Sketch Engine tutorial

English Linguistics Colloquium, LMU Munich

Quirin Würschinger

January 19, 2023

Table of contents

1 Tutorial materials

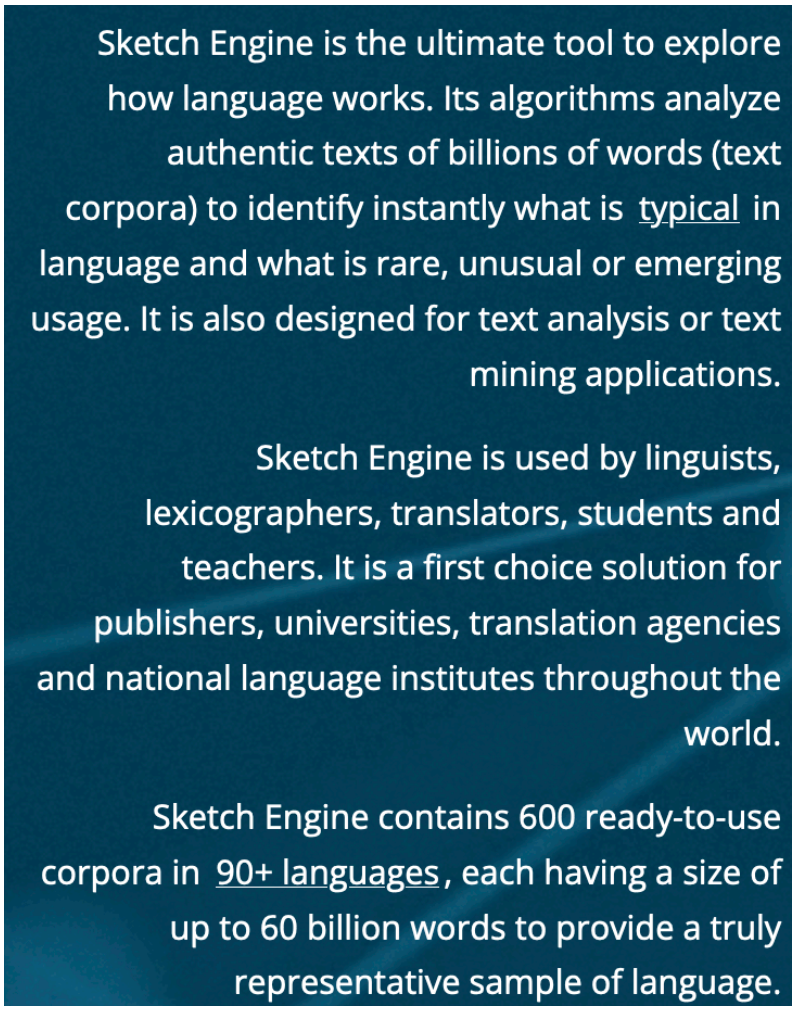
You can find all materials on this GitHub repo: https://github.com/wuqui/SkEtut

- website: https://wuqui.github.io/SkEtut/
- slides: https://wuqui.github.io/SkEtut/SkEtut_slides.html
- `pdf` handout: https://wuqui.github.io/SkEtut/SkEtut_handout.pdf
- data: https://github.com/wuqui/SkEtut/tree/main/data
- results: https://github.com/wuqui/SkEtut/tree/main/export

I will continue to work on these materials and would appreciate questions, comments, and suggestions via mail or on GitHub.

2 General information

2.1 What is Sketch Engine?



https://www.sketchengine.eu

2.2 LMU license

- funded by Projektfond des Nachwuchsförderungsprogramms der Fak. 13 (thanks to Susanne Handl for her help)
- funding secured for 2023 – hopefully longer?
- access: **all** LMU members (students and researchers) can access SkE using their LMU-Kennung

- please let me know if you use it

2.3 Related tools

- AntConc
- LancsBox
- WordSmith

2.4 Main features

corpus management

- **creating** corpora from your own data
- **hosting** these corpora online
- **annotating** corpora
- **sharing** your corpora

. . .

corpus analysis

- access to many **pre-loaded corpora**
- simple and complex **queries**
- **concordances**
- **collocation** analysis
- **text type** analysis

2.5 Resources

https://www.sketchengine.eu/quick-start-guide/

https://www.sketchengine.eu/guide/

3 Compiling corpora

3.1 Data format

https://www.sketchengine.eu/guide/create-corpus-from-files/

texts **without annotations**: most common

- structure: (ideally) use **1 document per file**

- file formats: **plain text**
  - `.txt`
  - `.csv`
  - …

annotated texts:

- `.xml`: powerful, but more involved

3.2 Raw data

Text files

```
845 CHAPTER 1. Loomings.
846
847 Call me Ishmael. Some years ago—never mind how long precisely—having
848 little or no money in my purse, and nothing particular to interest me
849 on shore, I thought I would sail about a little and see the watery part
850 of the world. It is a way I have of driving off the spleen and
851 regulating the circulation. Whenever I find myself growing grim about
852 the mouth; whenever it is a damp, drizzly November in my soul; whenever
853 I find myself involuntarily pausing before coffin warehouses, and
854 bringing up the rear of every funeral I meet; and especially whenever
855 my hypos get such an upper hand of me, that it requires a strong moral
856 principle to prevent me from deliberately stepping into the street, and
857 methodically knocking people's hats off—then, I account it high time to
858 get to sea as soon as I can. This is my substitute for pistol and ball.
859 With a philosophical flourish Cato throws himself upon his sword; I
860 quietly take to the ship. There is nothing surprising in this. If they
861 but knew it, almost all men in their degree, some time or other,
862 cherish very nearly the same feelings towards the ocean with me.
863
```

Melville: *Moby Dick* – available as part of Project Gutenberg

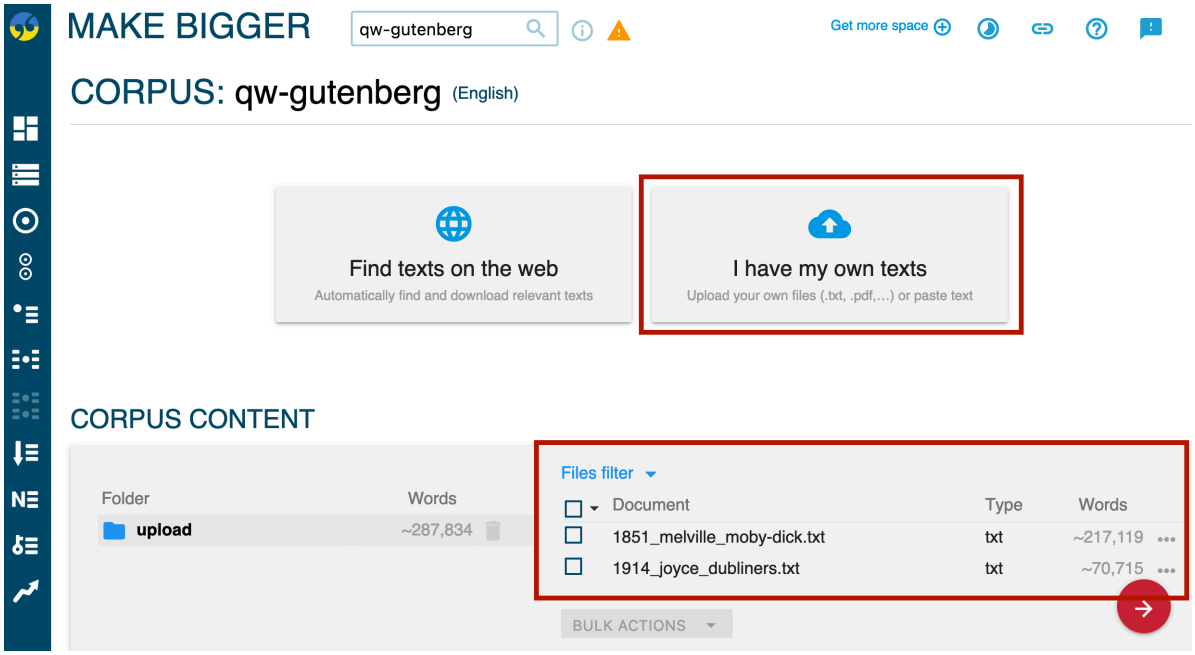
---

Tabular data

1. extract the column containing the text body from your spreadsheet (e.g. in new sheet)
2. export this column to `.csv`
3. you can then import this column in SkE just like a `.txt`

Note, however, that (meta)data in other columns will be lost[1].

3.3 Uploading files



3.4 Adding and editing metadata

3.5 Processed data

After compiling: `vert` ('vertical') format – word per line (WPL)[2]

[1] To preserve these data, you would have to convert your tabular data (`.xlsx` or `.csv`) into `xml` format before importing.

[2] More precisely: one *token* per line, including puctuation: e.g. `it`, `'s`, `,`.

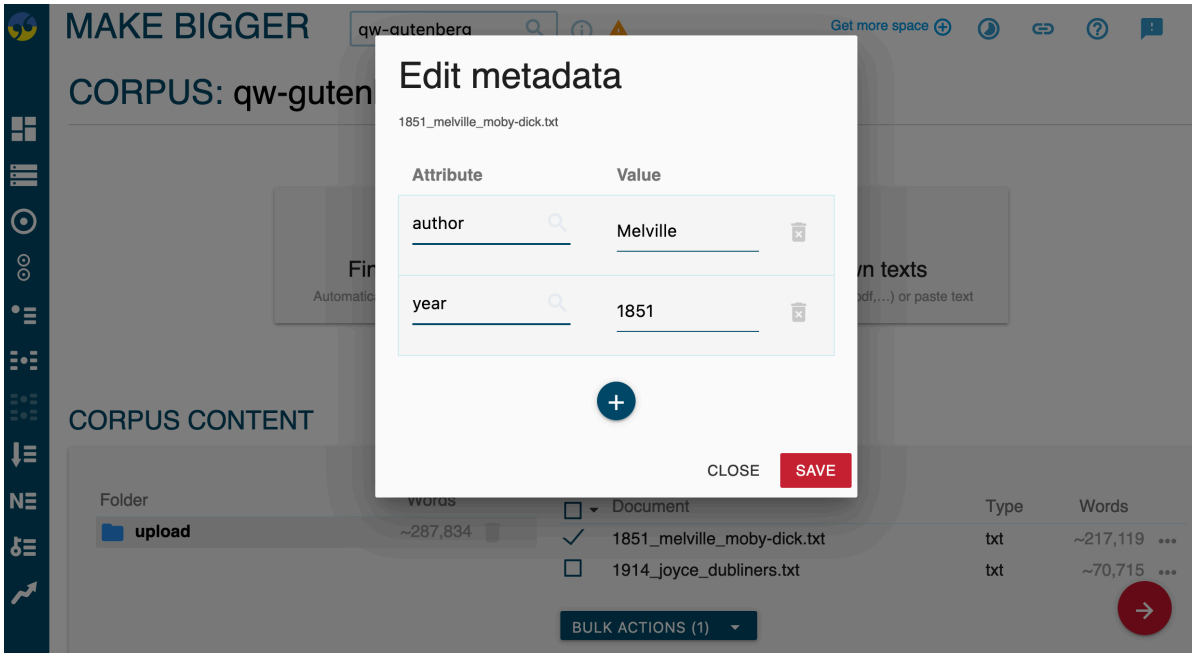
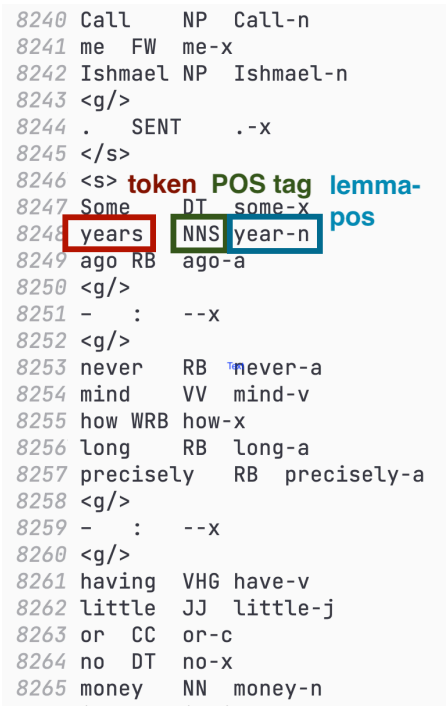


