PDEN 2019

Corpus design

Simon Gabay

11 décembre

Corpus Pragmatics

A Defintion

- Corpus pragmatic approaches typically adopt a quantitative perspective. Research questions often ask about the frequencies of certain elements in specific text samples and, crucially, about differences of these frequencies in different text samples.
- **a quantitative perspective requires a very solid foundation in the preparation of the data base and in the analysis and categorisation of the data.** (Jucker 2018)

Typology

Aarts (2011) proposes a typology:

- *Balanced corpora* provide a more or less representative mirror image of an entire language
- *Full-text corpora* contain one or more complete texts
- *Parallel corpora* contain texts of more than one language (or more than one variety of the same language)

There is a *big data caveat* according to O'Keeffe, a tension between

- small but richly contextualised sets of data
- large-scale corpora with a lot of quantifiable material but a very limited amount of context for each of the retrieved hits (Jucker 2018)

Definition

Aarts introduces in 2011 the idea of *balanced corpus*, but the creation of such corpora dates back to the 1960s:

- the *London-Lund Corpus of Spoken English* (LLC)
- the *Brown Corpus* of written American English
- the *Lancaster-Oslo-Bergen* (LOB) corpus of written British English

We now also have diachronic balanced corpora

- the *Early Modern English Medical Texts* (EMEMT)

Representativeness

> "A corpus is thought to be representative of the language variety it is supposed to represent if the findings based on its contents can be generalized to the said language variety" (Leech 1991)

But representativeness is a fluid concept closely related to your research questions

- A general corpus?
- A specialised corpus?

Any claim of corpus representativeness and balance must be interpreted in relative terms

●●●○○ Swisscom 15:08 100 %

Chats +49 175 5557944

Oh wie süß! 16:13

Wann kommst Du vorbei? Daisy möchte bestimmt mit dir spielen! 16:15

Wie wär's mit heute Abend? 16:15

Ça va. Tu viens à 19h00? 16:16

Yes, perfect. 16:16

Freu mich schon. Dann bis später! 👋 16:20

Bis später, bisous 💋 16:22

European Literary Text Collection (ELTeC)

First definition

Major deliverable of COST Action 16204, Distant Reading.
ELTeC is:

- a principled collection of literary text corpora
- uniformly encoded in TEI XML
- representing the production of novels in different European languages for the period 1840 to 1920.

First basic rules

- Corpus balanced with respect to language and publication date of the texts
- Excluding translation
- Using the first edition
- Usually from a book (not periodicals)

Language

- No distinction between local variations
- It assumes only European varieties
- The approach is language-based and not country based
- Preference for standard varieties over dialect varieties (if possible)

Cleaning

- Dealing with historical texts might require some cleaning up or normalizations.
- Merging all word forms which are separated by line breaks.

Eligibility

In order to be included, a text must…

- have been first published as a book (or or in case of need as series publication) between 1840 and 1920, sub-grouped and equally split by decades,
- have first been published in a European country. [maybe not "first": within that decade],
- be a novel, *i.e.* a fictional prose narrative,
- have originally been written in the language of the given subcollection.

Composition

- at least 10%-50% have been written by female authors for the language subcollection.
- 9 to 11 authors are represented with exact three novels.
- at least 30% are highly canonical novels and at least 30% should be novels of low canonicity; canonicity is assessed by investigating the number of times a title was reprinted during the period 1970-2009.
- at least 20% are short novels (10-50k word tokens), at least 20% are long novels (>200k word tokens).

Date : 1840 to 1920 (first iteration)

We will divide into four groups

- group A (1840-1859): code T1
- group B (1860-1879): code T2
- group C (1880-1899): code T3
- group D (1900-1920): code T4

Reprint count

We propose to use the number of times a work is reprinted as an objective measure of its reception during the period 1970-2009, using categories like the following:

- low: few or no reprints found during this period,
- high: many reprints found during this period

Note that the reprint count does not include digitizations of texts.

Author gender

We use the following three categories for actual (not claimed) author gender

- male
- female
- mixed (more than one author)
- undefined

Length

We include a variety of lengths

- short (10k-50k word tokens)
- medium (50k-100k word tokens)
- long (>100k word tokens)

Bibliography

- Aarts, Jan, "Corpus analysis", *Handbook of Pragmatics Manual*, Amsterdam/Philadelphia: John Benjamins, 2011.
- Andreas H. Jucker, "Corpus pragmatics", *Methods in Pragmatics*. Berlin: De Gruyter, 2018, 455-466. https://doi.org/10.5167/uzh-152158
- Jucker, Andreas H. and Irma Taavitsainen, "Diachronic corpus pragmatics: Intersections and interactions", *Diachronic Corpus Pragmatics*, 3-26.
- Merja Kytö, "Corpora and historical linguistics", *Rev. bras. linguist. apl.* vol.11 no.2 Belo Horizonte, 2011