Frequency List: Practice

Gede Primahadi Wijaya Rajeg

2024-07-20

Materials

- source files for all materials:
  - https://github.com/complexico/dipscorling2024
- pdf version as a handout here
- How to cite these materials:

> Rajeg, Gede Primahadi Wijaya. 2024. Materials for the *Diponegoro Summer Course in Corpus Linguistics* (*DipSCORLING 2024*) (22 - 27 July 2024). R Quarto. Zenodo. https://doi.org/10.5281/zenodo.12793922. (22 July, 2024).

Practice 1: Words beginning with certain strings

For this (and the remainder of the) practice, we will use the `Brown Family (CLAWS + TreeTagger tags)` corpus.

We will compare the results of retrieving words starting with certain strings/character using two approaches:

a. Retrieving all words then filter
b. Directly using BASIC's `starting with` approach

Pay attention to the results. Why do you think they differ?

- IMPLICATION 1: some limitation of SE regarding their result outputs.
- IMPLICATION 2: important for our aim to target/generate specific list with certain criteria using more targeted feature.

First approach: The basic, find all approach then filtering

1. In the `BASIC` tab (Figure 1), select: **words** > **all** > **GO**

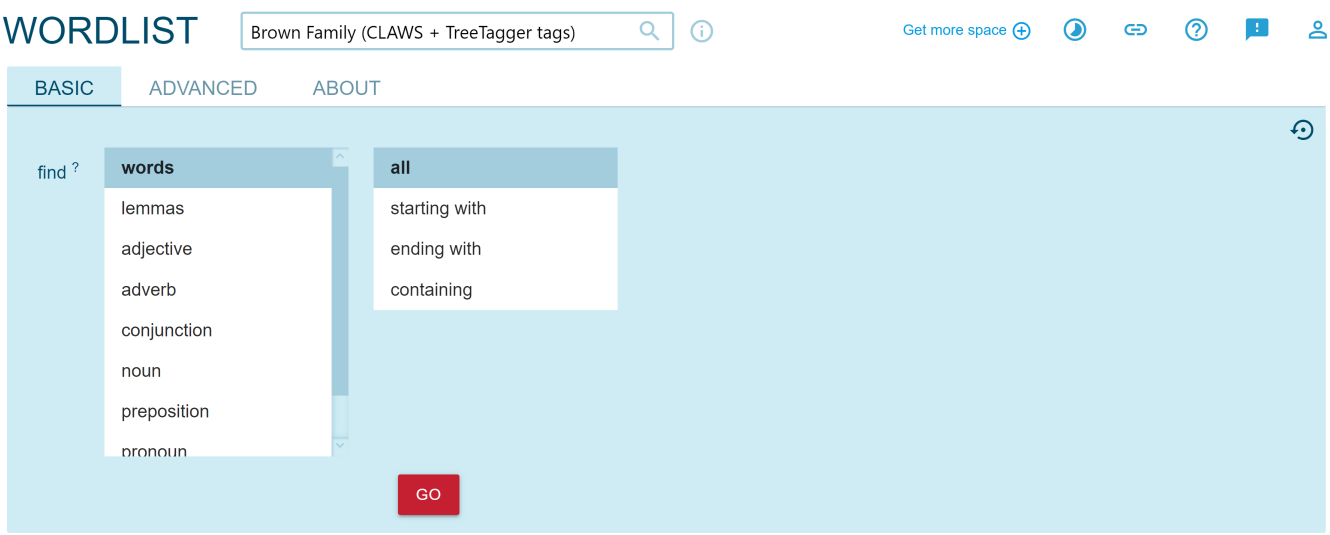

Figure 1: Wordlist `BASIC` interface searching for all words

2. On the output page of the Wordlist (see Figure 2), click on the `Filter` feature (i.e., the up-side-down triangle lines, to the right of the "eye-like" symbol on the upper right corner).
3. Select the option `Starting with` as shown in Figure 2 below.
4. Type in "kn" in the field
5. Press `Enter`; the result is shown in Figure 3

You will get five items.

