

Sketch Engine tutorial

English Linguistics Colloquium, LMU Munich

Quirin Würschinger

January 19, 2023

Table of contents

1 Tutorial materials	2
2 General information	3
2.1 What is Sketch Engine?	3
2.2 LMU license	3
2.3 Related tools	4
2.4 Main features	4
2.5 Resources	4
3 Compiling corpora	4
3.1 Data format	4
3.2 Raw data	5
3.3 Uploading files	6
3.4 Adding and editing metadata	6
3.5 Processed data	6
4 Analysing data	9
4.1 Dashboard	9
4.2 Available corpora	9
4.3 Subcorpora	9
4.4 Queries	10
4.5 Concordance view	16
4.6 Collocations	17
4.7 Word sketches	18
4.8 Annotating data	20
4.9 Exporting data	21

5 Use cases	21
5.1 Compiling a corpus: dead authors' minds	21
5.2 Studying syntactic constructions: <i>the N BE that</i>	22
5.3 Comparing collocational profiles	27
5.4 Investigating frequency over time: the rise of <i>whatever</i>	29

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?

Sketch Engine is the ultimate tool to explore how language works. Its algorithms analyze authentic texts of billions of words (text corpora) to identify instantly what is typical in language and what is rare, unusual or emerging usage. It is also designed for text analysis or text mining applications.

Sketch Engine is used by linguists, lexicographers, translators, students and teachers. It is a first choice solution for publishers, universities, translation agencies and national language institutes throughout the world.

Sketch Engine contains 600 ready-to-use corpora in 90+ languages, each having a size of up to 60 billion words to provide a truly representative sample of language.

<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, drizzling 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 lost¹.

3.3 Uploading files

The screenshot shows the 'CORPUS: qw-gutenberg' interface. On the left is a vertical toolbar with icons for search, corpus management, and analysis. The main area has a header 'MAK BIGGER' and a search bar 'qw-gutenberg'. Below the search bar are two buttons: 'Find texts on the web' (grey background) and 'I have my own texts' (white background with a red border). The 'I have my own texts' button is highlighted with a red box. The 'CORPUS CONTENT' section shows a list of files under a folder named 'upload'. A red box highlights the 'Files filter' dropdown and the list of files. The table below shows:

Folder	Words	Type	Words
upload	~287,834	Document	txt
		1851_melville_moby-dick.txt	~217,119 ...
		1914_joyce_dubliners.txt	txt ~70,715 ...

A red arrow points from the bottom right of the highlighted area towards a large red circular button with a white arrow.

3.4 Adding and editing metadata

3.5 Processed data

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

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

²More precisely: one *token* per line, including punctuation: e.g. it, 's, ,.

