

# Sketch Engine tutorial

English Linguistics Colloquium, LMU Munich

Quirin Würschinger

January 19, 2023

## Table of contents

<b>1 General information</b>	<b>2</b>
1.1 What is Sketch Engine? . . . . .	2
1.2 LMU license . . . . .	3
1.3 Related tools . . . . .	3
1.4 Main features . . . . .	3
1.5 Resources . . . . .	3
<b>2 Compiling corpora</b>	<b>4</b>
2.1 Data format . . . . .	4
2.2 Raw data . . . . .	4
2.3 Uploading files . . . . .	5
2.4 Adding and editing metadata . . . . .	5
2.5 Processed data . . . . .	5
<b>3 Analysing data</b>	<b>8</b>
3.1 Dashboard . . . . .	8
3.2 Available corpora . . . . .	8
3.3 Subcorpora . . . . .	8
3.4 Queries . . . . .	9
3.5 Concordance view . . . . .	15
3.6 Collocations . . . . .	16
3.7 Word sketches . . . . .	17
3.8 Annotating data . . . . .	19
3.9 Exporting data . . . . .	20
<b>4 Use cases</b>	<b>20</b>
4.1 Compiling a corpus: dead authors' minds . . . . .	20

4.2	Studying syntactic constructions: <i>the N BE that</i>	21
4.3	Comparing collocational profiles	26
4.4	Investigating frequency over time: the rise of <i>whatever</i>	28

## 1 General information

### 1.1 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>

## 1.2 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

## 1.3 Related tools

- [AntConc](#)
- [LancsBox](#)
- [WordSmith](#)

## 1.4 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

## 1.5 Resources

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

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

## 2 Compiling corpora

### 2.1 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

### 2.2 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](http://www.gutenberg.org/cache/epub/1/pg1.html)

---

## 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<sup>1</sup>.

## 2.3 Uploading files

The screenshot shows the SkE interface for the 'qw-gutenberg' corpus. On the left is a sidebar with various icons. The main area has a search bar with 'qw-gutenberg'. Below it, a section titled 'CORPUS: qw-gutenberg (English)' contains two buttons: 'Find texts on the web' and 'I have my own texts', with the second one highlighted by a red box. The 'CORPUS CONTENT' section shows a table of files under an 'upload' folder. A red box highlights this table, which includes columns for 'Folder', 'Words', 'Type', and 'Words'. The table lists three files: '1851\_melville\_moby-dick.txt' and '1914\_joyce\_dubliners.txt' (both txt type, ~217,119 and ~70,715 words respectively), and a 'Document' entry for 'upload' (~287,834 words). A 'BULK ACTIONS' button is at the bottom right of the table.

