Introduction to Corpora

LL6033: Using Corpora in Translation Studies

Description of the course

Assessment

What is a corpus?

What is a corpus?

Origin of the word

Use in other contexts

What is a corpus?

In corpus linguistics (and TS):

> [A] collection of texts in electronic format which are processed and analyzed using software specifically created for linguistic research (Zanettin 2012)

What can we study with corpora?

What can we study with corpora?

- Lexicography
- Computational linguistics
- Language variation
- Language pedagogy
- Contrastive linguistics

- ## Descriptive Translation Studies

  ### Key elements of corpora

Authenticity —

Key elements of corpora

Representativeness

> A corpus is not simplpy a collection of texts. Rather, a corpus seeks to represent a language or some part of a language. The appropriate design for a corpus therefore depends upon what it is meant to represent. (Biber et al. 1998: 246)

Key elements of Corpora

Representativeness

- Is representativeness possible?
- Why is it such a key issue?

Key issues in Corpora

Size

Types of corpora

Reference vs. specialised

Types of corpora

Synchronic vs. diachronic

Types of Corpora

Dynamic vs. static

Types of Corpora

Monolingual vs. Multilingual

Types of Corpora

Parallel vs. comparable

The Web as Corpus

- Web as an 'exhaustive' rather than 'representative' corpus of the language used in electronic written communication.
- Huge Size
- Uneven quality of the texts it contains
- **Issues of authenticity (authoritativeness)**

  > Everyday experience suggests that 'authentic' in the web often means inaccurate (misspelt words, grammar mistakes, improper usage by non-native speakers), i.e. texts that are almost certainly authentic in terms of their communicative intent but may be unreliable and lacking authority at the level of code. (Gatto 2014)

The web as/for Corpus

- Two options:
- Web as a macro-corpus: Application of parameters to restrict the results of searches (Bergh 2005)
- Web as a source of texts:
  - For large reference corpora (eg. COCA)

– For DIY disposable corpora

Key issues of the Web as corpus

- Authenticity
- Representativeness
- Size
- Dynamism
- Relevance and reliability

- **Access**

Web as macro-Corpus

- Internet searches: Google

- **Search operators**

  Web as macro-Corpus

  site:

  Web as macro-Corpus

  related:

  Web as macro-Corpus

  “exact match search”

  Web as macro-Corpus

  Boolean Operators: OR, AND, NOT

  Web as macro-Corpus

  -term

  Web as macro-Corpus

  Wildcards (*)

  Web as macro-Corpus

  ()

  Web as macro-Corpus

  AROUND(X)

  Web as macro-Corpus

  ..

  Web as macro-Corpus

  loc:placename

A Google a Day!

Uses of parameters

Phraseology and patterns

"To * a better future"

Gluten sensitive enteropathy (Gatto 2014) * gluten * enteropathie: * Entéropathie au gluten * Entéropathie d'intolerance au gluten * Entéropathie induite par le gluten. —

Uses of parameters

Disambiguation

Surgical removal or decompression of cysts

(Maniez 2007)

Useful web resources

Google Trends

Google Books N-Grams

- ---

Beyond Google

Concordancers, Web Corpora, Corpus Compilators

WebCorp

iWeb corpus

CQP Web
