Demo

LCD aggregation and data-analysis workbench (LADA)

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

Language Change Database (LCD) is an open resource for researchers of English historical linguistics that contains information on previously published articles in the field along with their quantitative results in a tabular format. The latest additions to the LCD 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**. We have developed a tool called LCD Aggregation and Data-Analysis workbench (LADA), which takes advantage of these new data and aims to make it easier for researchers to test and experiment with their research questions using meta-analytic methods.

Corpus compositions are data structures that describe the distribution of word frequencies in a corpus along dimensions such as time period, genre and variety. Many LCD publications report only absolute figures for their findings, and compositions can be used to normalise those values.

In our demo we will first show how to create an annotated version of the data table extracted from an article using Excel (**Figure 1**), and then demonstrate how the new annotation-driven data can be used together with new corpus compositions in LADA to filter, fuse and visualise word frequency data coming from multiple publications (**Figure 2**). LADA transforms annotated tables into RDF Data Cubes, which can be easily queried, manipulated and aggregated using semantic web and linked data tools.

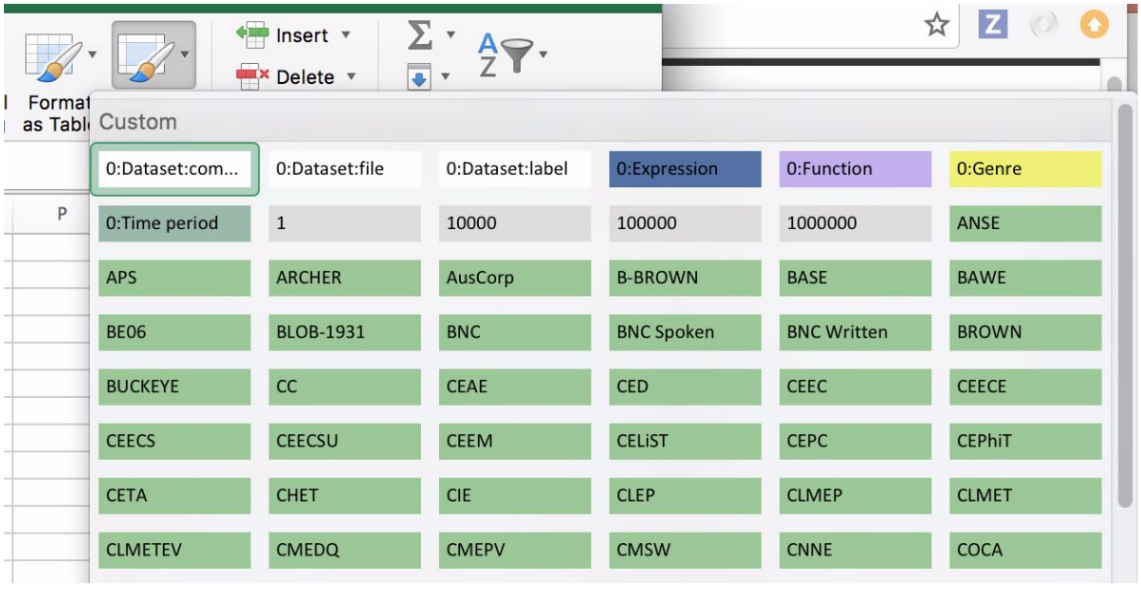

Figure 1. Annotating data tables with Excel styles. Cells can be annotated by Expression, Function, Genre, Time period, Value, and Corpus. The Value annotation (in grey) includes a normalisation base, and the Corpus annotation (in green) links the cell to the corpus entry in the LCD.

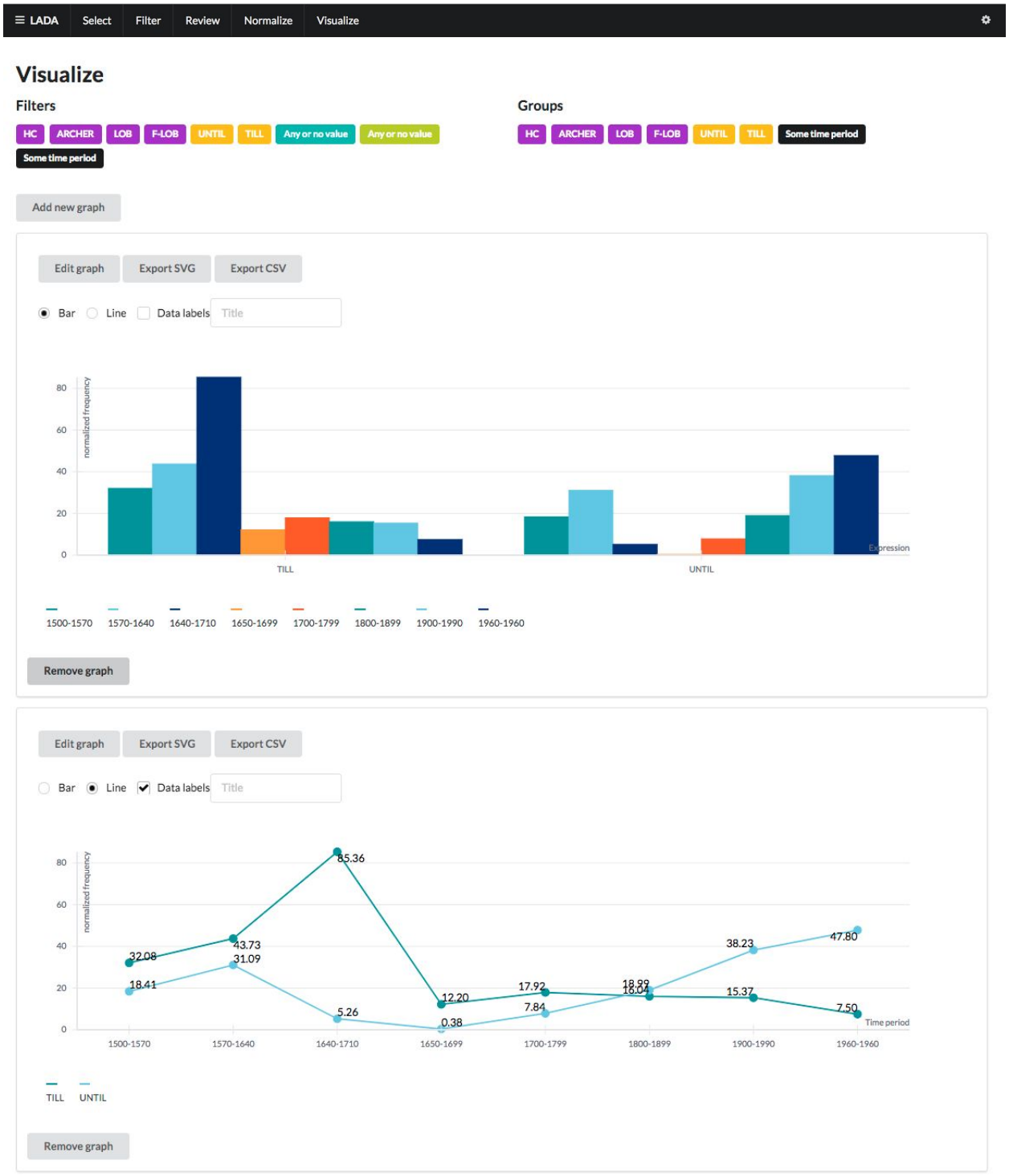

Figure 2. The LADA workflow is divided into five steps: Select, Filter, Review, Normalise, and Visualise (pictured here). Data illustrating the development of the English connectives TILL and UNTIL across multiple corpora over time.