Open Science for English Historical Corpus Linguistics:

INTRODUCING THE LANGUAGE CHANGE DATABASE

UNIVERSITY OF HELSINKI

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
    - Assistants: Agata Dominowska, Aatu Liimatta, Emily Öhman
    - Collaborating scholars: Peter Trudgill (sociolinguist), Jefrey Lijffijt, Jukka Suomela (computer scientists)

UNIVERSITY OF HELSINKI

WHERE ARE WE COMING FROM?

- The researchers in our project are experts in **historical corpus linguistics** and **corpus methodologies**.

- **Corpora** have been used to study language change since the early 1990s.

- Thousands of articles describing various kinds of grammatical and lexical change have since been written on the English language alone.

WHERE ARE WE COMING FROM?

- Much of the existing research is **fragmented**:
  - Many early corpus-based studies were published in edited volumes and festschrifts with low print runs.
    - Problems in **accessibility**
      - **Loss of information**
      - **Waste of resources** if earlier research is forgotten

WHERE ARE WE COMING FROM?

- **Desideratum 1**: make the field of historical corpus linguistics **more cumulative** by ensuring that researchers have **continuous access to the results of previous research** as well as to the **numerical data** reported in the articles.

  … and if achieved…

- **Desideratum 2**: advance the field by making it possible for researchers to undertake e.g. **meta-analyses** and **replication studies** based on existing data.

UNIVERSITY OF HELSINKI

WHERE ARE WE COMING FROM?

- Meta-analyses could uncover new and interesting information about the following questions:

  1. Are **some processes of change** more susceptible to **rapid developments** than others?
  2. Does **the rate of language** change correlate with **the type of community** in which the language is spoken?
  3. How do **changes in society, culture and contact situations** between speakers of different languages and dialects affect language change?

UNIVERSITY OF HELSINKI

WHERE ARE WE COMING FROM?

- Our solution:
  - **Language Change Database (LCD)**

  - An **open-access research database** that provides **real-time baseline data** for **modelling language change** in progress

  - **Summarises** the results of corpus-based research articles on variation and change in English

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.
- c. 300 entries at the moment.
- Currently in beta stage, published later in 2018.

UNIVERSITY OF HELSINKI

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

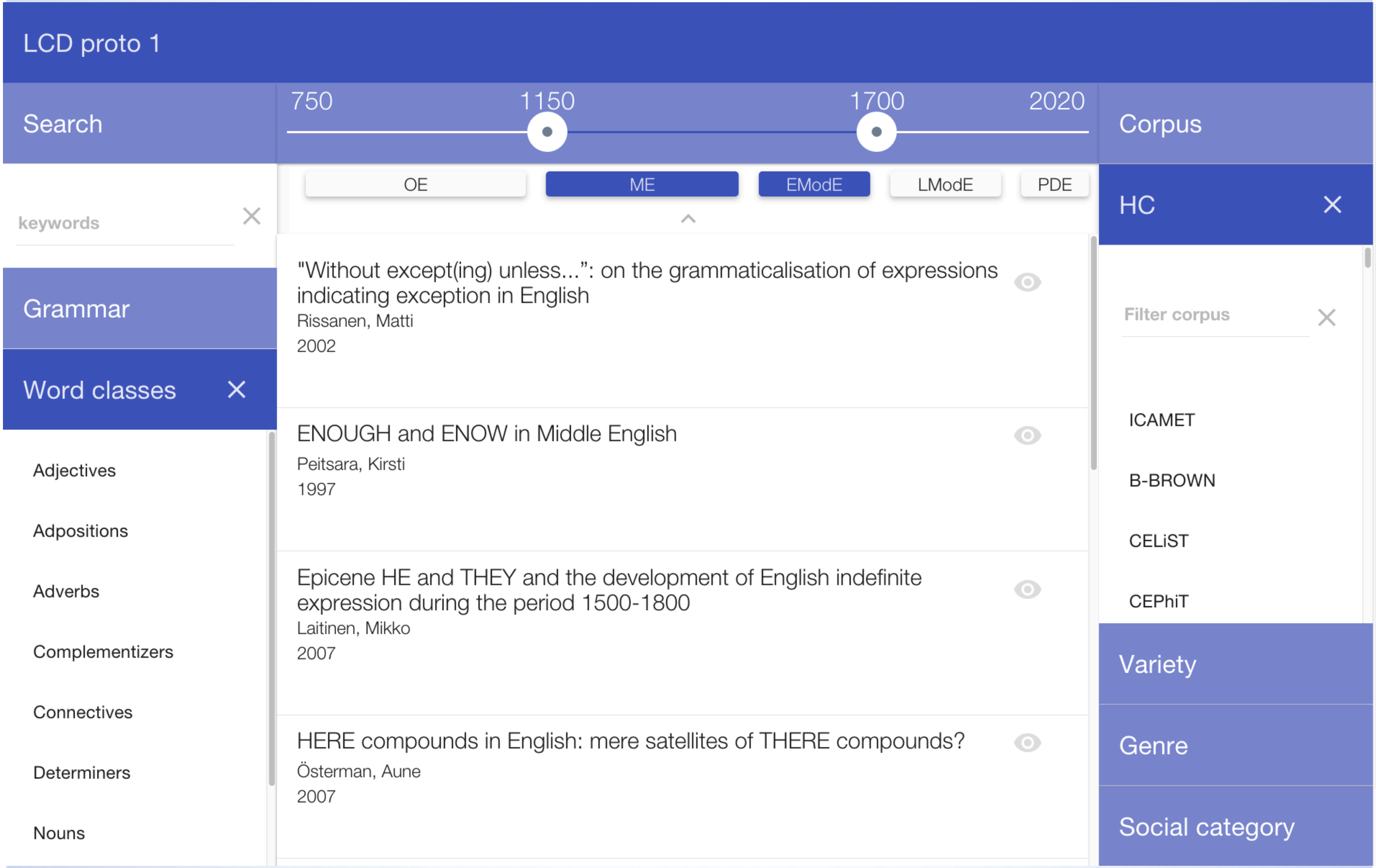
Annotated data file

Data file

Publication file

UNIVERSITY OF HELSINKI

SEARCH INTERFACE



LCD proto 1

Search

keywords ×

Grammar

Word classes ×

- Adjectives
- Adpositions
- Adverbs
- Complementizers
- Connectives
- Determiners
- Nouns