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

## 2.4 Adding and editing metadata

## 2.5 Processed data

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

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

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

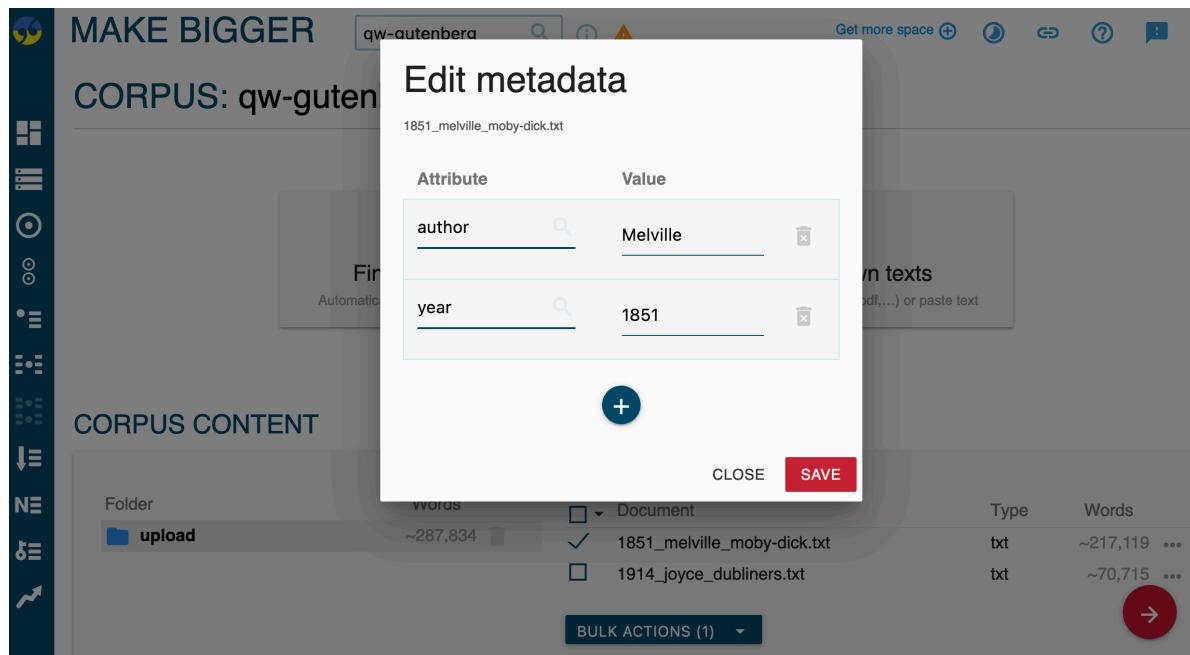


Figure 1: Adding and editing metadata

```

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
```

## 3 Analysing data

### 3.1 Dashboard

The screenshot shows the DASHiBOARD interface. On the left is a vertical toolbar with various icons. The main area is titled "DASHBOARD" and shows "ENGLISH WEB 2020 (ENTENTEN20)" selected. It features a "corpus selection" section with nine tools: Word Sketch, Word Sketch Difference, Thesaurus, Concordance, Parallel Concordance, Wordlist, N-grams, Trends, and OneClick Dictionary. To the right is a "RECENTLY USED CORPORA" section listing three corpora: English Web 2020 (enTenTen20), qw-gutenberg, and British National Corpus (BNC). At the bottom right is a promotional box for "Lexicom" with a registration button.

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

**RECENTLY USED CORPORA**

English Web 2020 (enTenTen20)    English    36,561,273,153   

qw-gutenberg    English    358,466   

British National Corpus (BNC)    English    96,134,547   

**Lexicom />**

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

**REGISTRATION**

### 3.2 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

### 3.3 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

### 3.4 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

result example ▾

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 the 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

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 to add more terms. On the right side, there is a red 'USE THIS CQL' button with a right-pointing arrow. Below the tokens, there is a result example section containing a large block of text from a document.

result example ▾

aken in the oven, and all that is dressed in the fryingpan, and in the pan, shall be the priest's that offereth it. </p><p> 7:10 And every meat offering, mingled with oil, and dry, si  
ne out of the whole oblation for an heave offering unto the LORD, and it shall be the priest's that sprinkleth the blood of the peace offerings. </p><p> 7:15 And the flesh of the dwelleth therein, whether 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; and what cities they be that they dw  
phni and Phinehas, the priests of the LORD, were there. </p><p> 1:4 And when the 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  
ss when it shall turn to the Lord, the vail shall be taken away. </p><p> 3:17 Now the time was that I hated thee; And yet it is not that I bear thee love; But since that thou cans  
irly? SILVIUS. I would have you. PHEBE. Why, that were covetousness. Silvius, the odds Was that it was for not being such a smile; The smile mocking the sigh that it would  
he her dieter. </p><p> ARVIRAGUS. Nobly he yokes A smiling with a sigh, as if the train Is that old noble lady, Duchess of Norfolk. FIRST GENTLEMAN. It is; and all the res  
N. Those men are happy; and so are all are near her. I take it she that carries up the

## Filtering by metadata

Options:

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

<sup>3</sup>‘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

### 3.5 Concordance view

The screenshot shows the CONCORDANCE interface with the following details:

- Search Query:** English Web 2020 (enTenTen20)
- Lemma:** Anglo-Saxon • 20,956
- Frequency:** 0.49 per million tokens • 0.000049%
- Tools:** annotate, shuffle, filter, Get more space, collocations, KWIC, GD EX, sort, Left context, Right context, display options, change query, download results.
- Results:** A list of 16 entries, each with a checkbox, a URL, and a snippet of text containing the lemma 'Anglo-Saxon'. The snippets illustrate various contexts of the word, such as its use in historical descriptions, its relation to other ethnic groups, and its presence in literature like Beowulf.