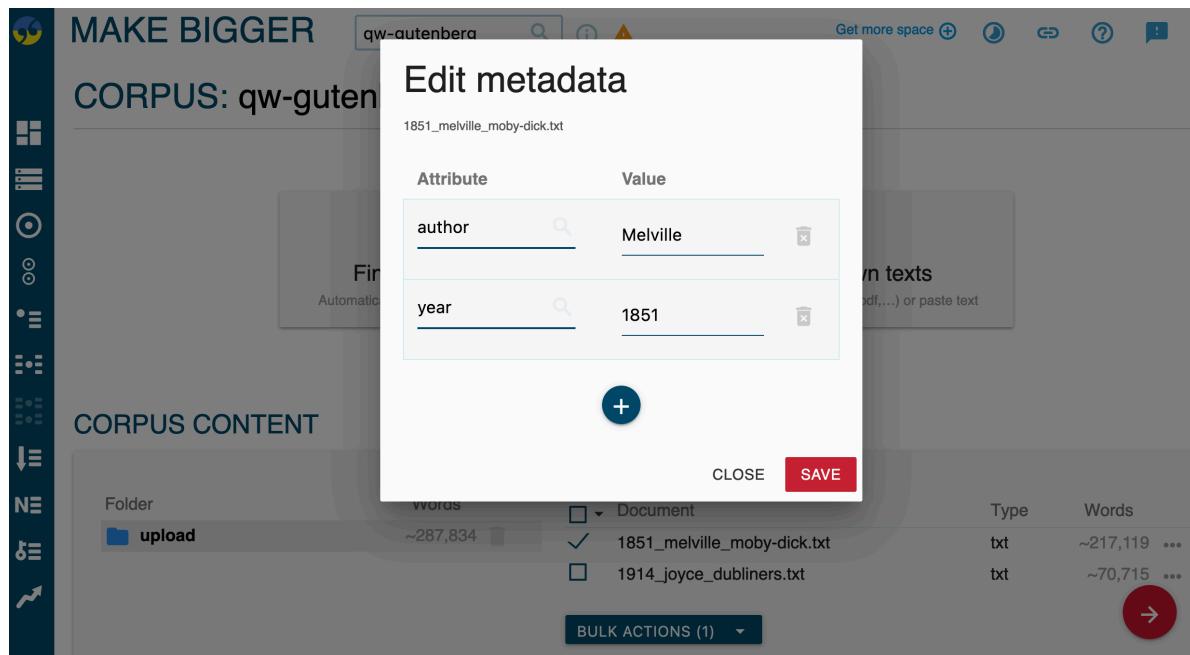


Figure 1: Adding and editing metadata

```

1 <doc id="file27506233"
      filename="1851_melville_moby-di
      ck.txt" parent_folder="upload"
      author="Melville" year="1851">
2 <s>           metadata
3 The DT the-x
4 Project NP Project-n
5 Gutenberg   NP Gutenberg-n
6 eBook   NN eBook-n
7 of IN of-i
8 Moby-Dick   NP Moby-Dick-n
9 <g/>
10 ; : ;-x
11 or CC or-c
12 The DT the-x
13 Whale   NN whale-n
14 <g/>
15 , , '-x
16 by IN by-i
17 Herman   NP Herman-n
18 Melville   NP Melville-n

