



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*

DASHBOARD Civil War Letters Corpus

CIVIL WAR LETTERS CORPUS **CORPUS INFO** **MANAGE CORPUS**

- Word Sketch**
Collocations and word combinations
- Word Sketch Difference**
Compare collocations of two words
- Thesaurus**
Synonyms and similar words
- Concordance**
Examples of use in context
- Parallel Concordance**
Translation search
- Wordlist**
Frequency list
- N-grams**
Multiword expressions (MWEs)
- Keywords**
Terminology extraction
- Trends**
Diachronic analysis, neologisms
- Text type analysis**
Statistics of the whole corpus
- OneClick Dictionary**
Automatic dictionary drafting

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

1. CREATE CORPUS > 2. ADD TEXTS > 3. COMPILE

Build your own private corpus from texts on the web or from your own documents.

Name required

Corpus type Single language corpus Multilingual corpus

Language 🔍

Description


Storage used: 10,268 of 1,000,000 words (1%)

Available features ▼

BACK NEXT

1. CREATE CORPUS > 2. ADD TEXTS > 3. COMPILE

← UPLOAD FILES or paste text



Choose a file or drag it here.
maximum files: 100
maximum file size: 500MB

You can upload: .csv, .doc, .docx, .htm, .html, .ods, .pdf, .tar.bz2, .tar.gz, .tei, .tgz, .tmx, .txt, .vert, .xif, .xliff, .xml, .zip

1. CREATE CORPUS > 2. ADD TEXTS > 3. COMPILE

Almost ready

Click COMPILE to finish.

ADD MORE TEXTS **COMPILE**

Expert settings ▼ Log ▼

Compilation Process:

Sketch Engine Searches...

DASHBOARD

Civil War Letters Corpus



CIVIL WAR LETTERS CORPUS

CORPUS INFO

MANAGE CORPUS

Word Sketch
Collocations and word combinations

Word Sketch Difference
Compare collocations of two words

Thesaurus
Synonyms and similar words

Concordance
Examples of use in context

Parallel Concordance
Translation search

Wordlist
Frequency list

N-grams
Multiword expressions (MWEs)

Keywords
Terminology extraction

Trends
Diachronic analysis, neologisms

Text type analysis
Statistics of the whole corpus

OneClick Dictionary
Automatic dictionary drafting

MY SEARCH HISTORY ANNOTATIONS

type to search

Only favourites Only history

☆	Open American National Corpus (written)	Wordlist	show "-d" • a = a "l" • include nonwords "1" • attribute "lempos"	12/18/2020, 7:59:37 PM	
☆	Civil War Letters Corpus	Wordlist	show "-d" • minimum frequency "0" • a = a "l" • include nonwords "1" • attribute "lempos"	12/18/2020, 7:59:23 PM	
☆	Civil War Letters Corpus	Concordance	col "[lempos=="so-d"]"	12/18/2020, 7:59:37 PM	
☆	Open American National Corpus (written)	Wordlist	show "-d" • a = a "l" • include nonwords "1" • attribute "lempos"	12/18/2020, 8:38:42 PM	
☆	Civil War Letters Corpus	Wordlist	show "-d" • minimum frequency "0" • a = a "l" • include nonwords "1" • attribute "lempos"	12/18/2020, 4:48:05 PM	
☆	Civil War Letters Corpus	Concordance	col "[lempos=="think-v"]"	12/18/2020, 2:44:37 PM	
☆	Civil War Letters Corpus	Wordlist	show "-d" • minimum frequency "0" • a = a "l" • include nonwords "1" • attribute "lempos"	12/18/2020, 2:44:31 PM	

