

# Syntax

*Seminar 'Corpus Linguistics'*

Quirin Würschinger, LMU Munich

June 26, 2025

# Session Overview

## 1. Theory and Research

- Theoretical framework (EC-Model)
- Case study: *N+BE+that* construction
- Research findings

## 2. Practice

- Corpus analysis with Sketch Engine
- Hands-on exercises

# The Paper: Schmid & Mantlik (2015)

## Abstract

**Abstract:** Data from eight historical corpora spanning the period between 1250 and 1871 are investigated with regard to occurrences of the ‘N+BE+*that*-construction’ (as in *my concern is that [...], the idea was that [...]*). The formal, semantic, and pragmatic changes of this construction are described on the basis of 1,588 attestations retrieved from the corpora. Following this, the usage profiles of individual authors are examined. It is shown that even authors who are comparable in terms of period and genre show significant differences with regard to the frequency of use of the construction, collocational ranges and preferences, the use of semi-fixed lexical expressions manifesting the construction, as well as their functional preferences. These differences are interpreted from the perspective of the so-called ‘Entrenchment-and-Conventionalization Model’ (Schmid 2014a and 2015). It is argued that the usage profiles of individual authors can provide insights into the ways in which the construction under investigation was represented in these authors’ minds, and that the observable collective long-term changes arise from the interaction of the cognitive processes in individual minds and the social processes taking place in the speech community.

Abstract from Schmid & Mantlik (2015, p. 583)

# Theoretical Framework

## Entrenchment-and-Conventionalisation Model (Schmid and Mantlik 2015; Schmid 2020)

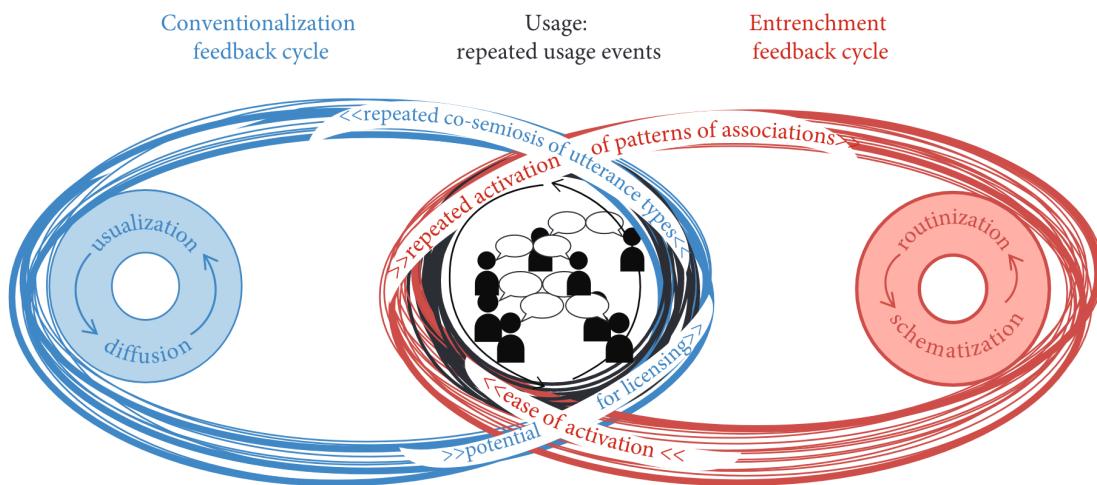


Figure 1.2 Language as a simple Tinguely machine

Entrenchment-and-Conventionalisation Model (Schmid 2020, p. 4)

# Research Question

How do syntactic constructions change over time?

**Case study:** The *N+BE+that* construction

- *my concern is that [...]*
- *the idea was that [...]*
- *the thing is that [...]*
- *the worry is that [...]*

# Construction Grammar Analysis

by *the talk*, by *is*, and by *that De Ruyter is come over-land* respectively. Following common practice in Construction Grammar (cf., e.g., Traugott and Trousdale 2013: 16), we will consider Pepys' utterance as a so-called 'construct' and assume that the production of this construct is licenced by one or more 'constructions' which were represented in Pepys' mind and activated while adding the entry to his diary. Several constructions on different levels of specificity could have motivated the construct in example (1):

- a. a very general, schematic construction of the type Det+N+copula+*that*-clause, whose meaning/function could be glossed as 'THING-concept (encoded by the noun) encapsulates proposition (encoded by the *that*-clause)';<sup>3</sup>
- b. a more specific but still schematic version, a sub-schema, which would take into account that *talk* is a linguistic noun; so the sub-schema would be Det+N<sub>linguistic</sub>+copula+*that*-clause, associated with the more specific meaning 'speech-reporting noun encapsulates message';
- c. a fixed expression, *all the talk is that*, i.e. a substantive, lexically filled construction, which could roughly be glossed as 'here is what people talk about these days'.

