LCD LANGUAGE CHANGE DATABASE

http://www.helsinki.fi/lcd/

META-ANALYSIS FOR HISTORICAL CORPUS LINGUISTICS

Using the Language Change Database

Terttu Nevalainen, Tanja Säily, Turo Vartiainen & Joonas Kesäniemi

WHO ARE WE?

- “**Reassessing Language Change: The Challenge of Real Time**”
  - Funded by the Academy of Finland, 2014–2018
  - Linguists: Terttu Nevalainen, Tanja Säily, Turo Vartiainen
  - Database architect: Joonas Kesäniemi
  - Assistant: Agata Dominowska
  - Collaborating scholars: Peter Trudgill (sociolinguist), Jefrey Lijffijt, Jukka Suomela (computer scientists), Aatu Liimatta (PhD student)

REASSESSING LANGUAGE CHANGE

- **Challenge**: Lack of access to data & previous corpus-based research
  → research on real-time change largely not cumulative
- **Our solution: Language Change Database (LCD)**
  - Comparative, real-time **baseline data** for modelling change in progress
  - **Summarises** the results of corpus-based research articles on variation and change in English
  - Can be used to carry out **meta-analyses** of a large number of linguistic changes
  - Open-access **linked data** web application, easy to **edit & query**

LANGUAGE CHANGE DATABASE (LCD)

- Each article is represented by one entry in the database
  - Each entry is **annotated** for linguistic and extra-linguistic features
  - **Numerical data** in a computer-readable form
  - Initially developed and updated in Helsinki, later each researcher can insert the information of their own work into the LCD
  - Currently in beta stage, c. 300 entries
- Search interface now available!
- **www.helsinki.fi/lcd**

LANGUAGE CHANGE DATABASE (LCD)

Study details
- Genre
- Variety
- Grammar
- Dialectology
- Language contact
- Sociolinguistics
- Pragmatics
- Discourse analysis
- Statistical methods

