Pesquisa de informação em *corpora*

Tópicos de análise lexical e sintáctica

Aula 2

Seminário de Mestrado, 2012

CLUL | Centro de Linguística da Universidade de Lisboa

Compilação

- gravar instâncias de língua falada

  situações "naturais"; situação mais ou menos formal; monólogo/ diálogo/ conversa; boas condições de captação de som; filmagem opcional (relação prosódia/gesto)

  diferentes tipos de oralidade:

  1º texto: oral espontâneo

  2º texto: oral preparado

  3º texto: sequências de leitura de texto escrito + sequências de oral espontâneo

- Seleccionar textos escritos
  - obtidos em formato electrónico
  - descarregados
  - em papel

Direitos de autor

- autorização dos falantes gravados
  exemplo1 exemplo2
- autorização dos detentores dos direitos de autor dos textos escritos
- pode determinar de que forma o *corpus* estará acessível (concordâncias / acesso integral)
- estabelecer a autoria e os direitos de autor do *corpus* (equipa ou instituição)
- direito à citação de um texto (máximo de 400 palavras) disponibilização de concordâncias com esse tamanho máximo

Compilação

- transcrição das instâncias de língua falada - normas
- digitalização dos textos em papel (OCR)

  digitalização, correcção, revisão
- limpeza dos ficheiros descarregados

  etiquetas html + eliminação de sequências que não pertencem ao texto com base em unidades multilexicais

  original

  limpo

Metadata

- informação bibliográfica
- informação sobre a estrutura do texto escrito: volume, capítulo, parágrafos, títulos, figuras, quadros, citações em língua estrangeira, etc.
- documentar cada gravação: situação, intervenientes
- informação sobre os ficheiros (texto ou texto/som/vídeo) domínio, local onde o ficheiro está armazenado, autor da gravação, autor da transcrição, etc.
  - analytic metadata
  - descriptive metadata
  - editorial metadata
  - administrative metadata

Exemplo metadata CRPC metadata META-NET

Metadata no *corpus* escrito CRPC exemplos

- ***Metadata* descritivos do texto: Informação bibliográfica**

Jornal/revista
  - Nome do jornal/revista
  - Secção
  - Número do jornal/revista

Livro
  - Número do volume
  - Colecção

Livro didáctico
  - Nome da disciplina curricular
  - Ano de escolaridade

Metadata no *corpus* escrito CRPC exemplos

- ***Metadata* descritivos do texto: autor**
  - Ano de nascimento do autor
  - Língua materna do autor
  - País de nascimento do autor
  - Local de nascimento do autor
  - País do autor (não é país de nascimento, mas país de cuja variedade é representativo)

- ***Metadata* descritivos do ficheiro**
  - Página (dimensão do excerto)
  - Número de linhas
  - Número de palavras

Metadata no *corpus* oral CRPC exemplos

- ***Metadata* descritivos da situação de gravação**
- ***Metadata* descritivos dos ficheiros de texto, som, alinhamento**
  - Nome do ficheiro de som
  - Nome do ficheiro de texto
  - Suporte físico original
- ***Metadata* descritivos do entrevistador**
- ***Metadata* descritivos do informante**

Metadata no *corpus* oral CRPC exemplos

- ***Metadata*** **analíticos (categorização):**
  - **Situação**
  - **Tema**
- ***Metadata*** **editoriais**
  - **Transcrição**
  - **Revisão**
- ***Metadata*** **administrativos**
  - **ID**
  - **Nº de utilizações**
  - **Projecto**

Representação da informação: *standards*

- Formatos: TXT, SGML, XML
- XML - Extensible Markup Language

  coralrom_XML coralrom_XML_ALG

- TEI Guidelines (Text Encoding Initiative)
  SGML (Standard Generalised Markup Language)
- CES – Corpus Encoding Standard

Representação da informação: *standards*

- DTD – Document Type Description
  representação formal que dá informação sobre os elementos que constituem o texto e que o descrevem
  coralrom
  - cabeçalho (header)
  - corpo do texto (body)

Armazenamento e preservação

- assegurar *backup* dos textos escritos
- gravações: armazenar cópia local, cópia em CD para uso e cópia em CD para armazenamento
- formatos rapidamente ultrapassados
- nova preocupação com *standards* e manutenção de bases de dados de grandes dimensões

Anotação

- acrescentar informação linguística ao *corpus* original
  - morfo-sintáctica
  - sintáctica
  - semântica
  - tipologia de erros num corpus de aprendizagem (L2) ou de aquisição
- problema da utilidade da anotação
- anotação manual > *corpus* de treino
- preparar um esquema de anotação e um manual que permitam consistência entre os vários anotadores exemplo manual cintil
- testar a consistência da anotação (concordâncias)
- Inter-annotator agreement

Anotação
Máximas de Leech

1. It should be possible to remove the annotation from an annotated corpus in order to revert to the raw corpus.
2. It should be possible to extract the annotations by themselves from the text.
3. The annotation scheme should be based on guidelines which are available to the end user.
4. It should be made clear how and by whom the annotation was carried out.