```

```
8240 Call      NP   Call-n
8241 me        FW   me-x
8242 Ishmael   NP   Ishmael-n
8243 <g/>
8244 .         SENT  .-x
8245 </s>
8246 <s> token POS tag lemma-
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

The screenshot shows the DASHiBOARD interface. On the left is a vertical toolbar with various icons. The main area is divided into two sections: 'DASHBOARD' and 'RECENTLY USED CORPORA'.

DASHBOARD: This section is titled 'ENGLISH WEB 2020 (ENTENTEN20)'. It features a 'corpus selection' bar at the top with a search icon and an information icon. Below are nine analysis tools:

- 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)
- A OneClick Dictionary (Automatic dictionary drafting)
- Bilingual terms (Bilingual terminology extraction)

RECENTLY USED CORPORA: This section lists three corpora:

Corpus	Language	Size	Action
English Web 2020 (enTenTen20)	English	36,561,273,153	trash
qw-gutenberg	English	358,466	trash
British National Corpus (BNC)	English	96,134,547	trash

At the bottom right of the dashboard is an advertisement for 'Lexicom'.

Lexicom
An intensive workshop in digital
lexicography and lexical computing
Cambridge, UK, 11–15 September 2023
REGISTRATION

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

Create subcorpus

Subcorpus name *

required

Subcorpus from text types
 Subcorpus from concordance

expand all collapse all

Top-level domain (e.g. com) ▾	Website (e.g. cnn.com) ▾	Web domain (e.g. news.blogs.cnn...) ▾
Heading ▾	Title ▾	Source ▾
Wikipedia categories ▾	Topic ▾	Genre ▾
	health home news recreation reference	blog discussion legal news

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 <i>m</i> -
*ing	words or lemmas ending <i>-ing</i>
have a * idea	unspecified word or lemma
have *** idea	three unspecified words or lemmas
???t	4-letter words or lemmas ending <i>-t</i>
multi--billion	find <i>multi-billion</i> or <i>multibillion</i> or <i>multi billion</i>
return go back	find <i>return</i> or <i>go back</i>

see also [lemma](#)

Advanced (CQL) queries

CONCORDANCE English Web 2020 (enTenTen20) Get more space + 🔍 ⌂ ⓘ

BASIC ADVANCED ABOUT

Query type ?

- simple
- lemma
- phrase
- word
- character
- CQL**

CQL [word="the"] [tag="N.*"] [lemma="be"] [word="that"]

Insert [] { } <> " " & \ | ~ # TAGS

CQL BUILDER ☰

Default attribute ? lemma

Subcorpus ? none (the whole corpus) 🔒 + Macro ? none

Filter context ? ▾

Text types ? ▾

Top-level domain (e.g. com) ▾ Website (e.g. cnn.com) ▾ Web domain (e.g. news.blogs.) ▾ GO

expand all collapse all

CQL manual ↗

Helpful: [manual](#) and [CQL builder](#).

CQL builder

CQL: [word="the"] [tag="N.*"] [lemma="be"] [word="that"]

normal token + [word="the"] + [tag = N.*] + [lemma="be"] + [word="that"]

Attribute ? Value

J.* (adjective)

RB.? (adverb)

CC (conjunction)

DT (determiner)

N.* (noun)

CD (numeral)

RP (particle)

IN (preposition)

PP.? (pronoun)

V.* (verb)

the UIA
the hell
the A846
the fact
the rule
the govern
the govern
the dis
the shudo
the lesson is that

generates revenues in excess of the bond costs, a net positive. So, re? And so it is with this battered copy of Tina Turner's se best when you pair it with something nice and easy to drive and w focus is on the ESA listed anadromous fish and the effects of t board must hire a candidate recommended by the superintende principle it asserted – "to the maximum extent he finds to be f Court's current justices have in recent years expressed constitutovkin doesn't want the third fight held in Las Vegas and especially oratic legislature and the Republican governor have wildly diverg the can never be just Sion, talking to his mates, on Twitter or any other part

result example ▾

USE THIS CQL ➤

Extracting parts of your query matches using **within**:

The screenshot shows the Sketch Engine CQL builder interface. At the top, it displays the CQL query: [tag="N.*"] within [word="the"] [tag="N.*"] [lemma="be"] [word="that"]. Below this, the query is broken down into tokens: a normal token [tag="N.*"], followed by a 'within' operator, then another normal token [word="the"], followed by another normal token [tag="N.*"], then a normal token [lemma="be"], and finally a normal token [word="that"]. There is a plus sign (+) and a '+' button between the first two tokens, and another plus sign (+) and a '+' button after the fourth token. To the right of the tokens is a red 'USE THIS CQL' button with a white arrow. Below the tokens, there is a result example section containing a large block of text from a Shakespearean play, with several words highlighted in red (priest, land, time, Lord, odds, train, knife) and some words in blue (sigh, odds). The text discusses offerings to the Lord and the flesh of the lamb.

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">.



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 X



↔ ▾



arts

business

4.5 Concordance view

CONCORDANCE English Web 2020 (enTenTen20) ⚡

Lemma Anglo-Saxon • 20,956
0.49 per million tokens • 0.000049%

change query ⚡ download results Left context KWIC sort Right context frequency

annotate shuffle filter Get more space collocations KWIC

Details

	Source	Text Excerpt
1	netlibrary.net	istoric Kingdom of Strathclyde.[</s><s>45] Most of the region settled by the Anglo-Saxons became unified as the Kingdom of England in the 10th century.[</s><s>46]
2	netlibrary.net	ethnic groups that settled there before the 11th century: the Celts, Romans, Anglo-Saxons , Norse and the Normans [</s><s>Welsh people could be the oldest ethnic gr
3	atangledweb.org	urrency, which allows them to export so competitively for instance), but the " anglo-saxon " model is based on free trade, unlike the social-market model, which predor
4	mixedracestudie...	ie market.</s><s>...The least proximate types of the human species are the Anglo-Saxon and the negro.</s><s>Their original homes are in widely separated regions </s>
5	mixedracestudie...	of his childlike vivacity."</s><s>they are mated.</s><s>The offspring of the Anglo-Saxon and the negro are fertile when bred inter se, but in a less degree than when </s>
6	mixedracestudie...	ions respecting the mulattoes, the offspring of the strictly white race-i.e., the Anglo-Saxon or Teuton-and the true negro" :</s><s>"That mulattoes are the shortest-live
7	bell-foundation...	the other resources in the series, which compare Mayan life with that of the Anglo-Saxons .</s><s>This resource was originally developed by L Webster and has been
8	doktorfrank.com	now, by the end, you had read and understood (and enjoyed) Beowulf in the Anglo-Saxon .</s><s>It's weird.</s><s>I don't know how it happened.</s><s>(One exam
9	colonialfilm.or...	grounding the dangers to the white man and the red-haired (therefore white Anglo-Saxon) but also red-hot woman. 1950 opened two decades of colonial confrontatio
10	s-gabriel.org	ndanavia and Anglo-Saxon Britain, a Harvard list of links.</s><s>Again, the Anglo-Saxons are at the end.</s><s>Questions?</s><s>E-mail Jodi McMaster.</s><s>(kn
11	icalteff.com	ialized use).</s><s>Old English was written in the runic alphabet known as Anglo-Saxon (or Anglo-Frisian).</s><s>This had between 26-33 letters and was used fro
12	druglibrary.org	><s>We began checking with our compatriots, she with Russians and I with Anglo-Saxons .</s><s>We quickly found that our individual attitudes characterized our resp
13	richarddenning...	ledge – such as Beowulf's funeral.</s><s>Above: a game of Hnefatafl The Anglo-Saxons were fond of dice games.</s><s>Dice were made from the knuckle bones of
14	richarddenning...	re the two sides were of different sizes and abilities – were very prevalent in Anglo-Saxon and later Viking cultures.</s><s>The Romans seemed to have brought Nine
15	richarddenning...	s in the oldest of old high German the name óstará This Ostará, like the [Anglo-Saxon] Éastré, must in heathen religion have denoted a higher being, whose wor
16	theosophy.wiki	ed period of time between incarnations.</s><s>Hell.</s><s>A term with the Anglo-Saxons , evidently derived from the name of the goddess Hela (q.v.), and by the Sc

4.6 Collocations

The screenshot shows the CONCORDANCE software interface. At the top, there's a search bar with 'English Web 2020 (enTenTen20)' and a magnifying glass icon. Below it, a message says 'CQL [lemma="commit" & tag="V.*"] • 3,037,094' and '70.43 per million tokens • 0.007%'. A tooltip for 'Sample 10000 • 10,000' shows '0.23 per million tokens • 0.000023%' with an info icon. On the right, there are buttons for 'Get more space +' and a circular refresh icon. Below the search bar is a toolbar with various icons. The main area is titled 'Collocations' with 'CHANGE CRITERIA' and 'BACK TO CONCORDANCE' buttons. To the left is a vertical toolbar with icons for different functions. The main content is a table with the following data:

	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 English Web 2020 (enTenTen20)   Get more space      

commit as verb 3,037,094x 

modifiers of "commit"	objects of "commit"	subjects of "commit"	"commit" and/or ...	prepositional phrases
allegedly allegedly committed	crime crimes committed	git git commit • especially: technology	commit commit , commit , commit , commit • especially: technology • usually: technology	"commit" to ... "commit" by ... "commit" in ... "commit" against ... "commit" on ... "commit" for ... "commit" during ... "commit" with ... "commit" at ... "commit" as ... "commit" under ... "commit" within ...
deeply deeply committed to	suicide commit suicide	person person commits	defendant defendant committed • especially: legal	
firmly is firmly committed to	act acts committed	offender offenders who commit	attempt committing or attempting	
fully fully committed to	murder commit murder	sin sins committed • especially: society	focus focused and committed to	
passionately passionately committed to	offence offences committed	crime crimes committed	involve involved and committed to	
verbally verbally committed to • especially: sports	atrocities committed	criminal criminals who commit	abort commit or abort • especially: technology	
strongly is strongly committed to	fraud commit fraud	Government Government has committed	motivate motivated and committed to	
involuntarily involuntarily committed to		individual individuals who commit		

Word sketch difference: between two words/phrases

WORD SKETCH DIFFERENCE English Web 2020 (enTenTen20)   Get more space      

deep 4,515,593x  profound 532,425x 

"deep/profound" and/or ...	subjects of "be deep/profound"	modifiers of "deep/profound"
blue 21,555 dark 27,374 wide 17,132 rich 19,136 broad 13,537 deep 48,784 spiritual 5,511 philosophical 2,326 lasting 3,780 subtle 959 severe 366 far-reaching 283	foot 33,614 inch 19,524 meter 5,999 knee 5,085 metre 4,914 m 4,910 influence 135 silence 157 insight 93 impact 84 consequence 13 implication 46	little 19,476 somewhere 4,656 much 33,220 moderately 1,000 too 30,706 surprisingly 1,505 extraordinarily 256 equally 573 spiritually 110 lyrically 34 theologically 59 philosophically 66

Word sketch difference: between two subcorpora

The screenshot shows the 'WORD SKETCH DIFFERENCE' interface with the following details:

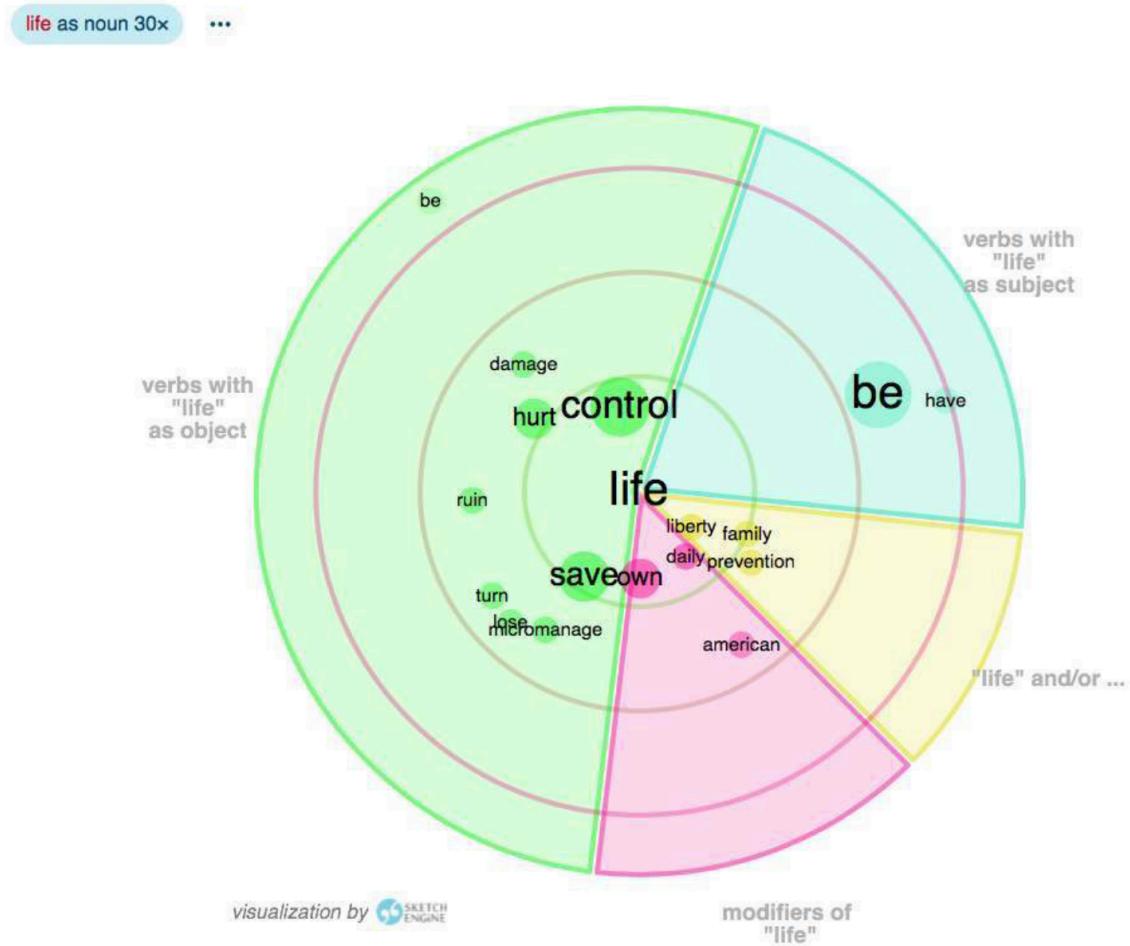
- Top Bar:** English Web 2020 (enTenTen20) with a search bar and various icons.
- Topic Selection:** 'bank' (Topic Business) 897,080x and '(Topic Recreation) 45,595x'.
- Panel 1: modifiers of "bank"**

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	...
- Panel 2: 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	...
- Panel 3: ... to "bank"**

lend	349	0	...
loan	587	0	...
lending	149	0	...
liquidity	118	0	...
deposit	132	0	...
bank	567	43	...
bridge	0	27	...
river	0	28	...
creek	0	10	...
cast	0	28	...
cross	0	87	...
tight	0	28	...

Visualizations



4.8 Annotating data

for metadata: see Figure 1 above

for concordance lines:

Label	Filter	Frequency
ethno-racial		0
historic		4
political		0

MANAGE ANNOTATIONS

SORT BY LABELS

EXIT ANNOTATION MODE

Details Left context KWIC Right context

- 1 researchgate.ne... English Literature was actually started with **Anglo-Saxon**
- 2 researchgate.ne... day part of England and Wales. The **Anglo-Saxon** **historic** works include genres such as epic poetry, h
- 3 researchgate.ne... d national epic status in Britain. The **Anglo-Saxon** **historic** Chronicle is a collection of early Engl
- 4 researchgate.ne... er of manuscripts remain from the 600 year **Anglo-Saxon** period, with most written during the l
- 5 researchgate.ne... works of the early Church Fathers; **Anglo-Saxon** **historic** chronicles and narrative history works; laws,

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 LMUers.

⁴When 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`).

CORPUS: qw-gutenberg (English)

NOW SHARED WITH

University of Munich (Ludwig-Maximilians-Universität) (MunichLMU_ELEXIS) Read only X

SHARE WITH

Users
User email address @ +

or group accounts or institutions
 🔍

University of Munich (Ludwig-Maximilians-Universität)

PRIVILEGES

Read only
Allows searching corpus.

Upload
Allows searching the corpus and adding data to the corpus.

Full access
Allows searching the corpus, adding data to the corpus, changing the corpus configuration, changing sketch grammar and recompiling the corpus.

CANCEL SHARE

5.2 Studying syntactic constructions: *the N BE that*

Select pre-loaded corpus: [Gutenberg English 2020](#)

Gutenberg English 2020 preloaded/gutenberg20_en X

[MANAGE CORPUS](#) [MANAGE SUBCORPORA](#) [COMPARE CORPORA](#) [TEXT TYPE ANALYSIS](#)

GENERAL INFO

Language: English

[CORPUS DESCRIPTION & BIBLIOGRAPHY](#)

[TAGSET](#)

[WORD SKETCH GRAMMAR](#)

[TERM GRAMMAR](#)

TEXT TYPES ⓘ

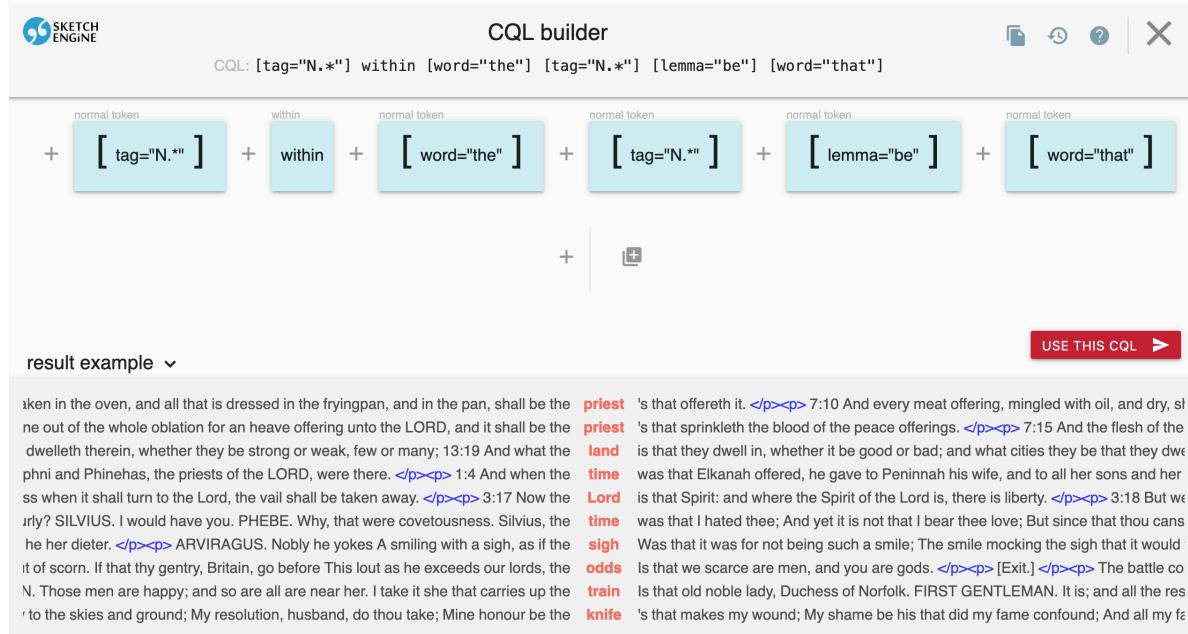
TEXT TYPE ANALYSIS	
<book> (8)	45,469 ▾
Author , book.author	15,173 ▾
Book id , book.id	45,469 ▾
Language , book.language	1 ▾
Title of the book , book.title	44,800 ▾
Topics of the book , book.topics	29,183 ▾
wordcount , book.wordcount	38,460 ▾
Year of author's birth , book.author_birth	487 ▾
Year of author's death , book.author_death	505 ▾
<g> (0)	583,084,741 ▾
<s> (0)	157,166,330 ▾
<p> (0)	57,002,871 ▾
<ll> (0)	474,413 ▾
<text> (0)	45,457 ▾

COUNTS ⓘ

Tokens	3,511,134,426
Words	2,903,177,585
Sentences	157,166,330
Paragraphs	57,002,871
Documents	45,469

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



The screenshot shows the Sketch Engine CQL builder interface. The top bar displays the title "CQL builder" and the query "CQL: [tag="N.*"] within [word="the"] [tag="N.*"] [lemma="be"] [word="that"]". Below the query, the search results are displayed in a grid format. The first result example is shown with the following text and highlighted tokens:

aken in the oven, and all that is dressed in the fryingpan, and in the pan, shall be the
ne out of the whole oblation for an heave offering unto the LORD, and it shall be the
dwelleth therein, whether they be strong or weak, few or many; 13:19 And what the
phni and Phinehas, the priests of the LORD, were there. </p><p> 1:4 And when the
ss when it shall turn to the Lord, the vail shall be taken away. </p><p> 3:17 Now the
irly? SILVIUS. I would have you. PHEBE. Why, that were covetousness. Silvius, the
he her dieter. </p><p> ARVIRAGUS. Nobly he yokes A smiling with a sigh, as if the
it of scorn. If that thy gentry, Britain, go before This lout as he exceeds our lords, the
N. Those men are happy; and so are all are near her. I take it she that carries up the
'to the skies and ground; My resolution, husband, do thou take; Mine honour be the
priest 's that offereth it. </p><p> 7:10 And every meat offering, mingled with oil, and dry, si
priest 's that sprinkleth the blood of the peace offerings. </p><p> 7:15 And the flesh of the
land is that they dwell in, whether it be good or bad; and what cities they be that they dw
time was that Elkanah offered, he gave to Peninnah his wife, and to all her sons and her
Lord is that Spirit; and where the Spirit of the Lord is, there is liberty. </p><p> 3:18 But we
time was that I hated thee; And yet it is not that I bear thee love; But since that thou cans
sigh Was that it was for not being such a smile; The smile mocking the sigh that it would
odds Is that we scarce are men, and you are gods. </p><p> [Exit.] </p><p> The battle co
train Is that old noble lady, Duchess of Norfolk. FIRST GENTLEMAN. It is; and all the res
knife 's that makes my wound; My shame be his that did my fame confound; And all my fa

Get frequency distribution of nouns in target construction:

CONCORDANCE

Gutenberg English 2020  

CQL [tag="N.*"] within [word="the"] [tag="N.*"] [lemma="be... • 52,398
14.92 per million tokens • 0.0015%

FREQUENCY

BASIC ADVANCED ABOUT

First word to the left  First word to the right More presets

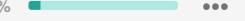
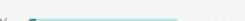
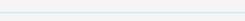
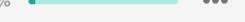
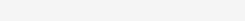
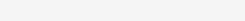
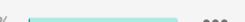
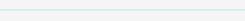
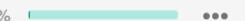
WORD FORMS  WORD FORMS  WORD FORMS 
PART OF SPEECH  PART OF SPEECH  PART OF SPEECH 
TAGS  TAGS  TAGS 
LEMMAS  LEMMAS  LEMMAS 
 

Details Left context KWIC Right context

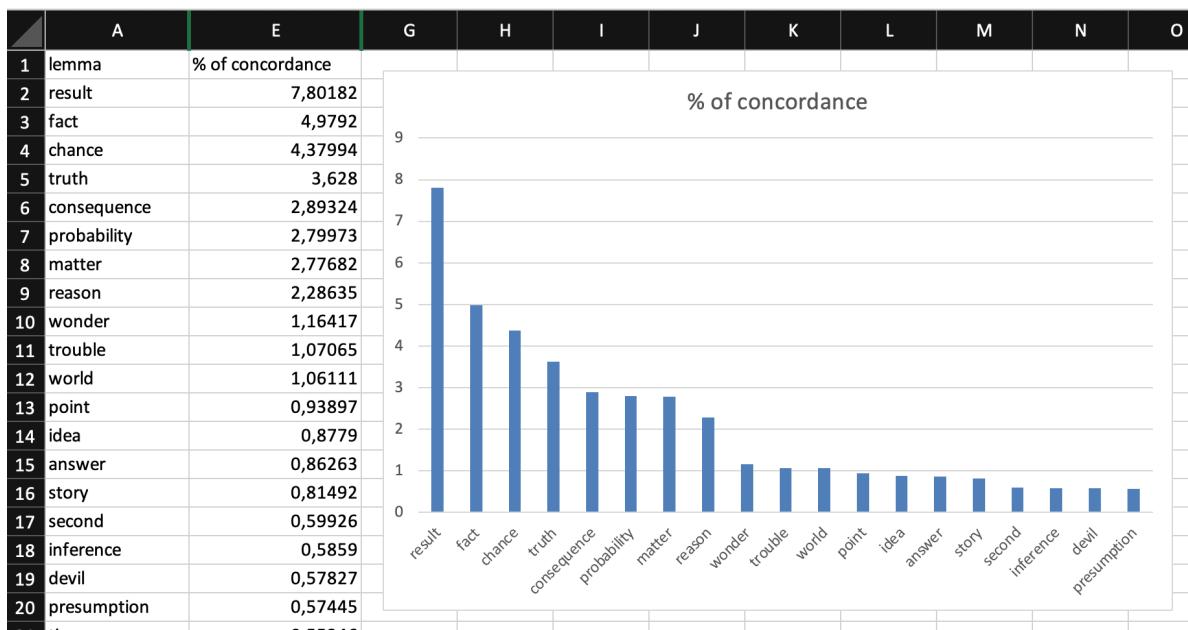
1  The King James ... at is dressed in the fryingpan, and in the pan, shall be the **priest** 's that offereth it.
2  The King James ... n for an heave offering unto the LORD, and it shall be the **priest** 's that sprinkleth the blood of the peace offering
