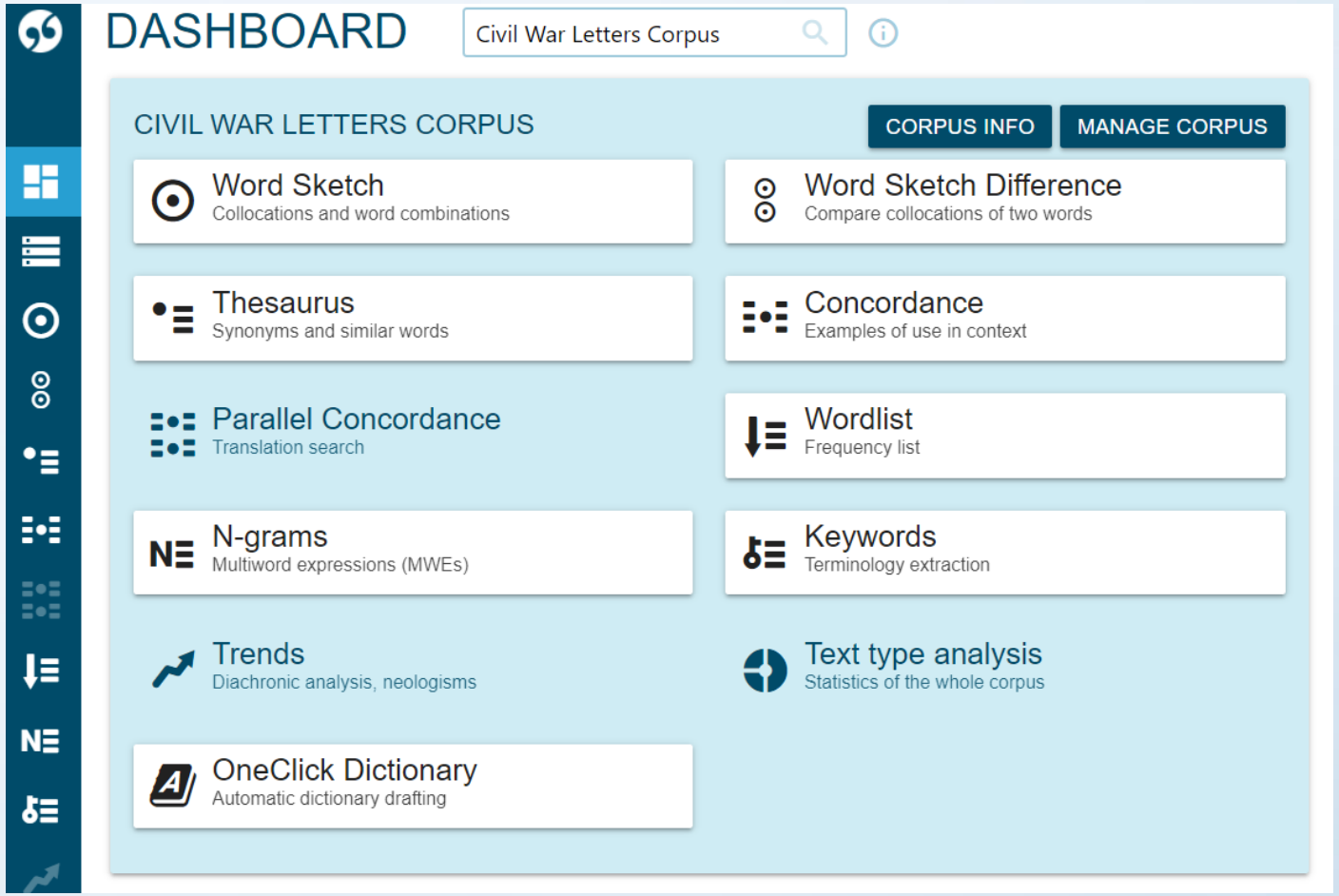
Research into the Methodology of Corpus Linguistics: Creating and Comparing Corpora

By: Isabella Guhl-Erdie
Faculty Sponsor: Dr. Felicia Jean Steele

What is a Corpus in Linguistics?