Anotação
Máximas de Leech

5. The end user should be made aware that the corpus annotation is not infallible, but simply a potentially useful tool.
6. Annotation schemes should be based as far as possible on widely agreed and theory-neutral principles.
7. No annotation scheme has the a priori right to be considered as a standard.

Anotação morfo-sintáctica

Categorias principais: N, V, ADJ, PREP, etc

Com/PREP as/DA desastrosas/ADJ notas/CN de_/PREP
os/DA exames/CN nacionais/ADJ ainda/ADV bem/ADV frescas/ADJ em_/PREP a/DA memória/CN ,*//PNT é/V altura/CN de/PREP
perguntar/INF :*//PNT o/DA que/REL se/CL passa/V com/PREP as/DA nossas/POSS crianças/CN .*//PNT

Corpus CINTIL (NLX-FCUL / CLUL)

Anotação morfo-sintáctica

Categorias principais: N, V, ADJ, PREP, etc

\+ Categorias secundárias: flexão nominal e verbal

Com/PREP as/DA#fp desastrosas/ADJ#fp notas/CN#fp de_/PREP os/DA#mp exames/CN#mp nacionais/ADJ#mp ainda/ADV bem/ADV frescas/ADJ#fp em_/PREP a/DA#fs memória/CN#fs ,*//PNT é/V#pi-3s altura/CN#fs de/PREP perguntar/INF#nifl :*//PNT o/DA#ms que/REL se/CL#gn3 passa/V#pi-3s com/PREP as/DA#fp nossas/POSS#fp crianças/CN#fp .*//PNT

Corpus CINTIL (NLX-FCUL / CLUL)

Anotação morfo-sintáctica e lematização

Categorias principais: N, V, ADJ, PREP, etc

\+ Categorias secundárias: flexão nominal e verbal

\+ Lematização

Com/PREP as/DA#fp desastrosas/**DESASTROSO**/ADJ#fp notas/**NOTA**/CN#fp de_/PREP os/DA#mp exames/**EXAME**/CN#mp nacionais/**NACIONAL**/ADJ#mp ainda/ADV bem/ADV frescas/**FRESCO**/ADJ#fp em_/PREP a/DA#fs memória/**MEMÓRIA**/CN#fs ,*//PNT é/**SER**/V#pi-3s altura/**ALTURA**/CN#fs de/PREP perguntar/**PERGUNTAR**/INF#nifl :*//PNT o/DA#ms que/REL se/CL#gn3 passa/**PASSAR**/V#pi-3s com/PREP as/DA#fp nossas/POSS#fp crianças/**CRIANÇA**/CN#fp .*//PNT

Corpus CINTIL (NLX-FCUL / CLUL)

Anotação

Nome próprios multi-lexicais (Entidades nomeadas - Named entities)

PER	designação de pessoa
ORG	de organização
LOC	de local
WRK	de obra (livros, filmes, quadros, etc)
MSC	restantes casos

(B - início da expressão; I - restantes palavras da expressão
O - tokens que não pertencem a expressões)

Washington/PNM[B-ORG] acompanhou/ACOMPANHAR/V#ppi-3s[O]
os/DA#mp[O] movimentos/MOVIMENTO/CN#mp[O] de/PREP[O]
Saddam/PNM[B-PER] desde/PREP[O] a/DA#fs[O] primeira/ORD#fs[O]
hora/HORA/CN#fs[O] .*//PNT[O]

Corpus CINTIL (NLX-FCUL / CLUL)
Lista de etiquetas

CLUL

Anotação

Predicados complexos

- CV predicado complexo formado por verbo + verbo
  CVR reestruturação **Ele não o queria ver.**
  CVC causativos **O professor fez comer a sopa aos meninos.**

- CN predicado complexo formado por verbo + nome
  CNB nome sem determinante **dar ajuda**
  CN com determinante **dar uma ajuda**

- Posição canónica e Posição no contexto
  depois de um[CN2_B] aviso[CN3_I] dado[CN1_E]

Anotação

Predicados complexos

- Ambiguidade: [CVC_VINF]
- Sobreposição: 2 etiquetas no elemento que pertence a dois predicados complexos

não o queriam[CVR1_B] deixar[CVR2_E][CVC_VINF1_B] fugir[CVC_VINF2_E]