750 1150 1700 2020

OE ME EModE LModE PDE

"Without except(ing) unless...": on the grammaticalisation of expressions indicating exception in English
Rissanen, Matti
2002

ENOUGH and ENOW in Middle English
Peitsara, Kirsti
1997

Epicene HE and THEY and the development of English indefinite expression during the period 1500-1800
Laitinen, Mikko
2007

HERE compounds in English: mere satellites of THERE compounds?
Österman, Aune
2007

Corpus

HC ×

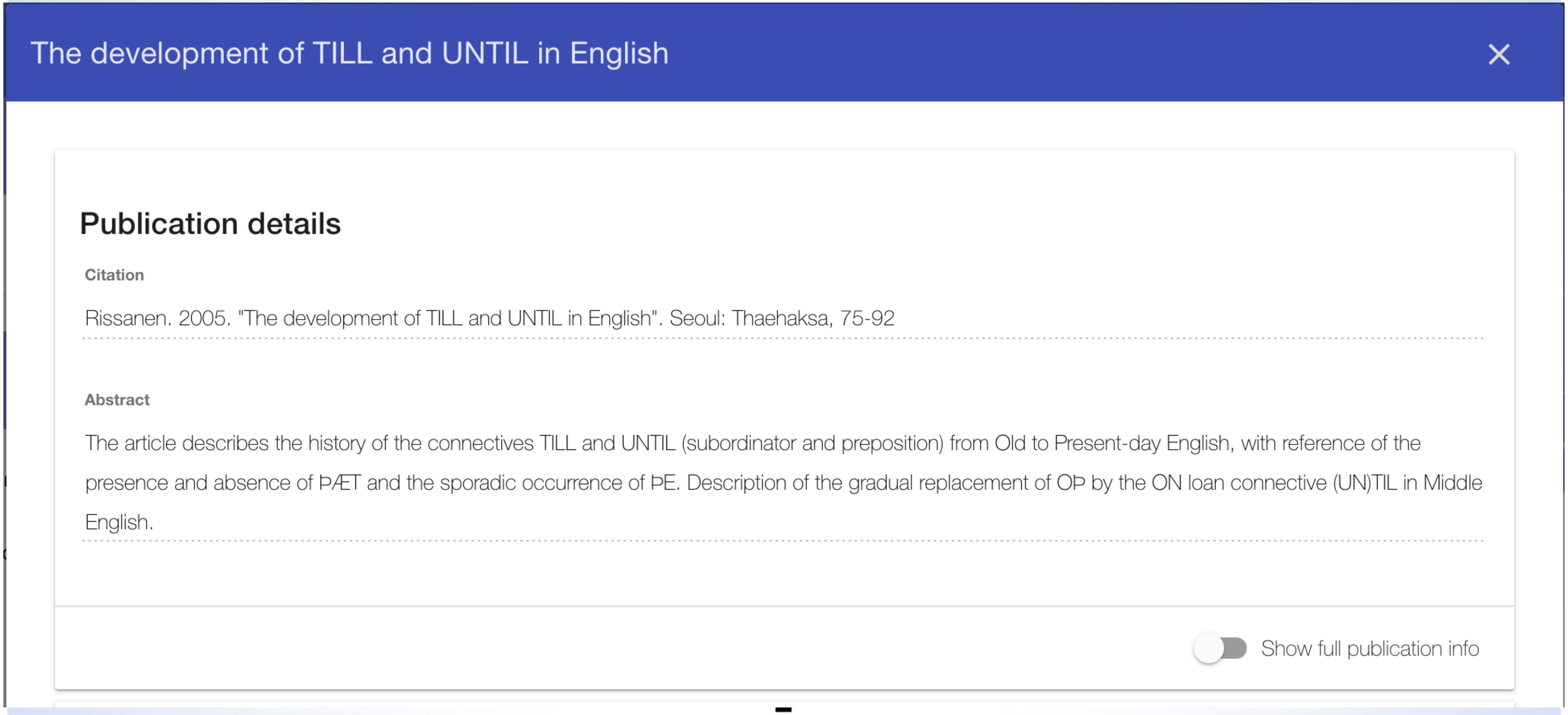
Filter corpus ×

- ICAMET
- B-BROWN
- CELiST
- CEPhiT

Variety

Genre

Social category



The development of TILL and UNTIL in English

Publication details

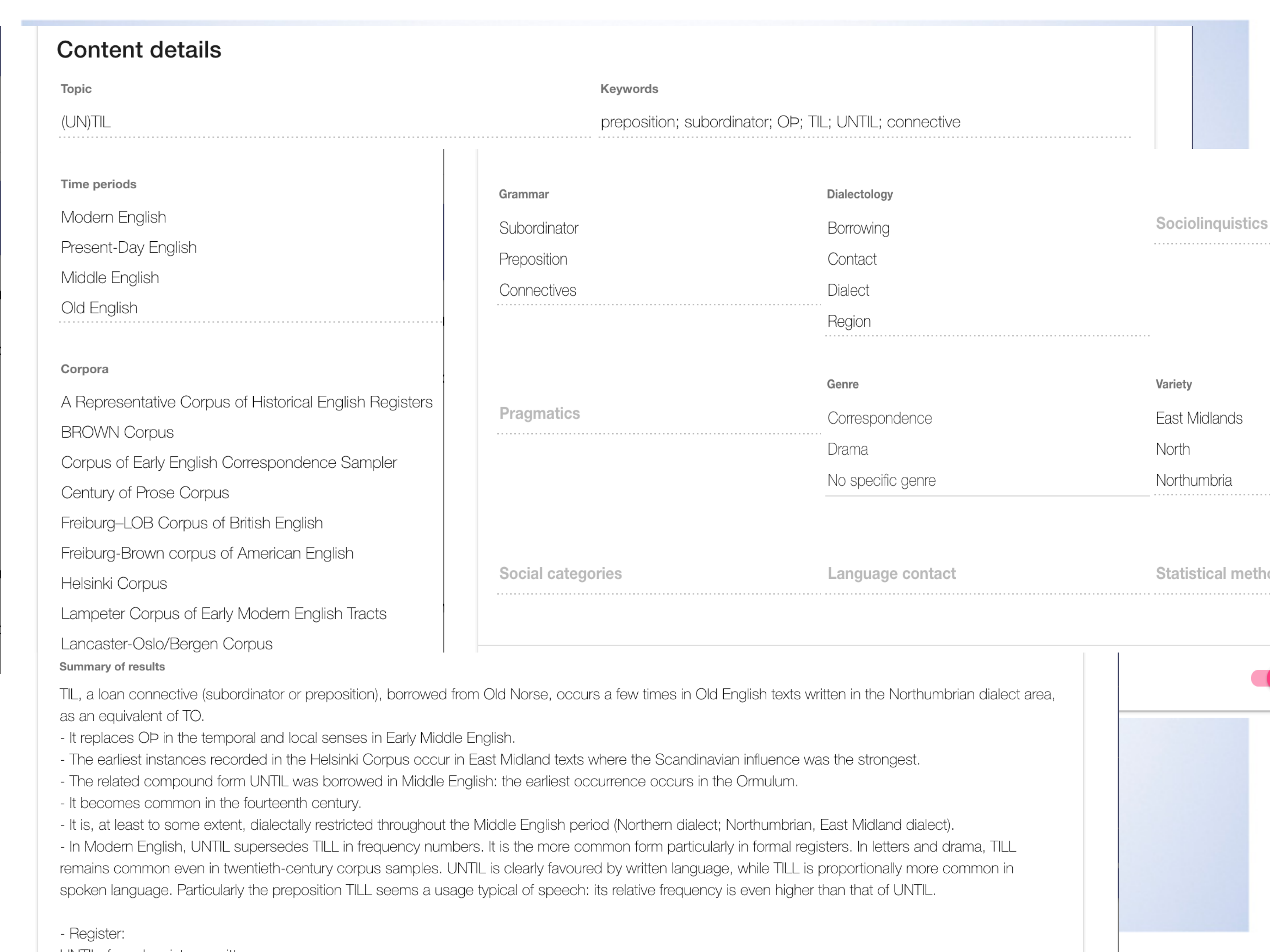
Citation

Rissanen. 2005. "The development of TILL and UNTIL in English". Seoul: Thaehaksa, 75-92

Abstract

The article describes the history of the connectives TILL and UNTIL (subordinator and preposition) from Old to Present-day English, with reference of the presence and absence of ÞÆT and the sporadic occurrence of ÞE. Description of the gradual replacement of OÞ by the ON loan connective (UN)TIL in Middle English.