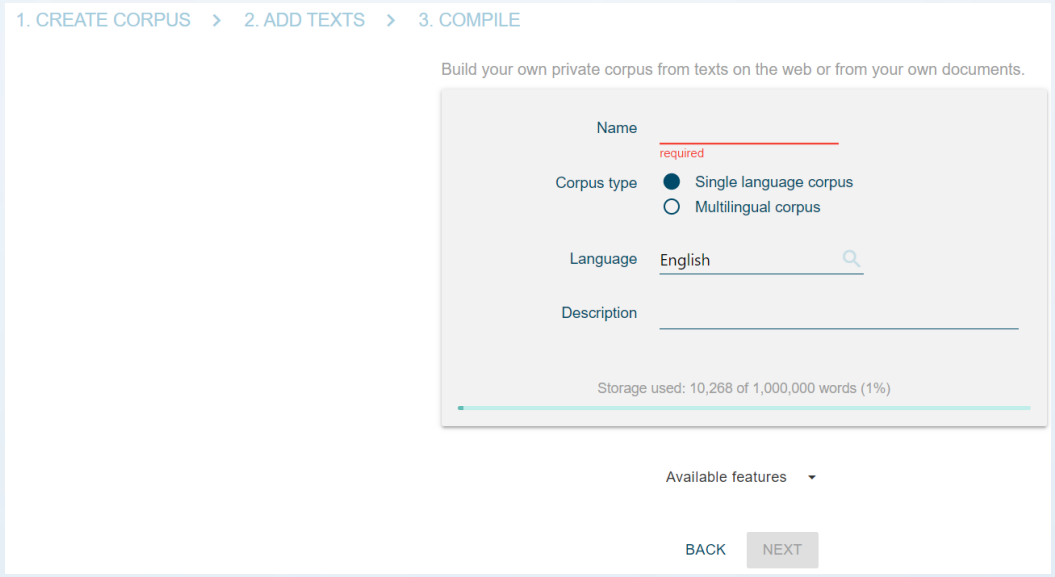
- “The body of written or spoken material upon which a linguistic analysis is based” (“corpus”, n.b).
- Examples:
  - AntConc
  - Wmatrix
  - Sketch Engine*

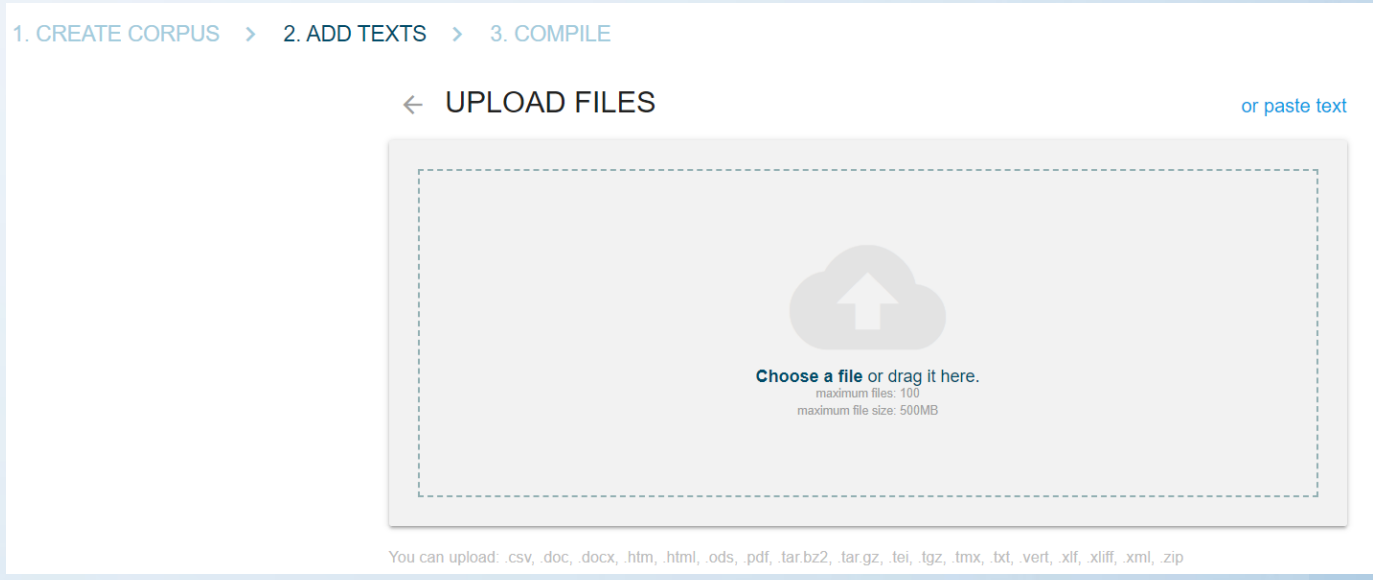


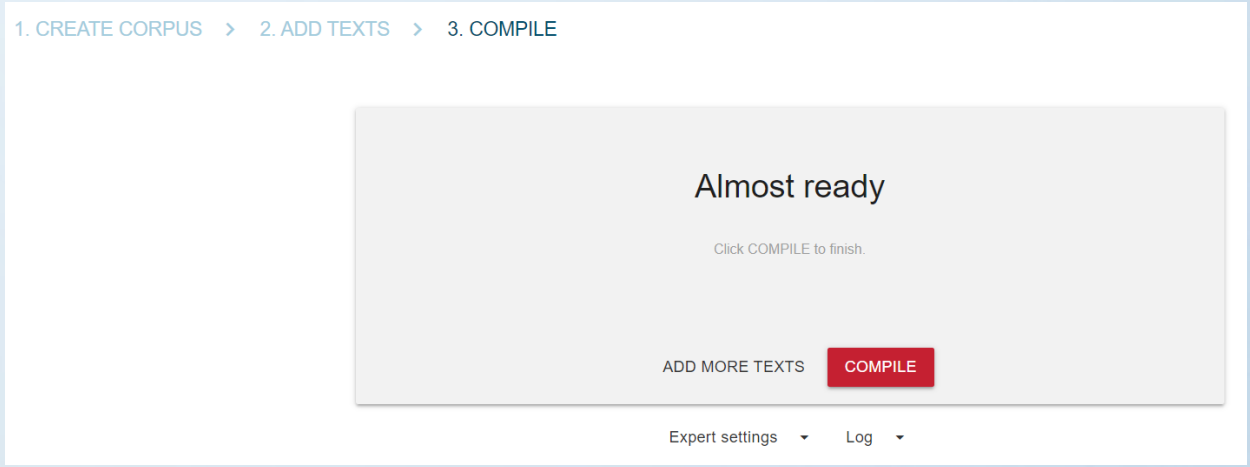
How does Sketch Engine work?

"Sketch Engine processes texts of billions of words and, within seconds, finds instances of the word, phrase or phenomenon and presents the results in the form of Word Sketches, concordances or word lists"

How did I use Sketch Engine in my research? As a body to house texts for analysis