### 3.6 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" indicates "0.23 per million tokens • 0.000023%". On the right, there are buttons for "Get more space" and a circular icon. The main area is titled "Collocations" with "CHANGE CRITERIA" and "BACK TO CONCORDANCE" buttons. To the left is a vertical toolbar with various icons. The results table has columns: Word, Cooccurrences ?, Candidates ?, T-score, MI, and LogDice ↓. The data is as follows:

	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.

### 3.7 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

WORD SKETCH DIFFERENCE

English Web 2020 (enTenTen20)

bank (Topic Business) 897,080x | (Topic Recreation) 45,595x |

"bank" and/or ... | verbs with "bank" as object | verbs with "bank" as subject | "bank" is a ... | ... is a "bank" | adjective predicates of "bank" | bank's ...

possessors of "bank" | pronominal possessors of "bank" | "bank" to ...

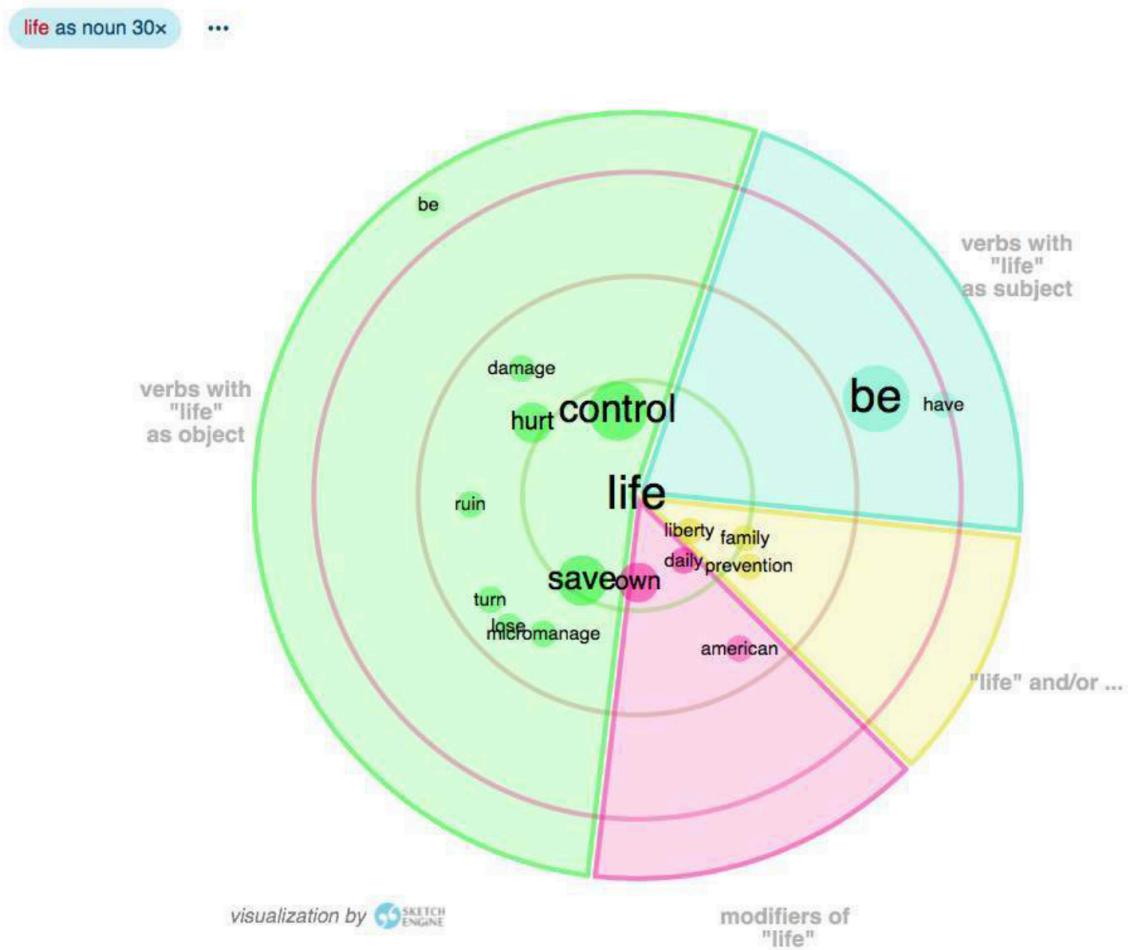
↔ **modifiers of "bank"** ↔ **nouns modified by "bank"** ↔ **... to "bank"**