Corpus

Study
- Summary of results
- Time period
- Topics

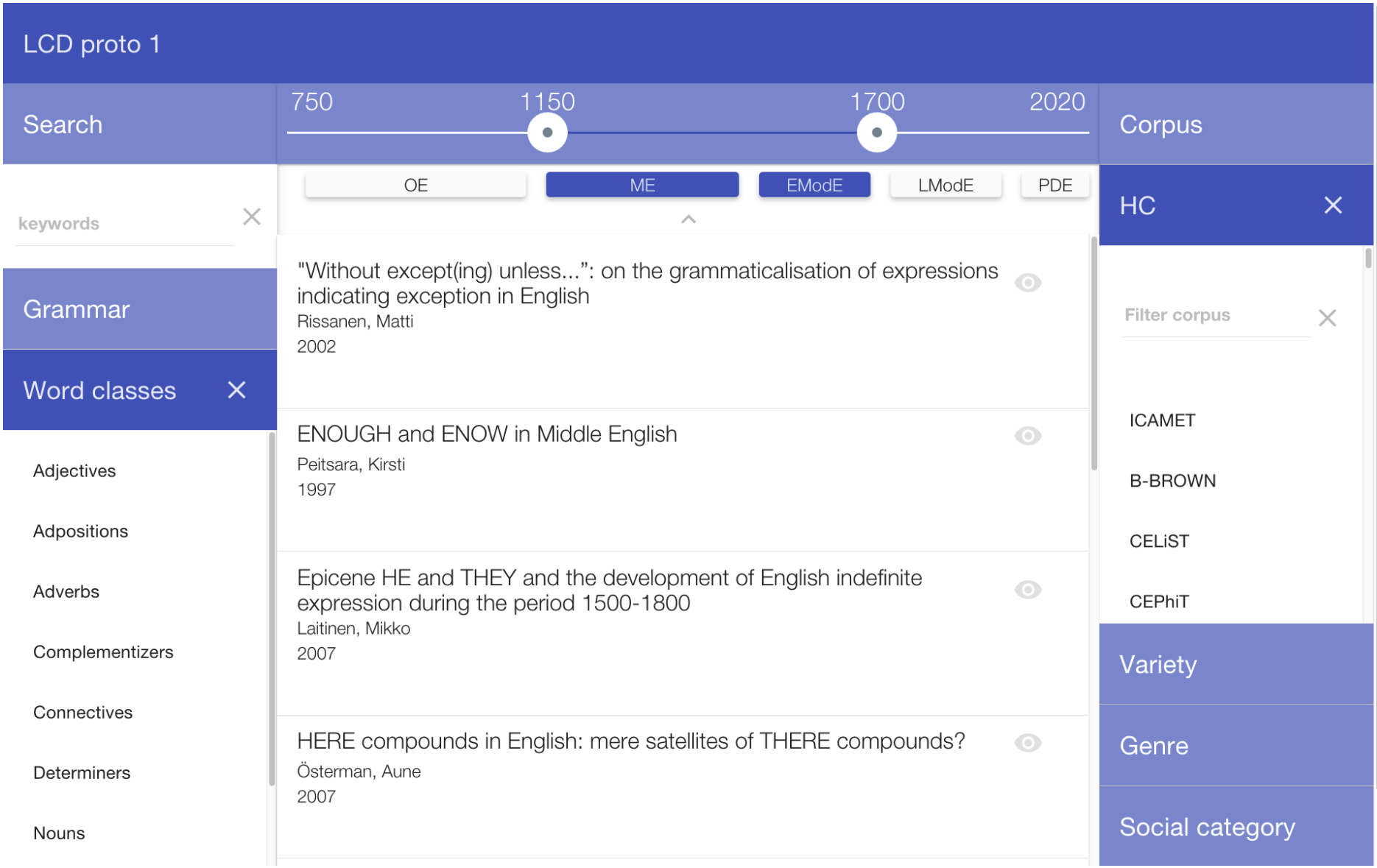
Publication
- Bibliographic data

Corpus composition file

Annotated data file

Data file

Publication file


LCD proto 1
Search
750
1150
1700
2020
Corpus
OE
ME
EModE
LModE
PDE
HC
keywords
Grammar
"Without except(ing) unless...": on the grammaticalisation of expressions indicating exception in English
Rissanen, Matti
2002
Filter corpus
Word classes
ENOUGH and ENOW in Middle English
Peitsara, Kirsti
1997
ICAMET
B-BROWN
CELiST
CEPhiT
Adjectives
Adpositions
Adverbs
Complementizers
Epicene HE and THEY and the development of English indefinite expression during the period 1500-1800
Laitinen, Mikko
2007
Variety
Connectives
Determiners
HERE compounds in English: mere satellites of THERE compounds?
Österman, Aune
2007
Genre
Nouns
Social category

Content details

Topic

(UN)TIL

Keywords

preposition; subordinator; OÞ; TIL; UNTIL; connective

Time periods

Modern English
Present-Day English
Middle English
Old English

Grammar

Subordinator
Preposition
Connectives

Dialectology

Borrowing
Contact
Dialect
Region

Sociolinquistics

Corpora

A Representative Corpus of Historical English Registers
BROWN Corpus
Corpus of Early English Correspondence Sampler
Century of Prose Corpus
Freiburg–LOB Corpus of British English
Freiburg-Brown corpus of American English
Helsinki Corpus
Lampeter Corpus of Early Modern English Tracts
Lancaster-Oslo/Bergen Corpus

