

Frequency List

Gede Primahadi Wijaya Rajeg

2024-07-20

Outline

1. What is a frequency list?
2. Two basic types of frequency
3. Frequency of different linguistic units
4. Examples of uses of frequency list
5. Demo & Practice

What is a frequency list?

- “The most basic corpus-linguistic tool” (Gries 2017: 12)
- How often a given linguistic unit occurs in a corpus
 - Often, this unit is a *word*

What is a frequency list?

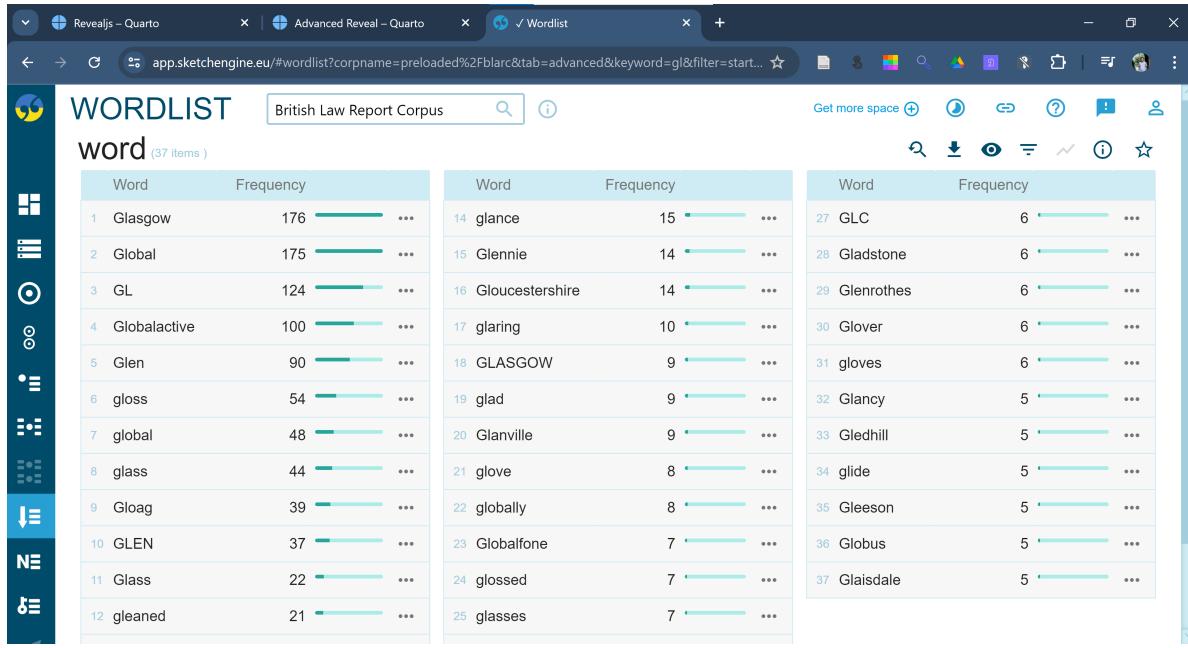


Figure 1: *Words* (starting with gl) and their frequency of occurrences

But, what's a word?

What is a word?

- “entities in text that are separated by either white-space or punctuation.” (Weisser 2016: 147)
 - how about: *can't*, *widely-held*, *co-operate*, or *white-space*?

What is a word?

English compound

- **written together:** *icecream* (14,590 matches)
- **hyphenated:** *ice-cream* (55,506 matches)
- **separated by white-space:** *ice cream* (676,402 matches)

Searches done in *English Web 2021* in Sketch Engine (SE)

What is a word?

