Pesquisa de informação em *corpora*

Tópicos de análise lexical e sintáctica

Aula 1

Amália Mendes
Iris Hendrickx
Michel Généreux

Seminário de Mestrado, 2012

CLUL | Centro de Linguística da Universidade de Lisboa

O *corpus*

- O que é?
- Qual é a sua utilidade?
- Como se desenha?
- Como se compila?
- Como documentar o conteúdo do *corpus*?
- Como preservá-lo?
- Como extrair informação?
- Como acrescentar informação ao *corpus*?

CLUL

O que é um *corpus*?

- *corpus*, sg; *corpora*, pl.
- "(…) a corpus is a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description." (Kennedy 98:1)
- Compilação planeada e estruturada
- *Corpus* electrónico (*machine-readable*)

CLUL

Linguística de *corpus*

- É uma metodologia?
- É uma teoria linguística?
- Dados do *corpus* vs. / e dados de introspecção

CLUL

Linguística de *corpus*

Crítica:

- não fornece todos os contextos necessários para a análise
- atenção centrada em dados de frequência

> "if natural scientists felt it necessary to portion out their time and attention to phenomena on the basis of their abundance and distribution in the universe, almost all of the scientific community would have to devote itself exclusively to the study of interstellar dust."
> Michael Polanyi, apud Fillmore, C. (1992)

why should I think that what you tell me is true

why should I think that what you tell me is interesting

CLUL

Dados do *corpus* e teoria linguística

Abordagens:

- **corpus-based (que se apoia nos dados do *corpus*)**
- **corpus-driven (que deriva dos dados do *corpus*)**

"The areas that traditional studies have neglected turn out to be the strengths of corpus-based studies of grammar."
(Biber et al. 98:56)

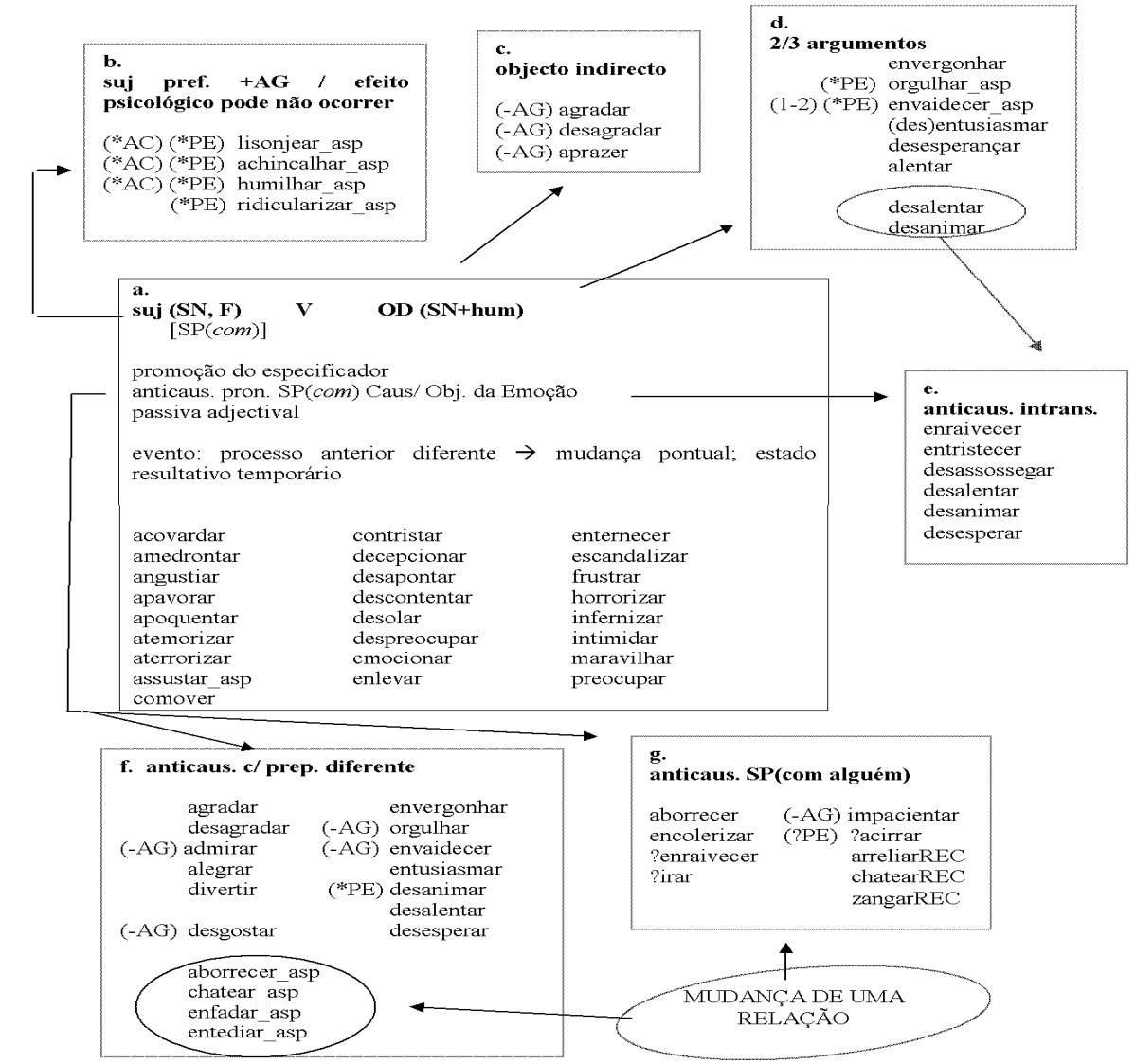
"Corpus linguistics are concerned typically not only with what words, structures or uses are possible in a language but also with what is probable – what is likely to occur in language use." (Kennedy 98:8)