Analysis of the *N+BE+that* construction in Construction Grammar terms (Schmid & Mantlik 2015, p. 586)

# Data Sources

**Table 1:** Data acquisition: eight corpora accessed online or retrieved from the Internet<sup>4</sup>

Corpus	Approx. Number of Words	Text Types	Time	Number of Tokens
Gutenberg	18,500,000	Fiction, Essays, Chronicles	1250–1871	1,039
OED3	20,000,000 (est.)	Mixed	1250–1871	194
Helsinki	1,500,000	Mixed	730–1710	13
Old Bailey	17,000,000	Court Proceedings	1720–1871	153
Paston	200,000	Letters	1422–1509	9
PCEEC	2,200,000	Letters	c. 1410–1695	140
Shakespeare	900,000	Drama, Poetry	1590–1612	17
Jane Austen's Letters	140,000	Letters	1796–1817	23
				<b>Total 1,588</b>

Historical corpora used in the study (p. 591)

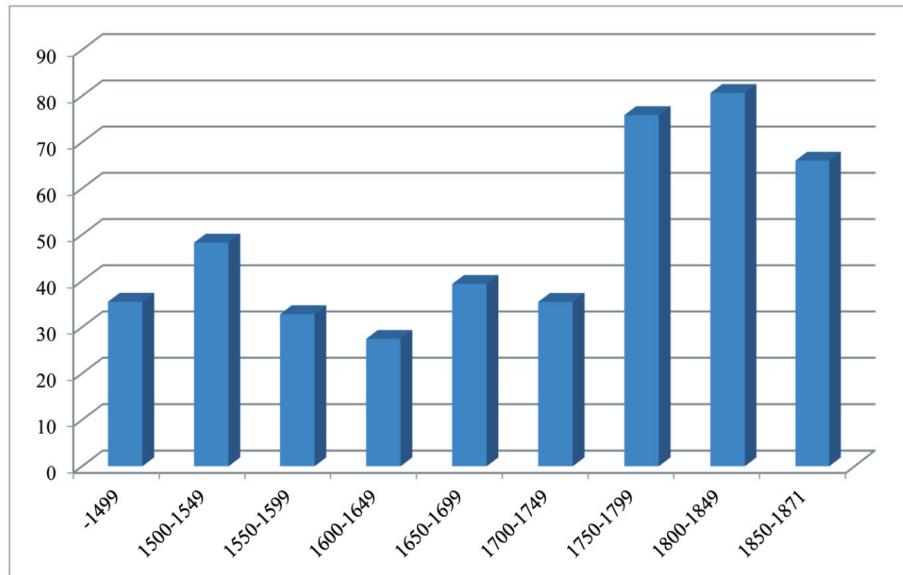
# Corpus Attestation

- (1) 29th. [October 1664] All **the talk is that** De Ruyter is come over-land home with six or eight of his captaines to command here at home, [...]. (*The Diary of Samuel Pepys*, kept from January 1660 to May 1669, first published 1825)

Examples from historical corpora (Schmid & Mantlik 2015, p. 586)

# Research Results

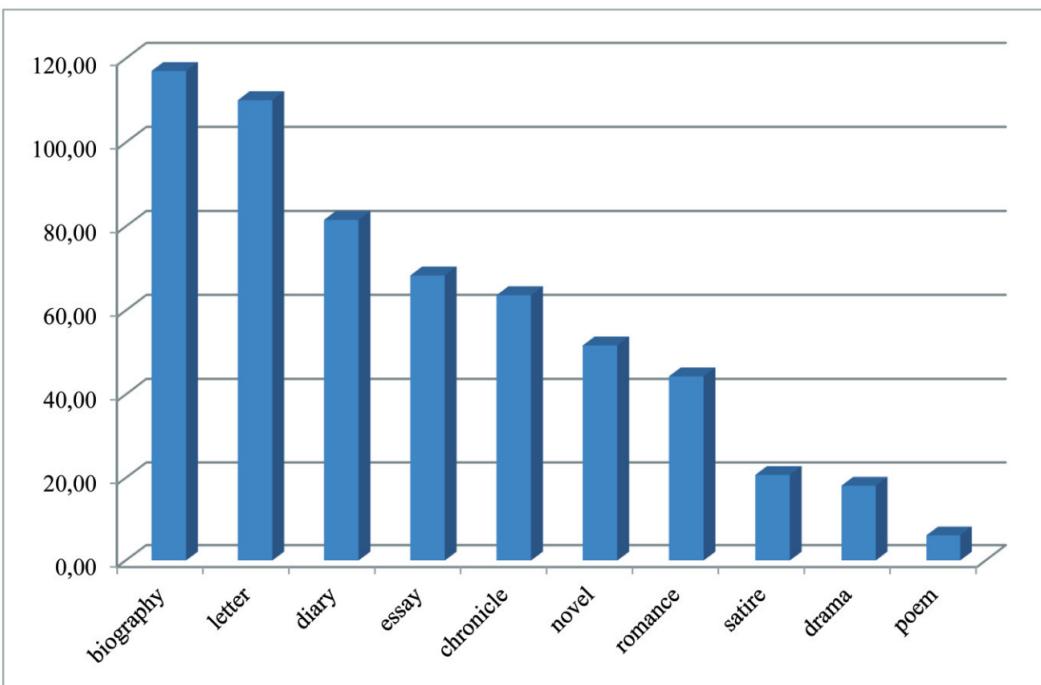
## Diachronic Frequency



**Figure 7:** Relative frequency per million words per 50-year period

Steady increase from medieval to modern English (Schmid & Mantlik 2015, p. 601)

# Text Type Distribution



**Figure 6:** Frequency per million words in different genres (n = 1069)

Distribution across different text types (Schmid & Mantlik 2015, p. 600)