Second approach (your turn): The basic, `Starting with` feature

1. In the `BASIC` tab, instead of using **words** > **all**, use **words** > **starting with**
2. Then, type `kn` then hit **GO**
3. Compare the current results with that in Figure 3. How many do you get? (Answer key (you need to LOGIN): https://ske.li/15i)

   3.1. Why do they differ?

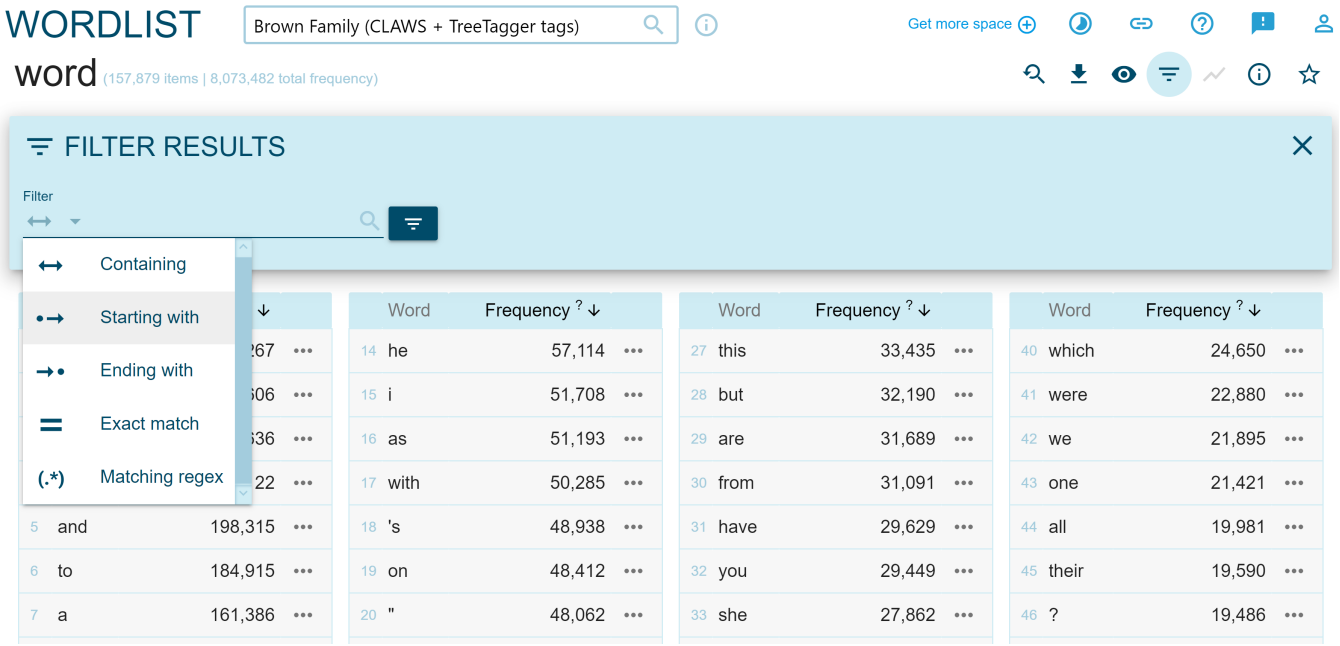

Figure 2: `Filter` option in the Wordlist output, choosing the `Starting with` condition.

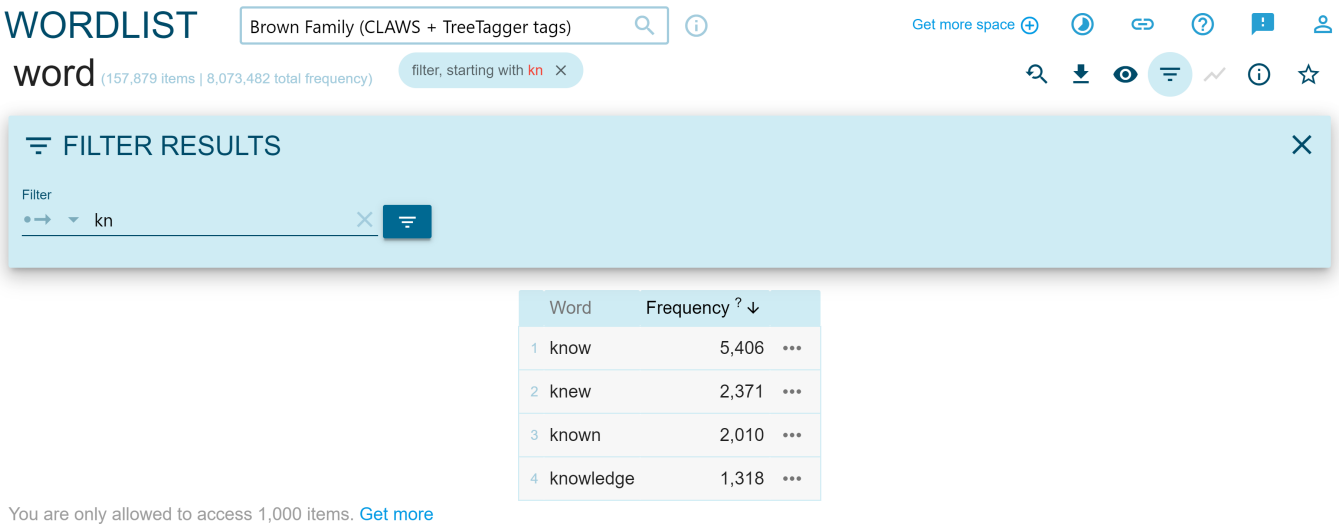

Figure 3: Output of filtering words starting with `kn` in the all word list.

Practice 2: Words of certain word-class/part-of-speech containing certain affixes

We are still in the `BASIC` tab. We will explore the productivity (the number of different items/type frequency) of English adjectives containing suffixes. We focus on suffixes meaning 'having a resemblance of', particularly comparing *-esque* (which has a more specialised meaning of 'in the style of ~') and *-ish* (see Bauer 2022: 55).

Requirement

- IMPORTANT: This involves results from two searches: one for each suffix
- Later, in the output interface, look at the upper left corner to find basic quantitative information:
  - the number of items (i.e., the type frequency)
  - the total frequency of these items (i.e., the token frequency especially of adjectives having these suffixes)

Task

- Conceptual aspect:
  - try to intuite which suffix would be more productive, in terms of their type and token frequency, in the corpus we use (later check this intuition with the results).
- Operationalisation:
  - How would you devise a targeted search using just the `BASIC` feature to retrieve **only adjectives** ending with these suffixes?
  - REMEMBER: you need to run two searches for each suffix
- Results:
  - How many items are there for *-esque* and what is the total frequency of these *-esque* adjectives? (Answer key: https://ske.li/15k)
  - How many items are there for *-ish* and what is the total frequency of these *-ish* adjectives? (Answer key: https://ske.li/15l)
  - Which suffix is more productive (in terms of the total number of items) in this `Brown Family (CLAWS + TreeTagger tags)` corpus? Is your intuition supported by the data?

Practice 3: *windshield* vs. *windscreen*

We are now in the `ADVANCED` tab of Wordlist.

Task

- Conceptual aspect:
  - these two words refer roughly to the same objects (see Google Image results for windscreen and windshield)
  - are they different in terms of their frequency?

Operationalisation:

- try to check their frequency in ALL Brown Family first (think about how you would devise the search)
- then, check their frequency in the combined region corpus.
  - American: BROWN, FROWN, AE06 (AmE)
  - British: FLOB, BLOB, LOB, BE06 (BrE)
  - Do you learn a pattern of use of these two words from their frequency searches? Let's discuss!

Results

- Overall frequency in the WHOLE BROWN FAMILY of
  - *windshield* (answer key: https://ske.li/15n)
  - *windscreen* (answer key: https://ske.li/15o)
- Frequency by the combined sub-corpus
  - *windsheild* in BROWN, FROWN, AE06 (AmE) (answer key: https://ske.li/15p)
  - *windsheild* in FLOB, BLOB, LOB, BE06 (BrE) (answer key: https://ske.li/15r)
  - *windscreen* in BROWN, FROWN, AE06 (AmE) (answer key: https://ske.li/15t)
  - *windscreen* in FLOB, BLOB, LOB, BE06 (BrE) (answer key: https://ske.li/15s)

- Do you learn a pattern of use of these two words from their frequency searches? Let's discuss!