UNIVERSITY OF HELSINKI



Content details

Topic

(UN)TIL

Keywords

preposition; subordinator; OÞ; TIL; UNTIL; connective

Time periods

Modern English

Present-Day English

Middle English

Old English

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

Grammar

Subordinator

Preposition

Connectives

Dialectology

Borrowing

Contact

Dialect

Region

Sociolinquistics

Pragmatics

Genre

Correspondence

Drama

No specific genre

Variety

East Midlands

North

Northumbria

Social categories

Language contact

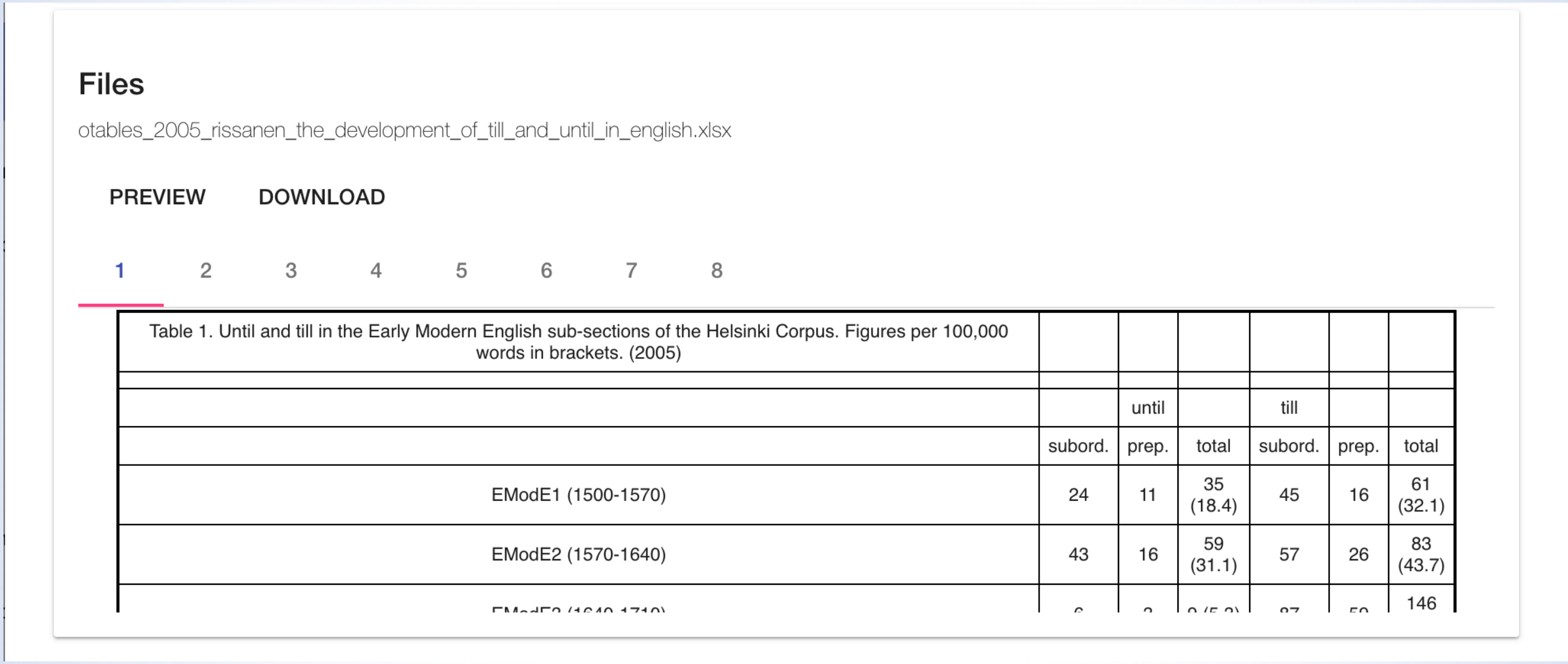
Statistical meth

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

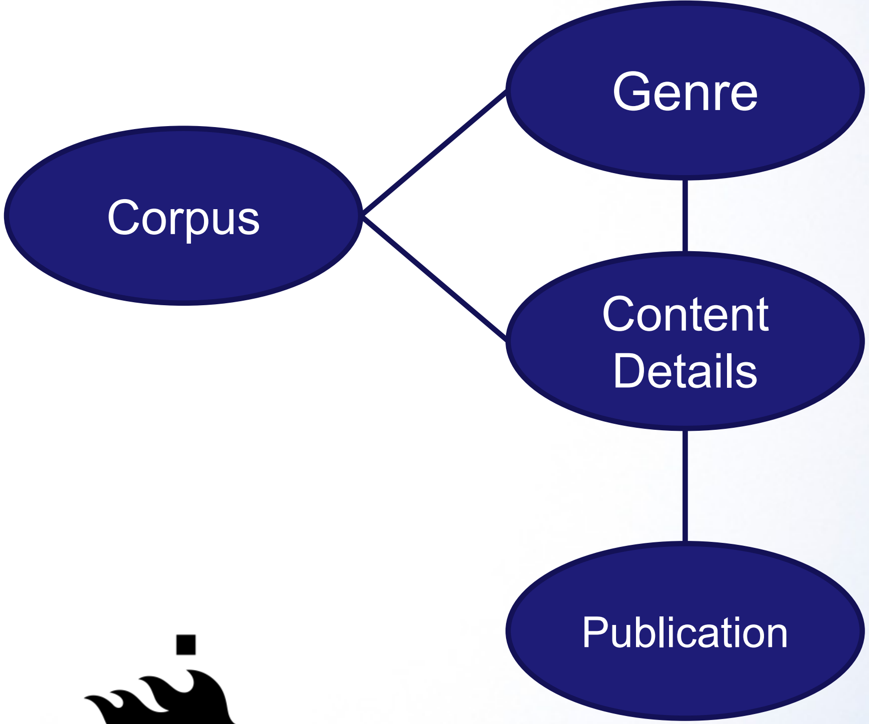
otables_2005_rissanen_the_development_of_till_and_until_in_english.xlsx

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

UNIVERSITY OF HELSINKI

LCD
RE-USABLE DATA

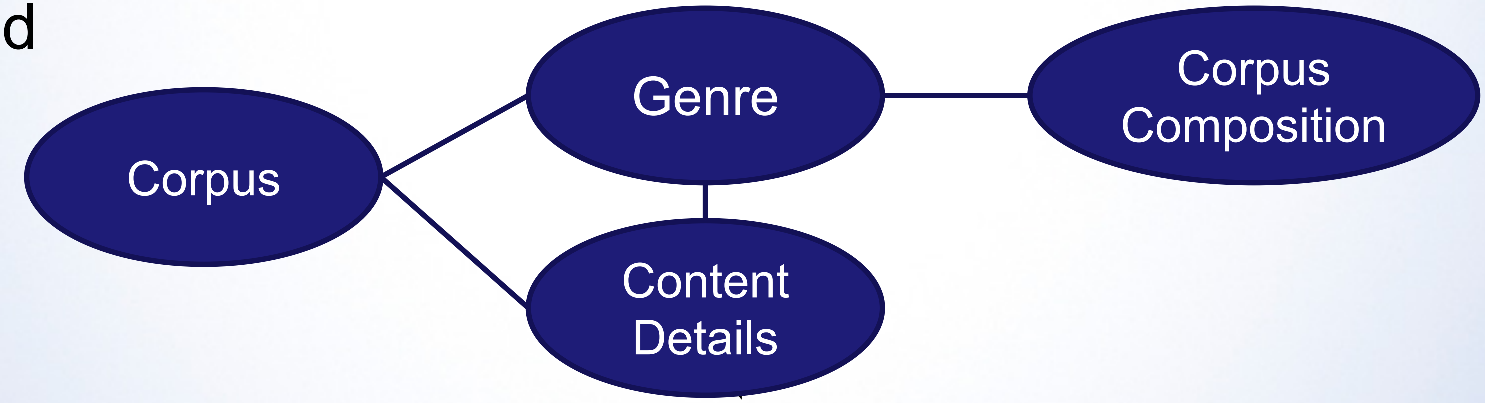
- Facilitates comparing and combining data for **meta-analysis**
- Everything (almost) is stored as **Linked Data**
  - Data
  - Metadata
  - Provenance
  - Ontology
- Re-use through **API**s
  - e.g. Search user interface