CLUL

Novos tópicos de investigação

- Expressões multi-lexicais
- Categorias vagas ‘fuzzy’ vs. categorias fechadas
  protótipo e grau de pertença
- Predicados verbais pertencentes a uma classe semântica: dados do *corpus* apontam para uma variação sintáctico-semântica muito superior à referida em estudos apenas teóricos
  S*lide* seguinte: variação na classe dos verbos psicológicos

CLUL



(*PE) inaceitabilidade da promoção do Especificador; (*AC) inaceitabilidade da construção anticausativa; (-AG) sujeito não agentivo
(1) estado [+temporário]; (2) estado [-temporário]; _asp diferenças aspectuais
REC interpretação recíproca possível

Tipos de *corpora*

- *corpus* geral

  equilibrado, com textos de diferentes géneros e temas, escritos e orais (BNC, ICE)
- *corpus* de língua escrita / de língua falada (*spoken*)
- *corpus* de fala (*speech*)
- *corpus* de referência

  Corpus de Referência do Português Contemporâneo (CRPC)
- *corpus* monitor (aberto ou dinâmico) vs *corpus* estático

  Bank of English
- *corpus* especializado

CLUL

Tipos de *corpora*

- *corpus* dialectal
  Atlas e CORDIAL (CLUL)
- *corpus* de variedades do português
  VAPOR (CLUL)
- *corpus* de aprendizagem
  corpus de L2 (Isabel Leiria)
- *corpus* diacrónico
- *corpora* comparáveis e *corpora* paralelos
- *corpus* de tradução
- *corpus* de aquisição

CLUL

Desenho do *corpus*

- definir o tipo de *corpus* que se pretende e planear a sua constituição

- para a selecção dos textos, é necessário estabelecer:
  - quantos géneros devem estar representados
  - quantos textos de cada género
  - o tamanho de cada amostragem de texto incluída

    CRPC: início, meio e fim do ficheiro

CLUL

Desenho do *corpus*

representatividade

- o *corpus* como amostragem da totalidade da variação textual de uma língua

- "representative in the sense that findings based on an analysis of it can be generalized to the language as a whole or a specified part of it" Leech (91)

CLUL

Desenho do *corpus*

representatividade

- selecção dos textos com base em critérios externos e não internos

- número variado de géneros: tipos textuais usados em determinado contexto social e com uma intenção comunicativa específica (escrito e oral)

- análise das propriedades de cada género pode ajudar a melhorar a representatividade do *corpus*
  (*factor analysis,* Biber)

CLUL

Desenho do *corpus*

equilíbrio

- peso dos diferentes géneros num *corpus* geral, estabelecido com base em critérios de representatividade
- tendência para uma maior secção de língua falada nos *corpora* mais recentes
- selecção de informantes: idade, sexo, nível de escolaridade, localidade, influências linguísticas

CLUL

Desenho do *corpus*

- *corpus* de variedades de uma língua: critérios para classificação dos falantes ou dos autores como nativos

  (certos autores podem não ser representativos de uma variedade (ex: Mia Couto))

- maior exigência em trabalhos de sociolinguística em relação à proporção de informantes

- repetições nos jornais > seleccionar jornais de dias de semanas diferentes

  distorção nos estudos lexicais e análise das expressões multilexicais

CLUL

Desenho do *corpus*

tipos de texto

- retórica: narrativo, descritivo, argumentativo, explicativo ou informativo (tipo de texto ou tipo de sequência, cf. Jean-Michel Adam)

- géneros (ou registos, Biber): tipos estabelecidos socialmente, que ocorrem num contexto específico (cartas, notícias, editoriais, testamento, leis, etc.)