3  The King James ... they be strong or weak, few or many; 13:19 And what the **land** is that they dwell in, whether it be good or bad;

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:



Individual analysis on Samuel Pepys' works:

BASIC ADVANCED ABOUT

CQL

Query type ⑦ simple

[tag="N.*"] within [word="the"] [tag="N.*"]
[lemma="be"] [word="that"]

Insert [] { } <> "" & \ | ~ # TAGS

CQL BUILDER ⑧

Default attribute ? lemma

Subcorpus ⑨ none (the whole corpus) 🔒 + Macro ? none

Filter context ⑩

Text types (1) ⑪

Book id ▾ Title of the book ▾

Author ▾

Pepys, Samuel ×

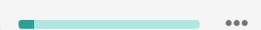
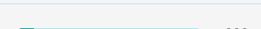
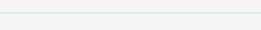
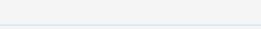
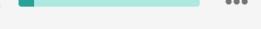
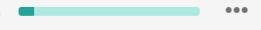
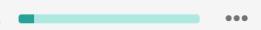
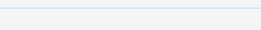
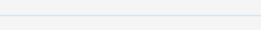
Jefferson, Thomas

Shakespeare, William

GO

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



Part of speech ?



auto



adjective



adverb



noun



verb



Subcorpus ?

Topic Recreation



Results:



WORD SKETCH DIFFERENCE

English Web 2020 (enTenTen20)



Get more space



bank

(Topic Business) 897,080x

(Topic Recreation) 45,595x



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	...

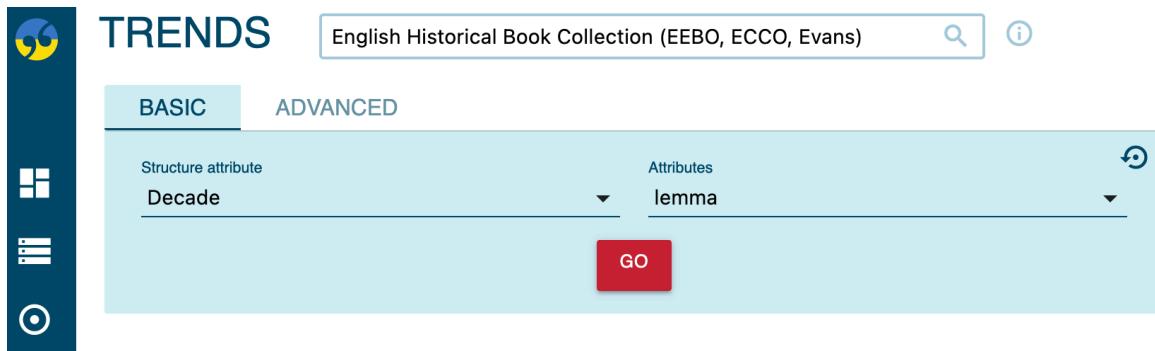
nouns modified by "bank"

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:

	Lemmas	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	knownen	↘	-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*:

The screenshot shows the KWIC interface with the 'FREQUENCY' tab selected. At the top, there's a search bar with 'word whatever • 62,314' and a note '63.12 per million tokens • 0.0063%'. Below the search bar are several buttons: a magnifying glass, a download icon, a list icon, an eye icon, a cross icon, a double arrow icon, a 'GEO' button, a 'KWIC' dropdown, a plus sign, an info icon, and a star icon. A red box highlights the 'KWIC' button.

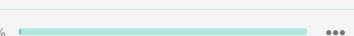
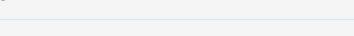
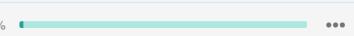
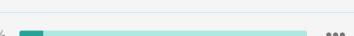
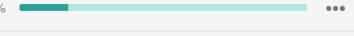
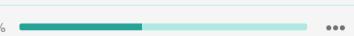
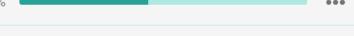
The main area has tabs for 'BASIC', 'ADVANCED', and 'ABOUT'. Under 'BASIC', there are sections for 'First word to the left' and 'First word to the right', each with four dropdown menus: 'WORD FORMS', 'PART OF SPEECH', 'TAGS', and 'LEMMAS'. To the right, under 'More presets', there are two more dropdown menus: 'TEXT TYPES' and 'LINE DETAILS', both with a red box around them. Below these are 'Left context' and 'Right context' buttons, also with a red box around them.

At the bottom, there's a list of three entries:

- 1 [checkbox] ① 1640 interpret all his actions in such a sense, as perswading our selvs, **whatever** things were amisse in Church or Common-weale, or whatever innr [file icon]
- 2 [checkbox] ① 1640 lvs, whatever things were amisse in Church or Common-weale, or **whatever** Innovations brought in, yea although under the name of Royall Au [file icon]
- 3 [checkbox] ① 1640 as is noted before, which least it be forgotten, we mention againe) **whatever** Conclusions or Orders are made at those Tables, or Boards / e th [file icon]

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:

