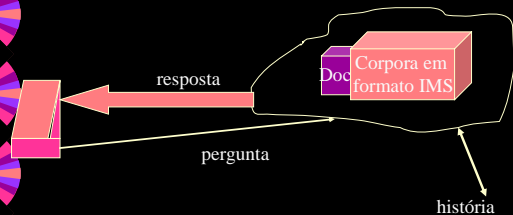
- ♦ AC/DC= Acesso a corpora/Disponibilização de corpora
- ♦ Três fases
 - Separação de frases e tokenização
 - Análise sintáctica (com Eckhard Bick)
 - Junção de outra informação (com outros investigadores)
- ♦ Futuro: “test suite” e “treebank”

AC/DC: o processo



Os dados variam de texto simples a SGML
A informação que juntamos vai desde marcar os limites dos parágrafos até analisar sintacticamente o texto

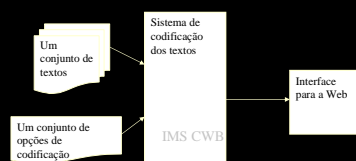
ACDC: o resultado



<http://cgi.portugues.mct.pt/aceso/>

Documentação dos corpora

Distribuir corpora na rede



CETEMPúblico

- ♦ Grande quantidade de texto do jornal Público distribuído aleatoriamente por extractos numerados
- ♦ Distribuído em
 - CD em formato texto
 - CDs em formato CQP
 - pelo AC/DC
- ♦ Esperança de vir a ser um corpus de referência para avaliação de sistemas

<http://cgi.portugues.mct.pt/cetempublico/>

O projecto COMPARA/DISPARA

- ♦ Iniciativa de Ana Frankenberg-Garcia, implementação do nosso projecto
- ♦ *COMPARA* é um corpus paralelo português-inglês
- ♦ *DISPARA* é um sistema de distribuição de corpora paralelos na rede, que foi desenvolvido numa primeira instância para o COMPARA, mas que é para ser utilizável por todos os interessados

<http://www.portugues.mct.pt/COMPARA/>

Características originais do DISPARA

- ♦ Inspecção das notas de tradução (N.T.)
- ♦ Procura por tipo de alinhamento (1:2, 1:1/2, etc.)
- ♦ Panorâmica quantitativa
- ♦ Procura nas duas direcções
- ♦ Procura arbitrariamente complexa nos dois lados

Alguns pormenores sobre o COMPARA

Textos com tamanhos diferentes

Unidade de alinhamento é sempre uma frase do original

Não segue o TEI