# Speaker Variation

**Table 4:** Comparison of authors from selected periods writing the same genre

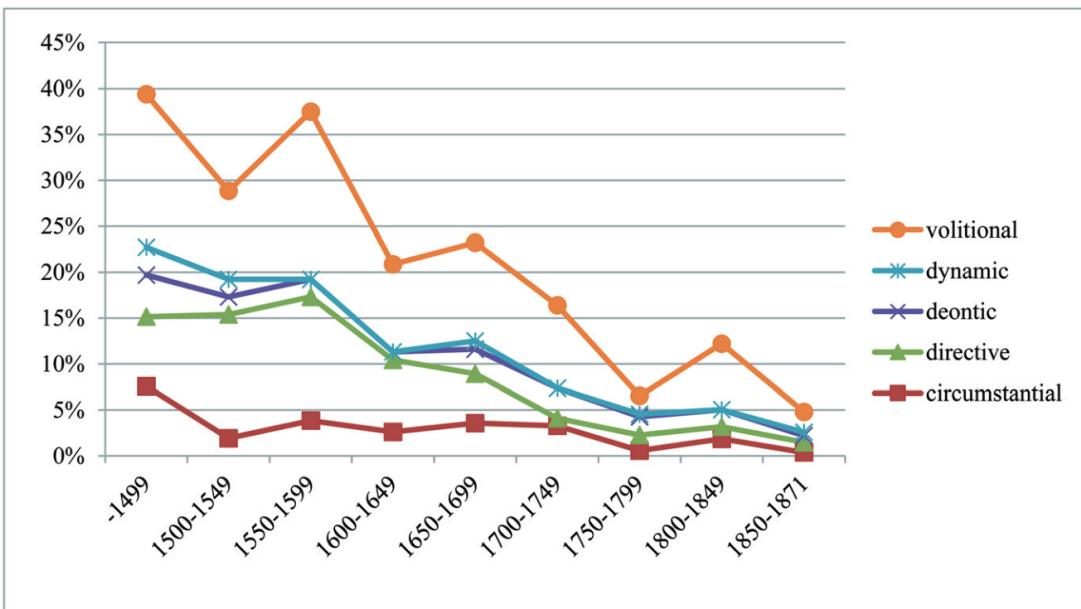
Author	Dates of Publications	Number of Words in Source	Number of Occurrences of N+BE+that	Relative Frequency per Million Words	$\chi^2$ Test Results
Mid-18 <sup>th</sup> -Century Fiction					
Richardson	1740–1748	1,406,495	42	29.86	$\chi^2 = 19.1447$ df = 3 p = 0.0002552
Fielding	1742–1751	691,821	33	47.70	
Haywood	1744–1748	170,858	7	40.97	
Smollett	1748–1771	659,690	5	7.58	
Early 19 <sup>th</sup> -Century Fiction					
Austen	1795–1818	887,680	71	79.98	$\chi^2 = 12.7824$ df = 2 p = 0.001676
Shelley	1818–1826	250,729	11	43.87	
Scott	1814–1820	983,630	41	41.68	

Individual author differences (Schmid & Mantlik 2015, p. 606)

Author	Dates of Publications	Number of Words in Source	Number of Occurrences of N+BE+that	Relative Frequency per Million Words	$\chi^2$ Test Results
<b>Mid-19<sup>th</sup>-Century Fiction</b>					
Brontë, E.	1847	117,276	6	51.16	$\chi^2 = 21.037$ df = 3 p = 0.0001034
Brontë, C.	1847–1856	395,351	12	27.82	
Dickens	1836–1861	1,409,404	144	102.17	
Thackeray	1848–1852	491,855	50	101.66	
<b>Late 18<sup>th</sup>-Century Philosophers</b>					
Hume	1739–1779	137,954	6	43.49	$\chi^2 = 58.3611$ df = 2 p = 2.124e-13
Paine	1776–1796	179,138	48	267.95	
Burke	1756–1796	1,515,549	129	85.12	

Usage profiles of different authors (Schmid & Mantlik 2015, p. 607)

# Semantic Change



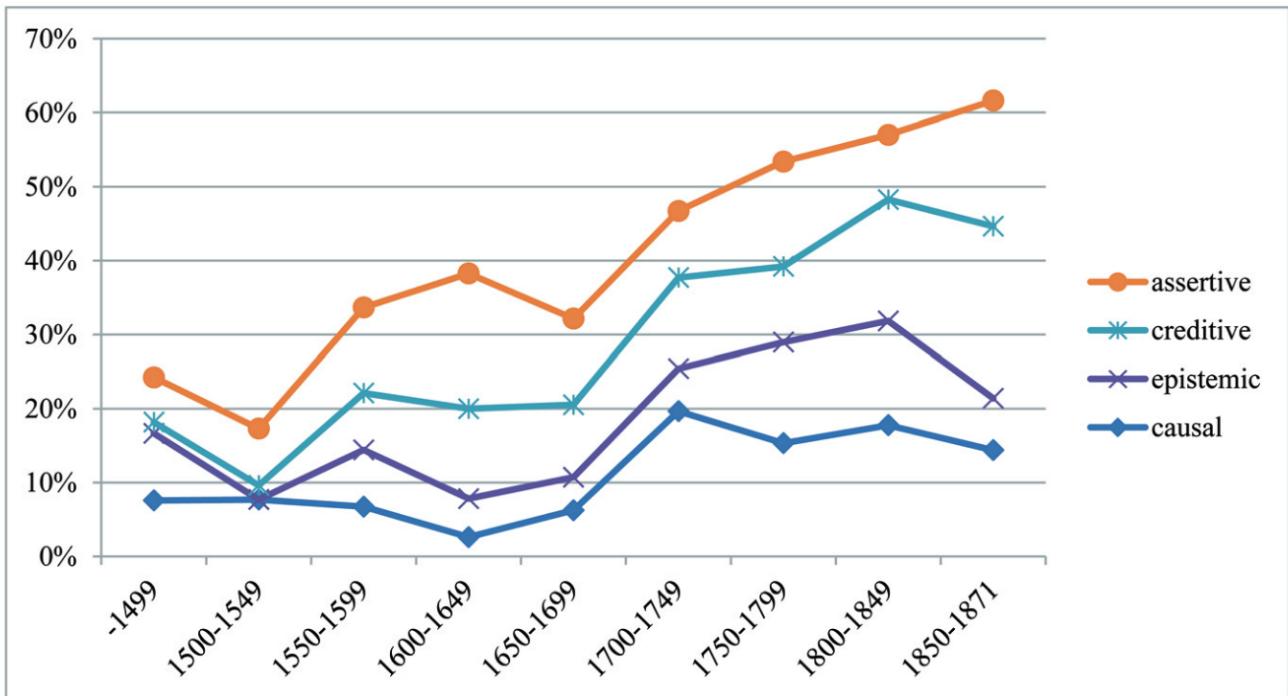
**Figure 3:** Decrease in relative frequency of nouns expressing volition, dynamic and deontic modality, directive illocutions, and circumstance

