Introduction to corpus linguistics

Adam J.R. Tallman

2023-10-20

Introduction

- Corpus Linguistics
  - a method for studying language using a document with recorded instances of naturally occurring utterances
  - could be contrasted with judgement-based linguistics or introspective linguistics
  - concern with frequencies of utterances
  - concern with variability (variable ways of saying things)

Introduction

- Why do corpus linguistics?
  - Introspection isn’t always reliable
    - controversial?
  - Introspection doesn’t tell us about relative frequencies
  - Frequencies tell us about language change and about variability in relation to different factors

Introduction

- Why is this a field?
  - (as opposed to the norm)

  - Because there has been a long tradition of intuiting data in linguistics
    - (e.g. McEnery & Wilson 2001; Stefanowitsch 2022)

  - Because drawing inferences from variable data is not trivial
    - (there are lots of ways frequency data can be misleadingly *presented* and lots of methodological issues that arise from analyzing quantative data)

  - Because developing usable corpora is not trivial

Introduction

- What you will learn
  - Main concepts of corpus linguistics
  - How to construct a corpus
  - How to analyze a corpus
  - The main copora that are available
  - Methodological issues that arise from using corpus data
  - (Some programming for corpus linguistics)

Course management

- First homework due on Novevmber 3rd, will be posted on October 27th
- Readings: to be completed before the class
- All homeworks are obligatory – you have to pass four out of five in order to be able to write the final project

Course management

- Final Project (two options)
  - Use available corpus to describe some linguistic phenomenon quantitatively
  - Construct a pilot corpus, describe how it is structured and what would be useful for.
  - 6-8 pages, 11 sized font Times New Roman, 1.5 spacing, 2.5 cm to 3cm margins
    - (p.s. you will not be penalized for going over the page limit)

Course management

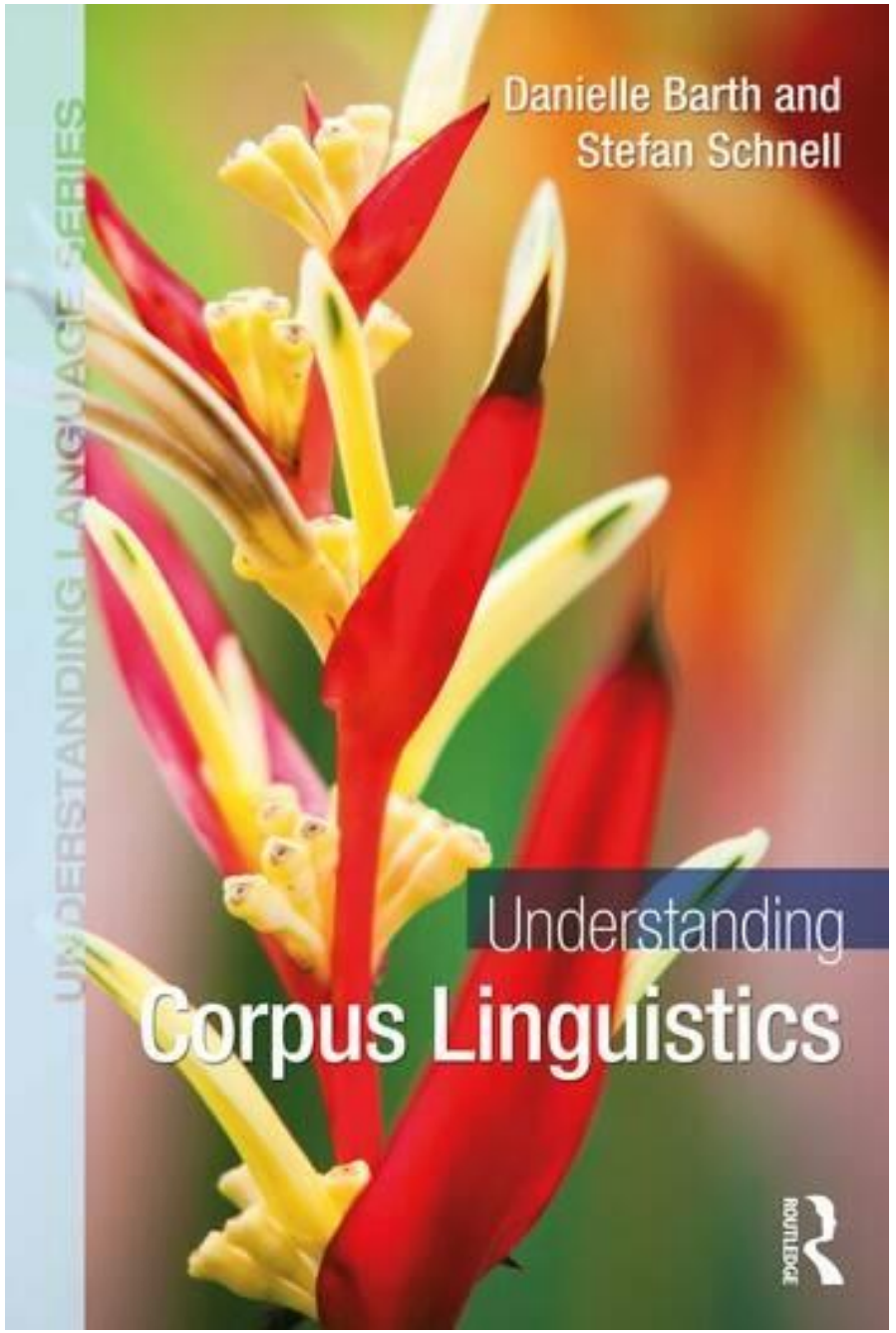
- **Final paper draft:**
  - A draft of your final project is due January 26th
- **Final paper due:**
  - Final paper is due February 14th
- **Extentions:**
  - You can write to me for extentions on homeworks over email adam.james.ross.tallman@uni-jena.de for homeworks and for the final Project, but keep in mind that for the final paper you need to talk to ASPA.

Readings

- My lectures will be based heavily on the textbook by Barth and Schnell

- By November 3rd you should have read Chapters 1 through 3 – I will scan the first three chapters and post them online for you next week, but after this you will have to find the textbook

- Recommended: read each chapter twice, once before the relevant class, once after



Course outline

- Next week: Basic concepts
- October 27: Corpus Types
- Nov. 3 and 10: Levels of Linguistic Representation
- November 17 and 24: Corpus Queries and Sketch engine

Course outline

- Sketch engine:

- Our University has an account:

https://auth.sketchengine.eu/#login?next=https%3A%2F%2Fapp.sketchengine.eu%2F

Course outline

- December 8th : Corpus annotation
- December 15 and January 5: Statistical description in corpus linguistics
- January 12 19 and 26: Different topics in corpus linguistics