- Practical consideration: *tool-specific*
- In SE:
 - “begin with a letter of the alphabet” (https://www.sketchengine.eu/my_keywords/word/)
 - Examples: *book*, *working*, *Mary*, *T-shirt*, *post-1945*, *mp3* or *CO2*
- Methodological consideration:
 - be explicit about word criteria (e.g., in the tool used)

Outline

1. What is a frequency list?
2. Two basic types of frequency
3. Frequency of different linguistic units
4. Examples of uses of frequency list
5. Demo & Practice

Two basic types of frequency

Types vs. Tokens (cf. Cheng 2012: 62; Gries 2017: 12)

- Types: the number of unique/distinct words in a corpus
- Tokens:
 - the total occurrences of all (unique) words in a corpus
 - the total occurrences of **a** (unique) word in a corpus

Two basic types of frequency

Types vs. Tokens (cf. Cheng 2012: 62; Gries 2017: 12)

The sky is sky blue while the estuary is turquoise.

- Tokens: 10 (2 tokens of *sky*, 2 tokens of *the*, 2 tokens of *is*, ...)
- Types: 7 (*the*, *sky*, *is*, *blue*, ...)



The estuary of the Bak Blau lake, on the Enggano island, Indonesia.

Two basic types of frequency

absolute vs. relative

- absolute frequency:
 - real, observed freq. of an item in the (sub)corpus
- relative frequency:
 - normalised frequency of an item on the basis of a base frequency (usually 1 million word-tokens) (cf. [SE's page here](#) for the formula)
 - often used in comparing frequency of the same word in two different corpus that are not equal in size
- SE allows both options.

Two basic types of frequency

absolute vs. relative

How to compute the relative frequency of a linguistic item

$$Rel.Freq = \frac{absolute\ frequency \times 1,000,000}{corpus\ size}$$

Two basic types of frequency

absolute vs. relative

Relative frequency: Examples

Say, in the ICE-GB corpus, you found the following (see Gries 2010: 271) :

- 297 tokens of *give* (in the **spoken** sub-corpus)
- 144 tokens of *give* (in the **written** sub-corpus)
- 128 tokens of *bring* (in the **spoken** sub-corpus)
- 69 tokens of *bring* (in the **written** sub-corpus)

The size of ICE-GB_{spoken} is 637,682 (word-tokens) while the size of ICE-GB_{written} is 423,581 (word-tokens). The relative frequencies of *give* and *bring* in the two sub-corpora become:

$$give_s : \frac{297 \times 1,000,000}{637,682} \approx 465.75$$

$$give_w : \frac{144 \times 1,000,000}{423,581} \approx 339.96$$

$$bring_s : \frac{128 \times 1,000,000}{637,682} \approx 200.73$$

$$bring_w : \frac{69 \times 1,000,000}{423,581} \approx 162.9$$

Two basic types of frequency

Take away

Important to know how to compute relative frequency!

Sometimes (most of the time?), the corpus-software tool we use **cannot** do want we want.