Changes in semantic preferences over time (Schmid & Mantlik 2015, p. 597)

The types of nouns collected in Figure 4 share semantic components related to actions and events: volition (see example 4), dynamic modality (5), obligation (6), giving orders (7), and circumstance (8):

- (4) volitional: **The wyl** of crist **was that** she sholde abyde. (Chaucer, *Canterbury Tales*, 1430)
- (5) dynamic: **My mete** ['means of support of strength', *OED3*, s.v. *mete*] **is that** I do the will of him that sente me. (*Wycliffe Bible*, John 4:34, 1384, *OED3*)
- (6) deontic: **The lawe** of Medis and Persis **is that** eche decree whiche the kyng ordeyneth be not leeful for to be chaungid. (*Wycliffe Bible*, Daniel 6:15, 1384, *OED3*)
- (7) directive: **Our request is, that** you would by the bearer of these presents, [...]. (Hakluyt, *The Principal Navigations, Voyages, Traffiques and Discoveries of the English Nation*, Vol. 1, 1598)
- (8) circumstantial: **The kind** of the mors **is that** they wole leefe noothing empty besides hem. (Mandeville, *Mandeville's Travels*, around 1425, *OED3*)

Development of semantic categories (Schmid & Mantlik 2015, p. 598)



**Figure 4:** Increase in relative frequency of nouns expressing assertion, belief, fact, and cause

Detailed semantic analysis (Schmid & Mantlik 2015, p. 598)

# Summary: What the Study Reveals

## Key Patterns (1250-1871):

- **Frequency** Steady growth over 600+ years
- **Authors** Individual cognitive differences
- **Meaning** Concrete → Abstract shift
- **Registers** Genre-specific patterns

## EC-Model Insights:

- **Usage** drives historical change
- **Entrenchment** varies by individual
- **Conventionalisation** creates norms
- **Corpus data** reveals cognitive processes

**Bottom line:** Syntactic constructions evolve through individual minds over historical time

# Practice: Corpus Study

**Tool:** English Historical Book Collection (EEBO, ECCO, Evans) in Sketch Engine

**Objective:** Replicate the syntactic construction analysis

# Your Tasks: Overview

## What you'll accomplish today:

1. **Search** for the *N+BE+that* construction using concordance tools
2. **Analyse** absolute and relative frequencies across time periods
3. **Export** results and create visualisations in Excel
4. **Use CQL** to investigate specific nouns in the construction
5. **Compare** patterns between historical periods (1400s vs 1800s)
6. **Interpret** findings using the EC-Model framework

**Goal:** Experience the full research workflow from corpus query to theoretical interpretation

# Part 1: Basic Corpus Analysis

## What you'll do in Steps 1–3:

1. **Search** for the *N+BE+that* construction in historical texts
2. **Analyse** absolute and relative frequencies
3. **Explore** how frequency changes over time (by century and decade)