Figure 1: Adding and editing metadata

```
<doc id="file27506233"
     filename="1851_melville_moby-di
     ck.txt" parent_folder="upload"
     author="Melville" year="1851">
<s>
The DT  the-x
Project NP  Project-n
Gutenberg   NP  Gutenberg-n
eBook   NN  eBook-n
of  IN  of-i
Moby-Dick   NP  Moby-Dick-n
<g/>
;   :   ;-x
or  CC  or-c
The DT  the-x
Whale   NN  whale-n
<g/>
,   ,   ,-x
by  IN  by-i
Herman  NP  Herman-n
Melville    NP  Melville-n
```

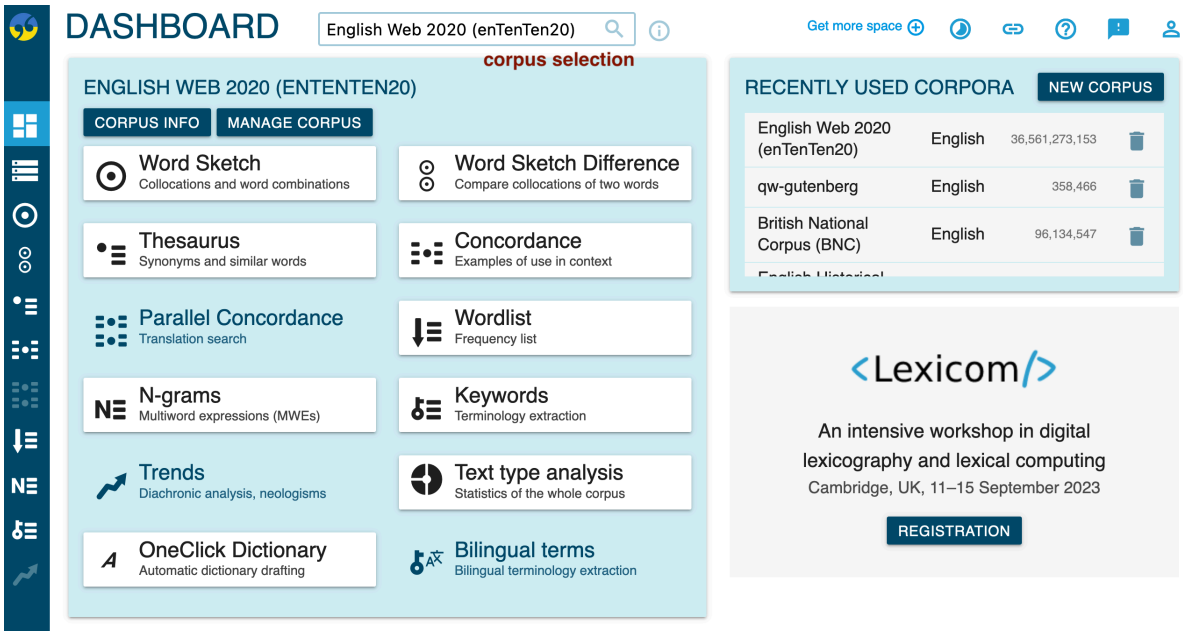
metadata




```
8240 Call    NP  Call-n
8241 me  FW  me-x
8242 Ishmael NP  Ishmael-n
8243 <g/>
8244 .   SENT    .-x
8245 </s>
8246 <s>
8247 Some    DT  some-x
8248 years   NNS year-n
8249 ago RB  ago-a
8250 <g/>
8251 -   :   --x
8252 <g/>
8253 never   RB  never-a
8254 mind    VV  mind-v
8255 how WRB how-x
8256 long    RB  long-a
8257 precisely   RB  precisely-a
8258 <g/>
8259 -   :   --x
8260 <g/>
8261 having  VHG have-v
8262 little  JJ  little-j
8263 or  CC  or-c
8264 no  DT  no-x
8265 money   NN  money-n
```

4 Analysing data

4.1 Dashboard



4.2 Available corpora

Browse full list of (English) corpora here.

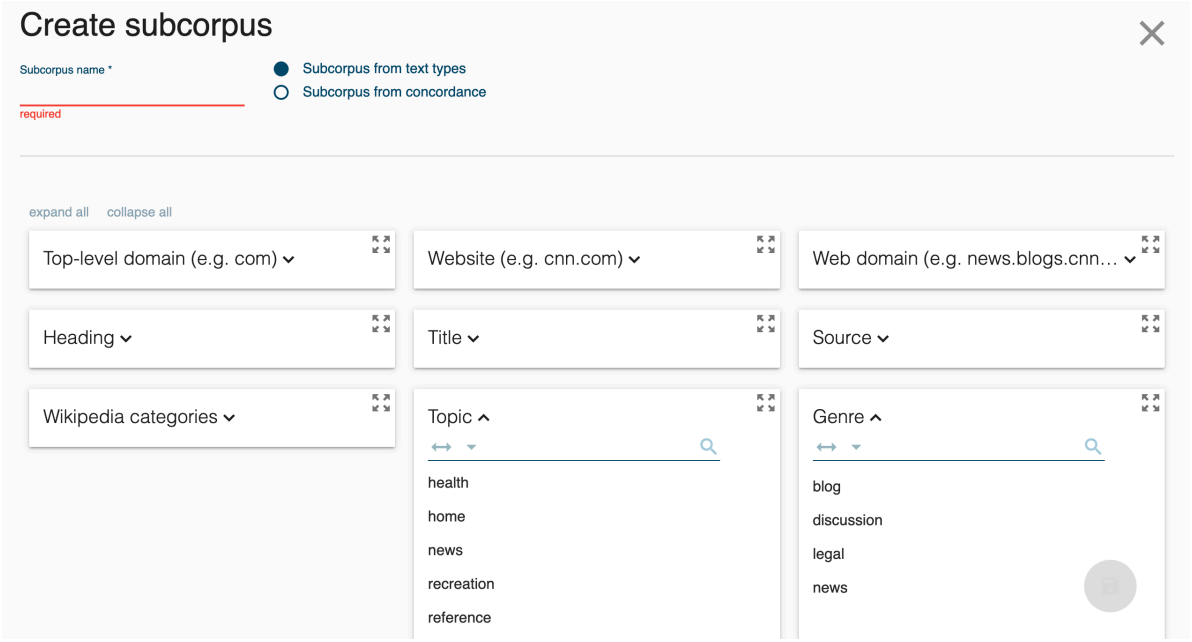
Among others, ...

- English Historical Book Collection (EEBO, ECCO, Evans): '**historical** corpus collection of English books published in the UK and the USA between 1473 and 1820'
- Gutenberg Project: large **book** corpus; also available for other languages
- British National Corpus (BNC): 'A balanced English corpus of samples of a written and spoken language of **British English** from the later part of the 20th century (1969–1994)'
- English Web 2020 (enTenTen20): big **web** corpus, including metadata about topics, genres, and web domains (e.g. `.com` vs `.co.uk`)
- Timestamped JSI web corpus 2014-2021 English: huge **diachronic web** corpus, annotated for topic, sentiment and more

4.3 Subcorpora

You can create subcorpora for pre-loaded and self-compiled corpora based on

- all available **metadata** categories (e.g. timestamps, topics, filenames)
- **concordance** searches

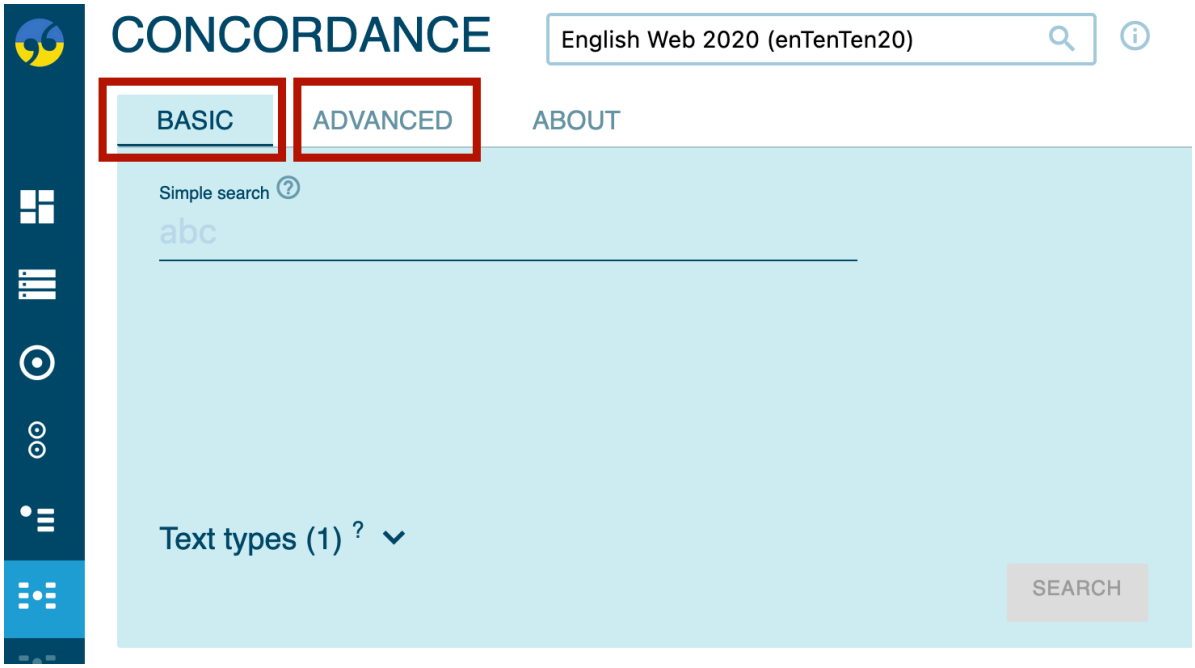


4.4 Queries

You run queries from the `Concordance` view.

