Frequency List: Practice

Gede Primahadi Wijaya Rajeg

2024-07-20

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

The source code files for all materials are available here: https://github.com/complexico/dipscorling2024

Practice 1: Words beginning with certain strings

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**

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

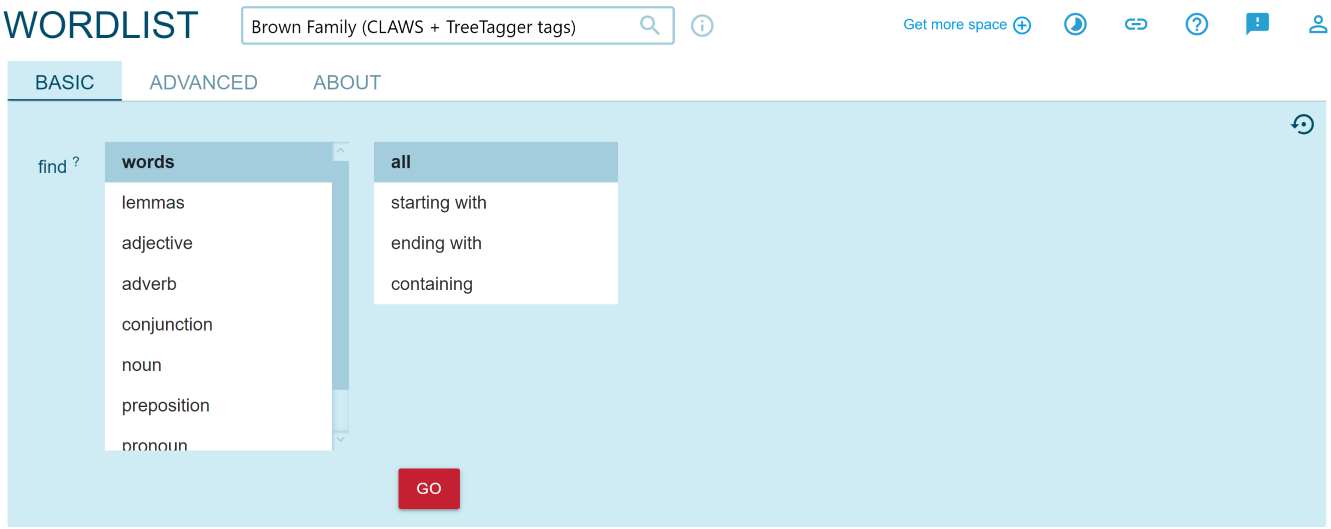
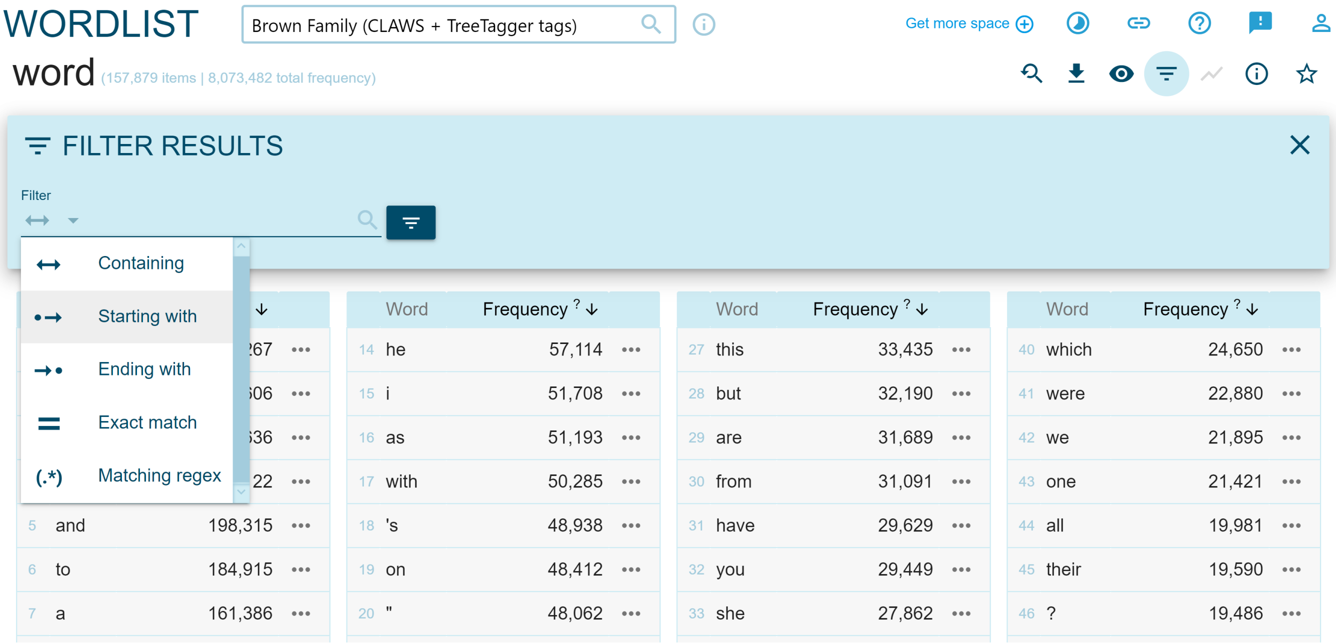