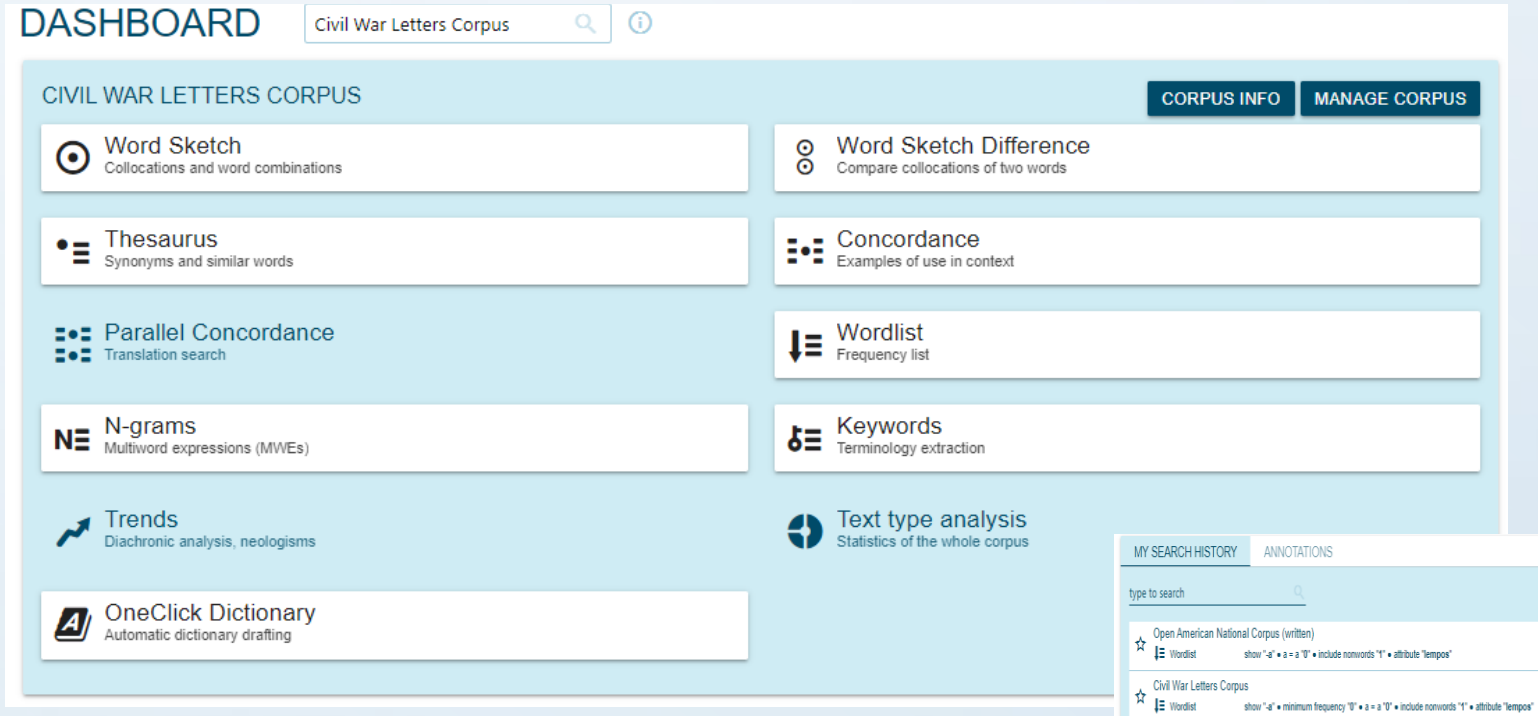


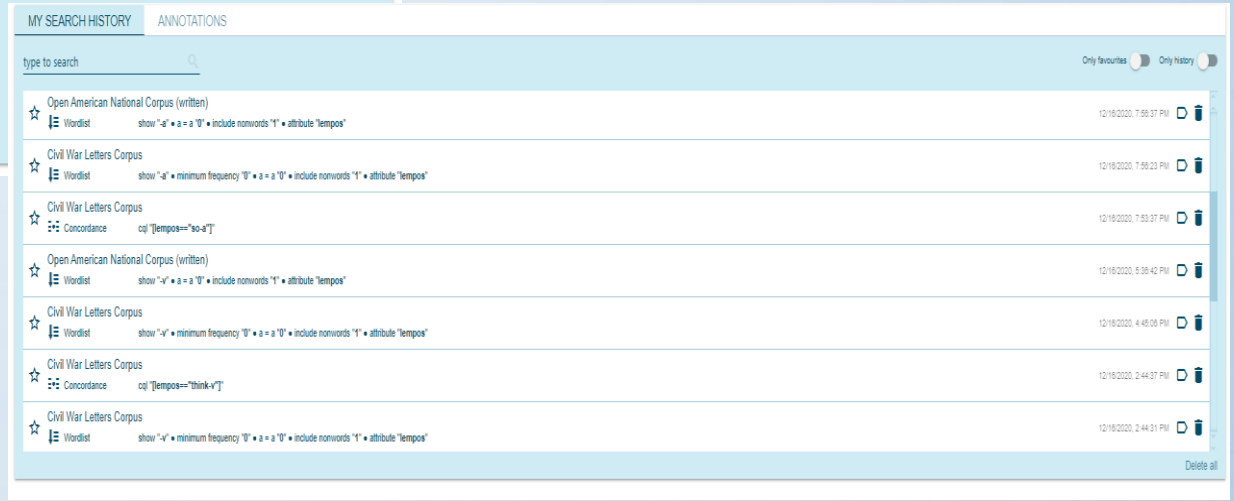


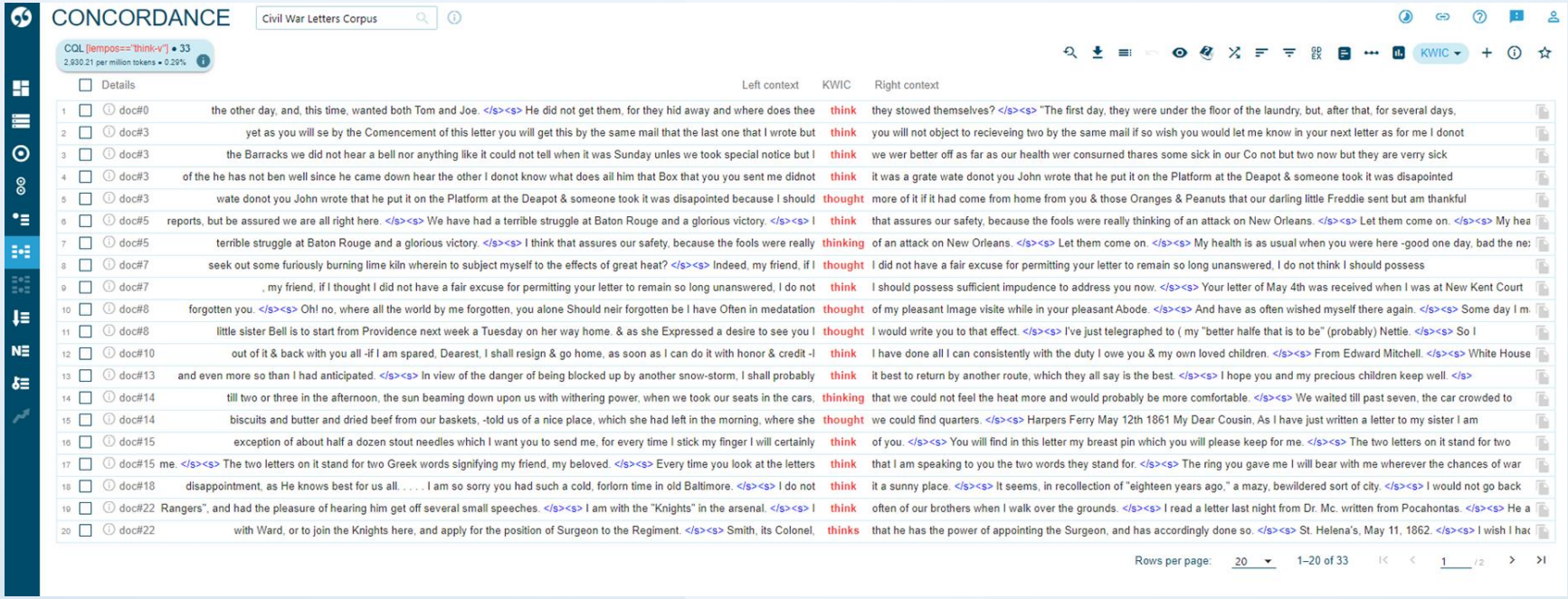


Compilation Process:

Sketch Engine Searches…



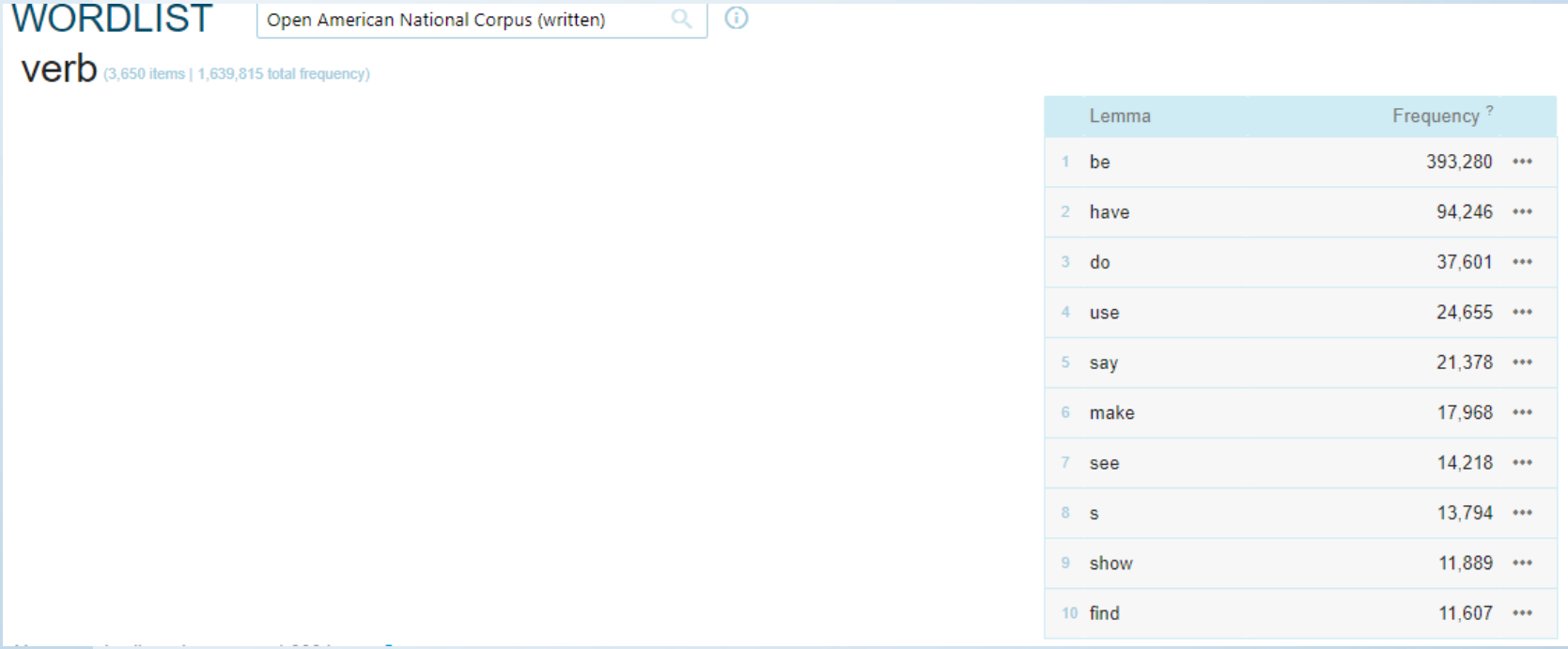




Search Results:

Comparing CWLC with Open American National Corpus

Initially concerned with hedging frequencies, but the verbs were too interesting!

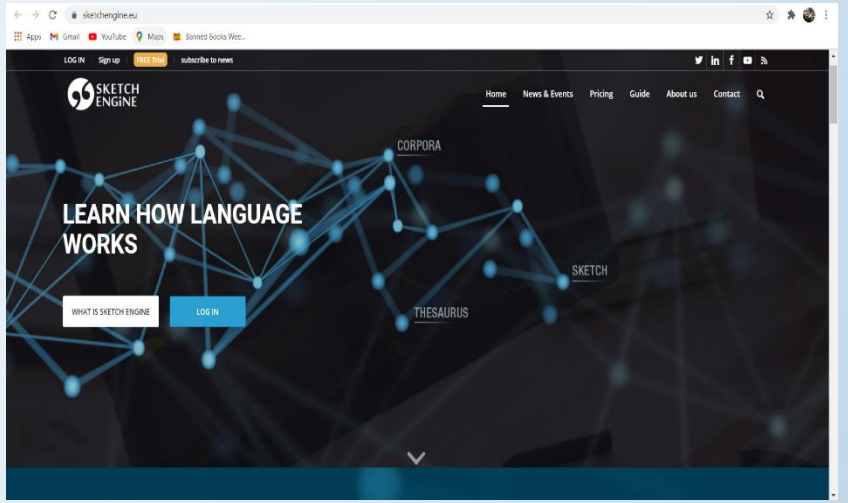


	Lemma	Frequency
1	be	393,280
2	have	94,246
3	do	37,601
4	use	24,655
5	say	21,378
6	make	17,968
7	see	14,218
8	s	13,794
9	show	11,889
10	find	11,607

Findings and Implications

- The unusual finding in comparison with OANCw were the frequencies of the verbs *think, come,* and *write*. The OANCw Corpus has over 11,000,000 words and so in comparing the frequencies I was surprised to find these 3 variants.
- The OANCw's verb frequency top 15 list is: *be*, *have*, *do*, *use*, *say*, *make*, *see*, '*s'*, *show*, *find*, *take*, *include*, *know*, *get*, *give*, etc.
- In Comparison, The CWLC has the same top three verbs followed by think, come, and write.
- In researching correspondence collection studies, I found that scholars such as Emma Moreton noted higher frequency in phrases such as "ich denke" which can be broadly translated as *I think* but noted no significant findings for *write* or *come* (619). Other Scholars including Marina Dossena have examined correspondences but have neglected an empirical approach.
- The importance of Corpus Linguistics is highlighted in these findings because it is an empirical approach which can verify or concretize linguistic phenomenon found through manual observation.

Works Cited



Textbook

Stefanowitsch, Anatol. Corpus Linguistics: A guide to methodology. Berlin: Language Science Press, 2020.

Online Sources

- Clinton, Catherine. "Southern Women and the Civil War." Journal of Women's History, vol. 8, no. 3, Johns Hopkins University Press, 1996, pp. 163–68

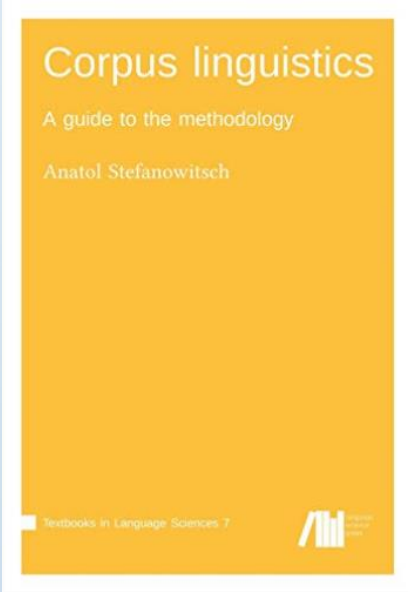
- Dossena, Marina. "'Many Strange and Peculiar Affairs': Description, Narration and Evaluation in Scottish Emigrants' Letters of the 19th Century." Scottish Language, vol. 27, Association for Scottish Literary Studies, 2008, p. 1–19.

- Moreton, Emma. 'Profiling the Female Emigrant: A Method of Linguistic Inquiry for Examining Correspondence Collections'. *Gender & History*, Vol.24 No.3 November 2012, pp. 617–646.

- *Sketch Engine*. Lexical Computing. 2003. https://www.sketchengine.eu/. Accessed 15 October 2020.

- "corpus, n." OED Online, Oxford University Press, March 2021, www.oed.com/view/Entry/41873. Accessed 10 April 2021.

- North American Women's Letters and Diaries Database. Web. Accessed 15 October 2020.



Questions & Thoughts