Bundles, clusters and *n*-grams

Introduction to Methods in Corpus Linguistics

Day 3

Kristopher Kyle

Yonsei University

Class updates

- Mini Project 2
- More corpora available

Questions, Mini project 1 issues?

Lemmatization in AntConc (quick tutorial)

- In AntConc, it is possible to group words based on formal characteristics
- One common method is lemmatization, though familization is also possible
- Please use the lemma lists available on the course website

Bundles, clusters, and *n*-grams

- What are some important characteristics of lexical bundles?
  - contiguous sequences of *n*-words
  - usually include a frequency cut-off
  - usually include a range cut-off
  - lists are generally smaller with larger *n*
  - can cut across phrasal boundaries
  - often have specific (or a small set of) functions
- What can a lexical bundle analysis tell us?

Extracting bundles, clusters, and *n*-grams with AntConc

Application activity

- Find all 4-word lexical bundles in the COCA sample that:
  - occur in at least 10% of the texts
  - occur 40 times per million words (67 times in 1.69 million words)
- Find all 4-word lexical bundles that:
  - occur in at least 10% of the texts
  - occur 40 times per million words (67 times in 1.69 million words)
  - occur in all five registers