There are two options:

- *basic* searches: basic
- *advanced* searches: more involved and powerful (e.g. searching for constructions based on lemmatized forms or word classes)


CONCORDANCE
English Web 2020 (enTenTen20)
BASIC
ADVANCED
ABOUT
Simple search
abc
Text types (1)
SEARCH

Basic queries

Simple search

Finds words which match the search word(s) or whose lemma (base form) matches the search word. It is not case sensitive.

Type a word or phrase. Use an asterisk (*) for any number of unspecified characters. Use a question mark (?) for exactly one unspecified character.

Use the pipe (|) to include more than one word or phrase.

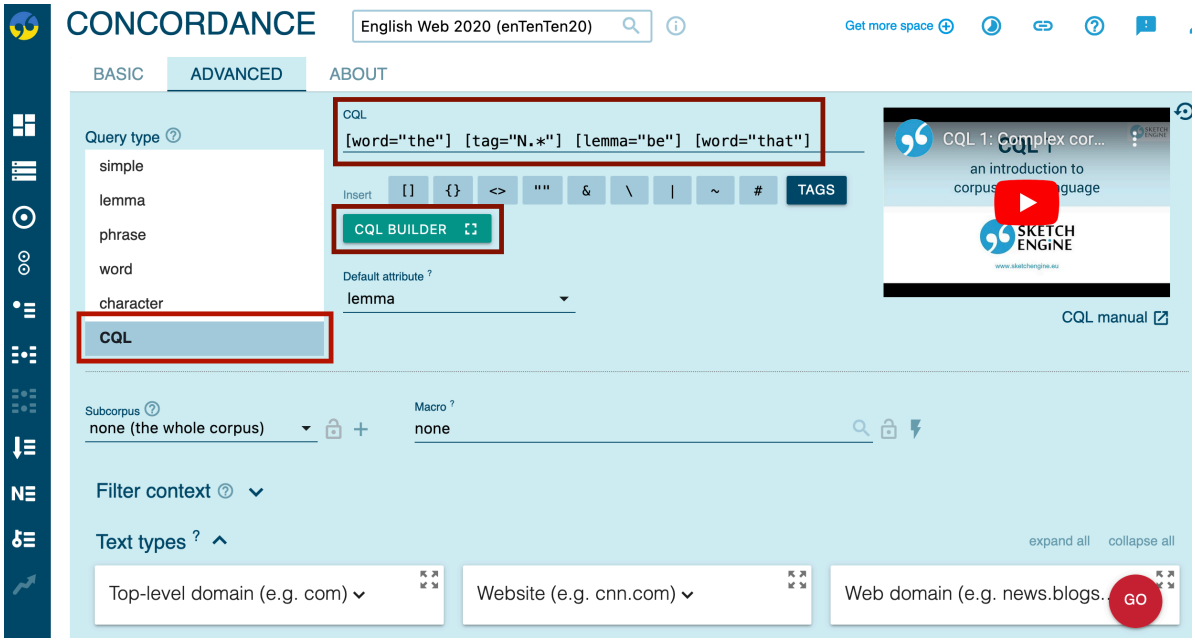
Use two hyphens (--) to search for a word that can be hyphenated, non-hyphenated or spelt as two separate words.

Examples

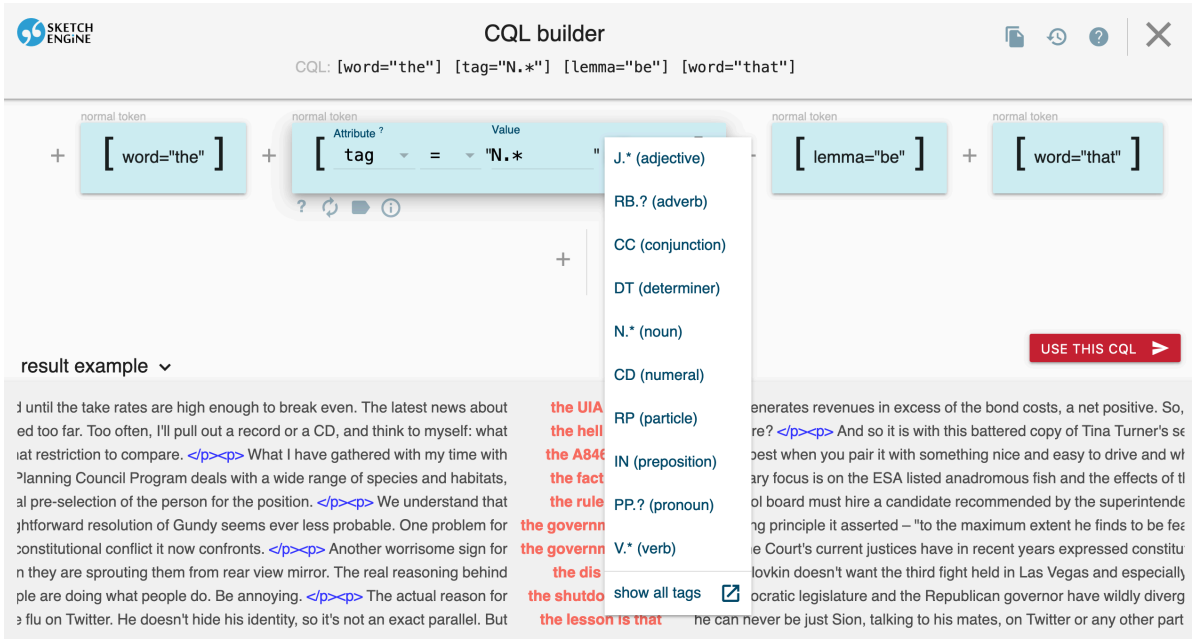
use	to find
m*	words or lemmas beginning *m-*
***ing**	words or lemmas ending *-ing*
have a * idea	unspecified word or lemma
have * * * idea	three unspecified words or lemmas
???t	4-letter words or lemmas ending *-t*
multi--billion	find *multi-billion* or *multibillion* or *multi billion*
return|go back	find *return* or *go back*

see also lemma

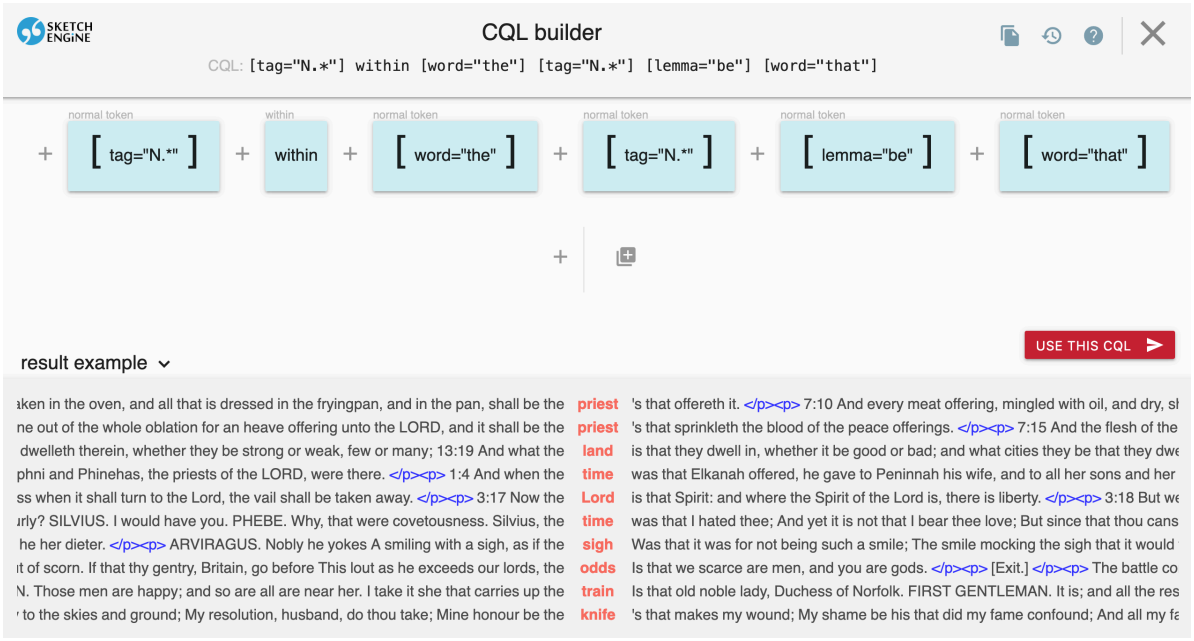
Advanced (CQL) queries



Helpful: manual and `CQL builder`.



Extracting parts of your query matches using `within`:

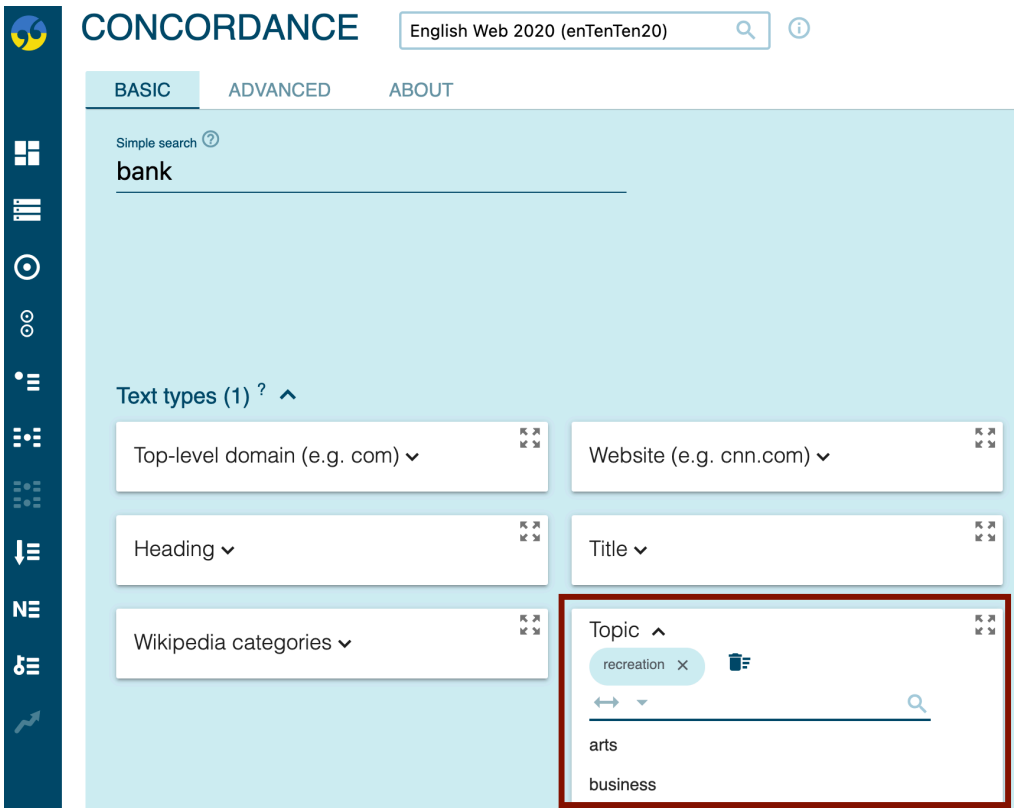


Filtering by metadata

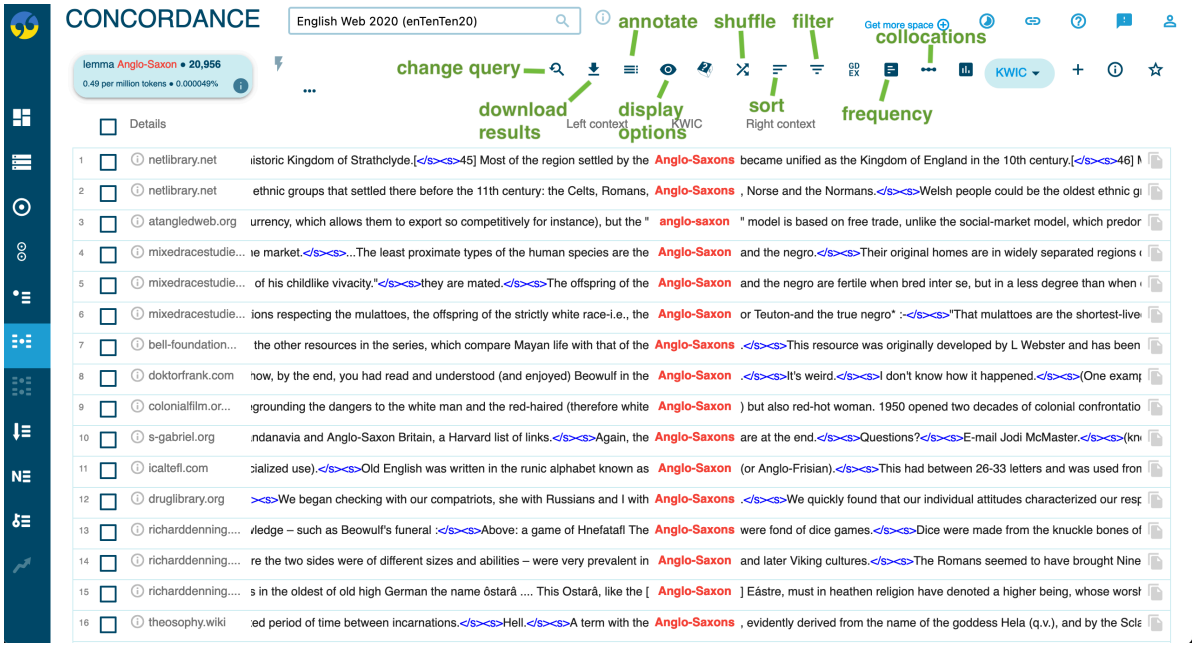
Options:

- query metadata within **CQL syntax** (e.g. `[word="bank"] within <doc topic="recreation" />`)
- perform 'text type'[3] filtering using the **dropdown menus**, which is also available for simple queries (see above).

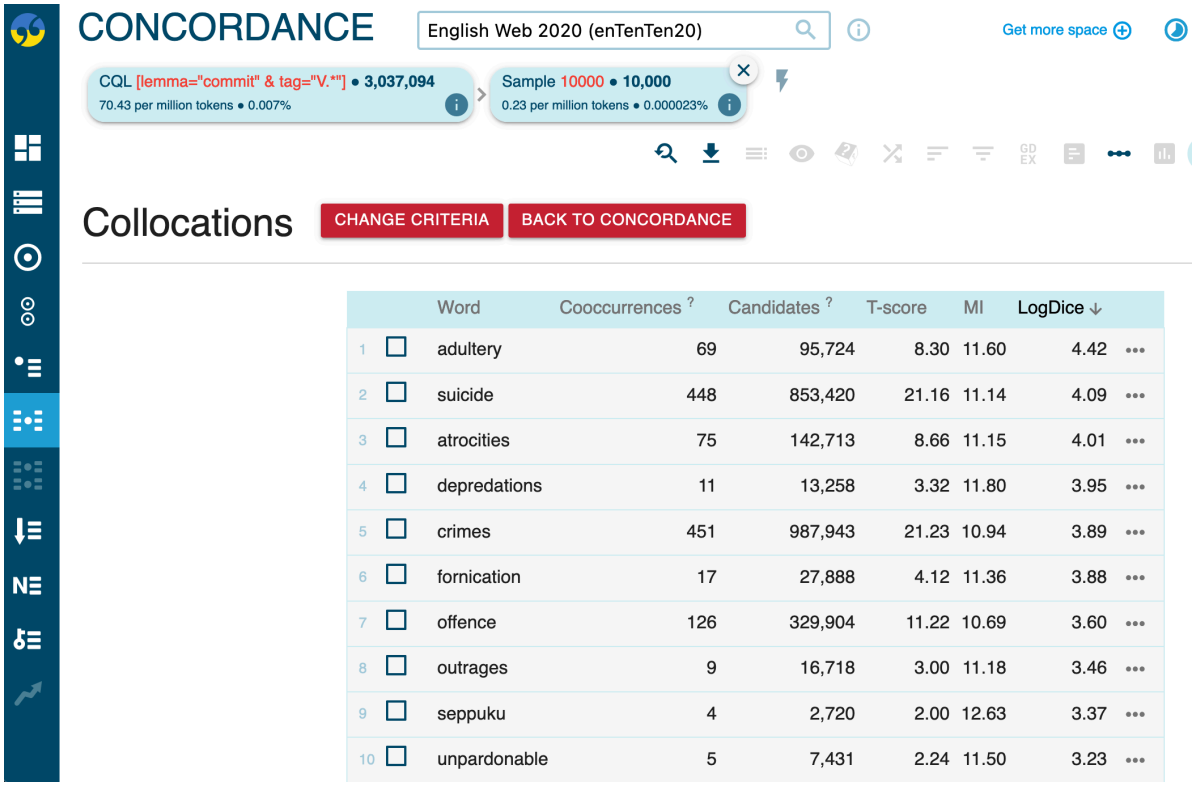
[3] 'Text types' in SkE are not text types in the linguistic sense, but in the technical sense: documents have different text types if they differ regarding any metadata category. For example, two 'types' could be texts tagged for `<doc year="1900">` vs `<doc year="2000">`.


CONCORDANCE
English Web 2020 (enTenTen20)
BASIC
ADVANCED
ABOUT
Simple search
bank
Text types (1)
Top-level domain (e.g. com)
Website (e.g. cnn.com)
Heading
Title
Wikipedia categories
Topic
recreation
arts
business

