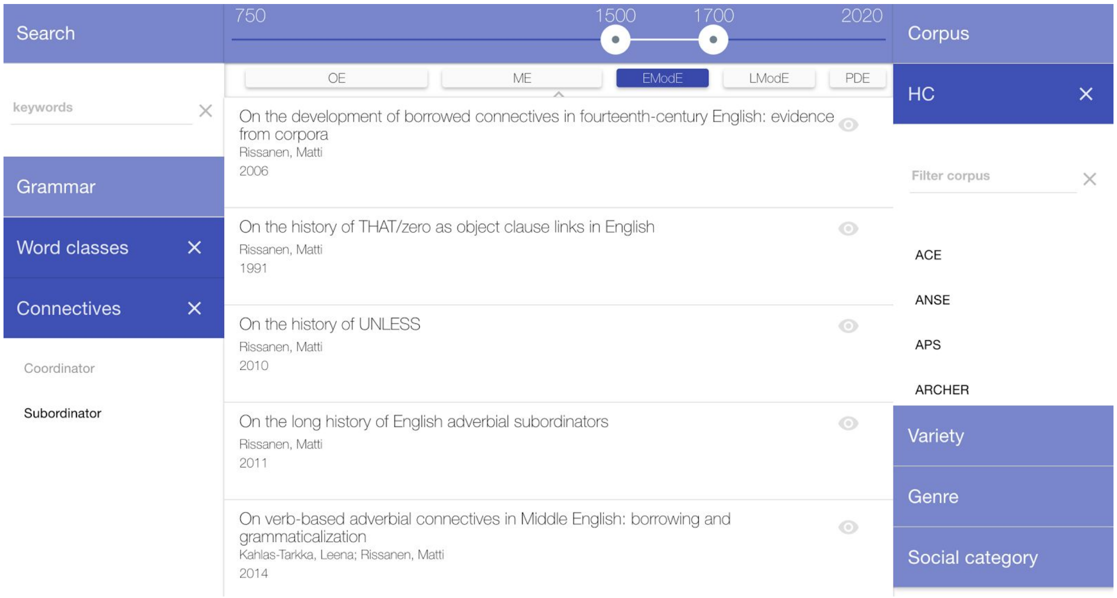
From linked open data to meta-analysis in historical linguistics (Presentation)

Joonas Kesäniemi, Turo Vartiainen, Tanja Säily, Agata Dominowska, Aatu Liimatta, Terttu Nevalainen

Empirical work on language change is fragmented, and some of the published works in the field are hard to come by. Furthermore, the data on which the research is based is seldom available to the research community. The **Language Change Database** (LCD) is an online resource which draws together earlier corpus-based research on English historical linguistics, with the goal of making it more accessible and cumulative by providing comparative baseline data from earlier studies (Nevalainen et al. 2016).

The information included in the LCD goes far beyond basic bibliographies in that it includes detailed descriptions of the results of the studies along with quantitative data in tabular format. In addition to historical linguistics in general, the data in the LCD is tailored to the needs of researchers interested in statistical modelling, systematic reviews, replication of earlier research and sociolinguistic typologies. The most recent additions to the data model, namely corpus composition data and annotated data tables, also make the LCD a valuable source of structured data for the purposes of meta-analysis.

The LCD data is published with permissive licensing and distributed via open interfaces, which makes it possible to integrate it to existing services or build new applications on top of it. For example, our search application prototype (**Figure 1**) takes advantage of the LCD's classification schemes and hierarchical grammar concepts, allowing the user to focus in on specific kind of content in the database.



On the development of borrowed connectives in fourteenth-century English: evidence from corpora
Rissanen, Matti
2006

On the history of THAT/zero as object clause links in English
Rissanen, Matti
1991

On the history of UNLESS
Rissanen, Matti
2010

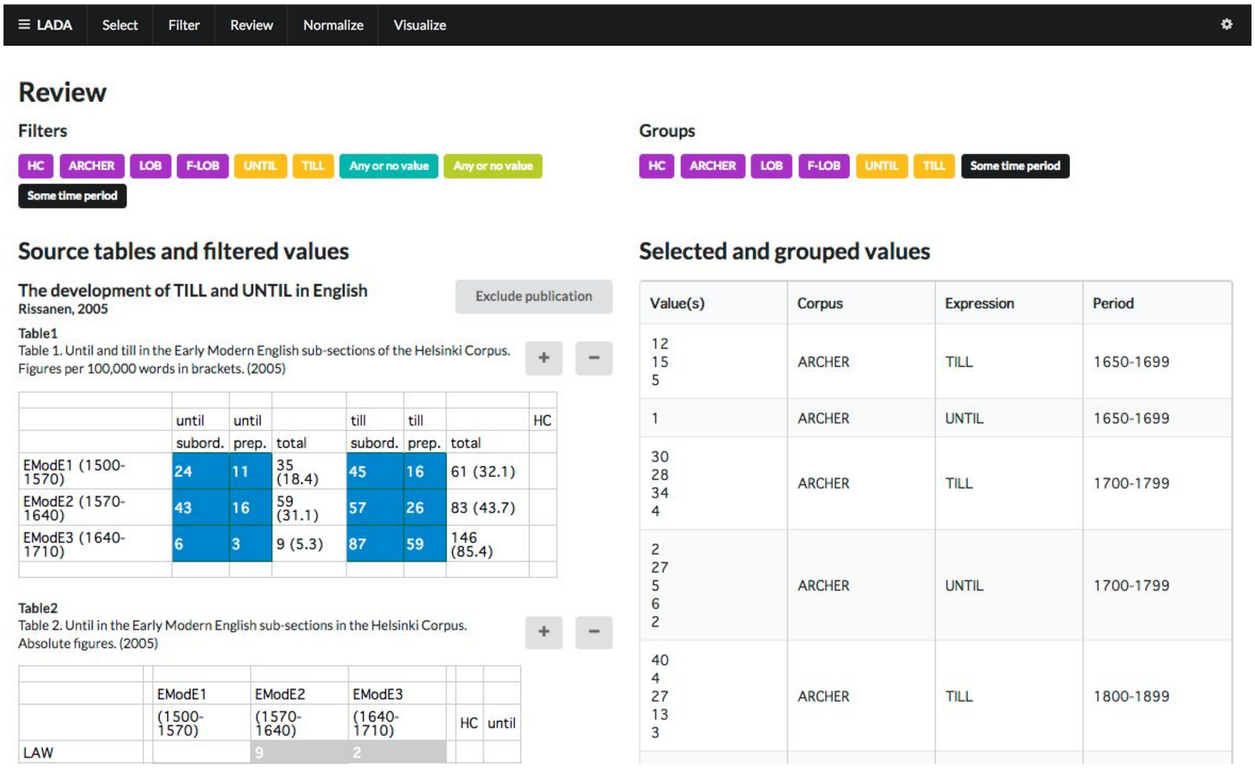
On the long history of English adverbial subordinators
Rissanen, Matti
2011

On verb-based adverbial connectives in Middle English: borrowing and grammaticalization
Kahlas-Tarkka, Leena; Rissanen, Matti
2014

Figure 1. Search interface prototype.

The data tables included in the LCD are currently being annotated to make the different kinds of information in them machine-readable. The annotated tables can be used by the **LCD**

Aggregated Data Analysis workbench (LADA), an application which provides researchers with a systematic workflow to perform exploratory meta-analyses based on earlier research results. The LADA workflow takes as its input a set of LCD publications collected via the aforementioned search application and deemed relevant to the research question at hand. This initial set of data is then filtered, reviewed (**Figure 2**) and normalized in order to create a new aggregated dataset, which can then be visualised or exported as raw data.



Review

Filters

Groups

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)	

Table2
Table 2. Until in the Early Modern English sub-sections in the Helsinki Corpus. Absolute figures. (2005)

		EModE1	EModE2	EModE3			
		(1500-1570)	(1570-1640)	(1640-1710)		HC	until
LAW			9	2			

Selected and grouped values

Value(s)	Corpus	Expression	Period
12 15 5	ARCHER	TILL	1650-1699
1	ARCHER	UNTIL	1650-1699
30 28 34 4	ARCHER	TILL	1700-1799
2 27 5 6 2	ARCHER	UNTIL	1700-1799
40 4 27 13 3	ARCHER	TILL	1800-1899

Figure 2. LADA review stage. The left-hand column illustrates the filtered tables and individual cells, while the right-hand column shows the new aggregated data table with user-defined groupings and dimensions.

All the original data as well as the datasets that are generated in the different stages of the workflow are stored and interlinked to create comprehensive provenance information about the experiment. The whole experiment can be exported and re-run by other researchers, for example, to validate, extend or comment on the results.

The LCD aims to be a collaborative, trusted and accumulative source of linked research data, and the LADA meta-analysis tool is an example of how this innovative approach to data-driven aggregation of empirical findings can be used in the context of historical linguistics. The LADA tool is currently at an early prototype stage, but we hope to expand it with more analytical functionality in the future.

Reference

Nevalainen, Terttu, Turo Vartiainen, Tanja Säily, Joonas Kesäniemi, Agata Dominowska and Emily Öhman. 2016. Language Change Database: A new online resource. *ICAME Journal* 40: 77–94. doi:10.1515/icame-2016-0006