Introduction to Methods in Corpus Linguistics

Dr. Kristopher Kyle

University of Oregon | Yonsei University

A quick overview of this session

- Discuss Syllabus
- Discuss Course Design
- Self-introduction
- Brief overview of CL
- Questions/troubleshooting
- Lets Analyze Some Data!

Discuss Syllabus

Course Design

Some words on AntConc and Python…

Self introduction

Kris' corpus related research interests

- Learner corpus analysis
  - lexical sophistication
  - lexical diversity
  - collocation
  - syntactic complexity
  - open-source tools
- Corpus collection
  - TMLE Project
- Corpus annotation
  - LL project
- And more!

Who are you?

- Name
- Department, etc.
- Favorite band
- Favorite hobby (non-academic)
- Research interest(s)
- Why are you in this class?

A brief introduction to corpus linguistics

- *Corpus* = collection of electronic texts
- *Corpus linguistics* = collection of methods for analyzing large electronic datasets comprised of language samples
- Corpus methods allow researchers to *systematically* investigate linguistic phenomena in a *reproducible* manner
- Corpus methods also allow researchers to analyze datasets that would be impractical to analyze by hand

Reference Corpora

- Reference corpus = a corpus intended to be representative of a particular language use domain.
- Some reference corpora are *intended* to represent a particular language within a particular society
  - the British National Corpus (BNC)
  - the Corpus of Contemporary American English (COCA)
- Others are intended to represent a particular mode of language in a particular context
  - Michigan Corpus of Upper-level Student Papers (MICUSP)
  - British Academic Written English (BAWE) corpus
  - Yonsei English Learner Corpus (YELC)

Lets check out some corpora

- Corpus of Contemporary American English (COCA)
  - https://www.english-corpora.org/coca/
  - (or just search for “COCA corpus”
- British National Corpus (BNC)
  - https://www.english-corpora.org/bnc/

Corpus sampling…

- A *balanced corpus* includes and equal number of samples across text types
- A *sample corpus* (also called a *representative corpus*) ideally includes a truly random sample of texts in a particular domain.

Important issues when using corpora

- The results of any corpus analysis can be evaluated by:
  - a) the representativeness of the corpus used
  - b) the appropriateness of the analysis method
    - including whether the analysis was conducted in a replicable way
- Corpus research tends to be primarily quantitative in nature but…
- Primarily qualitative research can also be conducted using corpora
  - i.e., using random samples from large corpora
- Frequency counts (and derivatives of frequency counts such as probabilities) are very important in corpus research
- This lends itself to usage-based theories of language learning, but is not limited to such theories

Questions about corpus linguistics?

Other questions/Troubleshooting?

Let’s Analyze Some Data!