4.5 Concordance view



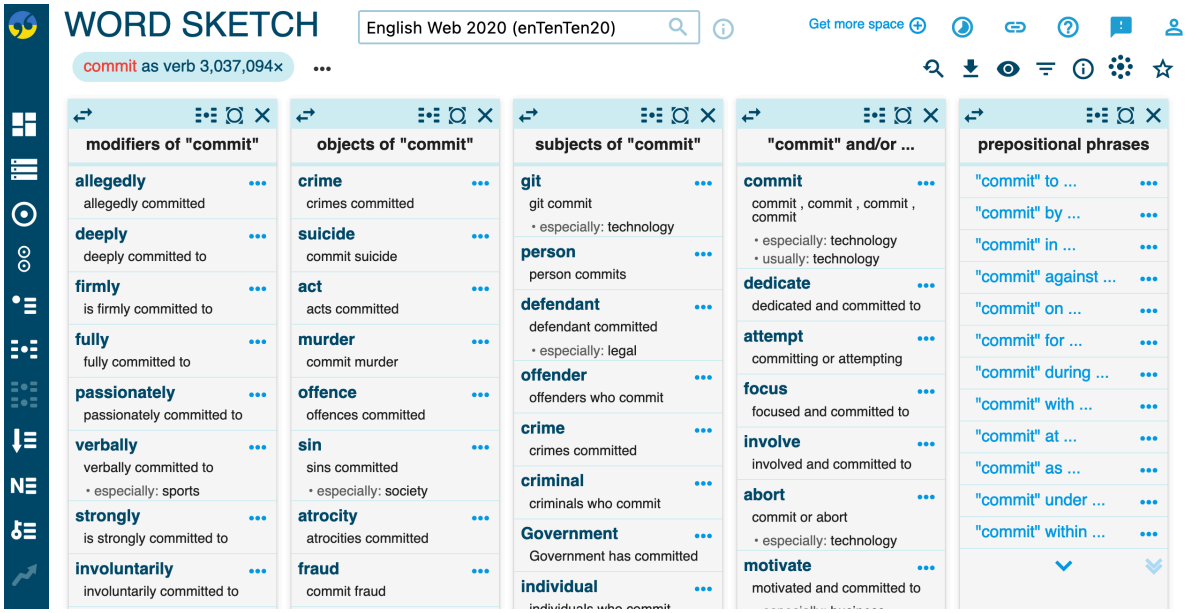
4.6 Collocations



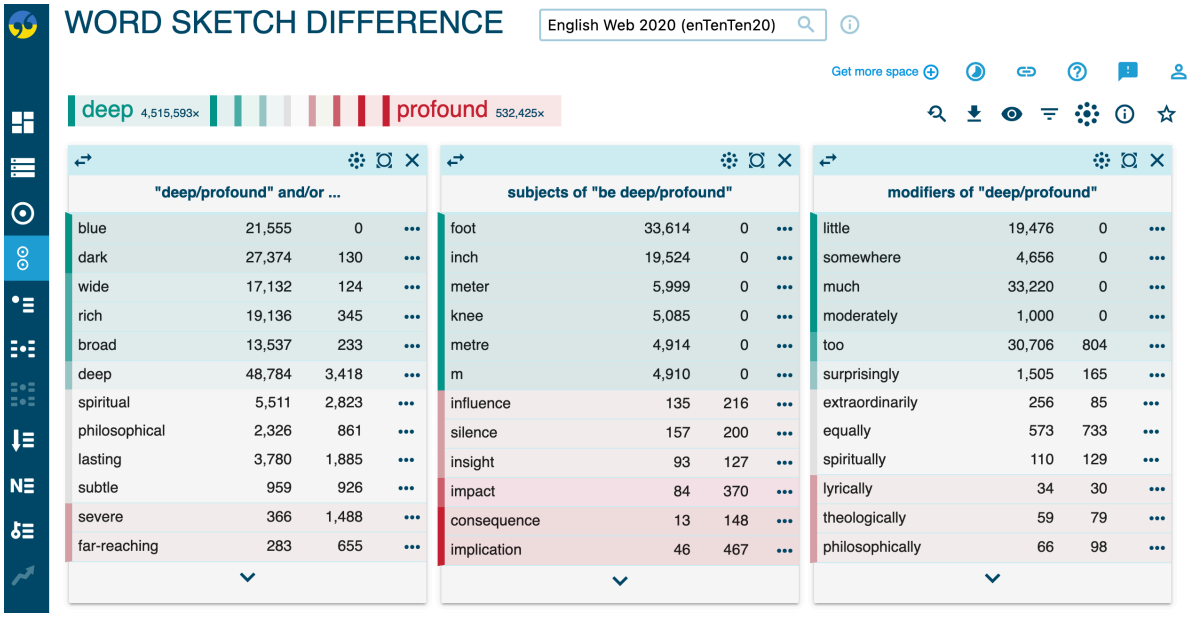
	Word	Cooccurrences	Candidates	T-score	MI	LogDice
1	adultery	69	95,724	8.30	11.60	4.42
2	suicide	448	853,420	21.16	11.14	4.09
3	atrocities	75	142,713	8.66	11.15	4.01
4	depredations	11	13,258	3.32	11.80	3.95
5	crimes	451	987,943	21.23	10.94	3.89
6	fornication	17	27,888	4.12	11.36	3.88
7	offence	126	329,904	11.22	10.69	3.60
8	outrages	9	16,718	3.00	11.18	3.46
9	seppuku	4	2,720	2.00	12.63	3.37
10	unpardonable	5	7,431	2.24	11.50	3.23

Additional measures (e.g. log likelihood) and other options are available in the advanced settings.

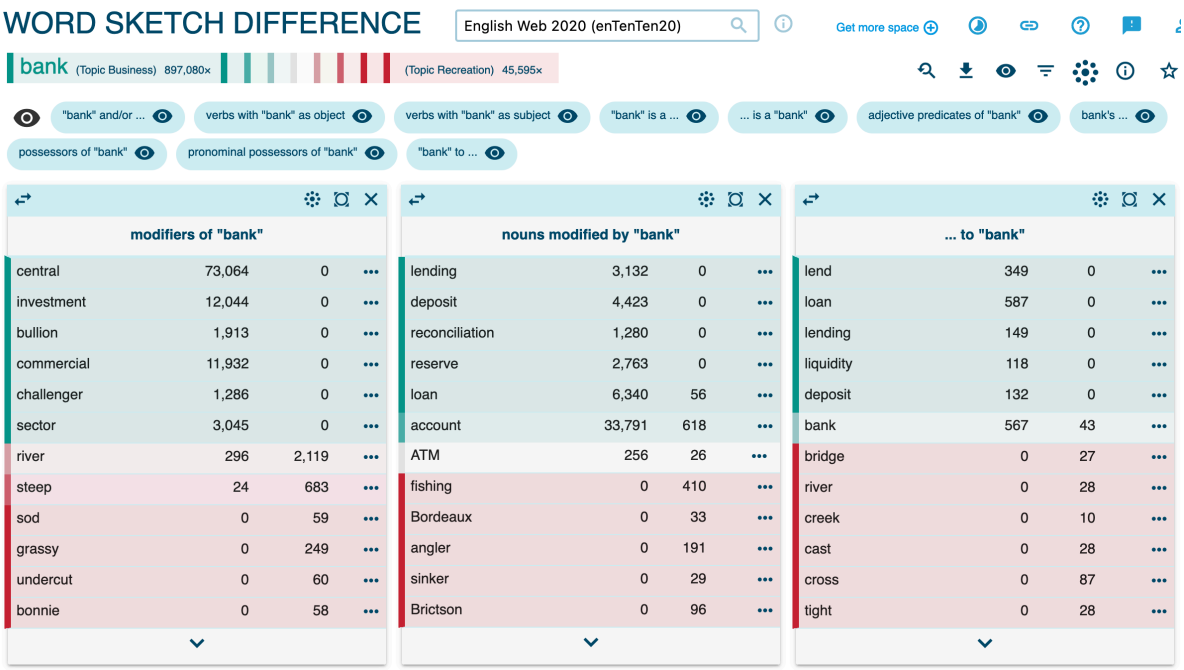
4.7 Word sketches



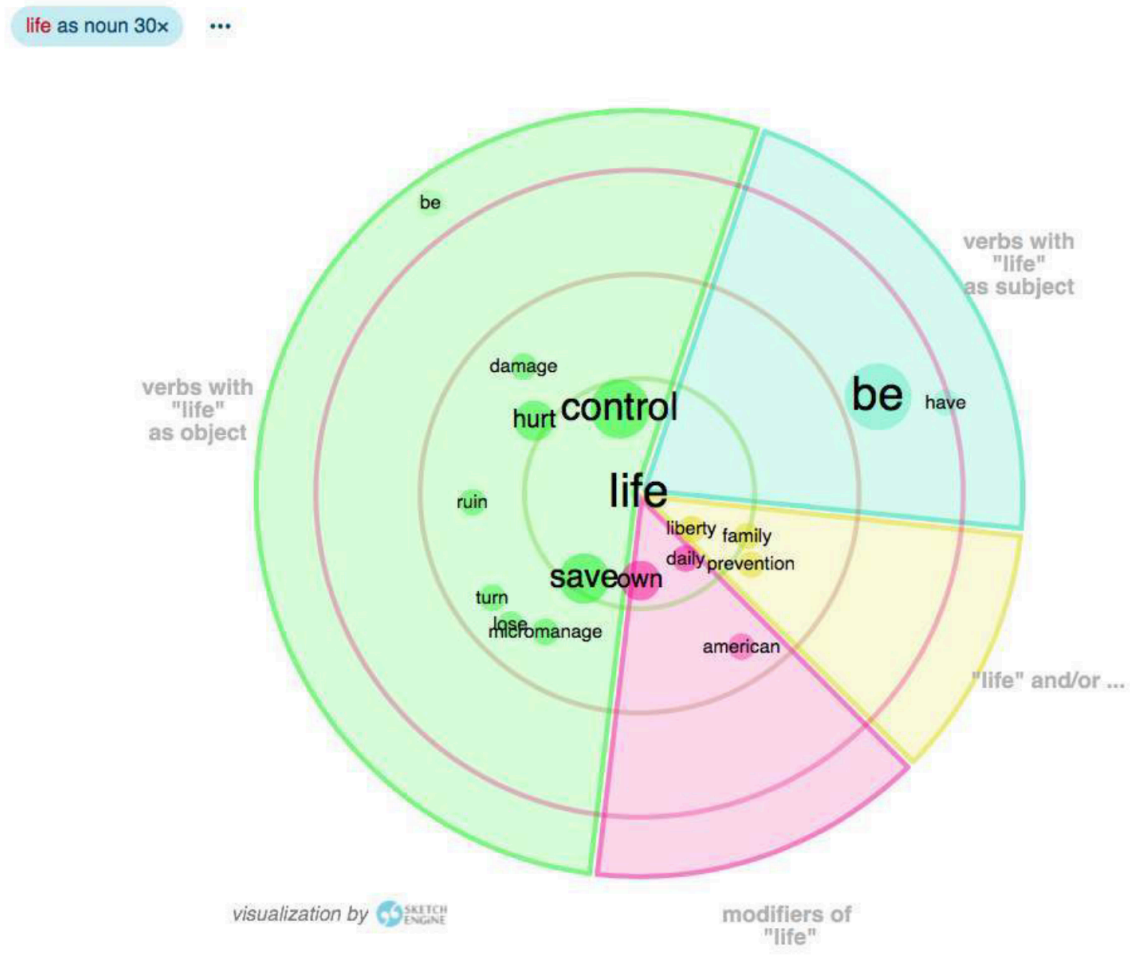
Word sketch difference: between two words/phrases



Word sketch difference: between two subcorpora



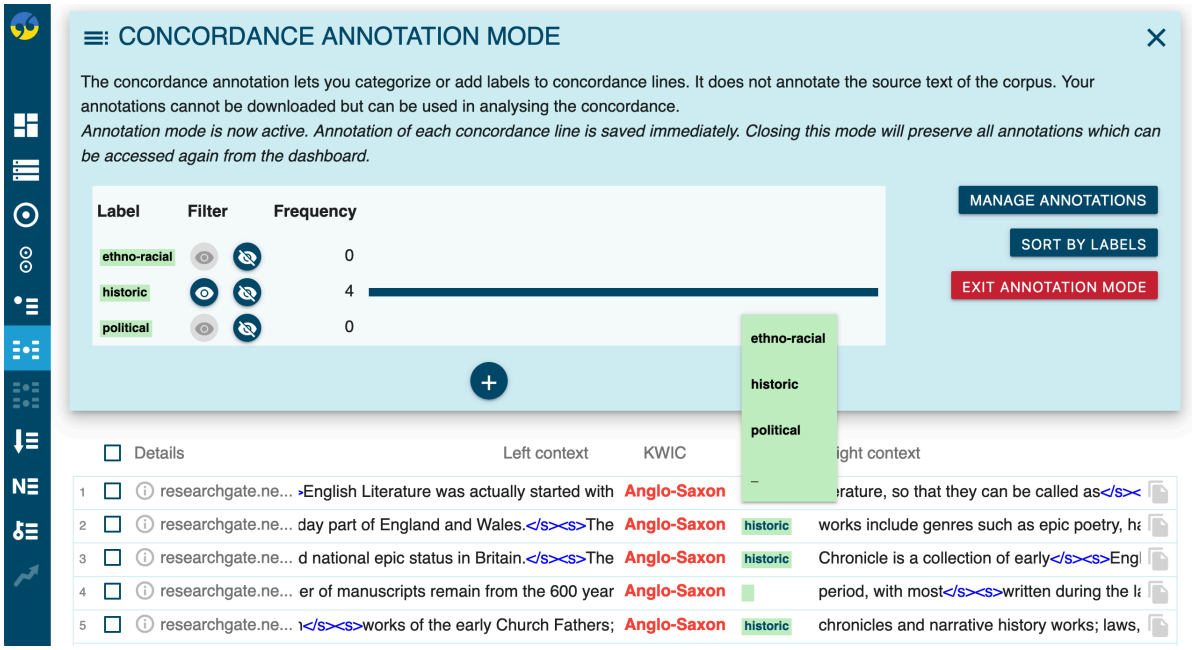
Visualizations



4.8 Annotating data

for metadata: see Figure 1 above

for concordance lines:



4.9 Exporting data

Almost everything can be exported:

- your entire annotated **corpora**
- results from **queries/concordances**
- results from **collocations**
- results from **word sketches**

I recommend exporting data in `.xlsx` format, since this seems to be best supported by SkE.[4]