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	...	

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	...	
Briston	0	96	...	

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



## 3.8 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** literature, so that they can be called as
- 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
- 4  researchgate.ne... er of manuscripts remain from the 600 year **Anglo-Saxon** period, with most
- 5  researchgate.ne... works of the early Church Fathers; **Anglo-Saxon** historic chronicles and narrative history works; laws,

### 3.9 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.<sup>4</sup>

## 4 Use cases

### 4.1 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.

---

<sup>4</sup>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).

## 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

## 4.2 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** i

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 ▾

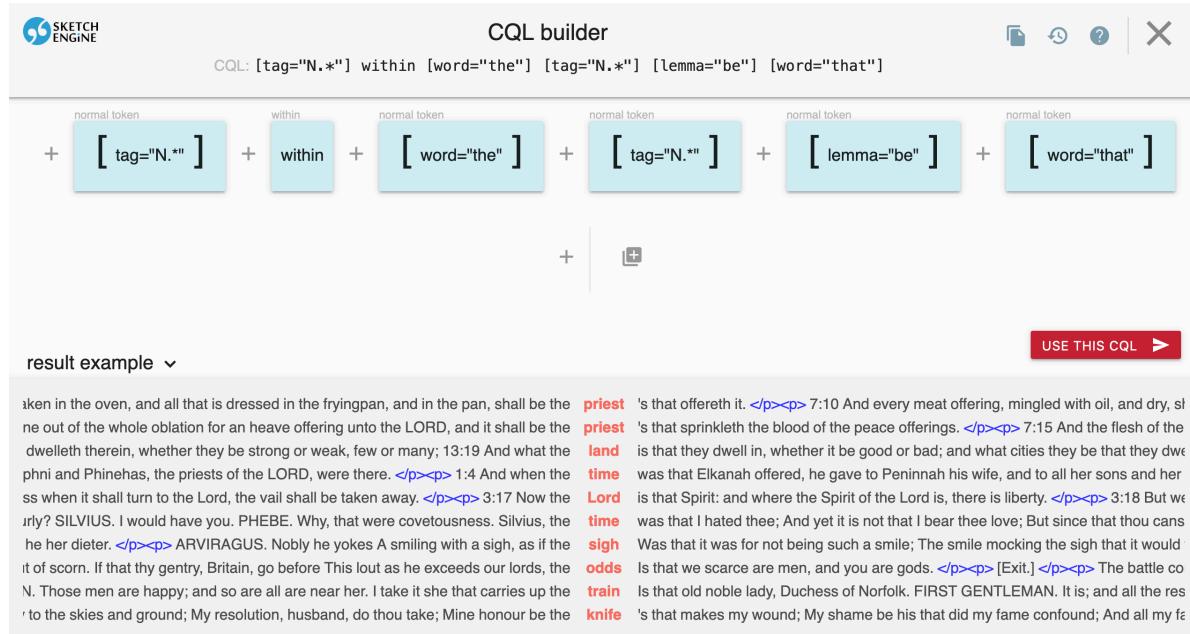
**COUNTS** i

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 query entered is:

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

The query is visualized as a sequence of tokens connected by operators:

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

Below the query, there is a "result example" section showing a snippet of text from a historical document. The text is as follows:

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 ury? 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