Pragmatics

Genre

Correspondence
Drama
No specific genre

Variety

East Midlands
North
Northumbria

Social categories

Language contact

Statistical meth

Summary of results

TIL, a loan connective (subordinator or preposition), borrowed from Old Norse, occurs a few times in Old English texts written in the Northumbrian dialect area, as an equivalent of TO.
- It replaces OÞ in the temporal and local senses in Early Middle English.
- The earliest instances recorded in the Helsinki Corpus occur in East Midland texts where the Scandinavian influence was the strongest.
- The related compound form UNTIL was borrowed in Middle English: the earliest occurrence occurs in the Ormulum.
- It becomes common in the fourteenth century.
- It is, at least to some extent, dialectally restricted throughout the Middle English period (Northern dialect; Northumbrian, East Midland dialect).
- In Modern English, UNTIL supersedes TILL in frequency numbers. It is the more common form particularly in formal registers. In letters and drama, TILL remains common even in twentieth-century corpus samples. UNTIL is clearly favoured by written language, while TILL is proportionally more common in spoken language. Particularly the preposition TILL seems a usage typical of speech: its relative frequency is even higher than that of UNTIL.

- Register:

Files

otables_2005_rissanen_the_development_of_till_and_until_in_english.xlsx

PREVIEW DOWNLOAD

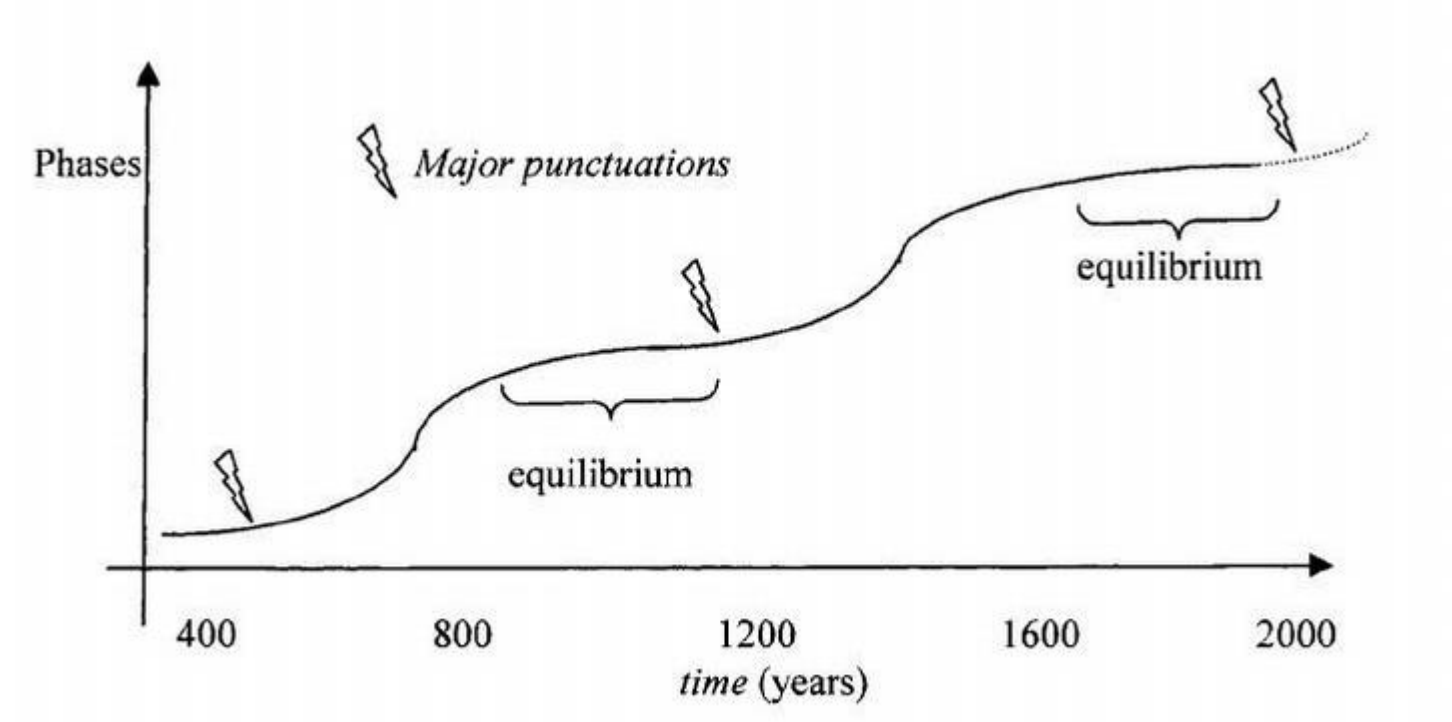
1 2 3 4 5 6 7 8

Table 1. Until and till in the Early Modern English sub-sections of the Helsinki Corpus. Figures per 100,000 words in brackets. (2005)						
		until		till		
	subord.	prep.	total	subord.	prep.	total
EModE1 (1500-1570)	24	11	35 (18.4)	45	16	61 (32.1)
EModE2 (1570-1640)	43	16	59 (31.1)	57	26	83 (43.7)
EModE3 (1640-1710)	6	3	9 (5.3)	87	59	146