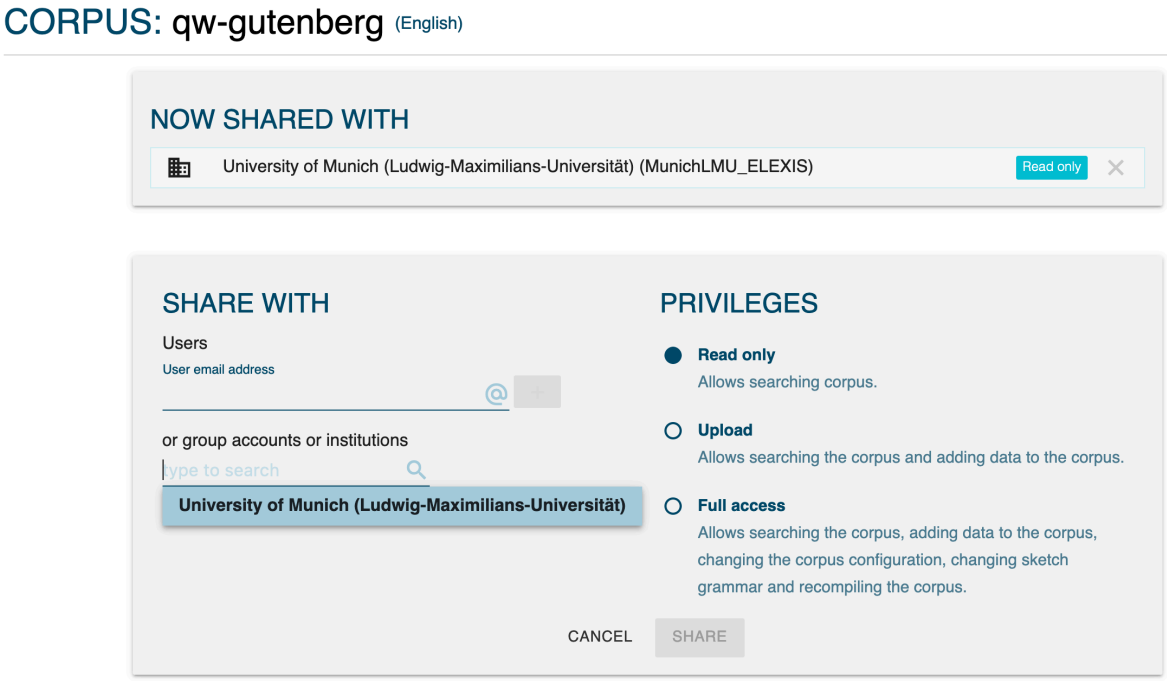
5 Use cases

5.1 Compiling a corpus: dead authors' minds

See Section *Compiling corpora* above.

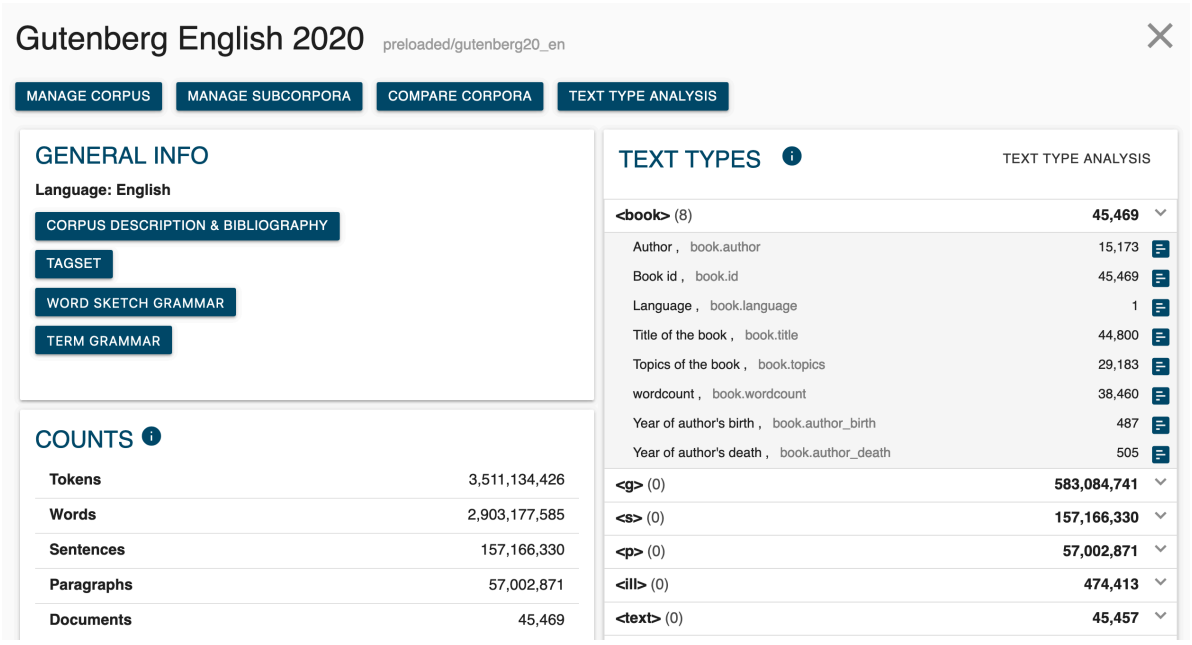
Sharing corpora: the toy corpus of Gutenberg books that I created for this tutorial is named `qw-gutenberg` and it should be accessible by all LMUlers.

[4]When exporting to `csv`, be careful with decimal/thousands separator: when using the `Text to columns` option in Excel, use `.` as decimal and `,` as thousands separator (e.g. one thousand point five: `1,000.5`).



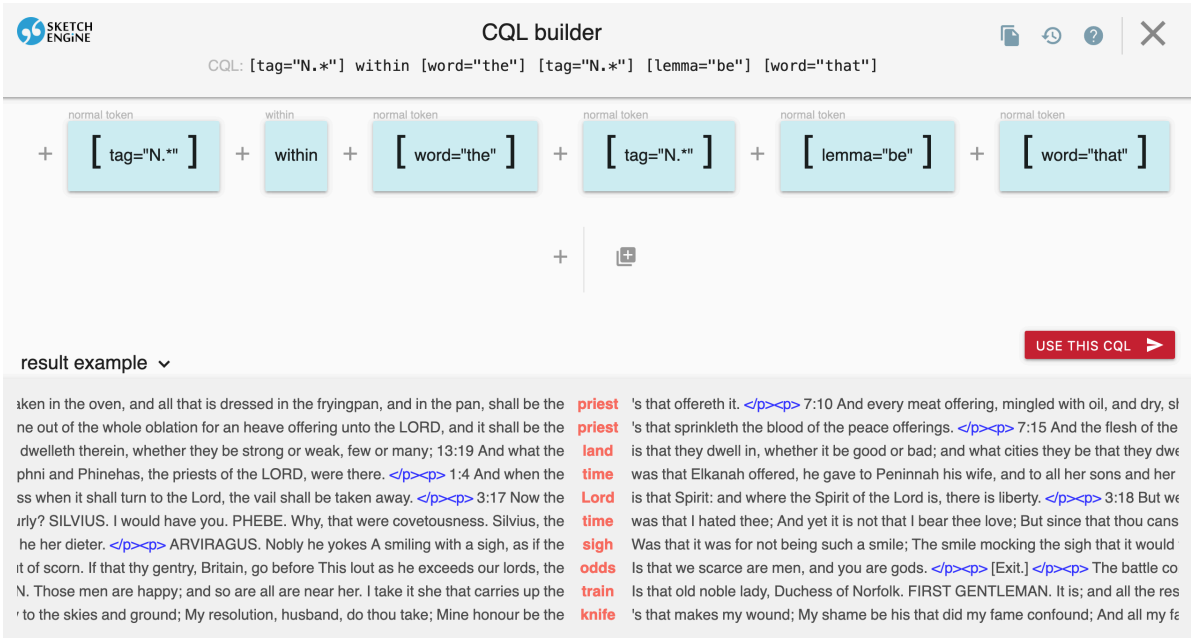
5.2 Studying syntactic constructions: *the N BE that*

Select pre-loaded corpus: Gutenberg English 2020

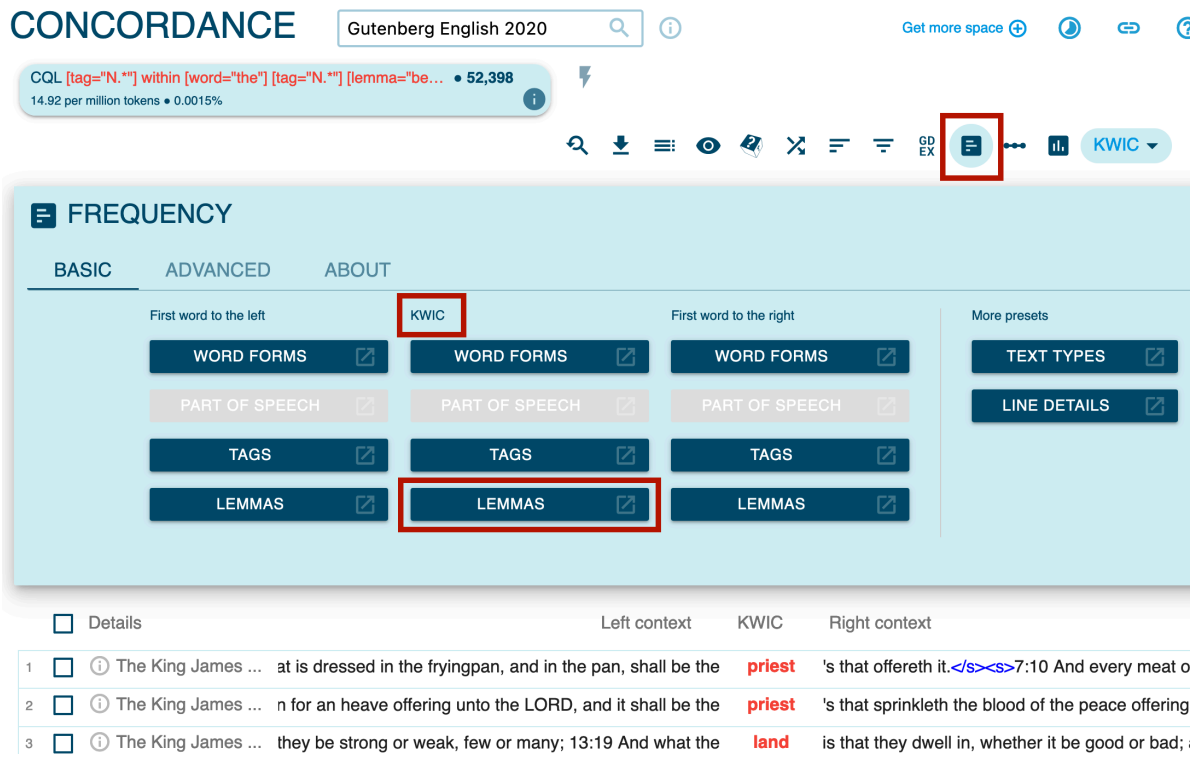


Query inspired by: Schmid, Hans-Jörg, and Annette Mantlik. 2015. 'Entrenchment in Historical Corpora? Reconstructing Dead Authors' Minds from Their Usage Profiles'. *Anglia* 133 (4): 583—623.

Search for target construction



Get frequency distribution of nouns in target construction:



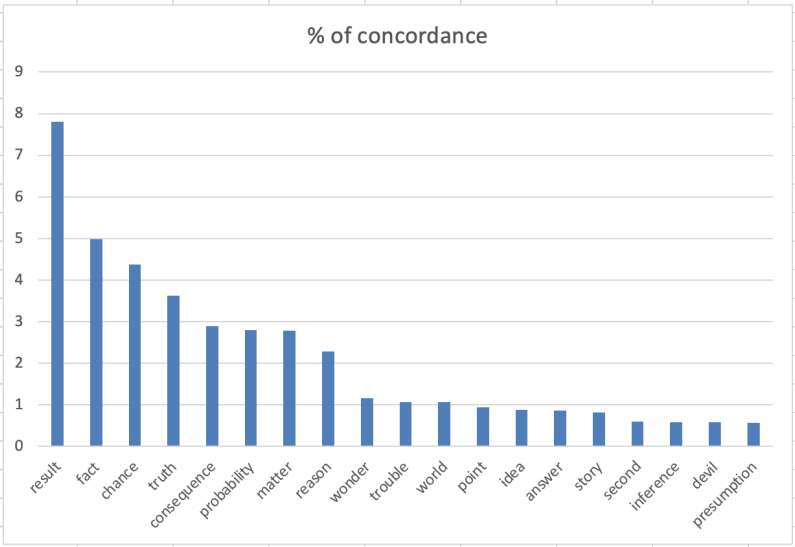
Distribution across all authors in SkE:

(5,917 items, 52,398 total frequency)

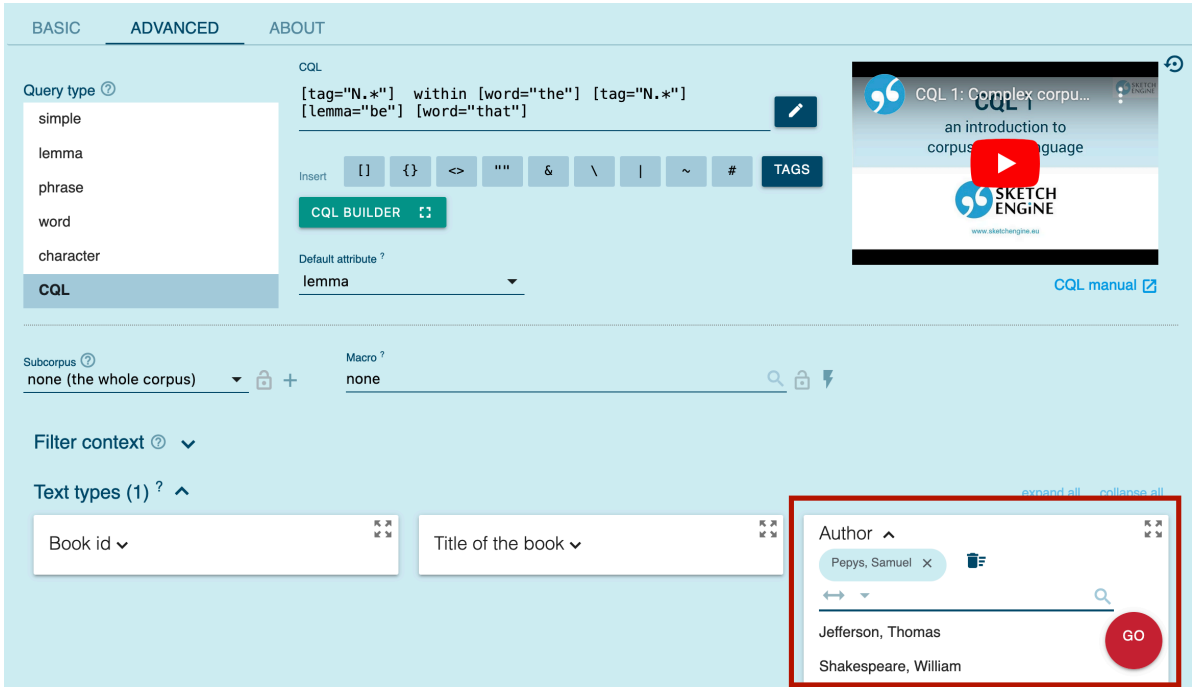
	Lemma	Frequency	Relative [?]	% of conc. [?]
1	result	4,088	1.16	7.80 %
2	fact	2,609	0.74	4.98 %
3	chance	2,295	0.65	4.38 %
4	truth	1,901	0.54	3.63 %
5	consequence	1,516	0.43	2.89 %
6	probability	1,467	0.42	2.80 %
7	matter	1,455	0.41	2.78 %
8	reason	1,198	0.34	2.29 %
9	wonder	610	0.17	1.16 %
10	trouble	561	0.16	1.07 %
11	world	556	0.16	1.06 %
12	point	492	0.14	0.94 %
13	idea	460	0.13	0.88 %
14	answer	452	0.13	0.86 %
15	story	427	0.12	0.81 %

Plot in exported Excel file:

	A	E
1	lemma	% of concordance
2	result	7,80182
3	fact	4,9792
4	chance	4,37994
5	truth	3,628
6	consequence	2,89324
7	probability	2,79973
8	matter	2,77682
9	reason	2,28635
10	wonder	1,16417
11	trouble	1,07065
12	world	1,06111
13	point	0,93897
14	idea	0,8779
15	answer	0,86263
16	story	0,81492
17	second	0,59926
18	inference	0,5859
19	devil	0,57827
20	presumption	0,57445



Individual analysis on Samuel Pepys' works:



Results for Samuel Pepys:

(11 items, 35 total frequency)

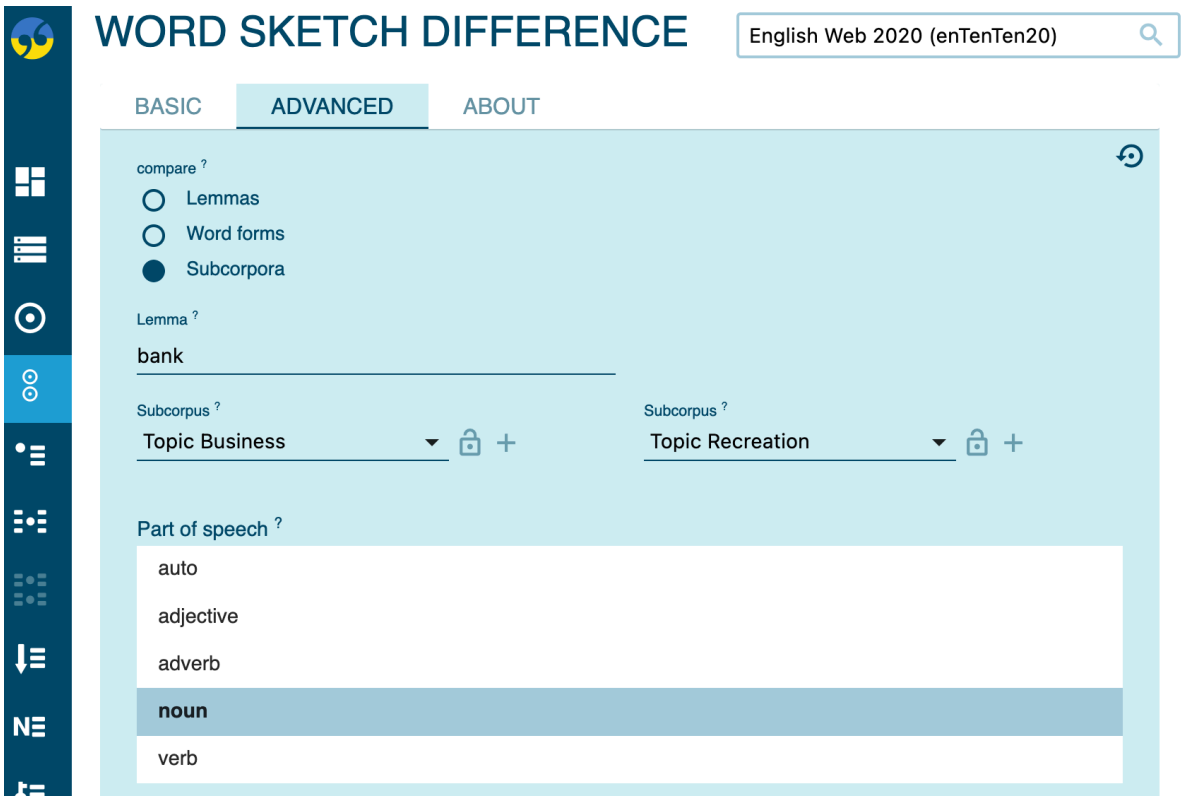
	Word	Frequency	Relative [?]	% of conc. [?]
1	talk	5	< 0.01	14.29 %
2	talke	3	< 0.01	8.57 %
3	news	3	< 0.01	8.57 %
4	evil	3	< 0.01	8.57 %
5	works	3	< 0.01	8.57 %
6	newes	3	< 0.01	8.57 %
7	matter	3	< 0.01	8.57 %
8	story	3	< 0.01	8.57 %
9	noise	3	< 0.01	8.57 %
10	business	3	< 0.01	8.57 %
11	hosier	3	< 0.01	8.57 %

Rows per page: 20 1–11 of 11 1 / 1

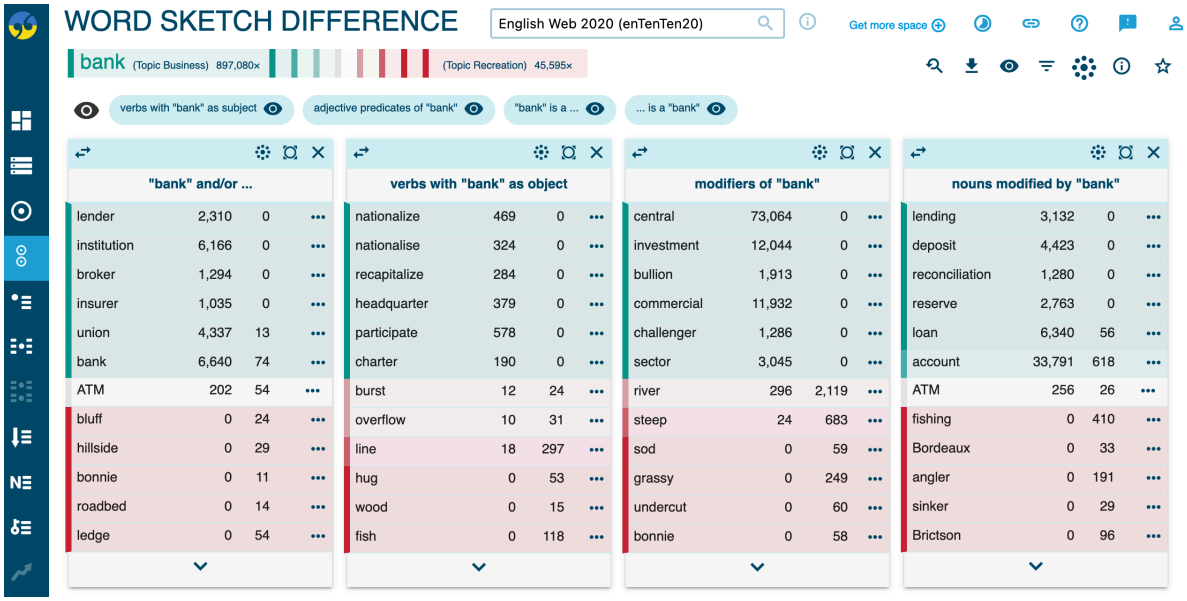
5.3 Comparing collocational profiles

corpus: enTenTen20

method: for the lemma *bank*^{n}, get word sketch differences between texts with `recreation` and `business` as topics


WORD SKETCH DIFFERENCE
English Web 2020 (enTenTen20)
BASIC
ADVANCED
ABOUT
compare ?
Lemmas
Word forms
Subcorpora
Lemma ?
bank
Subcorpus ?
Topic Business
Subcorpus ?
Topic Recreation
Part of speech ?
auto
adjective
adverb
noun
verb

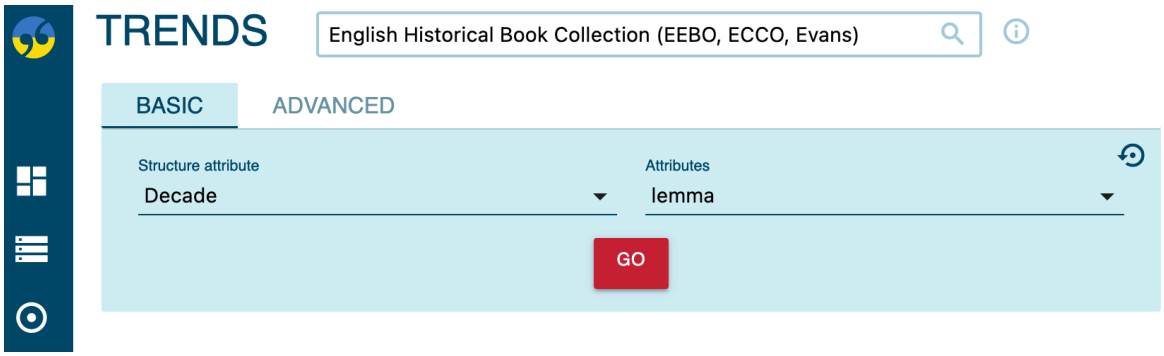
Results:


WORD SKETCH DIFFERENCE
English Web 2020 (enTenTen20)
Get more space
bank (Topic Business) 897,080×
(Topic Recreation) 45,595×
verbs with "bank" as subject
adjective predicates of "bank"
"bank" is a ...
... is a "bank"
"bank" and/or ...
lender 2,310 0
institution 6,166 0
broker 1,294 0
insurer 1,035 0
union 4,337 13
bank 6,640 74
ATM 202 54
bluff 0 24
hillside 0 29
bonnie 0 11
roadbed 0 14
ledge 0 54
verbs with "bank" as object
nationalize 469 0
nationalise 324 0
recapitalize 284 0
headquarter 379 0
participate 578 0
charter 190 0
burst 12 24
overflow 10 31
line 18 297
hug 0 53
wood 0 15
fish 0 118
modifiers of "bank"
central 73,064 0
investment 12,044 0
bullion 1,913 0
commercial 11,932 0
challenger 1,286 0
sector 3,045 0
river 296 2,119
steep 24 683
sod 0 59
grassy 0 249
undercut 0 60
bonnie 0 58
nouns modified by "bank"
lending 3,132 0
deposit 4,423 0
reconciliation 1,280 0
reserve 2,763 0
loan 6,340 56
account 33,791 618
ATM 256 26
fishing 0 410
Bordeaux 0 33
angler 0 191
sinker 0 29
Brictson 0 96

5.4 Investigating frequency over time: the rise of *whatever*

corpus: English Historical Book Collection (EEBO, ECCO, Evans)

1. Identify words that have significantly increased or decreased in frequency over time using the **trends** feature:



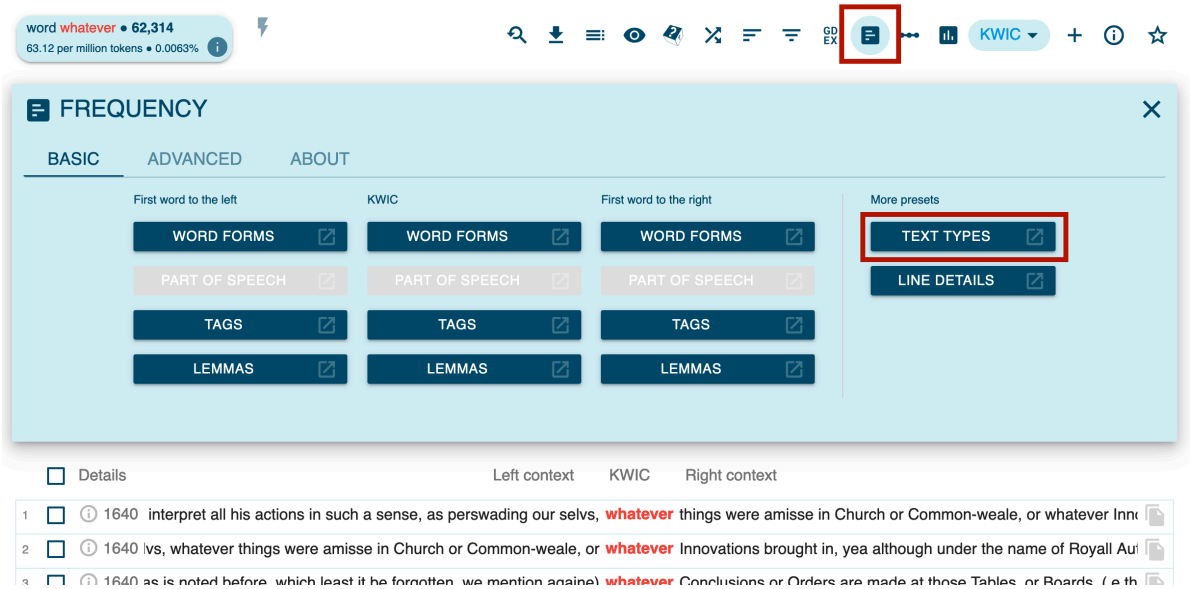
Results:

	Lemma	Trend ↓		Frequency	Sample	
1	whatever	↗	3.49	62,249		•••
2	frequently	↗	3.49	47,442		•••
3	reflect	↗	3.49	29,277		•••
4	enquiry	↗	3.49	16,295		•••
5	liable	↗	3.49	14,994		•••
6	powerful	↗	3.49	28,284		•••
7	enjoyment	↗	3.49	20,833		•••
8	palpitation	↗	3.49	685		•••
9	cautious	↗	3.49	5,316		•••
10	unravel	↗	3.49	1,036		•••
11	confinement	↗	3.49	4,471		•••
12	abhorrence	↗	3.49	3,021		•••
13	unluckily	↗	3.49	1,247		•••
14	endear	↗	3.27	4,600		•••
15	unite	↗	3.27	49,649		•••

	Lemma	Trend ↓		Frequency	Sample
1	knowe	↘	-3.73	41,199	
2	helpe	↘	-3.27	55,952	
3	euermore	↘	-3.27	8,195	
4	reuerence	↘	-3.27	14,545	
5	euery	↘	-3.27	150,240	
6	ende	↘	-3.27	39,025	
7	newe	↘	-3.27	18,652	
8	knowen	↘	-3.27	16,441	
9	olde	↘	-3.27	36,032	
10	holde	↘	-3.27	17,841	
11	deuoured	↘	-3.27	2,611	
12	drawe	↘	-3.27	7,818	
13	drawen	↘	-3.27	6,121	
14	thanke	↘	-3.27	9,406	
15	spende	↘	-3.27	1,030	

---

2. Investigating the frequency increase of *whatever*:



Results:

(23 items, 62,314 total frequency)

	Decade	Frequency	Relative in text type [?]	Relative density [?]
1	1590-1599	1	0.04	0.06 %
2	1600-1609	5	0.14	0.22 %
3	1610-1619	9	0.23	0.36 %
4	1620-1629	2	0.05	0.08 %
5	1630-1639	100	2.29	3.63 %
6	1640-1649	774	15.03	23.82 %
7	1650-1659	3,333	31.27	49.55 %
8	1660-1669	4,264	59.94	94.97 %
9	1670-1679	6,556	78.23	123.94 %
10	1680-1689	8,553	82.13	130.12 %
11	1690-1699	10,383	111.27	176.28 %
12	1700-1709	1,861	104.86	166.13 %
13	1710-1719	842	102.40	162.22 %
14	1720-1729	950	104.71	165.88 %
15	1730-1739	1,176	164.28	260.27 %

Plotting the exported version in Excel:

	A	B	C	D
1	Decade	Frequen	Relative density	Relative in text types
2	1590-1599	1	0,0625	0,03945
3	1600-1609	5	0,2243	0,14158
4	1610-1619	9	0,36081	0,22774
5	1620-1629	2	0,08132	0,05133
6	1630-1639	100	3,63352	2,29345
7	1640-1649	774	23,81597	15,03246
8	1650-1659	3333	49,5472	31,27383
9	1660-1669	4264	94,96611	59,94191
10	1670-1679	6556	123,94271	78,23172
11	1680-1689	8553	130,12427	82,13348
12	1690-1699	10383	176,27798	111,26536
13	1700-1709	1861	166,13154	104,861
14	1710-1719	842	162,22485	102,39512
15	1720-1729	950	165,8846	104,70513
16	1730-1739	1176	260,26644	164,27825
17	1740-1749	1789	210,20754	132,68144
18	1750-1759	1823	236,84605	149,49548
19	1760-1769	2446	237,86877	150,14101
20	1770-1779	3705	258,2669	163,01615
21	1780-1789	3738	225,38694	142,26257
22	1790-1799	8890	247,12564	155,98387
23	1800-1809	1105	240,25812	151,64915
24	1810-1819	9	292,41463	184,56995