**USE THIS CQL ➤**

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 

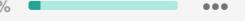
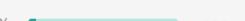
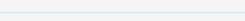
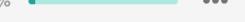
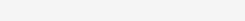
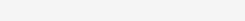
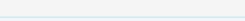
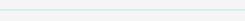
 TEXT TYPES   
 LINE DETAILS 

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;

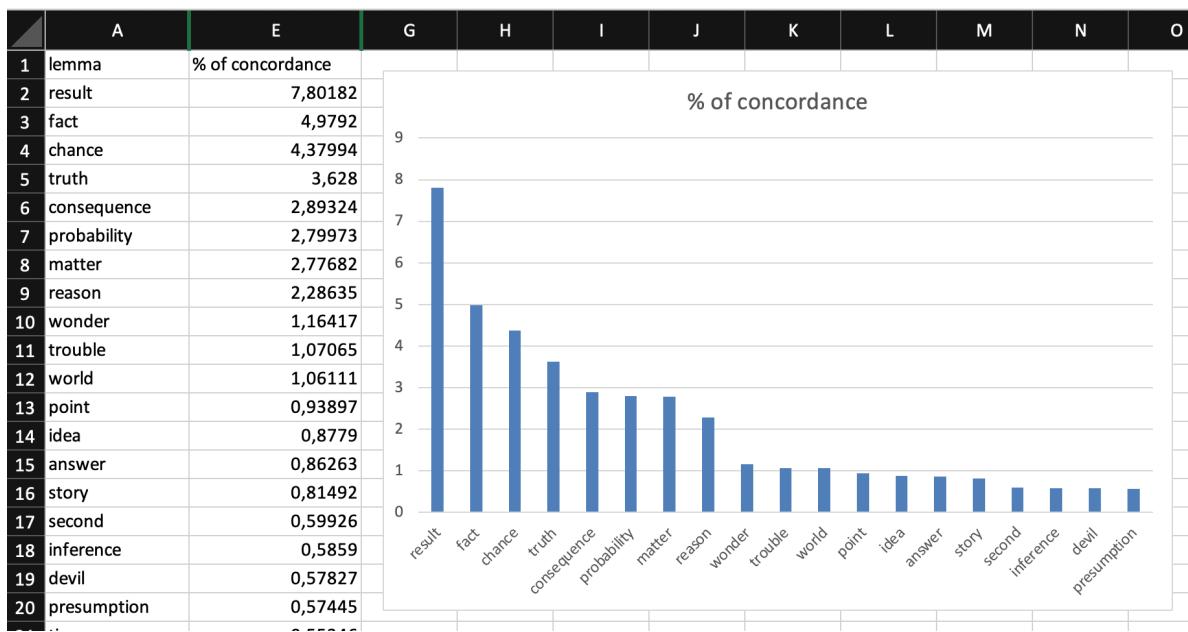
## Distribution across all authors in SkE:

(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 ② simple lemma phrase word character CQL

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

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

Default attribute ? lemma

Subcorpus ② none (the whole corpus) Macro ? none

Filter context ②

Text types (1) ②

Book id ▾ Title of the book ▾ Author ▾

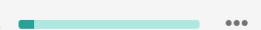
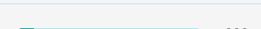
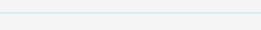
Pepys, Samuel × Jefferson, Thomas Shakespeare, William

expand all collapse all 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 >

### 4.3 Comparing collocational profiles

**corpus:** enTenTen20

**method:** for the lemma *bank*<sup>n</sup>, get word sketch differences between texts with **recreation** and **business** as topics

The screenshot shows the 'WORD SKETCH DIFFERENCE' interface. On the left is a vertical toolbar with various icons. The main area has tabs for 'BASIC', 'ADVANCED' (which is selected), and 'ABOUT'. A search bar at the top right says 'English Web 2020 (enTenTen20)' with a magnifying glass icon. Below the search bar are three selection boxes: 'compare ?' (radio buttons for Lemmas, Word forms, Subcorpora), 'Lemma ?' (text input 'bank'), 'Subcorpus ?' (dropdowns for 'Topic Business' and 'Topic Recreation'), and 'Part of speech ?' (radio buttons for auto, adjective, adverb, noun, verb, with 'noun' selected). The background is light blue.