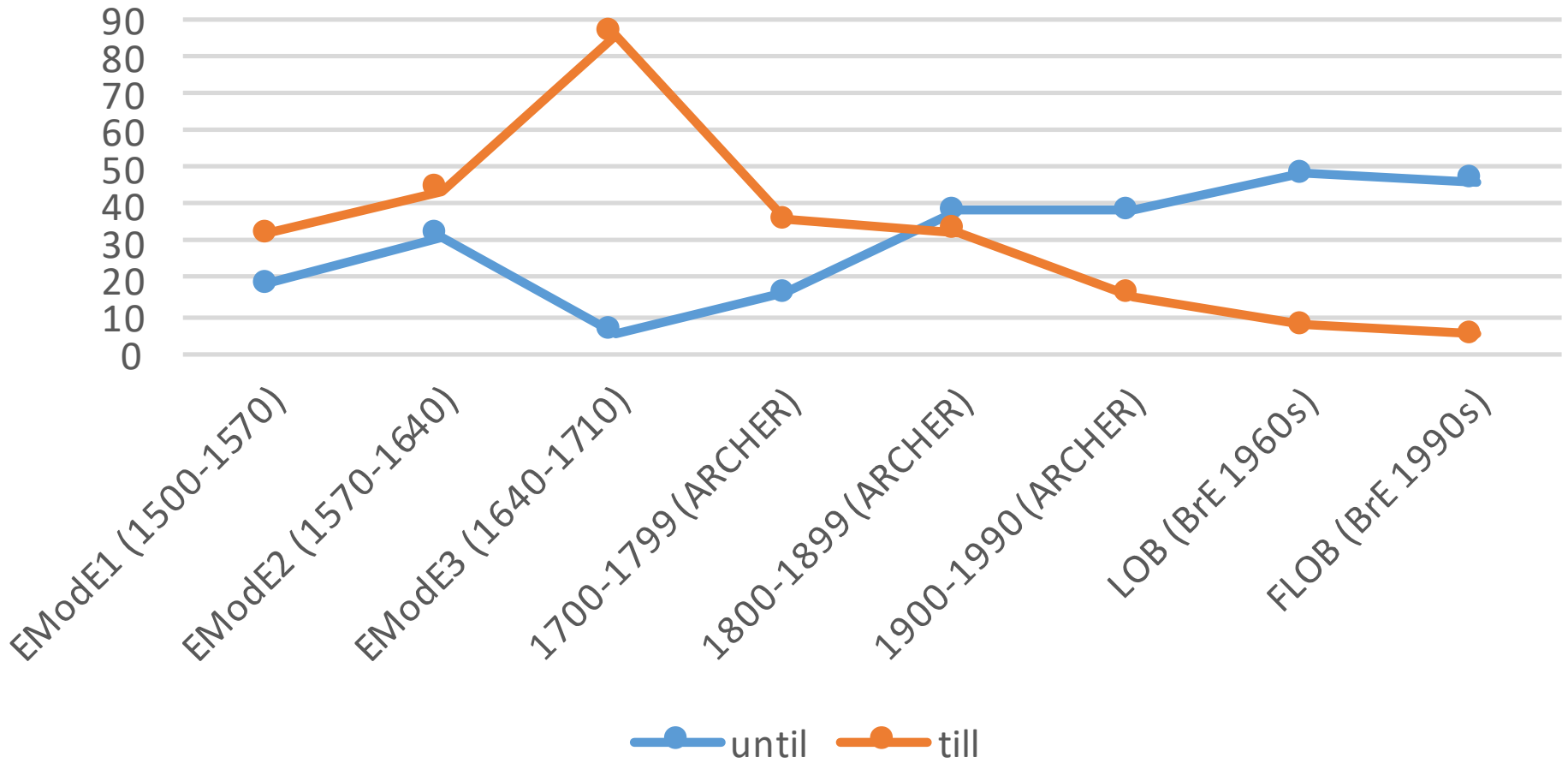
CASE STUDY

- At ISLE4, we presented a manual meta-analysis of **connectives** in the history of English
  - Data from Matti Rissanen's published research
  - Testing the notion of **punctuated equilibrium** using the LCD
  - The Civil War Effect? (Raumolin-Brunberg 1998, Trudgill 2010)
- Example: *till* and *until*

HISTORY OF ENGLISH AS PUNCTUATED EQUILIBRIA



(Bergs 2005: 54, based on Dixon 1997)


until and *till* in the long diachrony
90
80
70
60
50
40
30
20
10
0
EModE1 (1500-1570)
EModE2 (1570-1640)
EModE3 (1640-1710)
1700-1799 (ARCHER)
1800-1899 (ARCHER)
1900-1990 (ARCHER)
LOB (BrE 1960s)
FLOB (BrE 1990s)
until
till

FROM MANUAL TO SEMI-AUTOMATED META-ANALYSIS

- Annotated data tables implemented with **Excel styles**
- Connects spreadsheet cells to linked-data concepts in a machine-readable form

Expression

Function

Table 1. Until and till in the Early Modern English sub-sections of the Helsinki Corpus. Figures per 100,000 words in brackets. (2005)

A1

	until	until		till	till		HC
	subord.	prep.	total	subord.	prep.	total	
EModE1 (1500-1570)	24	11	35 (18.4)	45	16	61 (32.1)	
EModE2 (1570-1640)	43	16	59 (31.1)	57	26	83 (43.7)	
EModE3 (1640-1710)	6	3	9 (5.3)	87	59	146 (85.4)	