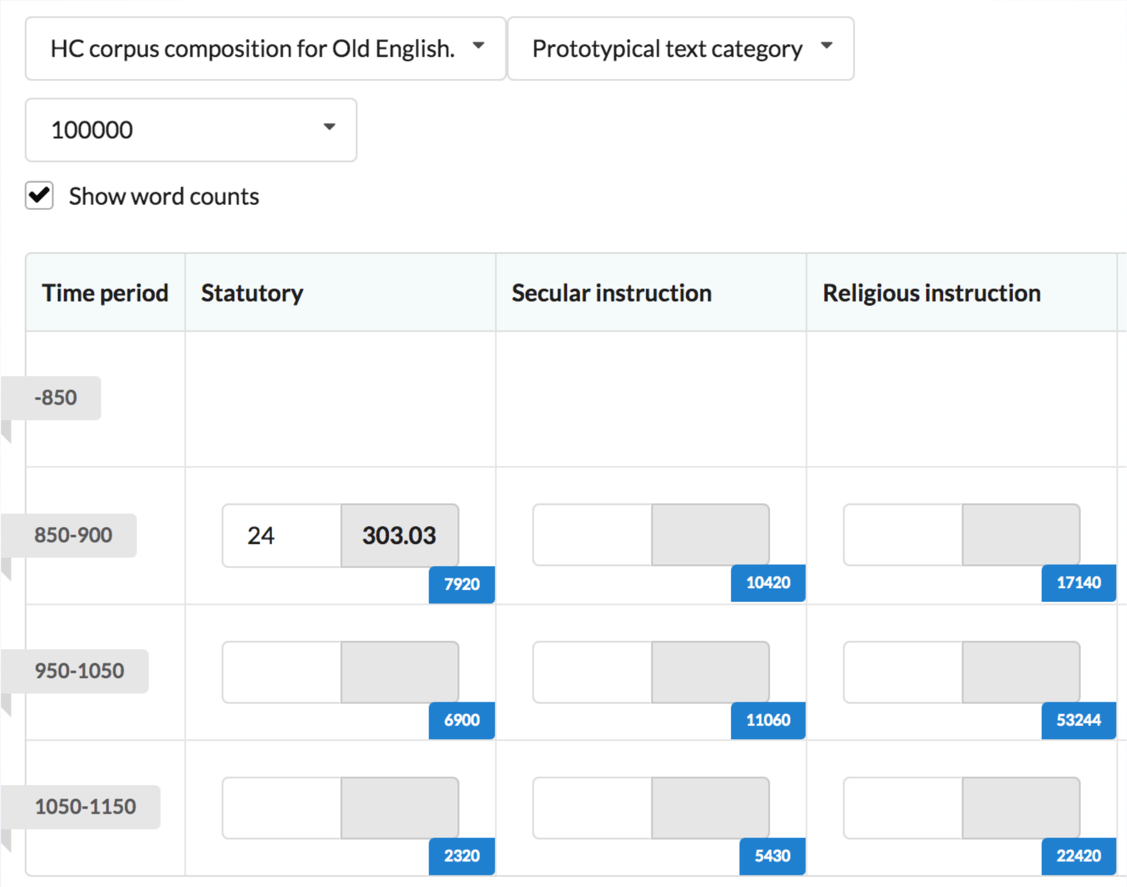
UNIVERSITY OF HELSINKI

LCD
CORPUS COMPOSITION

- Problem: **comparability** across studies
  - Some articles only report absolute values
  - Corpus compositions can be used for simple **normalization**
- Corpus **word count** partitions expressed as Linked Data
- Links corpus parts to LCD concepts such as genre and time period



CORPUS COMPOSITIONS: NORMALIZATION WIDGET



HC corpus composition for Old English. ▾

Prototypical text category ▾

100000 ▾

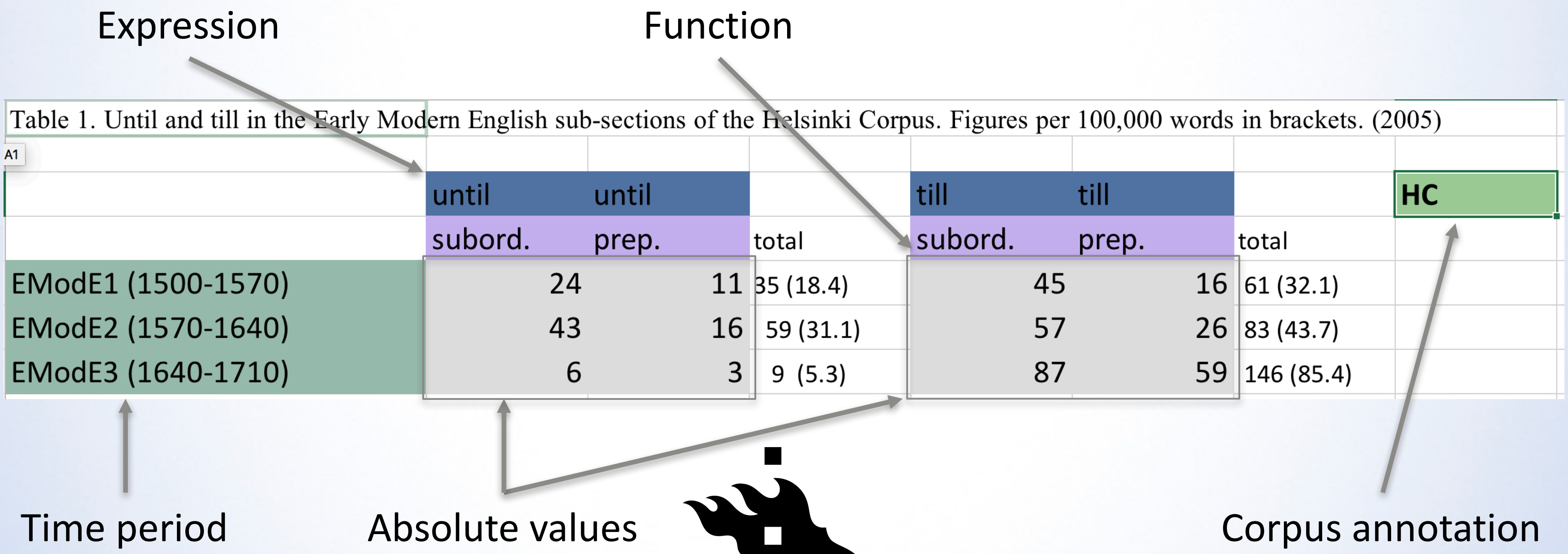
☑ Show word counts

Time period	Statutory	Secular instruction	Religious instruction
-850			
850-900	24 303.03 7920	10420	17140
950-1050	6900	11060	53244
1050-1150	2320	5430	22420

UNIVERSITY OF HELSINKI

LCD DATA TABLE ANNOTATIONS

- Implemented as **Excel styles**
- Links spreadsheet cells to the LCD concepts and ”Things”



