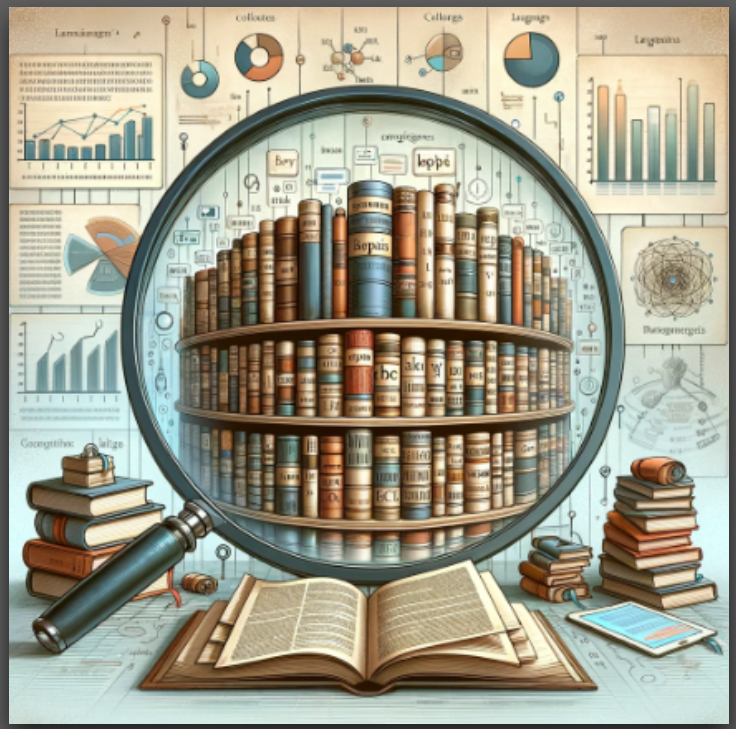
Corpus linguistics

Method, Theory, and Practice



Wednesdays, Spring 2024

Preface

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

Main questions in this book

- How corpus linguistics has developed?
- What are major theoretical issues that corpus linguists content with today?
- What are the problems that researchers using corpora must grapple with in practice, both within linguistics and across disciplines?

Preface

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

- Simply as a tool?
- Linguistic analysis
- 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/)

From *Corpus linguistics*

To *Corpus methods in linguistics*

Chapter 1.
What is corpus linguistics?

1.1 Introduction

- What is corpus linguistics?

  (cf. phonetics, phonology, phonology, syntax, semantics, pragmatics, semiotics, etc.)

  ➡ It is about **procedures** or **methods**.

  ➡ It may refine and redefine a range of theories of language. (Essentially, the growth of corpus linguistics has encouraged the development of linguistic theories that are grounded in the way language is genuinely used, as opposed to theories based on abstract or hypothetical uses of language.)

1.1 Introduction

- 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?

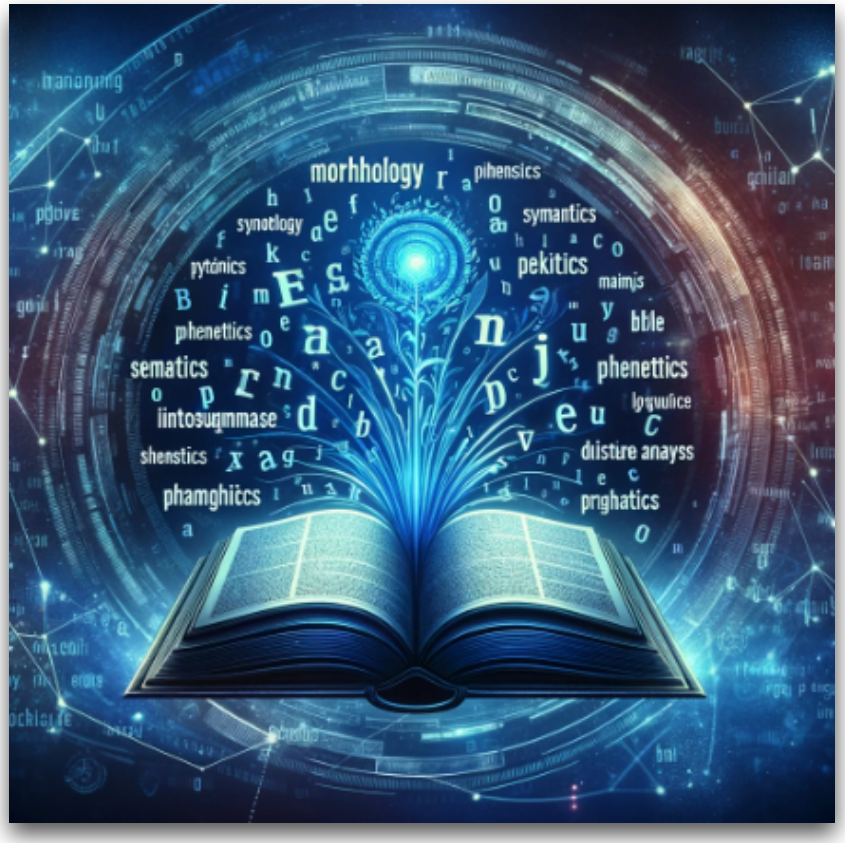
• Corpus linguistics is a heterogeneous field. (differences within corpus linguistics)

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.



Differences

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

1.2 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.

1.3 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)

Go to Github