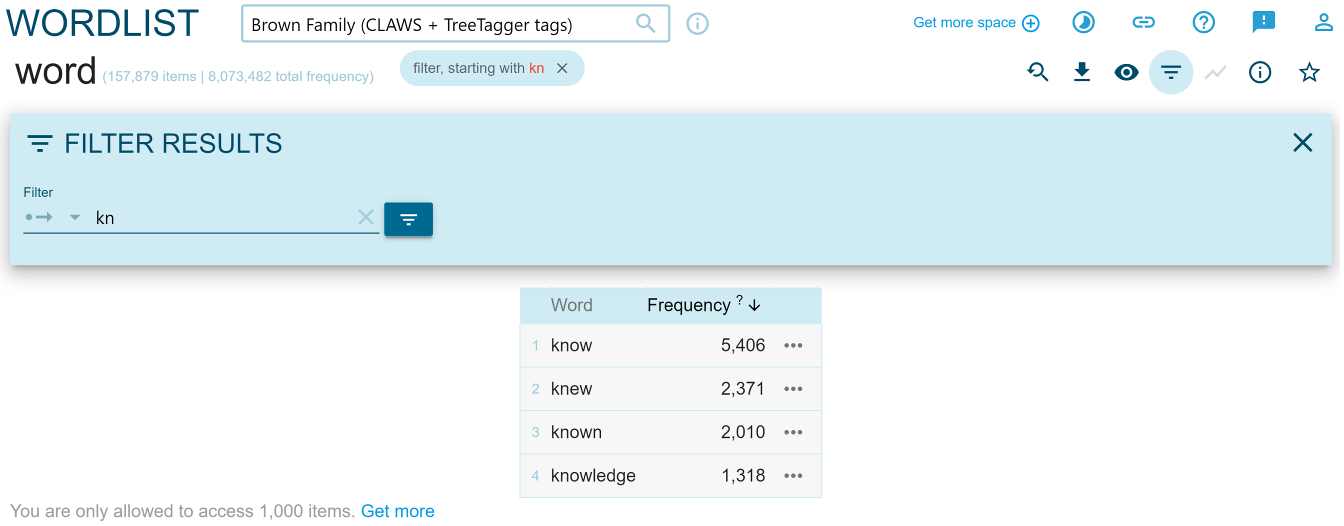
Figure 1: Wordlist `BASIC` interface searching for all words



		↓		Word	Frequency ? ↓		Word	Frequency ? ↓		Word	Frequency ? ↓
		267	14	he	57,114	27	this	33,435	40	which	24,650
		06	15	i	51,708	28	but	32,190	41	were	22,880
		36	16	as	51,193	29	are	31,689	42	we	21,895
		22	17	with	50,285	30	from	31,091	43	one	21,421
5	and	198,315	18	's	48,938	31	have	29,629	44	all	19,981
6	to	184,915	19	on	48,412	32	you	29,449	45	their	19,590
7	a	161,386	20	"	48,062	33	she	27,862	46	?	19,486

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

5. Press `Enter`; the result is shown in Figure 3



(157,879 items | 8,073,482 total frequency)

filter, starting with kn

kn

	Word	Frequency ? ↓
1	know	5,406
2	knew	2,371
3	known	2,010
4	knowledge	1,318

You are only allowed to access 1,000 items. Get more

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

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?

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 device 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 there are for -*esque* and what is the total frequency of these -*esque* adjectives? (Answer key: https://ske.li/15k)
    - How many items there are 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: Frequency of lexical-verb tags in American vs. British English for the Press: Editorial (AmE & BrE) vs. Press: Reportage (AmE & BrE)

- 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.

Reference(s)

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