HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
HUMANISTINEN TIEDEKUNTA
HUMANISTISKA FAKULTETEN
FACULTY OF ARTS

MAKING NEW USE OF OLD RESEARCH

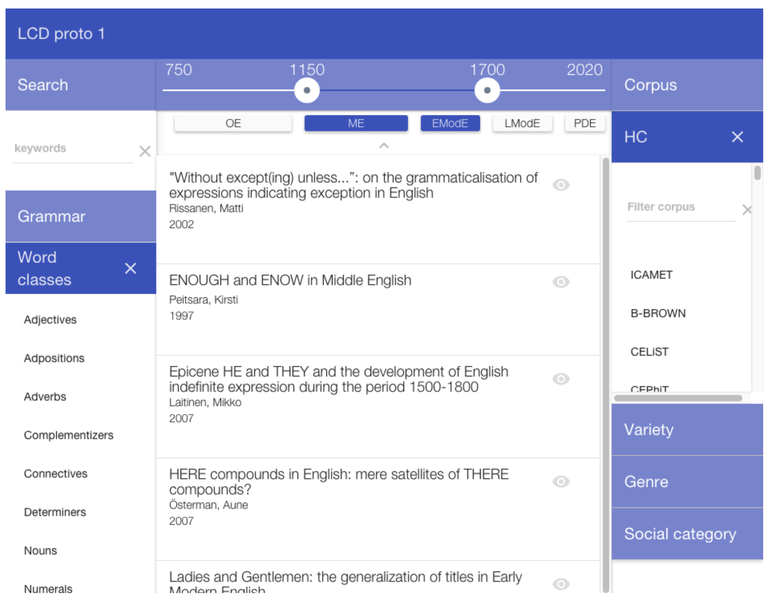
the Language Change Database

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

DISCOVER

The Language Change Database (LCD) draws together earlier **corpus-based research on English historical linguistics** and makes the data available (licenses permitting) through programmable interfaces (APIs) for anyone to utilize in their own applications or analyses.

Search interface



LCD's default search application currently allows users to **find relevant research articles** by making simple keyword searches, filtering results using facets such as corpus, variety and genre, and utilizing the hierarchically structured grammar terms to drill down to more specific terms.

The web application will provide a "shopping cart" type of functionality where users can create their own list of potentially interesting research data and finally **"checkout" related data** in a single package.

MAINTAIN

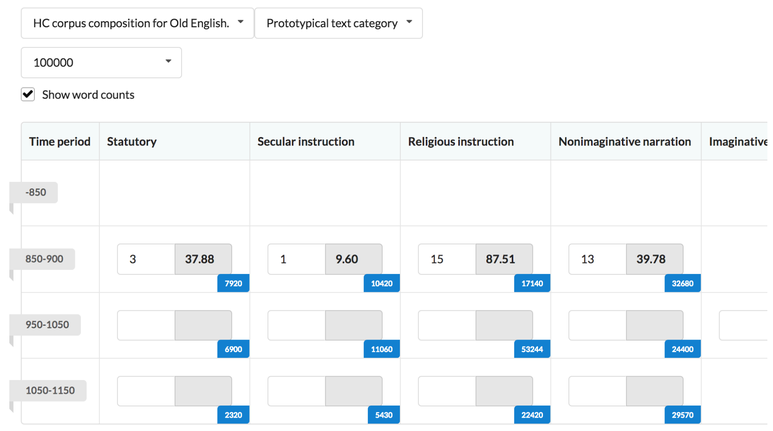
LCD data is administrated using a dedicated **web application**, which is visible to everyone, but where modifications require a user account. All input data is tagged with status (Draft, Under review, Final), and more detailed comments can be attached to any piece of data through annotations.

The goal is to make the maintenance of LCD a **collaborative** effort, where researchers can add, update and comment on LCD's content.

RE-USE

Normalization widget

Using the corpus composition files that describe the corpus structure we have created a simple web application that can be used to **normalize absolute frequencies of linguistic items** reported in previous research or observed by the user. The example on the right allows users to normalize values based on prototypical text category and text type within a time period in the Helsinki Corpus. The corpus structure data available in the LCD has chiefly been extracted from the Corpus Resource Database (CoRD).

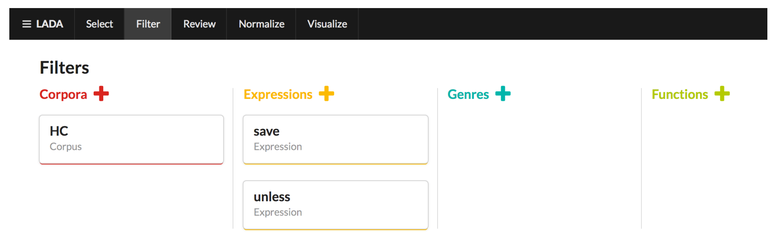


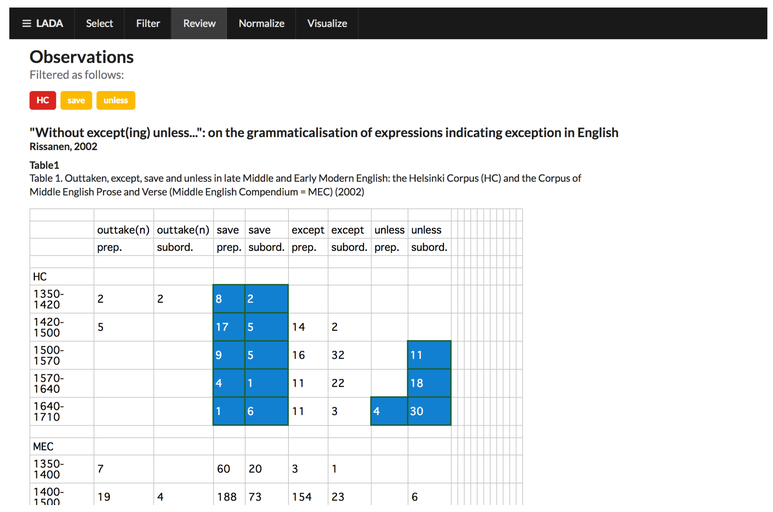
LADA

LCD Aggregated Data Analysis workbench

LADA is tool that facilitates **meta-analyses of previous research** using annotated versions of **Excel data files downloaded from the LCD**. In contrast to LCD, which is a centralized service run by VARIENG, LADA is a standalone Python software that can be downloaded and installed on the researcher's own computer.

LADA first translates spreadsheets into **RDF** graph data structure that can be combined and queried. The user can then reduce this dataset into potentially relevant values by **filtering** it based on certain corpora, genres and expressions. After **reviewing** the resulting dataset and possibly manually excluding further values, LADA uses corpus composition data to **normalize** all the values to a common base. Finally, the user can create a configurable **visualization** based on the filtered and normalized dataset. This new dataset can be then exported and **shared** with another researcher, who can reproduce all the steps on his/her own computer.





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