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

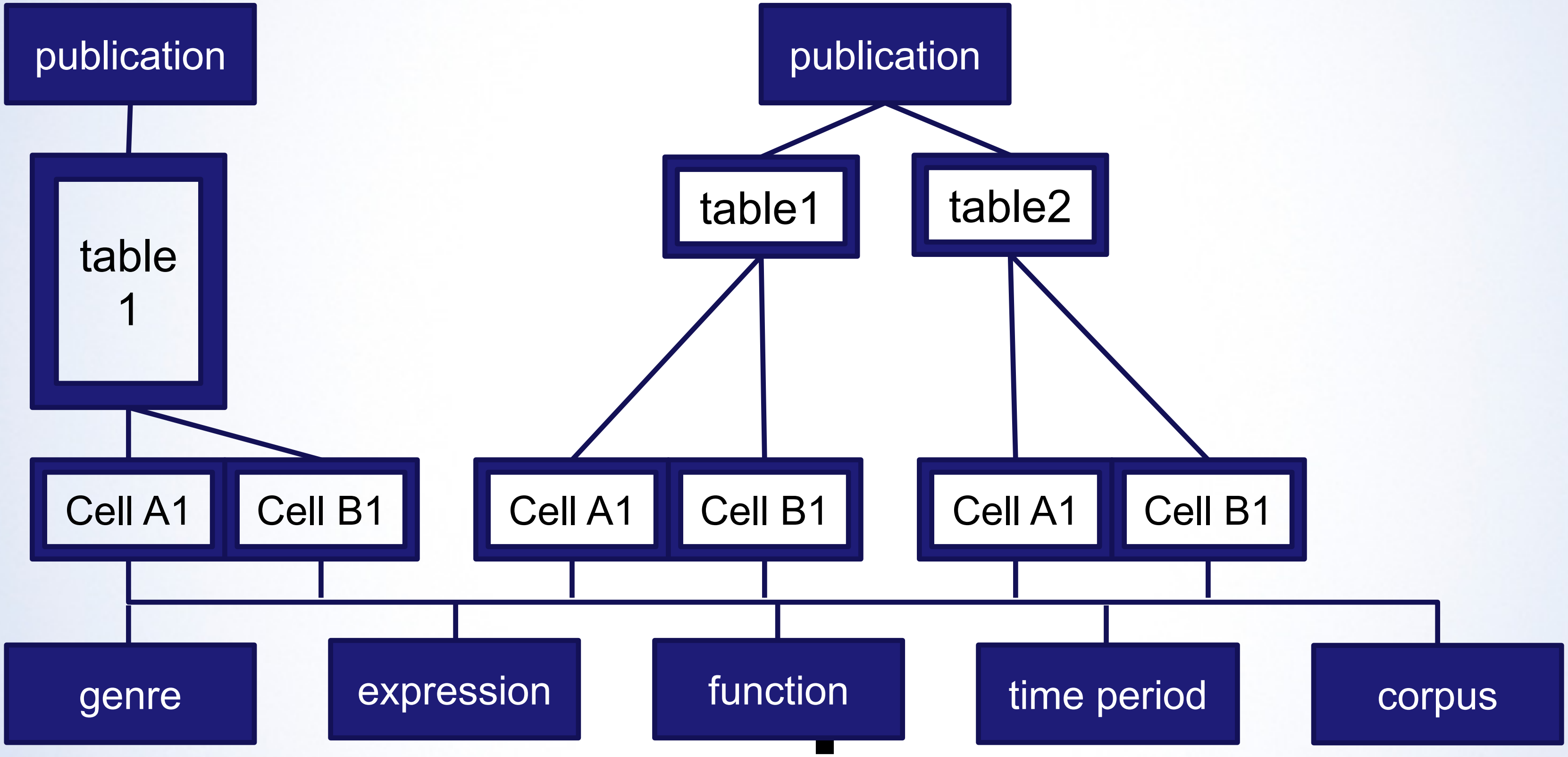
	until	until		till	till		HC
	subord.	prep.	total	subord.	prep.	total	
EModE1 (1500-1570)	24	11	35 (18.4)	45	16	61 (32.1)	
EModE2 (1570-1640)	43	16	59 (31.1)	57	26	83 (43.7)	
EModE3 (1640-1710)	6	3	9 (5.3)	87	59	146 (85.4)	

Time period

Absolute values

Corpus annotation

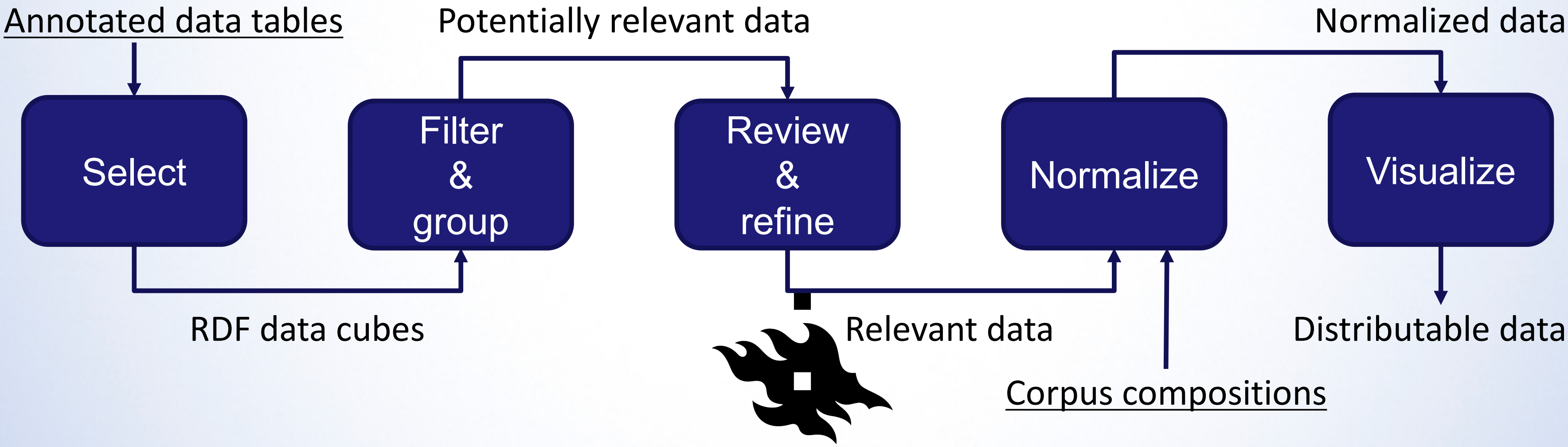
ANNOTATED DATA TABLE



UNIVERSITY OF HELSINKI

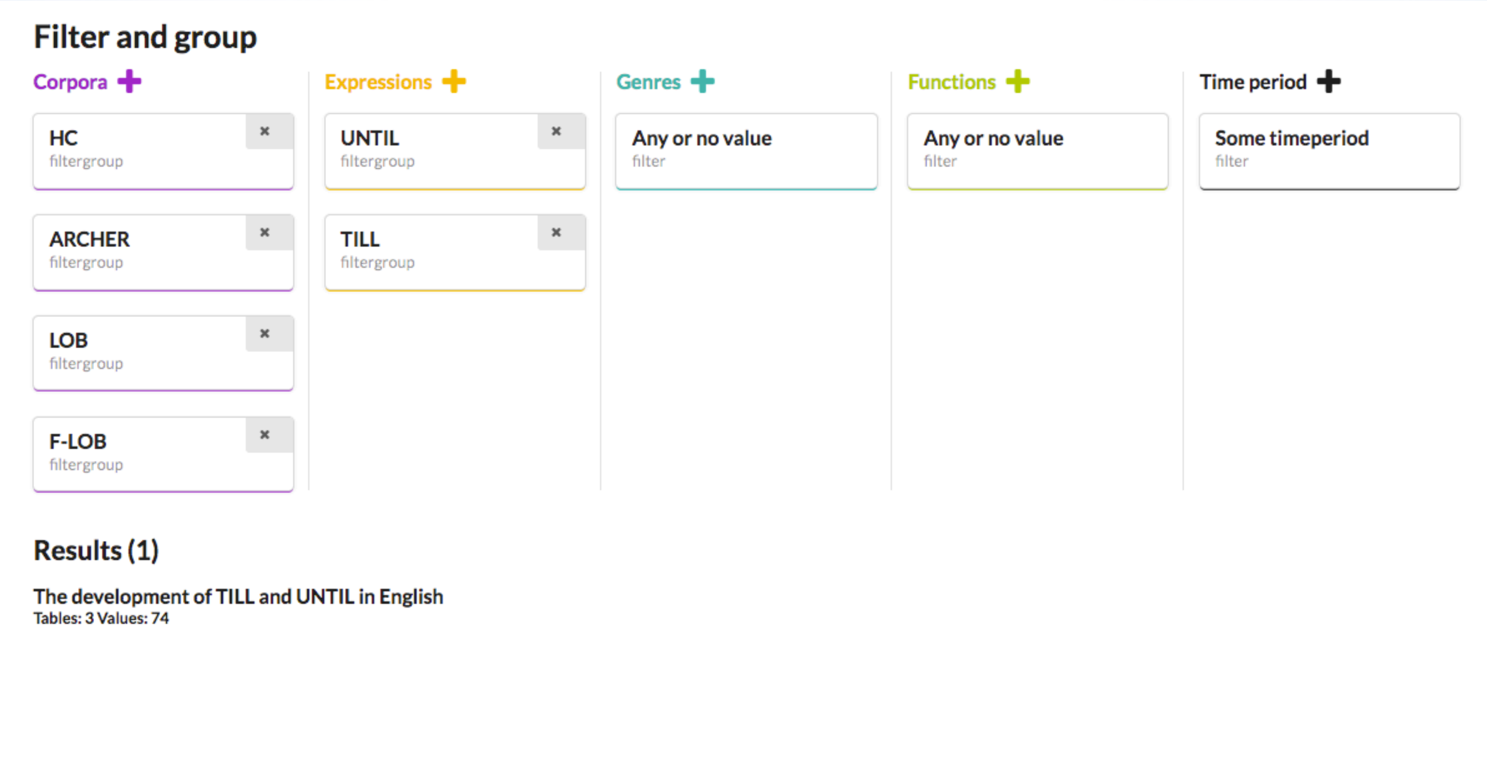
LADA

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



UNIVERSITY OF HELSINKI

LADA
FILTER & GROUP



Filter and group

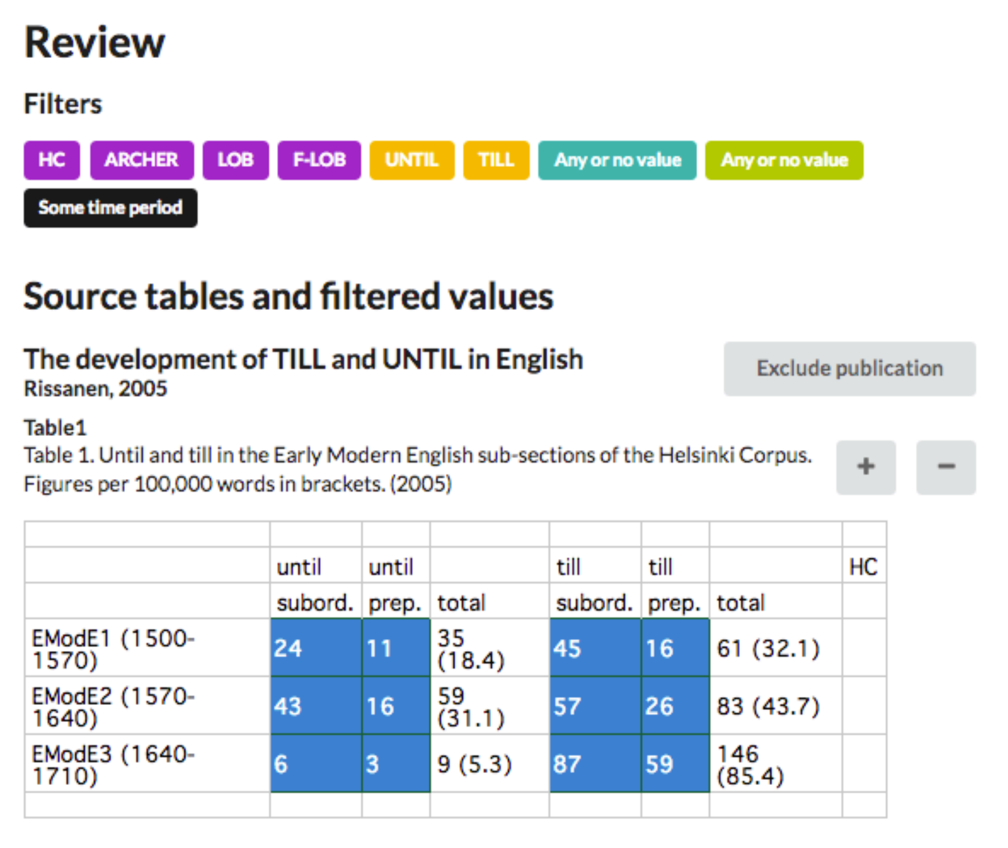
Corpora	Expressions	Genres	Functions	Time period
HC filtergroup	UNTIL filtergroup	Any or no value filter	Any or no value filter	Some timeperiod filter
ARCHER filtergroup	TILL filtergroup			
LOB filtergroup				
F-LOB filtergroup				

Results (1)

The development of TILL and UNTIL in English
Tables: 3 Values: 74

UNIVERSITY OF HELSINKI

LADA
REVIEW & REFINE



Review

Filters

Source tables and filtered values

The development of TILL and UNTIL in English

Rissanen, 2005

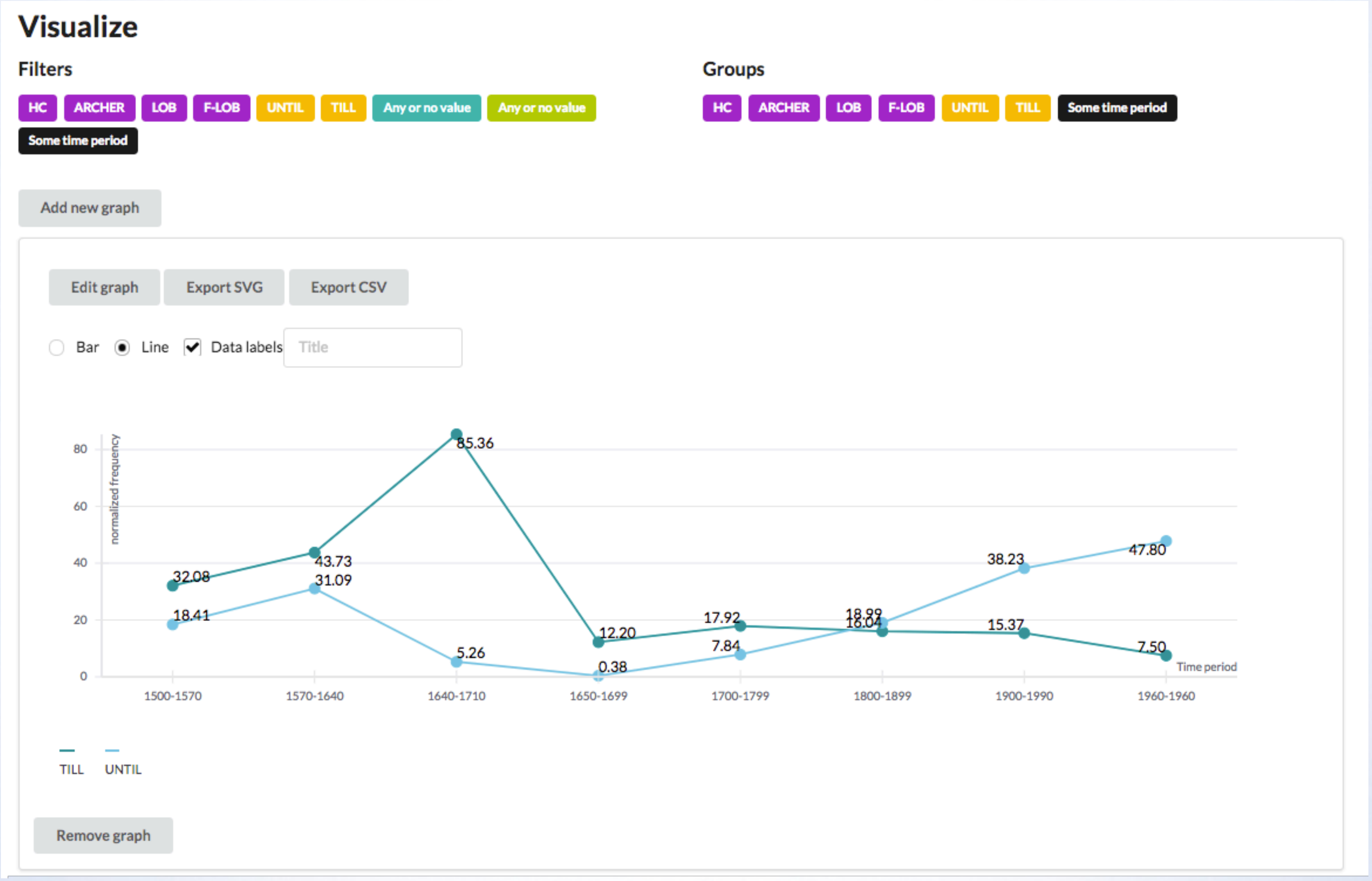
Table1

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

	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)	

UNIVERSITY OF HELSINKI

LADA VISUALIZE



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

○ Bar ● Line ☑ Data labels Title

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**.

UNIVERSITY OF HELSINKI

THANK YOU!

- To learn more about our project, please visit:
  - http://www.helsinki.fi/lcd/

UNIVERSITY OF HELSINKI