Examples

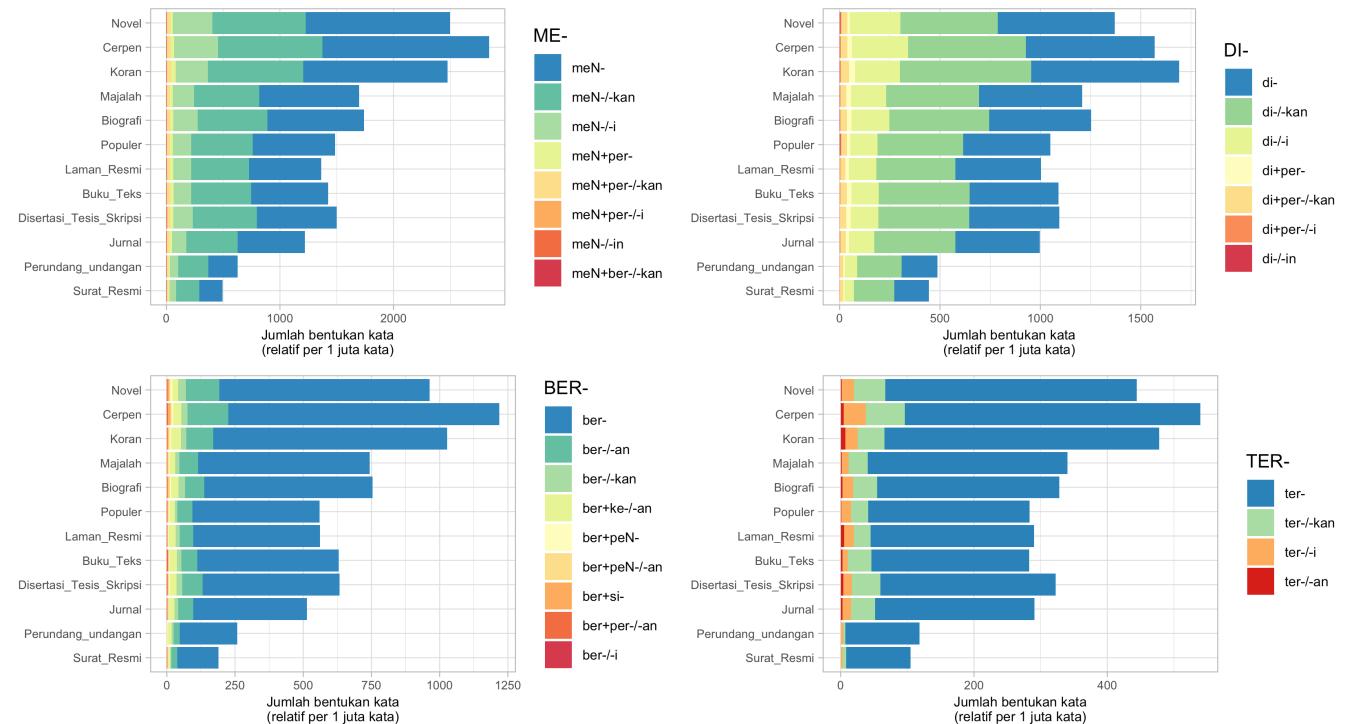


Figure 2: Productivity analysis (based on **relative type frequency**) of four Indonesian verbal prefixes across genres (Rajeg & Denistia 2023).

Outline

1. What is a frequency list?
2. Two basic types of frequency
3. Frequency of different linguistic units
4. Examples of uses of frequency list

5. Demo & Practice

Frequency of different linguistic units

- words
 - words and their word class (i.e., part-of-speech)
 - words containing particular strings/characters (e.g., prefixes/suffixes)
 - ...
- lemmas
 - the base/uninflected form of words from a given part-of-speech
 - * the verbs *go*, *went*, *going*, *gone*, *goes* are word-forms for the same lemma GO
- n-grams (multi-word units with n-number of components)
 - *part of the, for the purposes, on behalf of the, ...*
- word sequence, phrases
 - phrases containing a fixed word
 - ...

Outline

1. ~~What is a frequency list?~~
2. ~~Two basic types of frequency~~
3. ~~Frequency of different linguistic units~~
4. Examples of uses of frequency list
5. Demo & Practice

Uses of (word-)frequency list

- List of frequently occurring lexical items (e.g., General Service List [West 1953], Academic Word List [Coxhead 2000])
- Usage-based Cognitive Linguistics
 - degree of productivity (based on type frequency) and cognitive entrenchment (based on token frequency) of certain linguistic units
- Choosing experimental stimuli
- Spelling error correction
- Determining vocabulary sizes of learners
- Selection and ordering of language features in course textbooks
- In other corpus linguistic tools: keyword and collocation statistics
- ...