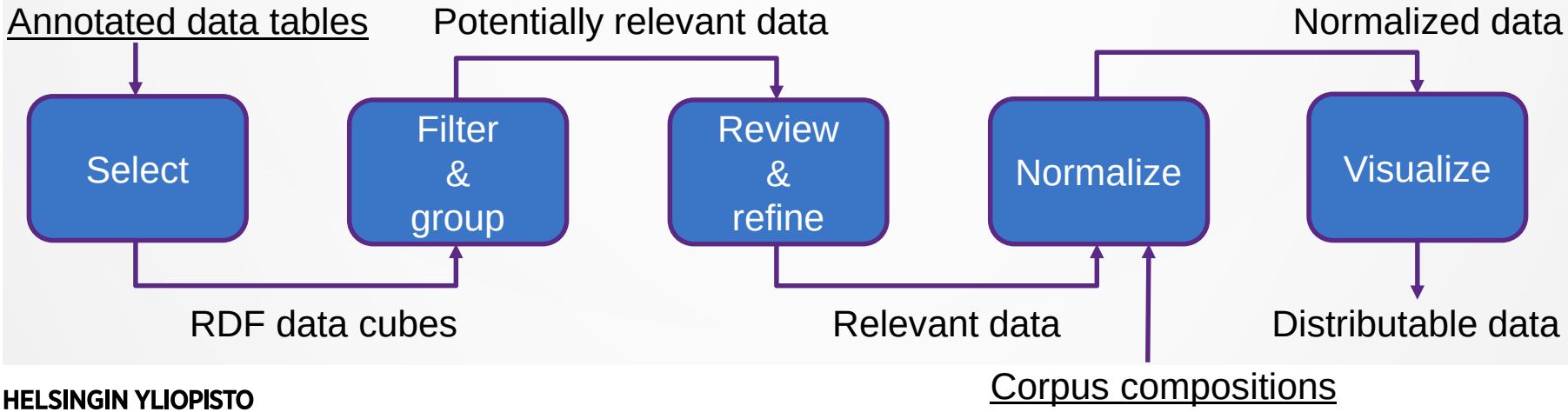
Time period

Absolute values

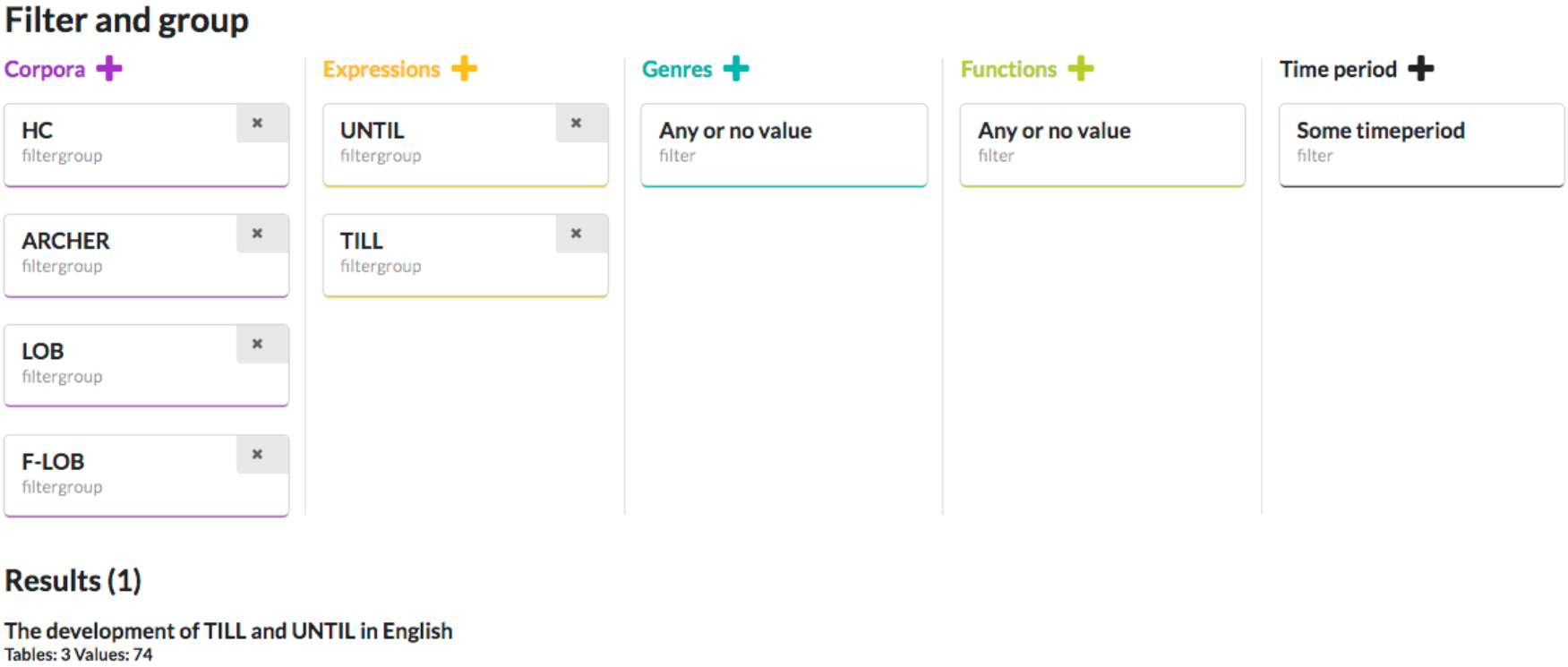
Corpus annotation

TOOL FOR RE-USING LCD DATA

- LADA = LCD Aggregated Data Analysis workbench
  - Tool for **experimenting** with LCD data
  - Small-scale **meta-analysis** across studies
- Generates RDF Data Cubes of the annotated data tables
- Provides a **workflow** for creating new aggregated datasets



LADA WORKFLOW



Review

Filters

HC | ARCHER | LOB | F-LOB | UNTIL | TILL | Any or no value | Any or no value | Some time period

Source tables and filtered values

The development of TILL and UNTIL in English

Rissanen, 2005

Exclude publication

Table1

Table 1. Until and till in the Early Modern English sub-sections of the Helsinki Corpus. Figures per 100,000 words in brackets. (2005)

+ −

	until	until		till	till		HC
	subord.	prep.	total	subord.	prep.	total	
EModE1 (1500-1570)	24	11	35 (18.4)	45	16	61 (32.1)	
EModE2 (1570-1640)	43	16	59 (31.1)	57	26	83 (43.7)	
EModE3 (1640-1710)	6	3	9 (5.3)	87	59	146 (85.4)	

Normalize

Filters

HC ARCHER LOB F-LOB UNTIL TILL Any or no value Any or no value Some time period

Groups

HC ARCHER LOB F-LOB UNTIL TILL Some time period

The development of TILL and UNTIL in English

Rissanen, 2005

Table1

Table 1. Until and till in the Early Modern English sub-sections of the Helsinki Corpus. Figures per 100,000 words in brackets. (2005)

Simple corpus composition of HC using time periods.

Table5

Table 5. Until and till in five genres of the Archer Corpus. Absolute figures. (2005)

ARCHER 3.2 British variant

Table6

Table 6. Until and till in Present-day English corpora. (2005)

LOB CC

F-LOB

Frequencies	Frequencies	Frequencies	Frequencies
absolute 24	absolute 45	absolute 43	absolute 57
absolute 11	absolute 16	absolute 16	absolute 26
generated 18.41 ✓	generated 32.08 ✓	generated 31.09 ✓	generated 43.73 ✓
HC UNTIL 1500-1570	HC TILL 1500-1570	HC UNTIL 1570-1640	HC TILL 1570-1640

Visualize

Filters

HC ARCHER LOB F-LOB UNTIL TILL Any or no value Any or no value Some time period

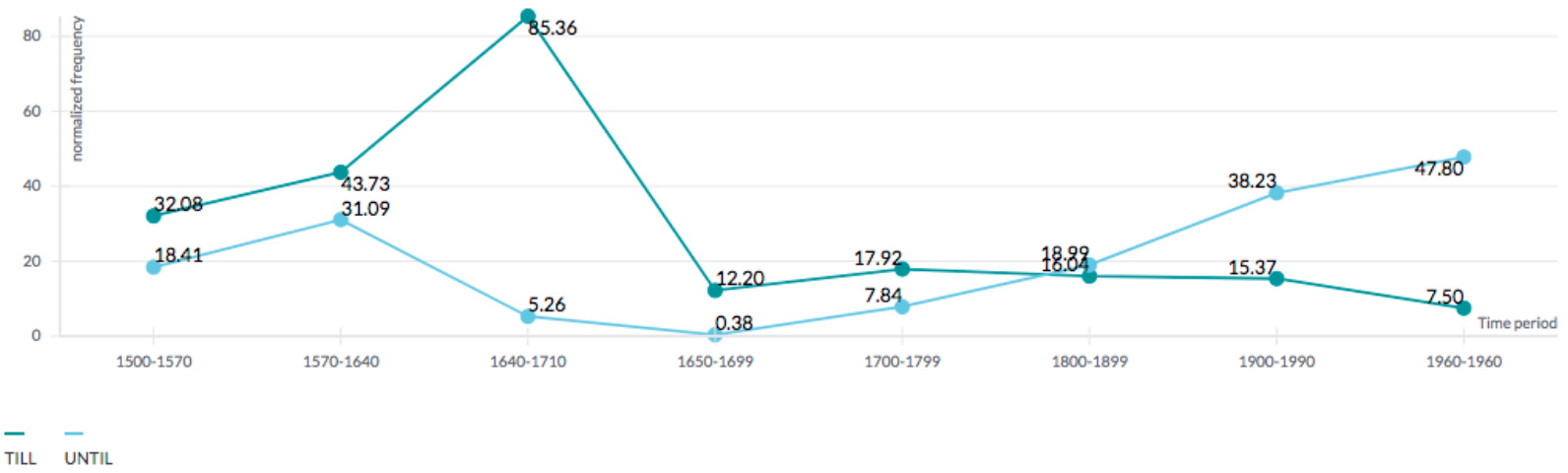
Groups

HC ARCHER LOB F-LOB UNTIL TILL Some time period

Add new graph

Edit graph Export SVG Export CSV

Bar Line Data labels Title



Remove graph

CONCLUSIONS

- The Language Change Database provides baseline data for **meta-analyses**, **replication studies** and **systematic reviews**.
- It is intended to make the field of English historical corpus linguistics **more cumulative** by ensuring **easy access** to the results of earlier research.
- The LADA tool can be used to **experiment** with the numerical data included in the LCD entries and to carry out small-scale meta-analyses.
- Both the LCD and LADA will be made available in accordance with **the best practices of open science**.

REFERENCES

http://www.helsinki.fi/lcd/

- Bergs, Alexander. 2005. *Social networks and historical sociolinguistics: Studies in morphosyntactic variation in the Paston Letters (1421–1503)*. Berlin & New York: Mouton de Gruyter.
- Dixon, R.M.W. 1997. *The rise and fall of languages*. Cambridge: Cambridge University Press.
- Francis, W. Nelson & Henry Kučera. 1979 [1964]. *Manual to accompany a standard sample of present-day edited American English, for use with digital computers*. Original edn. 1964, revised 1971, revised and augmented 1979. Providence, RI: Department of Linguistics, Brown University.
- HC = *The Helsinki Corpus of English Texts*. 1991. Compiled by Matti Rissanen (Project leader), Merja Kytö (Project secretary); Leena Kahlas-Tarkka, Matti Kilpiö (Old English); Saara Nevanlinna, Irma Taavitsainen (Middle English); Terttu Nevalainen, Helena Raumolin-Brunberg (Early Modern English). Helsinki: Department of Modern Languages, University of Helsinki. http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/

REFERENCES

- Kesäniemi, Joonas, Turo Vartiainen, Tanja Säily & Terttu Nevalainen. 2018. Open science for English historical corpus linguistics: Introducing the Language Change Database. *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference, Helsinki, Finland, March 7–9, 2018* (CEUR Workshop Proceedings 2084), ed. by Eetu Mäkelä, Mikko Tolonen & Jouni Tuominen, 51–62. CEUR-WS.org. http://ceur-ws.org/Vol-2084/paper4.pdf
- Nevalainen, Terttu, Turo Vartiainen, Tanja Säily, Joonas Kesäniemi, Agata Dominowska & Emily Öhman. 2016. Language Change Database: A new online resource. *ICAME Journal* 40: 77–94. doi:10.1515/icame-2016-0006
- Raumolin-Brunberg, Helena. 1998. Social factors and pronominal change in the seventeenth century: The Civil-War effect? In Jacek Fisiak & Marcin Krygier (eds.), *Advances in English historical linguistics (1996)*, 361–388. Berlin: Mouton de Gruyter.
- Trudgill, Peter. 2011. *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford: Oxford University Press.