**Goal:** Understand basic corpus query techniques and frequency analysis

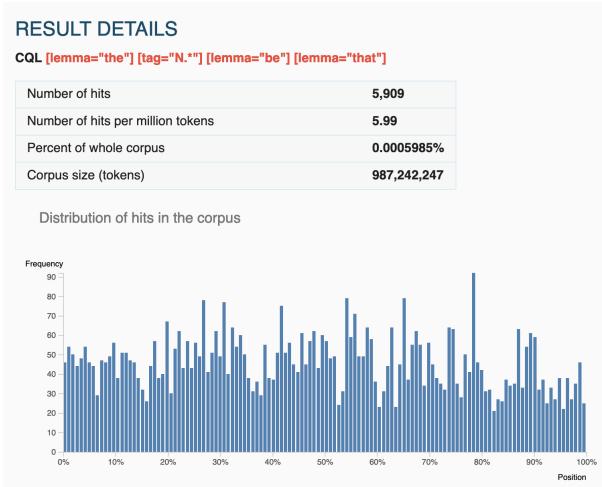
**Time:** ~15 minutes

## Step 1: Concordance Analysis

CONCORDANCE		English Historical Book Collection (EEBO, ECCO, Evans)	Get more space +	?	!	User icon
CQL [lemma="the"] [tag="N.*"] [lemma="be"] [lemma="that"] • 5,909 5.99 per million tokens • 0.0006%	i	⚡				
<input type="checkbox"/> Details	Left context	KWIC	Right context			
1 <input type="checkbox"/> i 1515 jht & put hym to warde the fee & came in to Englond to London there	<b>the kyng was that</b>	tyme and all the lordes of Englond / and helde a parlyament.</s><s>(				
2 <input type="checkbox"/> i 1531 o the owner againe.</s><s>But the more precious and delectable that	<b>the thyng was that</b>	I lente and let you haue at your pleasure / the more a great deale ye or				
3 <input type="checkbox"/> i 1529 om one that is not Christen.</s><s>Yet nevertheless the councelle of	<b>the apostle is that</b>	soche shulde not departe.</s><s>Inasmuche as his maryed mate why				
4 <input type="checkbox"/> i 1584 f the vworld being foolishenes before God, and by the contemnyng of	<b>the vworld is that</b>	other vvisdome gotten.</s><s>CHAP. 34.</s><s>IF any be wyse amor				
5 <input type="checkbox"/> i 1616 urth different sort of Waters is that of the Riuier.</s><s>The worst of all	<b>the rest is that</b>	of the Poole and Marish Grounds: and yet that which runneth not is wo				
6 <input type="checkbox"/> i 1616 as there shall be any of the occasions mentioned.</s><s>The worst of	<b>the Ciders is that</b>	which is made of wild Apples, stampf and cast into a vessell with fount				
7 <input type="checkbox"/> i 1570 nt lines which fall in ye concavitie or hollownes of the circumference of	<b>the circle is that</b>	which passeth by the centre.</s><s>And of all the other lines, that line				
8 <input type="checkbox"/> i 1609 s, and the Reformed Church are contrary in the fayd Councell.</s><s>	<b>The firft is that</b>	the faid Councell doth hold that through the workes which proceede of				
9 <input type="checkbox"/> i 1577 xecuted the office of Highpriefthode for one daye & no more.</s><s>	<b>The caufe was that</b>	Mathias the Highprieft dreamed the nyght before, that he had the comp				
10 <input type="checkbox"/> i 1568 :urall meane: of Musick to sing: of Arithmetick to number. &c.</s><s>	<b>The Accident is that</b>	,Accident is two wayes to be considered: eyther separable from the su				

Query for the construction and get a concordance view

## Step 2: Frequency Analysis



Frequency analysis results

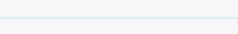
### Tasks:

- absolute frequency: 5,909 tokens
- relative frequency: 5.99 occurrences per million words (pmw)

# Step 3: Diachronic Analysis

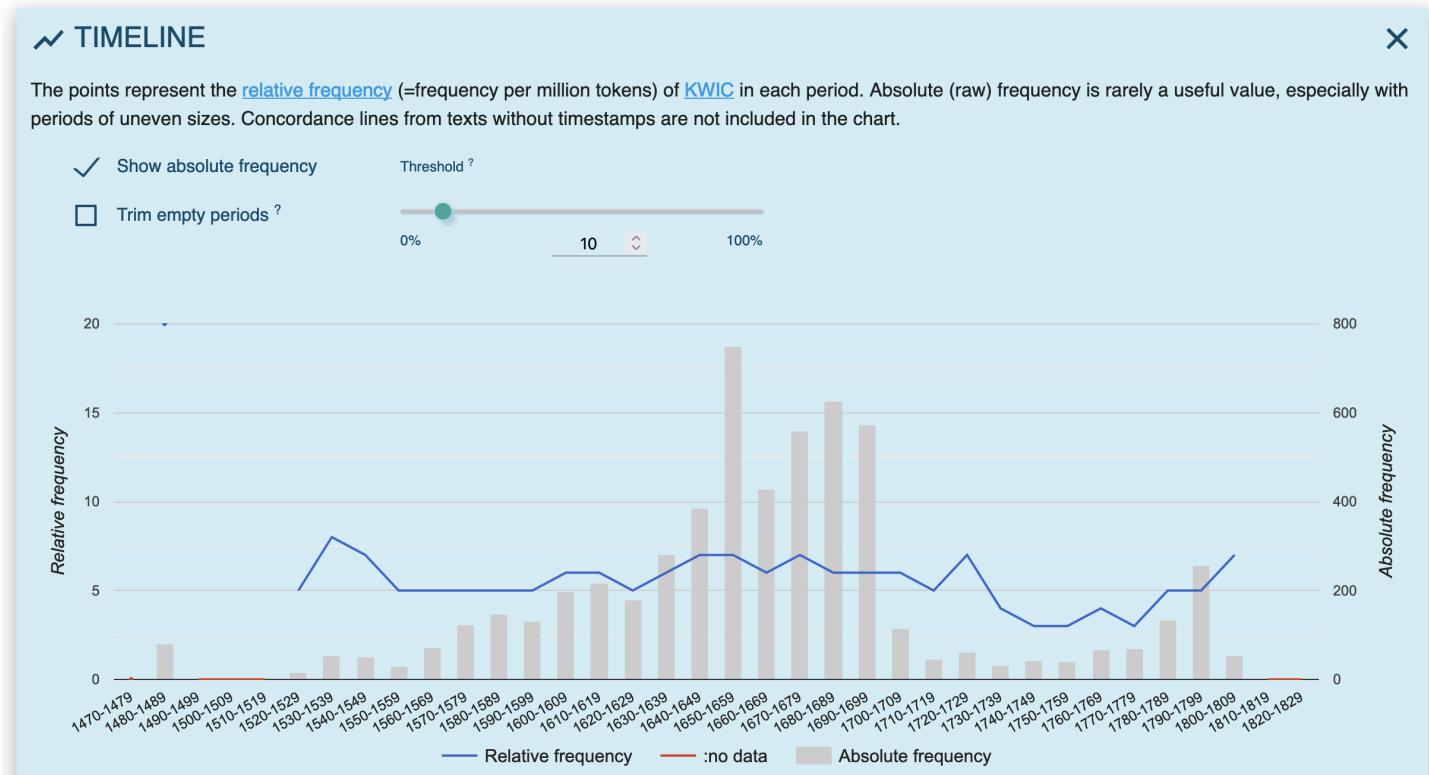
## Frequency by Century

(5 items, 5,909 total frequency)

	Century ↓	Frequency	Relative in text type ?	Relative density ?	
1	1400-1499	109	16.04	267.93 %	 ...
2	1500-1599	687	5.97	99.74 %	 ...
3	1600-1699	4,197	6.28	104.99 %	 ...
4	1700-1799	863	4.54	75.82 %	 ...
5	1800-1899	53	7.22	120.66 %	 ...

Grouped by text type [Century](#)

# Timeline by Decade



More detailed temporal analysis

## Part 2: Advanced Analysis & Visualisation

### What you'll do in Steps 4–5:

1. **Export** your results to Excel for further analysis
2. **Create** data tables and visualisation charts
3. **Use CQL** to query specific nouns in the construction
4. **Compare** patterns across different historical periods

**Goal:** Master data export, visualisation, and advanced corpus queries

**Time:** ~25 minutes

# Step 4: Excel Analysis

## CONCORDANCE

English Historical Book Collection (EEBO, ECCO, Evans)   Get more space 

CQL [lemma="the"] [tag="N.\*"] [lemma="be"] [lemma="that"] • 5,909  
5.99 per million tokens • 0.0006%

**DOWNLOAD**

Download first  rows    

Save the current view as 

See the [download limits](#). You may need to allow pop-ups in your browser when downloading.

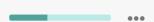
## Frequency

[CHANGE CRITERIA](#)

[BACK TO CONCORDANCE](#)

 Show relative in text types  Show relative density

(5 items, 5,909 total frequency)

Century	Frequency	Relative in text type ?	Relative density ?
1 	1400-1499	109	16.04 267.93 % 
2 	1500-1599	687	5.97 99.74 % 
3 	1600-1699	4,197	6.28 104.99 % 

Export results in Excel format (.xlsx file type)

# Model Solution Available

## **Excel file with sample analysis:**

[https://1drv.ms/x/s!AvkgNVI9yS6aokO4dB\\_h1\\_DiXKmw](https://1drv.ms/x/s!AvkgNVI9yS6aokO4dB_h1_DiXKmw)

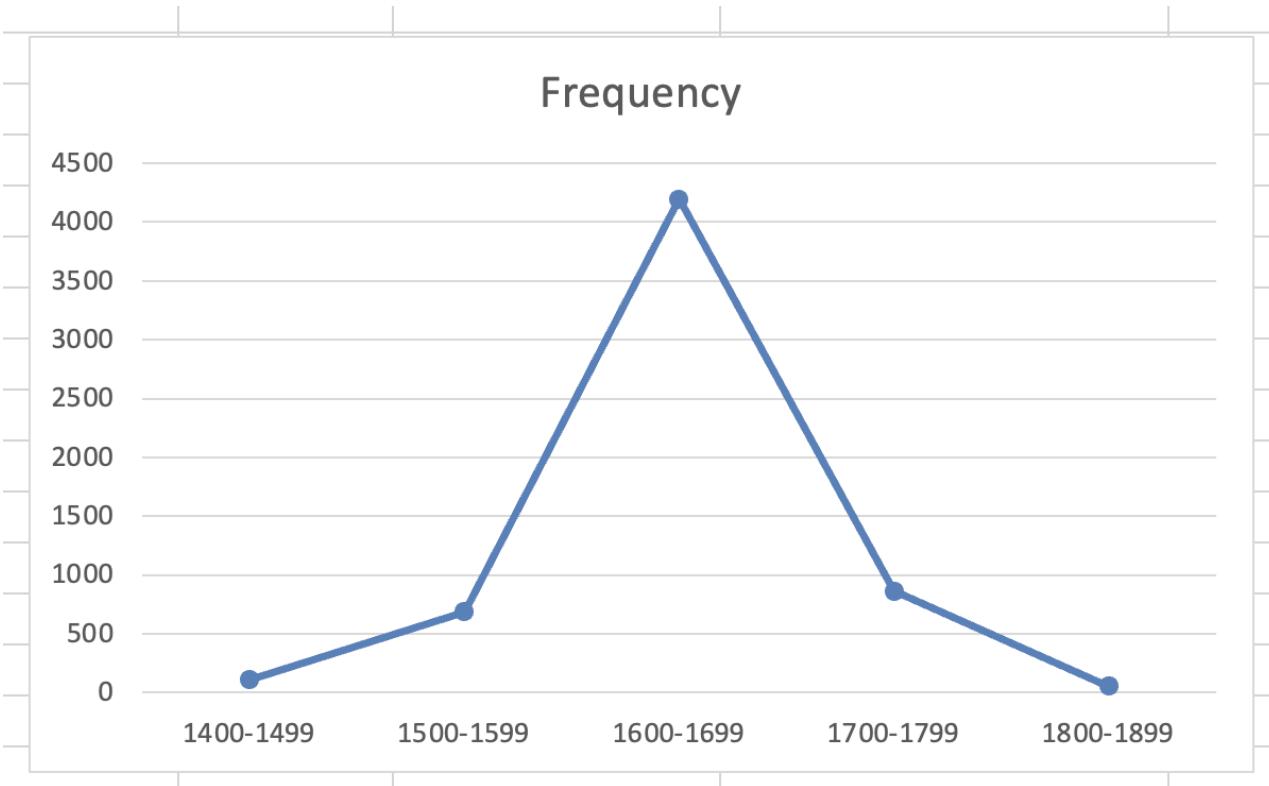
- export results in Excel format
- create data tables
- generate line charts for both absolute and relative frequencies

# Creating Data Tables

	A	B	C	D
1	Century	Frequency	Relative density	Relative in text types
2	1600-1699	4197	104.99451	6.2843
3	1700-1799	863	75.81926	4.53806
4	1500-1599	687	99.7377	5.96966
5	1400-1499	109	267.92554	16.03631
6	1800-1899	53	120.65968	7.22192
7				

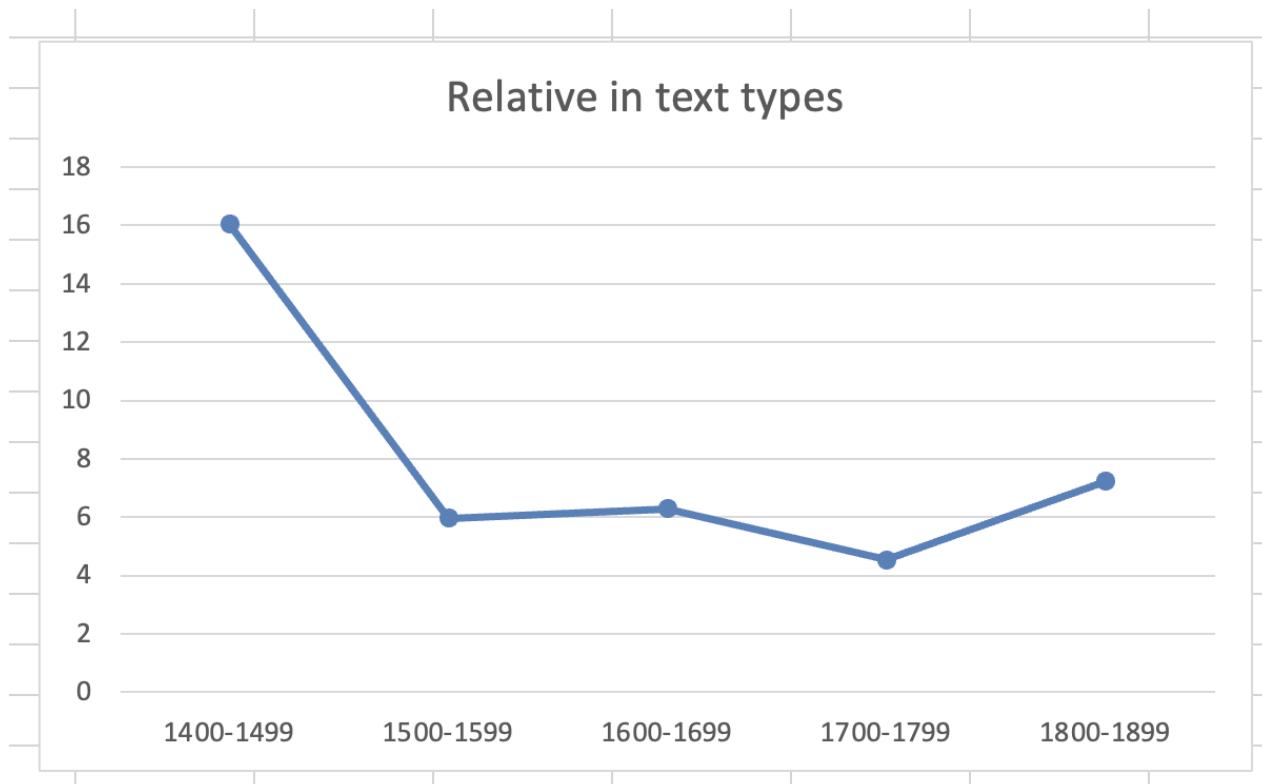
Make a table for further analysis

# Line Charts: Absolute Frequency



Visualising absolute frequency development

# Line Charts: Relative Frequency



Visualising relative frequency development

## Step 5: CQL Analysis

**Query specifically for the noun in the construction using CQL**

*Hint: use the operator **within** preceding the construction*

**Objective:** Identify the most frequent nouns used in the construction

# Most Frequent Nouns

SKETCH ENGINE

CQL builder

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

The screenshot shows the CQL builder interface with the following query structure:

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

Below the query, there is a "result example" section containing a block of Middle English text from Chaucer's Canterbury Tales. The text discusses the four elements of the query: kynges, thynge, apostle, and vworld. The text is annotated with red highlights for these words.

USE THIS CQL ➤

result example ▾

as moche as he myght & put hym to warde the fee & came in to Englonde to London there the  
o be loste / that is restored to the owner againe. But the more precious and delectable that the  
item parsonse may departe from one that is not Christen. Yet neverthelesse the councelle of the  
ome, the vvisdome of the vworld being foolishenes before God, and by the contemyngne of the  
bottome of a Valley. The fourth different sort of Waters is that of the Riuer. The worst of all the  
it, and according as there shall be any of the occasions mentioned. </p><p> The worst of the  
greatest of those right lines which fall in ye concaultie or hollownes of the circumference of the  
ce now to my deare God, I am deuoted. Moreouer by his death it doth appeare, How great the  
by his own teftimonie, it is alwayes lawfull to double of a generall trueth, howe good foever the  
fome doe hold, that thoſe things which feme here to dye. doe paffe or goe, ad non eus: But the

**kynges** was that tyme and all the lordes of Englonde / and helde a parlyament. Godewin fente to hym  
**thynge** was that lente and let you haue at your pleasure / the more a great deale ye oughte to haue ti  
**apostle** is that soche shulde not departe. Inasmuche as his maryed mate whyche ys not Christen doth  
**vworld** is that other vvisdome gotten. CHAP. 34. </p><p> If any be wyse amongst you, </p><p> Cor. 3.  
**rest** is that of the Poole and Marish Grounds: and yet that which runneth not is worse than all the re  
**Ciders** is that which is made of wild Apples, stampf and cast into a vessell with fountaine water in suffi  
**circle** is that which passeth by the centre. And of all the other lines, that line which is higher to the lin  
**loue** is that my God doth bear To me sinnes Monster and most worthy blame; The badge of ignom  
**man** be that affirmeth the fame. </p><p> 2. & vrb. cond. 1. 2. ca 24. Hip. de virg. morb. </p><p> And wherea  
**trueth** is that the thynges whiche feeme to dye haue their beyng: For feyng life is to haue beyng, wh

## Analysis of lexical preferences

## FREQUENCY

BASIC ADVANCED ABOUT

Select an attribute and its position in the concordance: ?

KWIC
1 2 3 4 5 6

left context

6

5

4

3

2

1

KWIC

right context

5

6



Group by first column

GO

More presets

- KWIC WORD FORMS**
- KWIC TAGS**
- KWIC LEMMAS**
- TEXT TYPES**
- LINE DETAILS**

	<input type="checkbox"/> Details	Left context	KWIC	Right context
1	<input type="checkbox"/> ① 1515 yght & put hym to warde the fee & came in to Englonde to London there the		<b>kyng</b>	was that tyme and all the lordes of l
2	<input type="checkbox"/> ① 1531 † to the owner againe.</s><s>But the more precious and delectable that the		<b>thyng</b>	was that I lente and let you haue at
3	<input type="checkbox"/> ① 1529 from one that is not Christen.</s><s>Yet nevertheless the councelle of the		<b>apostle</b>	is that soche shulde not departe.</s>
4	<input type="checkbox"/> ① 1584 of the vworld being foolishenes before God, and by the contemnyng of the		<b>vworld</b>	is that other vvisdome gotten.</s><

Extended frequency analysis

(1,623 items, 4,194 total frequency)

	Lemma	Frequency	Relative ?	
1	truth	101	0.10	...
2	Lord	79	0.08	...
3	spirit	66	0.07	...
4	meaning	62	0.06	...
5	firft	58	0.06	...
6	thing	56	0.06	...
7	law	54	0.05	...
8	cause	50	0.05	...
9	Pope	47	0.05	...
10	reason	43	0.04	...

Detailed breakdown of noun frequencies

# Diachronic Noun Analysis

CHANGE CRITERIA

BASIC ADVANCED ABOUT

Query type ②

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

CQL

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

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

CQL BUILDER [ ]

Default attribute ? lemma

Subcorpus ② none (the whole corpus) ▾ 🔒 + Macro ② none

Filter context ② ▾

Text types (1) ② ▾

Type of paragraph ▾

Century ▾

- 1600-1699 X
- 1400-1499
- 1500-1599

Decade ▾

GO

expand all collapse all

The screenshot shows the 'Change Criteria' interface of Sketch Engine. The 'CQL' tab is selected. A query is entered: '[tag="N.\*"] within [word="the"] [tag="N.\*"] [lemma="be"] [word="that"]'. Below the query is a toolbar with operators like [], {}, <>, "", &, \, |, ~, #, and TAGS. A 'CQL BUILDER' button is also present. The 'Default attribute?' dropdown is set to 'lemma'. Under 'Subcorpus', 'none (the whole corpus)' is selected. A 'Macro' dropdown is set to 'none'. A 'Filter context' dropdown is open. A 'Text types' section shows a single entry: 'Type of paragraph'. A 'Century' filter dropdown is open, showing three options: '1600-1699' (selected), '1400-1499', and '1500-1599'. A green box highlights the 'Century' dropdown. To the right, a 'Decade' dropdown is shown. At the bottom right is a red 'GO' button. On the right side of the interface, there's a sidebar with a video player showing a 'CQL 1: Complex cor...' video from Sketch Engine, and links to 'CQL manual' and 'Sketch Engine website'.

How noun preferences change over time

# Historical Comparison: 1400–1499

(55 items, 109 total frequency)

	Lemma (lowercase)	Frequency	Relative ?	
1	fyrst	11	0.01	...
2	thyrd	9	< 0.01	...
3	second	9	< 0.01	...
4	cause	7	< 0.01	...
5	kyng	5	< 0.01	...
6	fyrft	4	< 0.01	...
7	fyrfte	3	< 0.01	...
8	maner	3	< 0.01	...
9	caufe	3	< 0.01	...
10	firft	3	< 0.01	...

Most frequent nouns in early period

# Historical Comparison: 1800–1899

(37 items, 53 total frequency)

	Lemma (lowercase)	Frequency	Relative ?	
1	fact	6	< 0.01	...
2	confequence	5	< 0.01	...
3	probability	4	< 0.01	...
4	perfons	2	< 0.01	...
5	misfortune	2	< 0.01	...
6	devil	2	< 0.01	...
7	story	2	< 0.01	...
8	truth	1	< 0.01	...
9	bay	1	< 0.01	...
10	ecliptic	1	< 0.01	...

Most frequent nouns in later period

# Summary

- **Theoretical framework:** EC-Model for understanding syntactic change
- **Methodological skills:** Corpus querying, frequency analysis, CQL
- **Research workflow:** From hypothesis to interpretation
- **Historical perspective:** How constructions evolve over centuries
- **Cognitive perspective:** How individual authors' minds differ and shape language change

# References

- Schmid, Hans-Jörg. 2020. *The Dynamics of the Linguistic System: Usage, Conventionalization, and Entrenchment*. Oxford: Oxford University Press.
- Schmid, Hans-Jörg, and Annette Mantlik. 2015. "Entrenchment in Historical Corpora? Reconstructing Dead Authors' Minds from Their Usage Profiles." *Anglia* 133 (4): 583–623.