See Gries (2017: 13–14) and Miller (2020: 77–78) for details

Outline

1. ~~What is a frequency list?~~
2. ~~Two basic types of frequency~~
3. ~~Frequency of different linguistic units~~
4. ~~Examples of uses of frequency list~~
5. Demo & Practice

Demo & Practice

- Demo:
 - Queries: Basic & Advanced features of SE's *Wordlist*
 - Outputs: Options for exploring outputs
- Corpus: Brown Family (CLAWS + TreeTagger tags)
- Practices
 - also need a spreadsheet software (e.g., Excel, LibreOffice Calc, Google Spreadsheet)

Demo & Practice

- Demo:
 - Queries: Basic & Advanced features of SE's *Wordlist*

The screenshot shows the Stanford CoreNLP Wordlist interface. At the top, there is a header with the text "WORDLIST" and "Brown Family (CLAWS + TreeTagger tags)" followed by a search bar and a help icon. To the right are links for "Get more space", "Help", "Feedback", "FAQ", and "Log in". Below the header, there are three tabs: "BASIC" (which is selected), "ADVANCED", and "ABOUT". The "BASIC" tab contains a "find ?" input field and a dropdown menu labeled "words" with options: lemmas, adjective, adverb, conjunction, noun, preposition, and pronoun. The "ADVANCED" tab contains a "find ?" input field and a dropdown menu labeled "all" with options: starting with, ending with, and containing. A large red "GO" button is located at the bottom center of the interface.

Layer 1:

Various restricted searches (words, lemmas, POS)

Layer 2:

Capturing all or parts of the units restricted in Layer 1

Demo & Practice

- Demo:
 - Queries: Basic & Advanced features of SE's *Wordlist*

The screenshot shows the Sketch Engine Wordlist interface with the following configuration:

- Words:** Lemmas, adjective, adverb, conjunction, noun, preposition, nnnnoun.
- Find?**: words
- Text types?**: doc.date, doc.region, doc.title, doc.corpus, doc.genre (selected).
- Exclude these words:** (empty)
- Include nonwords?**: checked
- A = a?**: checked
- Frequency min?**: 0
- Frequency max?**: 0
- result format:** Simple list (selected)
- Subcorpus?**: none (the whole corp...)
- GO** button

Demo & Practice

- Demo: Advanced feature
- Question:
 - contrasting the list of nouns in the Mystery sub-genre of the Fiction genre of the American vs. British variety of the Brown Family
- Requirement(s):
 - the preloaded Brown Family has provided the (combined) subcorpora category for “American” and “British”
 - two searches: one for each variety
 - save each output into .csv
 - Explore the first 30 items: are there non-overlapping nouns? How many of them?
- Operationalisation (demo):
 - Find?: **noun** (layer 1) ; **all** (layer 2)
 - Display as:
 - * check **tag**
 - * check **lemma** (check the **A = a** of **lemma**)
 - Text types:

- * doc.genre: Mystery sub-genre of Fiction
- * doc.region: American
- Results (demo)
 - save into .csv and call it `noun-in-mystery-brownfam-AmE.csv`
 - run the search for British by changing only one criteria: doc.region (DEMO)

End of Frequency List

(slides: <https://github.com/complexico/dipscorling2024>)

References

- Cheng, Winnie. 2012. *Exploring corpus linguistics: Language in action* (Routledge Introductions to Applied Linguistics). London ; New York, NY: Routledge.
- Gries, Stefan Th. 2010. Useful statistics for corpus linguistics. In Aquilino Sánchez & Moisés Almela (eds.), *A mosaic of corpus linguistics: Selected approaches*, 269–291. Frankfurt am Main: Peter Lang. http://www.linguistics.ucsb.edu/faculty/stgries/research/2010_STG_UsefulStats4CorpLing_MosaicCorpLing.pdf.
- Gries, Stefan Th. 2017. *Quantitative corpus linguistics with R: A practical introduction*. Second edition. New York: Routledge.
- Miller, Don. 2020. Analysing Frequency Lists. In Magali Paquot & Stefan Th. Gries (eds.), *A Practical Handbook of Corpus Linguistics*, 77–97. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_4.
- Rajeg, Gede Primahadi Wijaya & Karlina Denistia. 2023. Afiksasi Verba dalam Bahasa Indonesia. Bali, Indonesia. <https://doi.org/10.6084/m9.figshare.22336729>.
- Weisser, Martin. 2016. *Practical corpus linguistics: An introduction to corpus-based language analysis*. First edition. Hoboken, NJ: Wiley-Blackwell.