Delete all

CONCORDANCE Civil War Letters Corpus

CQL [lempos=="think-v"] • 33
2,930.21 per million tokens • 0.26%

Left context KWIC Right context

1	doc#0	the other day, and, this time, wanted both Tom and Joe. </s><s> He did not get them, for they hid away and where does thee	think	they stowed themselves? </s><s> "The first day, they were under the floor of the laundry, but, after that, for several days,
2	doc#3	yet as you will se by the Comencement of this letter you will get this by the same mail that the last one that I wrote but	think	you will not object to recieving two by the same mail if so wish you would let me know in your next letter as for me I donot
3	doc#3	the Barracks we did not hear a bell nor anything like it could not tell when it was Sunday unles we took special notice but I	think	we wer better off as far as our health wer consumed thares some sick in our Co not but two now but they are very sick
4	doc#3	of the he has not ben well since he came down hear the other I donot know what does all him that Box that you you sent me didnot	think	it was a grate wate donot you John wrote that he put it on the Platform at the Deapot & someone took it was disappointed
5	doc#3	wate donot you John wrote that he put it on the Platform at the Deapot & someone took it was disappointed because I should	thought	more of it if it had come from home from you & those Oranges & Peanuts that our darling little Freddie sent but am thankful
6	doc#5	reports, but be assured we are all right here. </s><s> We have had a terrible struggle at Baton Rouge and a glorious victory. </s><s> I	think	that assures our safety, because the fools were really thinking of an attack on New Orleans. </s><s> Let them come on. </s><s> My hea
7	doc#5	terrible struggle at Baton Rouge and a glorious victory. </s><s> I think that assures our safety, because the fools were really	thinking	of an attack on New Orleans. </s><s> Let them come on. </s><s> My health is as usual when you were here -good one day, bad the ne:
8	doc#7	seek out some furiously burning lime kiln wherein to subject myself to the effects of great heat? </s><s> Indeed, my friend, if I	thought	I did not have a fair excuse for permitting your letter to remain so long unanswered, I do not think I should possess
9	doc#7	, my friend, if I thought I did not have a fair excuse for permitting your letter to remain so long unanswered, I do not	think	I should possess sufficient impudence to address you now. </s><s> Your letter of May 4th was received when I was at New Kent Court
10	doc#8	forgotten you </s><s> Oh! no, where all the world by me forgotten, you alone Should neir forgotten be I have Ofen in medatation	thought	of my pleasant image visite while in your pleasant Abode. </s><s> And have as often wished myself there again. </s><s> Some day I m
11	doc#8	little sister Bell is to start from Providence next week a Tuesday on her way home. & as she Expressed a desire to see you I	thought	I would write you to that effect. </s><s> I've just telegraphed to (my "better halfe that is to be" (probably) Nettie. </s><s> So I
12	doc#10	out of it & back with you all -if I am spared, Dearest, I shall resign & go home, as soon as I can do it with honor & credit -I	think	I have done all I can consistently with the duty I owe you & my own loved children. </s><s> From Edward Mitchell. </s><s> White House
13	doc#13	and even more so than I had anticipated. </s><s> In view of the danger of being blocked up by another snow-storm, I shall probably	think	it best to return by another route, which they all say is the best. </s><s> I hope you and my precious children keep well. </s>
14	doc#14	till two or three in the afternoon, the sun beaming down upon us with withering power, when we took our seats in the cars,	thinking	that we could not feel the heat more and would probably be more comfortable. </s><s> We waited till past seven, the car crowded to
15	doc#14	biscuits and butter and dried beef from our baskets, -told us of a nice place, which she had left in the morning, where she	thought	we could find quarters. </s><s> Harpers Ferry May 12th 1861 My Dear Cousin, As I have just written a letter to my sister I am
16	doc#15	exception of about half a dozen stout needles which I want you to send me, for every time I stick my finger I will certainly	think	of you. </s><s> You will find in this letter my breast pin which you will please keep for me. </s><s> The two letters on it stand for two
17	doc#15	me. </s><s> The two letters on it stand for two Greek words signifying my friend, my beloved. </s><s> Every time you look at the letters	think	that I am speaking to you the two words they stand for. </s><s> The ring you gave me I will bear with me wherever the chances of war
18	doc#18	disappointment, as He knows best for us all. I am so sorry you had such a cold, forlorn time in old Baltimore. </s><s> I do not	think	it a sunny place. </s><s> It seems, in recollection of "eighteen years ago," a mazy, bewildered sort of city. </s><s> I would not go back
19	doc#22	Rangers", and had the pleasure of hearing him get off several small speeches. </s><s> I am with the "Knights" in the arsenal. </s><s> I	think	often of our brothers when I walk over the grounds. </s><s> I read a letter last night from Dr. Mc. written from Pocahontas. </s><s> He a
20	doc#22	with Ward, or to join the Knights here, and apply for the position of Surgeon to the Regiment. </s><s> Smith, its Colonel,	thinks	that he has the power of appointing the Surgeon, and has accordingly done so. </s><s> St. Helena's, May 11, 1862. </s><s> I wish I hac

Rows per page: 20 1-20 of 33 < > >>

Search Results:

Comparing CWLC with Open American National Corpus

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

WORDLIST

Open American National Corpus (written)

verb (3,650 items | 1,639,815 total frequency)

	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.

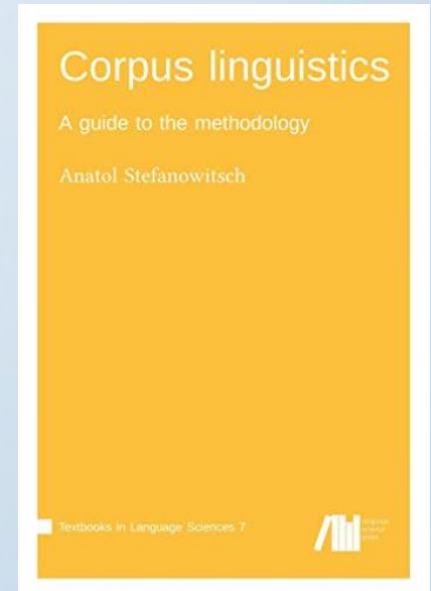
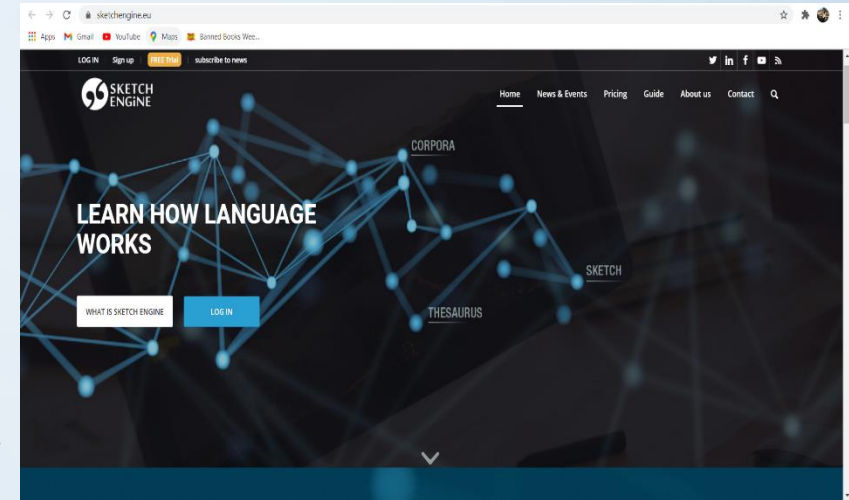
Works Cited

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

