Bundles, clusters and *n*-grams

Introduction to Methods in Corpus Linguistics

Day 3

Kristopher Kyle

Yonsei University

Class outline

- Discuss Corpus Design
- Discuss Stop Lists
- Discuss Lemma and Family Lists
- Bundles, clusters, and n-grams

Corpus Design

Some things to consider:

- Possible research questions (this is very important)
- Language use domain (this is very important)
- Included modes (written? spoken? chat?)
- Included registers/genres
- Necessary annotation

What features will your corpus have?

https://docs.google.com/presentation/d/1PWTRQKCJ_Ti8dTwoD0c2TWoCTeQQcAFjoSe4QBbus2g/edit?usp=sharing

Questions, Mini project 1 issues?

Some important terms…

content words	**open class words** nouns, verbs, adjectives, adverbs* (some adverbs are open class, while others aren't)
function words	**closed class words** articles, conjunctions, prepositions, etc.
tokens	**running words in a text** Note: If you have converted a text into a list of words, you can obtain the number of tokens using the function len()
types	**unique words in a text** Note: the operational definition of a type depends on whether words, lemmas, or families are considered Note 2: The function set() converts a list into a list that includes unique items only

Using stop lists to identify content words

- Apply the function-word stop list on the course web page to identify the most frequent content words in the Brown corpus

Some important terms

word	**a string of letters separated by white space (or punctuation)** freq_dict = {‘nation’ : 3, ‘nations’ : 4, ‘national’ : 2, ‘nationalize’ : 6}
lemma	**all inflected forms of a word** lemma_dict = {‘nation’ : [‘nation’, ‘nations’]} lemma_freq_dict = {‘nation’ : 7}
family	**all inflected and derived forms of a word** family_dict = {‘nation’ : [‘nation’, ‘nations’, ‘national’ , ‘nationalize’]} family_freq_dict = {‘nation’ : 15}

List-based lemmatization

- What are the most frequent content word lemmas in the Brown corpus?
- What are the most frequent content word families in the Brown corpus?
- https://docs.google.com/document/d/1Gj5L4ilbcIHulnezZOyXH2T7ohx1IgdjVKW02hxls-0/edit?usp=sharing

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 parameters did Biber et al (2004) use to identify lexical bundles?
- 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