This practice is inspired from Stefanowitsch (2020).

Practice 4: Frequency of lexical-verb tags in American (AmE) vs. British (BrE) English for the Press: Editorial (AmE & BrE) vs. Press: Reportage (AmE & BrE)

Requirements

- We need to run four searches (and hence download four search results):
  - lexical verb tags for Editorial in AmE
  - lexical verb tags for Reportage in BrE
  - lexical verb tags for Editorial in BrE
  - lexical verb tags for Reportage in AmE

Task

- Conceptual aspect:
  - Do these sub-genres differ in the use of certain classes of lexical verbs across the two English varieties?
    - We can look at certain tag, for instance the simple past tag
- Operationalisation:
  - Advanced
  - Select `tags` (Layer 1)
  - `matching regex` (Layer 2)
  - click the `TAGS` to reveal the pattern to get tag for "lexical verb" (i.e., the `VV.*`)
  - Text types[?] select:
    - doc.genre: Press > Editorial
    - doc.region: American
  - Hit **GO**
- Results:

- NOTE: it takes a while, even for this small, less than 10 million tokens BROWN Family.
- In the `View options`, check the Frequency per million words to see the relative frequency
- Answer key: https://ske.li/15m
- There are several list of verb tags, focus on comparing a given tag between Editorial and Reportage (within variety) and between the same sub-genre across variety.
- For more insightful analyses, we need to further process these datasets in Excel or in statistical programming language such as R.
    * Most of the time, corpus tools like Sketch Engine or AntConc are just part of the means (e.g., providing raw data) to an end. For example, we might want to directly visualise the distribution (i.e., relative frequency) of selected tags by genres between varieties as shown in Figure 4. This requires processing the output of corpus tool further that these two corpus tools cannot do. The graph in Figure 4 is produced in R (code file) based on the data in the repository here (see the .csv files starting with `lex-verb-...`).

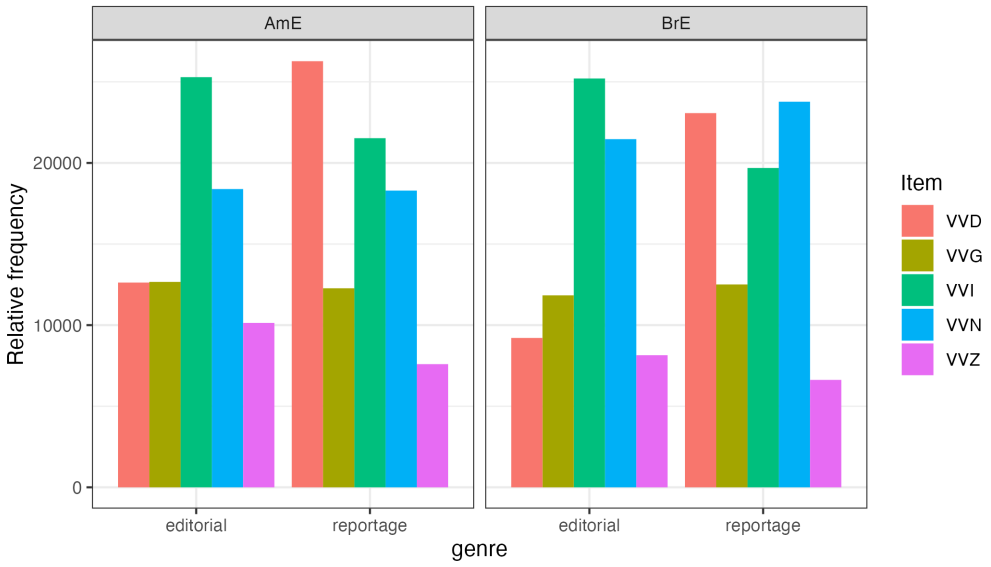

Figure 4: Relative frequency of selected verb tags with absolute frequency greater than 1,000

Reference(s)

Bauer, Laurie. 2022. *An introduction to English lexicology* (Edinburgh textbooks on the English language). Edinburgh: Edinburgh University Press.

Stefanowitsch, Anatol. 2020. *Corpus linguistics: A guide to the methodology.* Berlin: Language Science Press.