Corpus linguistics

Chapter 1. What is corpus linguistics?

Wednesdays, Spring 2024

Preface

http://www.cambridge.org/mcenery-hardie

- Simply as a tool?
- Linguistic analysis
- Corpus linguistics to corpus analysis in linguistic research
- Skills to access and handle corpus or using online interface
  - ➡ BNCweb (https://www.english-corpora.org/bnc/)
  - ➡ Brigham Young University (http://corpus.byu.edu)
  - ➡ Michigan Corpus of Academic Spoken English: MICASE (https://quod.lib.umich.edu/m/micase/)

Ch.1 What is corpus linguistics?

- It is not directly about the study of any particular aspect of language. Rather, it is an area which focuses upon a set of **procedures**, or **methods**, for studying language.
- **Impact**: The development of corpus linguistics has at least facilitated the exploration of, new theories of language - theories which draw their inspiration from attested language use and the findings drawn from it.

  >> What impacts corpus linguistics has on linguistics?

- Corpus linguistics is a heterogeneous field.

Generalization: Set of texts or corpus

- Machine-readable ***text***
- Corpora are invariably exploited using tools
- Users of a corpus must be aware of its internal variations (degree of homogeneity)

  e.g., Concordance (qualitative) and frequency (quantitative) data

  The computer files within a corpus do not need to be textual.

Different types / features of studies in corpus linguistics:

1. Mode of communication

2. Corpus-based versus corpus-driven linguistics

3. Data collection regime

4. Annotated vs. unannotated corpora

5. Total accountability vs. data selection

6. Multilingual vs. monolingual corpora

[1] Mode of communication

- Corpora may encode language produced in any mode: spoken or written. (Video, sign language, etc.)

  ➡ Written corpora and encoding problem > Unicode

  ➡ Spoken corpus is time-consuming to gather and transcribe; serious hazards involved if transcripts made by non-linguists; phonemically transcribed material is of much more use

  ➡ Gesture corpora (e.g., sign language)

BNC: written and spoken texts

- Written vs. Spoken language: Biber et al. (1999), Carter & McCarthy (1995)
- Thinking about corpora in terms of mode of production is not just a matter of different data collection and technical issues; it is, rather, linguistically a very real distinction.

[2] Corpus-based vs. corpus-driven linguistics

- Corpus-based: Studies typically use corpus data in order to explore a theory or hypothesis (to validate, refute or refine theories)

  ➡ Corpus approach as **a method**.

- Corpus-driven: This approach rejects the characterization of corpus linguistics as a method. It claims instead that the corpus itself should be the sole source of our hypotheses about language. (~ neo-Firthians; extreme end point of corpus linguistics)

[3] Data collection regimes

- How can we ensure that the match is good enough?
- Corpus construction (and data collection) emerges as a critical issue for corpus linguistics.

  ➡ The **monitor corpus** approach (Sinclair, 1991): the corpus **continually expands** to include more and more texts over time

  ➡ The **balanced corpus** or sample corpus approach (Biber, 1993; Leech 2007): a **careful sample corpus** is constructed according to a specific sampling frame.

A. Monitor corpora approach (J. Sinclair)

- This approach seeks to develop a dataset which grows in size over time and which contains a variety of materials.
- As the corpus grows, the assumption is that any skew in the data naturally self-corrects, since there is no consistent key in the data input.

➡ The Bank of English (BoE, Univ. of Birmingham, 1980s): half a billion words (450 million words).

➡ The Corpus of Contemporary American English (COCA; Davies 2009): via regular sampling

B. The Web as Corpus (Kilgrarriff & Grefenstette 2003)

- e.g., Google search (or Interface WebCorp, Renouf 2003)
- Limitations: 1) The web is a mixture of carefully prepared and edited texts, and what might charitably be termed 'casually prepared' material across genres. 2) Undifferentiated mass, which may require a great deal of processing to sort into meaningful groups of texts. 3) Many errors of all sorts.

  e.g., receive (300M hits) vs. receive (8.6 M.) > maybe interesting for studying spelling reform, but can be noise in the data.

  e.g., 'swanning around' (BNC displays 13 examples; Google finds 32,300)
- Words with high frequency is overwhelming, others can be discarded.
- The web is forever changing, and it is difficult to replicate a study. (**Replicability** issue)

C. The sample (balanced) corpus approach

- This type of corpus represent a particular type of language over a specific span of time. This makes corpora balanced and representative within a particular sampling frame.

  e.g., Service interactions in UK in the 1990s. (If a specific shop data is gathered, then relatively context-specific lexis would appear.)