CLUL

Desenho do *corpus*

Alguns critérios de selecção de textos:

- ficção vs. não ficção
- fonte: livro, revista, jornal
- formal ou informal
- idade, origem dos autores
- ano de publicação / produção
- menor ou maior divulgação de textos
- frequência de leitura de um texto
- disponibilidade dos textos
- tópico

CLUL

Desenho do *corpus*

tamanho dos textos ou amostragens

- Biber: uma amostra de 2000-5000 palavras é suficientemente grande para representar determinada categoria de texto

- o número típico de amostras de cada género (20-80 textos) num *corpus* como o LOB é suficiente para estudos sobre variação

  Corpus Lancaster-Oslo/Bergen (LOB) 1M escrito

CLUL

Desenho
exemplos de *corpora*

British National Corpus (BNC) - 100M

International Corpus of English (ICE) – 1M

Corpus de Referencia del Español Actual – 160M

CLUL

Constituição do BNC
British National Corpus

- **British National corpus**
  - modelo seguido para outros *corpora* nacionais
  - Synchronic (informative from 1975, imaginative from 1960)
  - Not subject-specific
  - Monolingual British English
  - Spoken (10%) and written (90%)
- **BNC - Written Texts**
  - Reception and production
  - Published text is representative of written language that is received
  - Some selected from Books in Print for 1992
  - Others according to selection features
- **Target Selection Features**
  - Domain - 75% informative, 25% imaginative
  - Time - informative from 1975-, imaginative from 1960-
  - Medium - 60% books, 30% periodicals, 10% miscellaneous

CLUL

Constituição do BNC

- **Samples**
  - Maximum 45,000 words
  - Continuous stretch of discourse from the whole
  - Only one sample from any one text
  - Taken randomly from beginning, middle or end of long texts
  - Items in composite material (e.g. newspapers) marked by domain
- **Descriptive Features for Written Texts**
  - Author details: type, age, domicile, gender
  - Target audience: age, gender, level
  - Place of publication - regional
  - Reception status: low - high
- **Miscellaneous Written Material**
  - 7,000,000 words
  - Published - brochures, fact sheets etc
  - Unpublished - school essays, company memoranda etc
  - Written to be spoken - scripted material for broadcasts

CLUL

Constituição do BNC

- **BNC - Spoken Texts**

  4,000,000 words of spontaneous conversational English obtained by demographic sampling
  - 124 adults chosen by age group, gender, social class, region
  - Recorded all their conversations over a 2-7 day period
  - Date, time, setting, other participants
  - Began with a pilot to test procedures
- **Context-governed** samples - e.g. from lectures, legal proceedings etc
  - Total of over 6,000,000 words
  - Four categories: educational, business, public/institutional, leisure
  - Each divided into monologue (40%) and dialogue (60%)
  - Three geographic regions and some national material (broadcast categories etc)

CLUL

International Corpus of English (ICE)

Números entre parênteses indicam nº de textos com 2000 palavras

Spoken (300)	**Dialogues** (180)	**Private** (100)	Conversations (90) Phonecalls (10)
		Public (80)	Class Lessons (20) Broadcast Discussions (20) Broadcast Interviews (10) Parliamentary Debates (10) Cross-examinations (10) Business Transactions (10)
	Monologues (120)	**Unscripted** (70)	Commentaries (20) Unscripted Speeches (30) Demonstrations (10) Legal Presentations (10)
		Scripted (50)	Broadcast News (20) Broadcast Talks (20) Non-broadcast Talks (10)
Written (200)	**Non-printed** (50)	**Student Writing** (20)	Student Essays (10) Exam Scripts (10)
		Letters (30)	Social Letters (15) Business Letters (15)
	Printed (150)	**Academic** (40)	Humanities (10) Social Sciences (10) Natural Sciences (10) Technology (10)
		Popular (40)	Humanities (10) Social Sciences (10) Natural Sciences (10) Technology (10)
		Reportage (20)	Press reports (20)
		Instructional (20)	Administrative Writing (10) Skills/hobbies (10)
		Persuasive (10)	Editorials (10)
		Creative (20)	Novels (20)

CLUL

Desenho do *corpus*

tamanho global

- *corpus* de 5M > *hapax legomena* = 40%
- uma nova palavra aparece em cada 30 palavras em média
- pelo menos metade dos significados atestados das palavras polissémicas ocorre apenas uma vez
- quanto mais especializado é o *corpus* mais palavras lexicais (e não gramaticais) aparecem na lista das mais frequentes
- as palavras lexicais mais frequentes são também as mais polissémicas

CLUL

Desenho do *corpus*

"while any linguistic corpus increases linearly in tokens in a completely regular or stable shape, its increase in types – though close to that of tokens at the beginning – starts declining the more the corpus grows, as it contributes fewer new types. The cumulative words – tokens – are distributed linearly, while the cumulative word forms – types – are distributed curvilinearly (Biber 1993: 259). Similarly, this fall or gradual decline, as regards types, would be more dramatic with respect to lemmas."

Aquilino Sanchez & Pascual Cantos, "Predictability in linguistic corpora", *International Journal of Corpus Linguistics*

> tokens, types, lemmas

CLUL