Corpora Comparáveis

Belinda Maia
FLUP

Citações de:

- EAGLES - **Expert Advisory Group on Language Engineering Standards**
- **Guidelines – 1996 – at:**
- http://www.ilc.pi.cnr.it/EAGLES96/browse.html

Comparable corpora - definition

- "A comparable corpus is one which selects similar texts in more than one language or variety. There is as yet no agreement on the nature of the similarity, because there are very few examples of comparable corpora".

Comparable corpora - uses

- "The possibilities of a comparable corpus are to compare different languages or varieties in similar circumstances of communication, but avoiding the inevitable distortion introduced by the translations of a parallel corpus".

Corpora Comparáveis (CC)

- Textos que reflectem as convenções da cultura subjacente
- Textos legais do sistema legal local
- Textos socialmente convencionados: participações de falecimentos, anúncios de casas ou empregos
- Textos académicos / científicos – convenções diferentes em culturas e domínios diferentes

CC – Vantagens

- Disponibilidade de mais textos – e mais variedade
- Versatilidade para investigação em:
  - Análise do discurso
  - Pragmática
  - Pesquisa de informação
  - Engenharia do conhecimento

O que torna textos/corpora COMPARÁVEIS?

EAGLES - quotes

- "A comparable corpus is one which selects similar texts in more than one language or variety".
  - Similar – em mais do que uma língua

E/OU

  - Similar – em variedade

"... circunstâncias similares de communicação.."

Similares em – Forma/conteúdo?

- Forma
  - Tamanho> num. de palavras, frases, parágrafos
  - Tamanho dos textos
  - Formato - .txt, .doc, .html, .xml
- Conteúdo
  - Linguagem geral
  - Domínios especializados

Similares em – Estrutura / Função?

- Estrutura
  - Textos formais e bem construídos – ex. Textos legais
  - Discurso informal – ex. Transcrições de conversas
- Função
  - Social
  - Cultural

Similares - Registo?

- Registo
  - 'Field' – situação, assunto, etc
  - 'Tenor' – relações interpessoais
    - ex. formal/informal, cortesia, etc
  - Modo
    - Falado: ex. diálogo formal / informal
    - Escrito: ex. livro, artigo, manual de instruções
    - Multimedia: ex. Encarta, cinema, internet

Similar - Dialecto?

- Dialecto
  - Geográfico > ex. Áreas urbanas/rurais, países desenvolvidos / em desenvolvimento
  - Temporal > ex. Período histórico, grupos etários diferentes
  - Social > ex. Classes sociais, níveis de educação

Comparabilidade de Corpora Muito Grandes

- Corpora Muito Grandes são comparáveis se:
  - Similares em tamanho
  - Construídos segundo os mesmos critérios > ex. quantidade e qualidade de tipos de texto
- Por exemplo?
  - British National Corpus
  - Mannheimer Corpus

Comparabilidade de corpora jornalísticos

- Corpora jornalísticos variam em:
  - Tipo: qualidade/popular, conteúdo geral/especializado
  - Data de publicação: o mesmo dia/mês/ano
- Por exemplo?
  - CETEMPúblico – Português
  - Corpus da Reuter – Inglês

Comparabilidade em Corpora Literários

- Período:
  - Medieval, Século XVIII, Pós-guerra
- Escola:
  - Romantismo, Realismo, Pós-modernismo
- Género (Genre):
  - Romance, ficção científica, teatro, poesia

Comparabilidade em corpora técnicos e científicos - forma

- Panfletos
- Manuais
- Livros de ensino
- Artigos e comunicações
- Dissertações, teses

Comparabilidade em corpora técnicos e científicos - conteúdo

- Informação geral
- Informação enciclopédica
- Instruções
- Educação
- Comunicação entre peritos

Construir CC – linguagem geral

- Começar?
- Corpora Muito Grandes comparáveis em 2 ou mais línguas >só a página da Comissão Europeia!
- Corpora gerais, cuidadosamente seleccionados – ex. ICAME corpora (Brown, LOB etc) = possível mas limitado

Utilizar CC de linguagem geral

- Vantagens:
  - Investigação comparativa a todos os níveis
  - Úteis para investigação do léxico e das estruturas sintácticas
- Desvantagens:
  - Dificuldades para análise mais cuidadosa
  - Desnecessários para certos tipos de análise

Construir CC – Textos de jornais

- Fáceis de adquirir
- Grande variedade de temas
- Comparáveis a vários níveis
- Possível arranjar versões diferentes da 'mesma' notícia – 'concurrent corpora'

Utilizar CC – Textos de jornais

- Comparação de tratamento de notícias:
- ex.
  - Política – campanhas eleitoriais
  - Futebol durante a Taça Mundial
- OU > estilos de jornalistas individuais

Construir CC – linguagem geral + textos semelhantes

- Tipos de texto semelhante: ex. Entradas em enciclopédias, publicidade turística
- Textos literários do mesmo autor, período, escola ou 'genre'
- Textos técnicos e científicos com uma forma ou função semelhante – ex. Livros de ensino

Utilizar CC – linguagem geral + textos semelhantes

- Pragmática
- Análise do discurso
- Análise de 'genre'
- Análise sociolinguístico
- Análise cultural

Construir CC – domínios especializados

- Domínios especializados a níveis diferentes – ex.
  - Geografia > demografia da população > minorias étnicas ...
  - Engenharia > engenharia mecânica > tribologia ...
  - Medicina > oncologia > cancro da mama

Utilizar CC – domínios especializados

- Análise de 'genre'
- Extracção de terminologia
- Pesquisa de informação
- Tecnologia de 'browsers'
- Engenharia do conhecimento

A construcção de CC - deve

- Estabelecer uma política geral em relação:
  - À forma – estrutura computacional
  - Ao conteúdo dos sub-corpora
  - À possibilidade de partilhar os recursos com um público geral ou restrito
- Especificar os objectivos para a construção dos sub-corpora

A construção de CC - deve

- Respeitar os direitos de autor ('copyright')
- Lembrar factores contextuais e/ou exteriores ao texto
  - O estilo e particularidades do autor individual
  - As convenções da escrita numa situação específica cultural/social
  - Efeito homogenizante da internacionalização
    - 'Eurospeak'
    - Anglicismos em terminologia científica e técnica

CC - limitações

- Qualquer corpus pequeno é construído para um fim específico – ex. extracção de terminologia, análise de 'genre', estudo contrastivo de línguas
- Estes corpora têm um uso específico e um 'prazo de validade'

Corpógrafo - permite

- Construir CC em várias línguas
- Analisar a linguagem dos textos com concordâncias, n-gramas, etc
- Extrair terminologia semi-automaticamente
- Criar bases de dados terminológicas
- Extrair definições e relações semânticas semi-automaticamente