  ➡ Even when we are interested in specific shops, we still need some sense of ‘typical’ to avoid skewness. We need representativeness for the data in a corpus.

  e.g., the Lancaster-Oslo/Bergen (LOB) corpus: a ‘snapshot’ of the standard written form of modern British English in the early 1960s. (Comparable to The Brown Corpus (AmE))

D. Balance, representativeness and comparability

- The measures of balance and representativeness are matters of degree.
- Even though these are hard to attain, we should not give up. We should aim at a gradual approximation to these goals, as crucial desiderata of corpus design.
- There is a scale of representativity, of balancedness, of comparability. We should seek to define realistically attainable positions on these scales, rather than abandon them altogether.

E. ‘Opportunistic’ corpora and minority and endangered languages

- There are many collections of data, reasonably described as corpora, which do not necessarily match the description of either a monitor or a snapshot corpus.

  ➡ These are best described as ‘opportunistic’ corpora.

- Limitation: some corpora were built using whatever relevant material could be accessed in electronic form. This problem clearly no longer generally applies to English or most other major languages, but it still persists for some languages.

- Official majority languages, Official minority languages, Unofficial languages, and Endangered languages. (Financial support varies; in particular, spoken data)

- The use and construct must sometimes be determined by pragmatic considerations.

[4] Annotated vs. unannotated corpora

- **Corpus annotation**: e.g., POS tagging (N, mnemonic code; talk_N)
- Inline annotations (tags directly into the text) vs. stand-off annotations (tags separately)
- Some linguists object to annotation (when undertaken manually rather than automatically by a computer): due to accuracy and consistency

[5] Total accountability vs. data selection

- Corpora may vary in how they are used by the analysts who exploit them.

  ➡ Total accountability, falsifiability, and replicability:

A. Total accountability, falsifiability and replicability

- **Total accountability**: the necessity of employing rigorous, unbiased methods and maintaining a commitment to objectivity.

  ➡ If you approach a corpus with a specific theory in mind, it can be easy to unintentionally focus on and pull out only the examples from the corpus that support the theory (**confirmation bias**)

  ➡ The principle of total accountability: we ***must not*** select a favorable subset of the data in this way. (Via randomized subsample; avoid conscious selection of data)

- **Falsifiability** (Poppe, 1934): a criterion for demarcating science from non-science. It posits that for a theory or hypothesis to be scientifically valid, it must be testable and potentially refutable by empirical evidence. There must be a conceivable observation or experiment that could prove the theory false.

- **Replicability:** In all the sciences, new results are typically considered provisional until they are known to be replicable — and in many cases, it is precisely through that process of continuous checking of results as theories develop and expand that falsifiability is achieved.

- Corpus linguistics appealed to the notion of the replicable result for credibility. (This helps us address the problem of the limited dataset. In sum, total accountability to the data at hand ensures that our claims meet the standard of falsifiability; total accountability to other data in the process of checking and rechecking ensures that they meet the standard of replicability; and the combination of falsifiability and replication can make us increasingly confident in the validity of corpus linguistics as an empirical, scientific enterprise.

- In sum, total accountability to the data at hand ensures that our claims meet the standard of falsifiability;
- total accountability to other data in the process of checking and rechecking ensures that they meet the standard of replicability;
- and the combination of falsifiability and replication can make us increasingly confident in the validity of corpus linguistics as an empirical, scientific enterprise.

B. Data selection - not (necessarily) a bad thing

- **Corpus-informed** research: when researchers use the corpus simply as a bank of examples to illustrate a theory they are developing.

  Cf.) **CDA** (Critical Discourse Analysis): minimally uses corpus data (a detailed analysis of a small amount of data to fully investigate); it pursues qualitative analyses using conformances.