AC/DC: acesso a corpora / disponibilização de corpora

Diana Santos
Paulo Rocha
Linguateca
www.linguateca.pt

Linguateca

- O projecto processamento computacional do português lançou a Linguateca (centro de recursos distribuído)
  - informação e visibilidade
  - aumento do número de recursos
  - avaliação

Recursos: corpora

- Identificação do que existia e qual a forma de acesso (Catálogo)
- Agilização e facilitação do acesso
- Aumento da informação associada
  - género; estrutura (frases, parágrafos, etc.)
  - anotação morfossintáctica (com Eckhard Bick)
- Criação de novos corpora

Criação de novos corpora

- Para (também) distribuição física
  - CETEMPúblico
  - CETENFolha
- Para ter mais géneros: literário, correspondência electrónica, semanário, local e partidário
- Para ter outras funcionalidades:
  - COMPARA (com Ana Frankenberg-Garcia)
  - Floresta Sintá(c)tica (com Eckhard Bick ++)

O projecto AC/DC

- Por razões históricas consideramos, estritamente, o projecto AC/DC como a parte que dá acesso aos corpora monolingues através da rede (excluindo o COMPARA e a Floresta)
- Dois pontos de acesso
  - http://acdc.linguateca.pt/acesso/
  - http://cetempublico.linguateca.pt
- Vários serviços

Vários serviços

- Acesso a concordâncias
- Acesso a distribuição (formas, fontes, características morfossintácticas)
  - por corpus
  - em todos os corpora
- Acesso a ”ordem” de frequência (ranking)
- Acesso a informação quantitativa detalhada sobre cada corpus

Mapa do artigo

- Panorama dos corpora do projecto AC/DC
- Estudo sobre os utilizadores do projecto AC/DC (usando os logs)
- Nova funcionalidade: ”Alfaiate de corpora”, corpus personalizado: escolha da constituição do corpus por tipo de texto (já usado no ensaio das Morfolimpíadas)

Estudo dos utilizadores

- A maior parte dos utilizadores de serviços electrónicos não fornecem informação sobre a sua identidade ou motivos
- Mas é possível observar o seu comportamento através das pegadas que deixam
- Esta é uma área em grande expansão na informática, que nós pretendemos aplicar aos corpora

Estudo dos utilizadores 2

- Três tipos de utilizadores
  - Utilizadores casuais
  - Utilizadores desencorajados
  - Utilizadores persistentes
- Queremos diminuir o número dos segundos e dar melhor serviço aos terceiros
- Idealmente até cativar os primeiros

Alfaiate de corpora

- A maior parte dos corpora a que damos acesso são monogénero
- Os tamanhos de cada corpus e de cada género são diferentes
- É portanto necessário aplicar “regras de três”
- É possível com um mínimo de processamento criar corpora à medida: dada uma proporção de géneros e de tamanho final desejado

Alfaiate de corpora 2

- Exemplo de corpus usado para o ensaio das morfolimpíadas
  - Igual proporção entre as duas variantes (PE e PB)
  - Texto literário original e traduzido em igual proporção
  - 5% de linguagem oral
  - variação máxima de género e origem
- Serviço na rede, no pólo de Braga, muito brevemente

Comentários finais

- É indiscutível que a Linguateca tem vindo a aumentar significativamente (o acesso a)os recursos de corpora para o português
- Ainda há uma comunidade muito pequena de investigadores e linguistas que saibam usar corpora
- É preciso ter contacto com os utilizadores para melhorar o serviço (respondemos através de projecto@informatics.sintef.no)