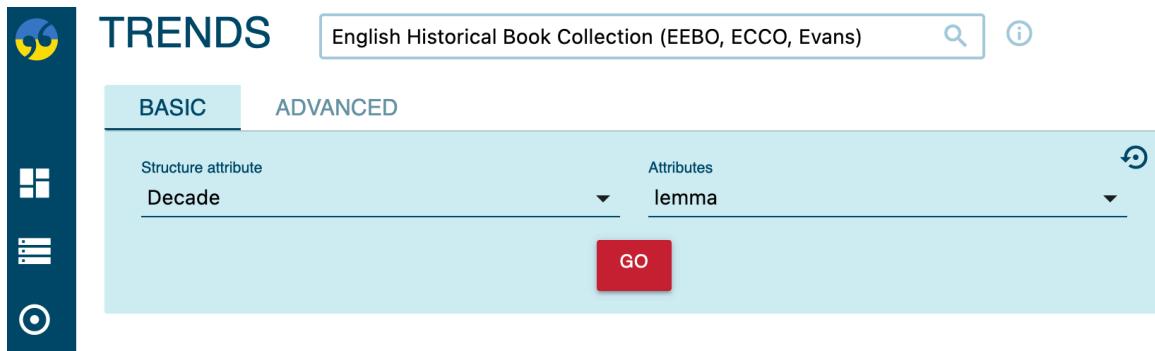
Results:

The results page for the word 'bank' shows four tables side-by-side. The first table, 'verbs with "bank" as subject', lists words like lender, institution, broker, insurer, union, bank, ATM, bluff, hillside, bonnie, roadbed, and ledge. The second table, 'verbs with "bank" as object', lists nationalize, nationalise, recapitalize, headquarter, participate, charter, burst, overflow, line, hug, wood, and fish. The third table, 'modifiers of "bank"', lists central, investment, bullion, commercial, challenger, sector, river, steep, sod, grassy, undercut, and bonnie. The fourth table, 'nouns modified by "bank"', lists lending, deposit, reconciliation, reserve, loan, account, ATM, fishing, Bordeaux, angler, sinker, and Brictson. Each table has columns for frequency (e.g., 897,080x) and other metrics. The background is white with colored bars above each table.

#### 4.4 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	endarse	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*:

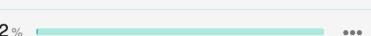
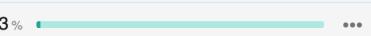
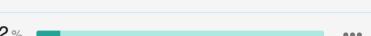
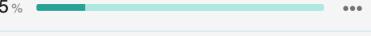
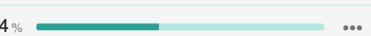
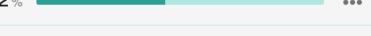
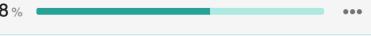
The screenshot shows the 'FREQUENCY' tool in the Corpus Explorer interface. At the top, there's a search bar with 'word whatever • 62,314' and a note '63.12 per million tokens • 0.0063%'. To the right are various icons for search, download, and export. A red box highlights the 'TEXT TYPES' button in the 'More presets' section. The main area shows frequency filters for 'WORD FORMS', 'PART OF SPEECH', 'TAGS', and 'LEMMAS' both for the left and right words. Below the tool, a list of three historical documents from 1640 is displayed, each with a checkbox and a snippet of text containing the word 'whatever'.

Details	Left context	KWIC	Right context
1 <input type="checkbox"/> ① 1640 interpret all his actions in such a sense, as perswading our selvs, <b>whatever</b> things were amisse in Church or Common-weale, or whatever inn			
2 <input type="checkbox"/> ① 1640 lvs, whatever things were amisse in Church or Common-weale, or <b>whatever</b> Innovations brought in, yea although under the name of Royall Au			
3 <input type="checkbox"/> ① 1640 as is noted before, which least it be forgotten, we mention againe) <b>whatever</b> Conclusions or Orders are made at those Tables, or Boards / e th			

---

## 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:

