Exploring meta-analysis for historical corpus linguistics based on linked data

Joonas Kesäniemi, Turo Vartiainen, Tanja Säily, Terttu Nevalainen

Note: This is a post-print version of the following paper: Kesäniemi, Joonas, Turo Vartiainen, Tanja Säily and Terttu Nevalainen. 2018. "Exploring meta-analysis for historical corpus linguistics based on linked data". *Journal of Research Design and Statistics in Linguistics and Communication Science* 5(1–2): 4–47. doi:10.1558/jrds.36709, https://journals.equinoxpub.com/JRDS

Abstract

Empirical work on English historical corpus linguistics is plentiful but fragmented, and some of it is hard to come by. This paper proposes a solution for making it more accessible and reusable for meta-analysis. We present an online Language Change Database (LCD), which provides comparative, real-time baseline data from earlier corpus-based studies. LCD entries summarize the findings and include numerical data from the articles. We discuss the LCD from the perspective of database design and linked data management. Furthermore, we illustrate the reuse of LCD data through a meta-analysis of the history of English connectives. For this purpose, we have developed an application called the LCD Aggregated Data Analysis workbench (LADA). We show how researchers can use LADA to filter, refine and visualize LCD data. Thus we are paving the way for a future where both research results and research data are regularly available for verification, validation and re-use.

1. Introduction

Linguistic research has undergone significant changes in the past few decades. The advent of digital corpora has revolutionized the field of empirical linguistics and enabled the investigation of research questions that used to be either impossible or impractical to study. Researchers of English have been at the forefront of corpus-based research, and while the early corpora, such as the Brown (Francis and Kučera 1979) and the LOB (Johansson et al. 1978), represented Present-day English usage, research into the earlier stages of English has been possible from a corpus linguistic perspective on a larger scale since the publication of the *Helsinki Corpus of English Texts* (HC) in 1991. Since then, many other historical corpora have been published, and thousands of articles that make use of quantitative corpus data have been written about the variation and change of the English language. In this paper, we argue that the variety of research questions studied in the past twenty-five years, and the large amounts of data on which the research is based, can now open new avenues of research, just like the introduction of corpora did decades ago. In particular, we suggest that a research database which not only includes detailed information about published articles on the history of English, but also the published data in annotated form, can be used as a basis for the replication of earlier research and for meta-analysis: by making use of existing data, we can explore questions that would otherwise be too labour-intensive to study.

Meta-analyses are commonly performed in other fields, such as medicine, where they are used to obtain a more comprehensive understanding of complex phenomena that have previously been studied on a smaller scale and with different data sets. Typically, meta-analyses synthesize the results of dozens or even hundreds of individual studies. For instance, Reynolds et al. (2003) studied the association between alcohol consumption and risk of stroke by examining 122 reports and 35 observational studies published in 1966–2002. Renehan et al. (2008), on the other hand, studied the association between body mass index and cancer risk based on 221 data sets from 141 research articles, and Wang et al. (2016) based their meta-analysis of coffee and cancer risk on 105 articles selected from a total of 69,495 potential articles. These kinds of large-scale analyses can be used to establish the current state-of-the-art in the field studied, and they also have the potential to result in new and significant findings: the results of meta-analyses are often more than just the sum of their parts. Importantly, it would have been impossible to carry out such analyses without access to the results gained in earlier research as well as to the actual research data.

Similarly to medicine, meta-analyses in linguistics are based on a large body of quantitative results obtained in earlier research. However, not all fields of linguistics have traditionally made use of quantitative data, and even in some of the more quantitatively-oriented fields, such as historical corpus linguistics, the research results may be reported by way of illustrative examples instead of frequency tables. Consequently, meta-analyses in many fields of linguistics have been relatively uncommon. A notable exception is applied linguistics, which has a long tradition in quantitative methods. Here, the beginnings of research synthesis and meta-analysis extend back to the 1970s (Chaudron 2006). The articles in Norris and Ortega (2000) provide examples of meta-analyses on effective teaching practices, effectiveness of corrective feedback, and adult second-language learners' access to Universal Grammar. Durrant (2014), on the other hand, studied the use of corpora in test design by carrying out a meta-analysis based on 19 different tests with 1,568 individual test takers. His goal was to determine the correlation between second-language learners' knowledge of collocations and the corpus frequency of these collocations.

The lack of meta-analyses on language change can in part be explained by the under-reporting of quantitative results, but there is also a more pragmatic reason for it: many of the early corpus-based studies of language change that are still relevant today were published in edited volumes and festschrifts that can be hard to find even in a well-stocked university library. One unfortunate consequence of this lack of access is that individual researchers and research groups may spend valuable time and resources on questions that have already been thoroughly investigated in the past. For instance, in a recent publication Newberry et al. (2017) tested the hypothesis that at least some linguistic changes represent stochastic drift rather than selection. Their data included some well-known and comprehensively studied changes in the history of English, such as the rise of *do*-support and the regularization of past tense verbs. However, the authors made almost no reference to this large body of research, possibly because much of it was not readily available in an electronic form. This unfortunate oversight also had significant consequences on the reliability of their findings: in the case of *do*-support, for example, their data did not include the affirmative declarative *do*, which casts serious doubt on their general conclusions.

Although not an excuse for doing sloppy research, the fact that the information included in research articles is typically not available in a form that would facilitate its re-use is a serious problem both from the perspective of meta-analysis and the accumulation of knowledge. For instance, numerical information in linguistic articles is typically expressed as figures or tables that are printed in accordance with the publisher's stylesheet. Importantly, the data is not available in a machine-readable form, which means that in order to re-use the data in any automated way, the data first need to be copied into a program such as Excel either manually or by using optical character recognition software (OCR). This step is very labour-intensive, and therefore expensive, and every individual researcher, or research group, needs to do this on their own; at present, there are no data repositories that would effectively mitigate the problems associated with data acquisition.[1] Furthermore, even when the data has been transferred to a spreadsheet application or a database management system, the information presented in the tables and figures is typically too complex to allow computer-assisted analysis. For instance, in addition to the frequency of the linguistic item studied, a single number in a table may be simultaneously associated with features like **time period**, **corpus**, **genre** and **grammatical function**. In order for this data to be useful for efficient and automated processing for research purposes, it needs to be annotated and linked to relevant features.

To address these issues, this paper introduces the **Language Change Database (LCD)**, a new linguistic resource currently under compilation, which is designed to facilitate the dissemination, verification and re-use of linguistic research and research data. Moreover, we present a tool that is specifically designed to be used with the data acquired from the LCD for experimentation purposes, **LCD Aggregated Data Analysis Workbench** or **LADA**. This solution is based on the principles of linked data management (Bizer et al. 2009), and it allows the existing data to be combined and queried to study new and original research questions. We focus on the workflow of data processing and management from the perspective of meta-analysis. To illustrate the linguistic motivation behind our multidisciplinary research approach, we will carry out a sample study of the frequency variation of *till* and *until* over a long period of time. The study, which is part of a larger research topic on the history of English connectives, is used as an example throughout the paper to facilitate the understanding of the more technical aspects of our proposed solution.

The rest of the paper is organized as follows. In section 2, we introduce the LCD. The focus of the discussion in section 2.1 is on the general architecture and design of the database, while in sections 2.2–2.4 we elaborate on the benefits of using a model for managing linguistic data based on the principles of linked data. Section 3 describes the additional features that were added to the LCD data model in order to make the data better suited for meta-analysis. Section 4 introduces the LADA tool, while section 5 provides a concrete example of how the LCD and LADA can be used to carry out a small-scale meta-analysis. Sections 6 and 7 conclude the paper with a discussion and some future prospects.

2. Language Change Database (LCD)

2.1. Overview of the LCD

Preservation of knowledge and ease of access are at the heart of the project introduced in this paper. We are currently compiling a database of corpus-based research on the history of the English language, called the **Language Change Database**, or the **LCD** for short (Nevalainen et al. 2016). On the one hand, the LCD will serve as a source of knowledge and a means of disseminating information and research findings within the linguistic community. On the other hand, the numerical data included in the database can be used as a baseline for replication studies, meta-analyses and statistical modelling (e.g. Blythe and Croft 2012). Although at this point we, as the developers of the LCD, have taken on the responsibility of feeding information into the database, once the beta-testing phase has been completed, researchers will be able to insert information on their own research articles into the LCD. By contributing to the LCD, they can also gain visibility for their work and increase the impact of their research: as the LCD will be available online

[1] We are only aware of one repository that is specifically designed for the storage and re-use of linguistic data: the Tromsø Repository of Language and Linguistics (TROLLing, https://dataverse.no/dataverse/trolling). However, as TROLLing includes data from all areas of linguistics, and is very modest in size, it is not well-suited for meta-analyses from the perspective of historical corpus linguistics. All URLs in footnotes retrieved on 15 November 2018.

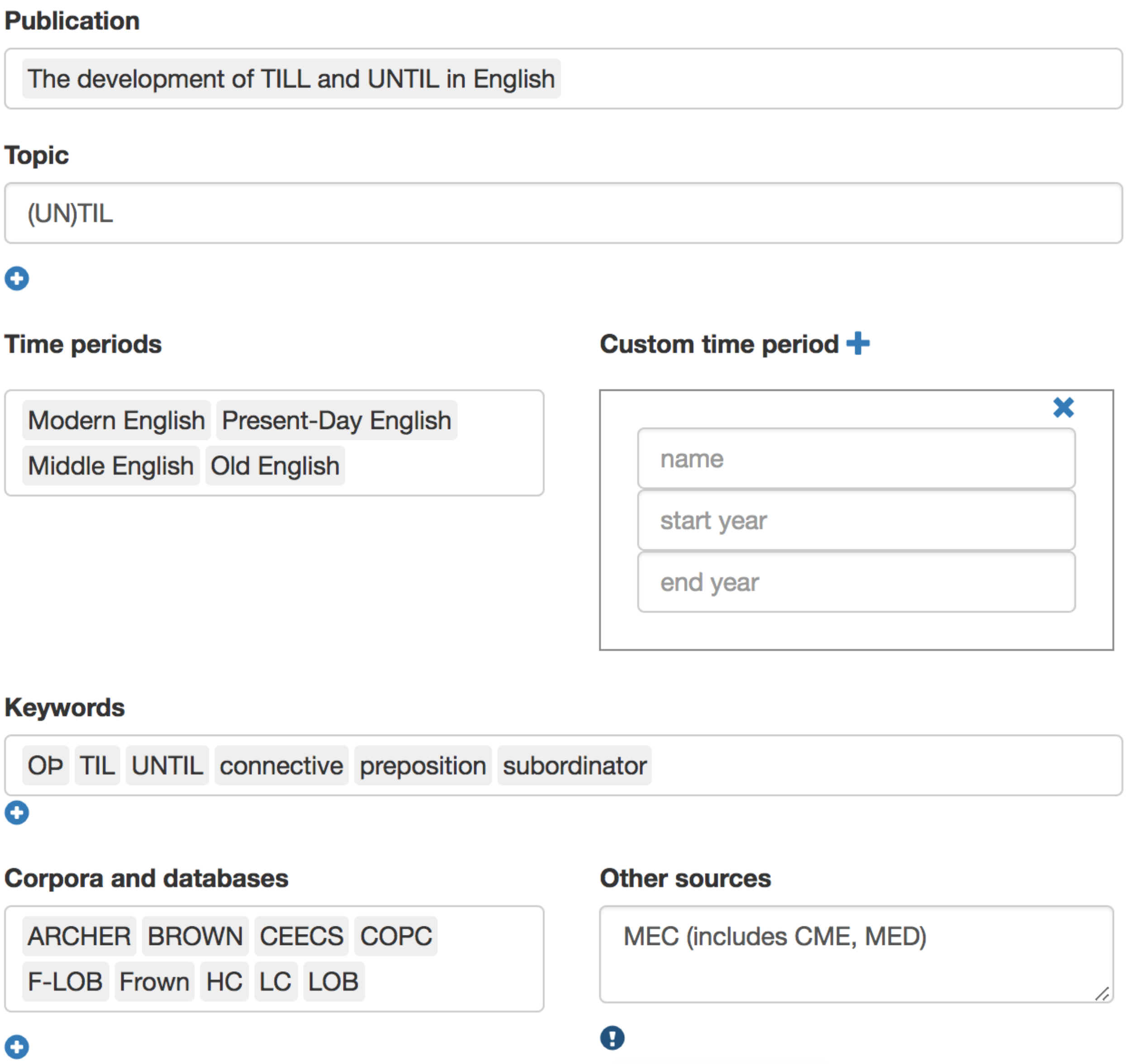
and open access, we envision it to be useful to scholars, teachers and students of English all around the world.

Each research article is represented by one entry in the LCD. Each entry includes the basic **publication details** of the article (title, authors, name of the journal or the volume where it has been published, page numbers, abstract, etc.), and detailed information about the contents of the article. These **content details** include the linguistic items and the time periods studied, the corpora and other databases used as data, as well as information about the variety of English and genres studied, if relevant. Crucially, the main results of the article are summarized as bullet points, which gives the end user of the database an idea of whether the research discussed in the article is relevant to their own purposes. Each article is also annotated for the grammatical categories that are studied in the articles, such as **modal verbs**, **word order**, **grammatical case** or **progressive aspect**. The LCD includes a grammar component which has been designed for the purposes of research on the history of English, but the users of the database can also add new grammatical categories as keywords (see further Nevalainen et al. 2016: 80–82). Finally, if the article discusses language change from a sociolinguistic or pragmatic perspective, the entry can be annotated accordingly by selecting keywords such as **gender**, **age** and **politeness**. We have already included many such keywords in the LCD based on our close reading of hundreds of research articles, but researchers can add new ones according to need.

The basic architecture of the LCD can be divided in two parts: the **front-end** and the **back-end**. The back-end includes tools related to database management and administration as well as the web forms used for the inclusion of new database entries. Much of the back-end will only be available to the administrators of the LCD: researchers willing to create new entries based on their own research will be given restricted access to some of the search functions in the back-end as well as to the tools necessary for making a new entry. Furthermore, all contributions and keyword suggestions will be reviewed by the administrators before the entries are made public.

Figure 1 shows a partial view of the web form, available at the back-end, that is used to feed information into the LCD. All information is inserted manually, but most fields are equipped with an autocomplete function to ensure that the categories that are already stored in the database will not be needlessly included twice.[2] Autocomplete also helps the users of the database to connect a new entry to an author whose work is already included in the LCD. This not only facilitates data insertion but also helps in data management by ensuring that the new data is correctly linked to already existing categories.

[2] We have taken spelling variation into account by having alternate terms for some well-known variants, such as *grammaticalization*, *grammaticalisation* and *grammaticisation*.



Publication

The development of TILL and UNTIL in English

Topic

(UN)TIL

Time periods

Modern English | Present-Day English | Middle English | Old English

Custom time period

name

start year

end year

Keywords

OÞ | TIL | UNTIL | connective | preposition | subordinator

Corpora and databases

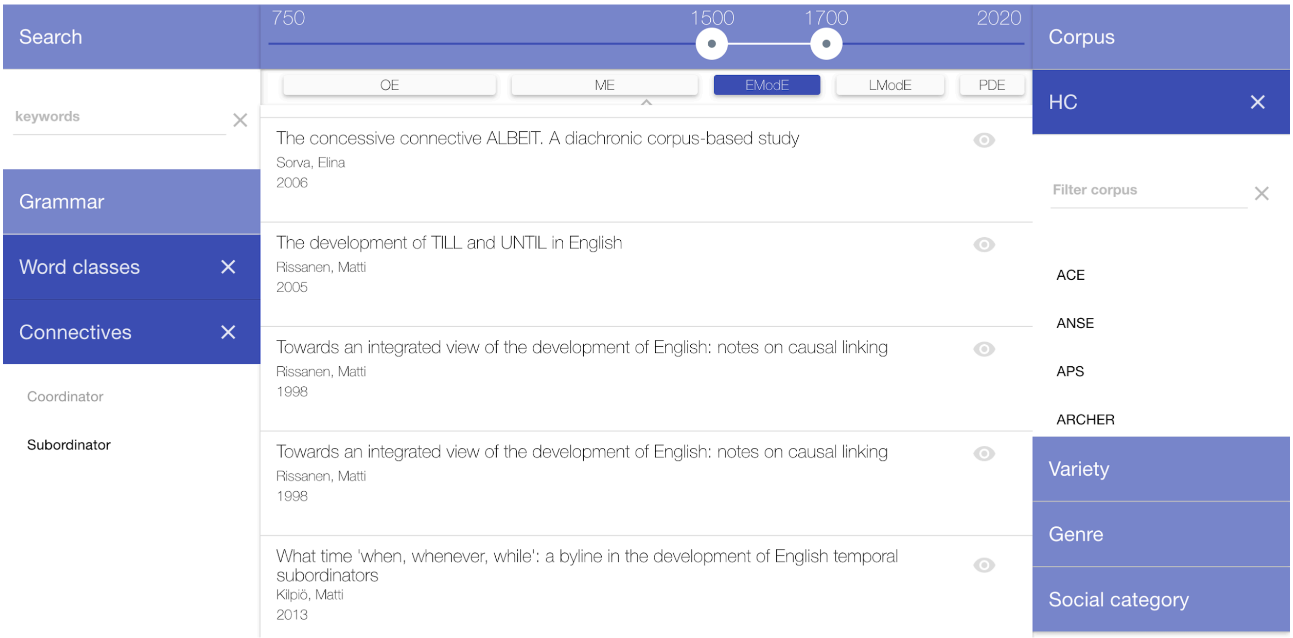
ARCHER | BROWN | CEECS | COPC | F-LOB | Frown | HC | LC | LOB

Other sources

MEC (includes CME, MED)

Figure 1. Web form for inputting a new study (partial view).

The front-end of the LCD consists of an online search tool which is used to query the database (Figure 2). The end users can search the database according to all the features for which the articles have been annotated. This can be done in two ways: i) by performing a simple keyword search, or ii) by using filters (represented as light- and dark-blue boxes in Figure 2). The filters on the left target the grammatical categories explored in the articles, while the ones on the right can be used to query the database according to the corpora used for data, the variety studied, genre-based information and social categories. In Figure 2, the database is searched for research articles that focus on connectives in the early modern period (EModE, 1500–1700), and which draw their data from the *Helsinki Corpus* (HC). The results are presented in the centre column, and by clicking on an entry, the user will be able to read the publication and content details associated with that entry.



The concessive connective ALBEIT. A diachronic corpus-based study
Sorva, Elina
2006

The development of TILL and UNTIL in English
Rissanen, Matti
2005

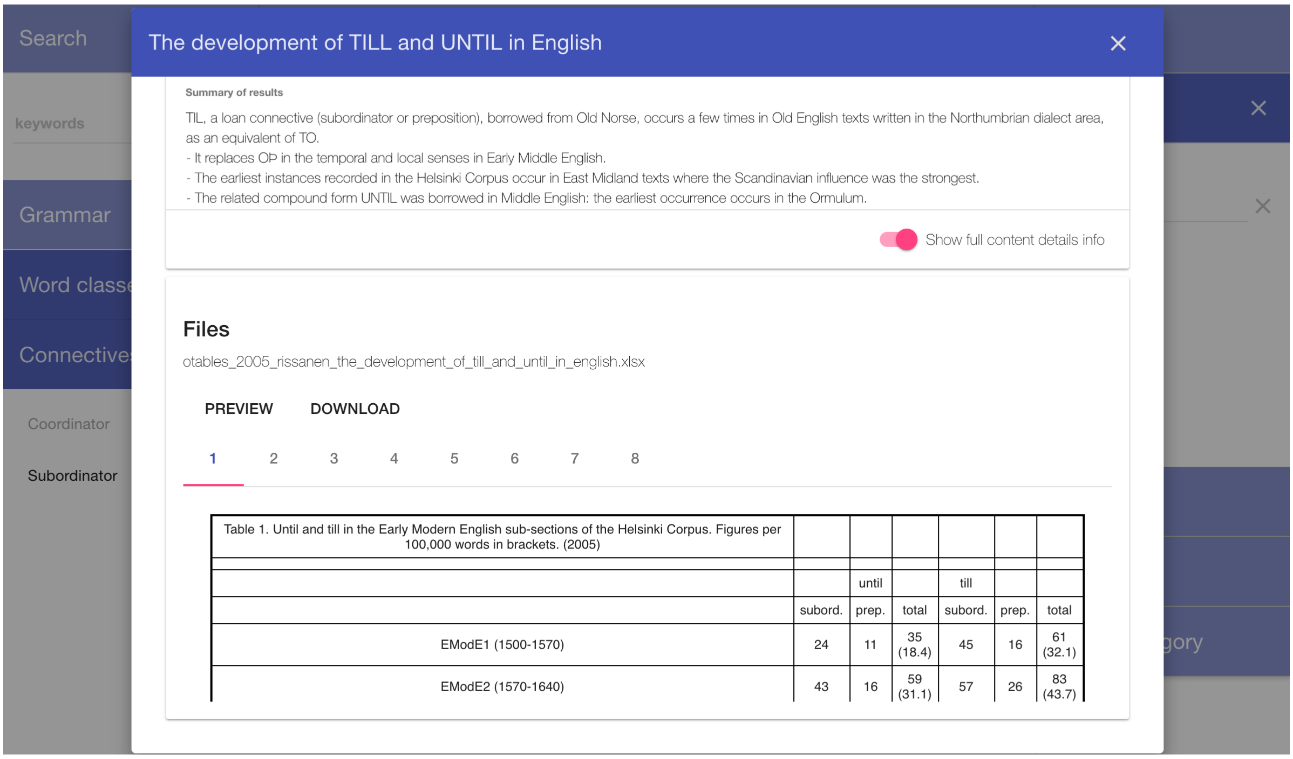
Towards an integrated view of the development of English: notes on causal linking
Rissanen, Matti
1998

Towards an integrated view of the development of English: notes on causal linking
Rissanen, Matti
1998

What time 'when, whenever, while': a byline in the development of English temporal subordinators
Kilpiö, Matti
2013

Figure 2. The front-end of the LCD. The filtering shows publications that are linked to Connectives, and therefore potentially relevant to our sample study.

Figure 3 shows the (partial) content details of an LCD entry in the search front-end, including the summary of the main findings of the research article expressed as bullet points. The Excel files that include the numerical information extracted from the article are listed at the bottom of the entry. The end users can either view these tables online or download them on their personal computers for later examination and re-use. It will also be possible to add the publication to a "shopping basket" for later retrieval. The users of the database can add multiple publications to the basket and then use a "checkout" functionality to download all the publications and their related data files as a single zip file.



The development of TILL and UNTIL in English

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.

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)

Figure 3. Partial content details of an LCD entry with a preview of the tabular data extracted from the article (Rissanen 2005). We can see that the first table contains research results pertinent to our sample study.

2.2. Database implementation

The LCD is run on an integrated linked-data application development environment called **Callimachus**.[3] Callimachus was chosen as the preferred environment for the LCD because of its versatility and flexibility; as Callimachus is specifically designed for the integration of new data with existing data, it is perfectly suited for accomplishing the general goals of the project: the accumulation and dissemination of knowledge and research data. The scalability and robustness of Callimachus have been thoroughly tested in large-scale projects by several companies and government organizations in the United States (e.g. Wood 2015). Moreover, the open source code, which is regularly updated by the development team of Callimachus, not only facilitates database development in the future but also conforms to the spirit of open science promoted by the LCD project. One example of an open-science project that makes use of Callimachus is the US Environmental Protection Agency's open data portal,[4] which is used to maintain various kinds of information related to c. 1.3 million facilities regulated by the agency.

Callimachus combines several key properties in one package: it functions as an RDF database,[5] a web application framework and a data publishing and integration interface. As a web-based solution, it provides a user-friendly interface both for database developers and the end users of the resource. Callimachus also comes with authentication and role-based authorization features, which can be used, for example, to restrict access to certain functionalities or data (e.g. resource deletion) for specific users or groups. It is true that these functionalities are also available in other technologies, such as Semantic MediaWiki (SMW), but Callimachus was chosen as the basis for the LCD for practical reasons: while an SMW-based solution would have required a lot of study before getting to the actual implementation stage, Callimachus allowed us to start constructing the database at a very early stage in the design process (see Nevalainen et al. 2016: 83–85 for further discussion).

With the exception of externally created files (e.g. image and spreadsheet files), all data, configurations and the application structure in Callimachus are stored and manipulated using a graph-based data model implemented in RDF. All Callimachus applications are therefore **linked data applications**, where the data layer conceptually consists of resources identified by Universal Resource Identifiers (URIs) that are connected with labeled, directed properties, as opposed to relational database or document-based approaches. For example, a corpus with an identifier `<lcd:hc>` that has a property `<dct:title>` with a value "Helsinki Corpus" can be expressed in RDF using the following construct: `<lcd:hc> <dct:title> "Helsinki Corpus"`. This is called a triple, which consists of a subject (`<lcd:hc>`), a predicate (`<dct:title>`) and an object (`"Helsinki Corpus"`). The identifiers are here presented using a compact format, where the part before the colon refers to a namespace that functions as a context to the name that follows it. For example, `dct` is usually used to refer to the Metadata Terms vocabulary of the Dublin Core Metadata Initiative[6], whereas the namespace `lcd` is used to denote terms in the context of the Language Change Database.

[3] https://www.w3.org/2001/sw/wiki/Callimachus

[4] https://opendata.epa.gov

[5] Resource Description Framework; https://www.w3.org/TR/rdf11-concepts/

[6] http://dublincore.org/documents/dcmi-terms/

Once Callimachus is up and running, all application development is done via the web browser. Making changes to the application data and the user interface is easy and can even be done “on the fly” while inserting new data. While this kind of flexibility can also prove problematic for the maintenance of the application’s source code, especially if there are several developers involved, the source code of the LCD has been developed and managed by a single software architect right from the start, which has allowed us to keep track of all the changes and developments concerning the database in real time (for more information on the design principles of Callimachus, see Battle et al. 2012).

2.3. LCD data model

In this section, we use the term *data model* to refer to a specification that defines the concepts, their relationships and other associated properties needed to capture the data required by the domain of the LCD application, i.e. research results in the field of corpus linguistics. The core classes of the LCD data model are `Publication`, `Corpus` and `ContentDetails` (Figure 4). The figures in this section have been created using a web-based tool called webVOWL (Lohmann et al. 2016).[7]

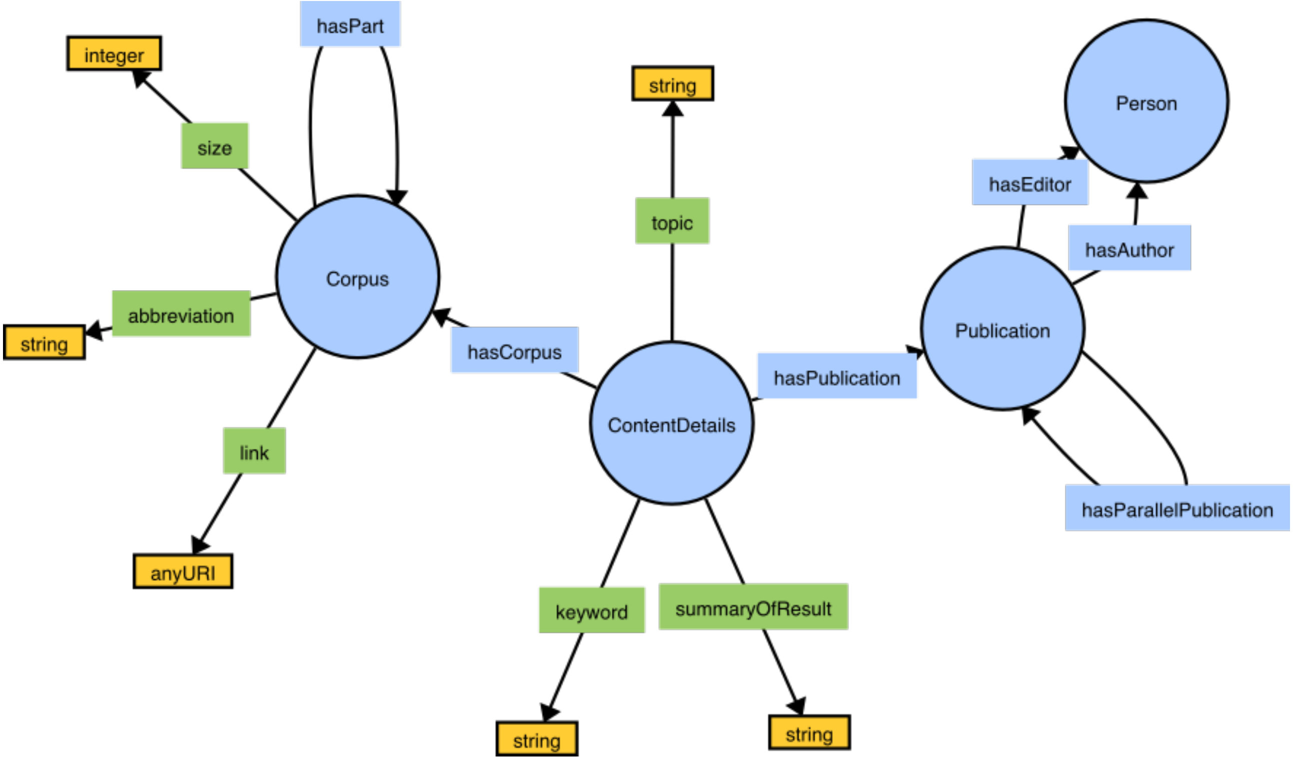


Figure 4. Main concepts of the LCD data model and some of their properties.

`Publication` works as the starting point for data insertion, and it is therefore a natural part of the LCD data model. The publication is described with comprehensive bibliographic metadata, including, for example, the title, publication year, authors and different kind of identifiers. The purpose of these publication details is to identify the exact version of the publication accurately. The LCD can contain data related to multiple published versions of the same article, in which case they can be linked together as parallel publications. This information can then be used, for example, to cluster together the search results from all the parallel versions. The authors and editors of publications are modelled as instances of the

[7] http://vowl.visualdataweb.org/webvowl.html

Person class in order to reliably group together publications by a certain individual. However, we decided not to include any people-related external identifiers, such as ORCIDs[8], in the model in an effort to minimize the amount of personal information in the database.

Corpus represents the second essential part of the model, since every publication added to the LCD contains references to one or more corpora. The model is based on the information available from the Corpus Resource Database (CoRD)[9], which contains detailed descriptions of a large number of English language corpora. The LCD corpus model contains a subset of the information submitted to CoRD in a structured format. It includes the abbreviation, description and total word count of the corpora as well as information about their temporal coverage, language variety, genre distribution and social categories, if applicable. All of this information can be utilized when searching for a particular corpus.

ContentDetails hosts the information extracted from the publication through close reading, and it can be related to one or more corpora used in the study. ContentDetails allows the users of the LCD to target their search to the potentially interesting research results and work their way to the publications and data instead of going the other way around, which is the case when the search functionality relies on just the bibliographic metadata.

ContentDetails also links the study to a set of shared vocabularies/taxonomies that are maintained as part of the LCD (see Figure 5). ContentDetails can be classified based on related concepts from vocabularies related to grammar, genre, discourse analysis, pragmatics, sociolinguistics, dialectology, language contact, statistical method and variety. The controlled vocabularies are modelled as ConceptSchemes using the Simple Knowledge Organization System[10] (SKOS) model, which provides a standardized means of adding preferred, alternative and hidden labels, semantic relationships between concepts (e.g. hierarchies) and mapping between different classification systems. Different types of labels can be utilized for example in the search interface to accommodate spelling variations such as *standardization* and *standardisation*.

[8] https://orcid.org/

[9] http://www.helsinki.fi/varieng/CoRD/

[10] http://www.w3.org/2004/02/skos/core.html

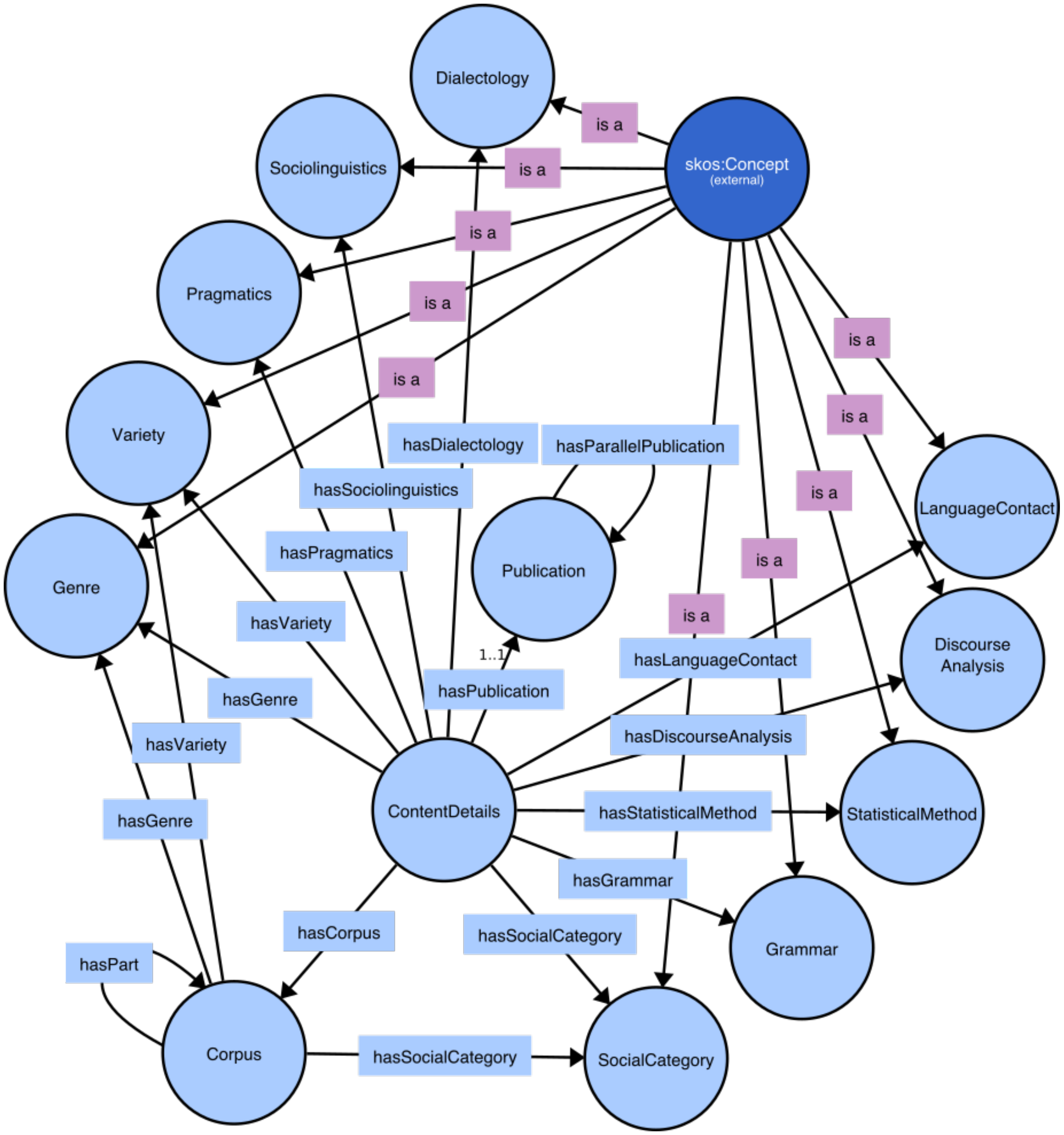


Figure 5. The main classes and their relationships in the LCD ontology.

In addition to unique study descriptions and metadata about the publications and corpora, the LCD also allows users to add and link files to the publications (see Figure 6). Each publication can be linked to exactly one `PublicationFile`, which represents some version of the published article in PDF format and includes a link for downloading the actual file.

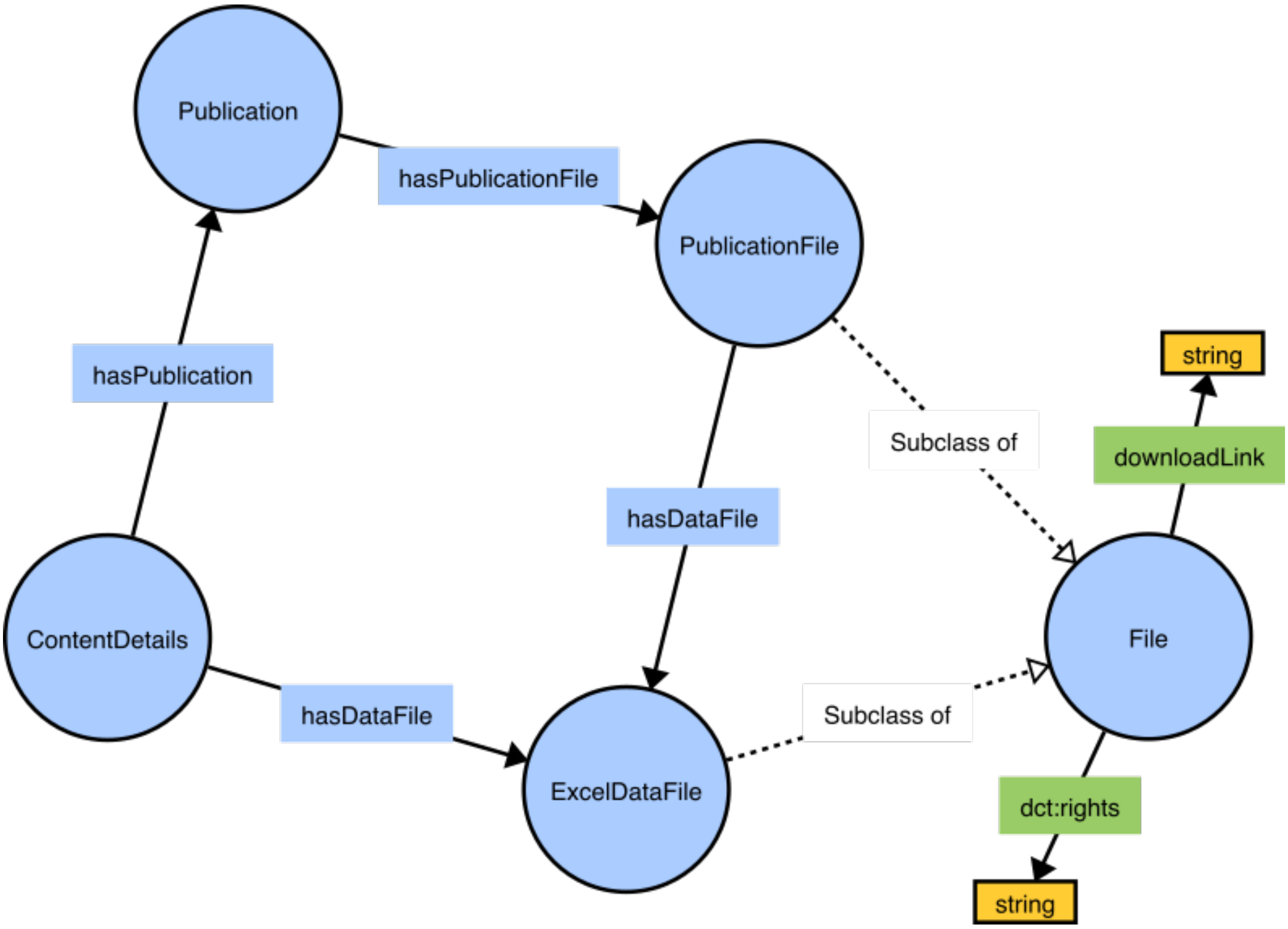


Figure 6. LCD files and related concepts.

ExcelDataFile denotes a file that contains one or more data tables extracted from a given PublicationFile version of a certain Publication, as expressed in the data model by the hasPublicationfile and hasDatafile relationships (see Figure 6). If a PublicationFile is linked to multiple ExcelDataFile resources, it is possible to add an additional link directly from ContentDetails to ExcelDataFile to make explicit which of the data files are related to a certain ContentDetails. It should be noted that whereas the LCD data is openly available to anyone, access to the PDF versions of articles may be limited due to copyright restrictions. Such conditions are described as part of the File resource using standard Dublin Core properties.

Figure 7 illustrates what the LCD data looks like for a single article. The figure shows the main resource types, i.e. publication, corpus and content details, and how they are connected and described using the shared vocabularies for linguistic properties such as grammar concepts, genres and varieties. The graph in Figure 7 visualizes the concrete data stored in the LCD. It is a manifestation of the LCD's data model, which defines how the data can be organized into types of things, how these things relate to each other and what kinds of properties different types can possess.

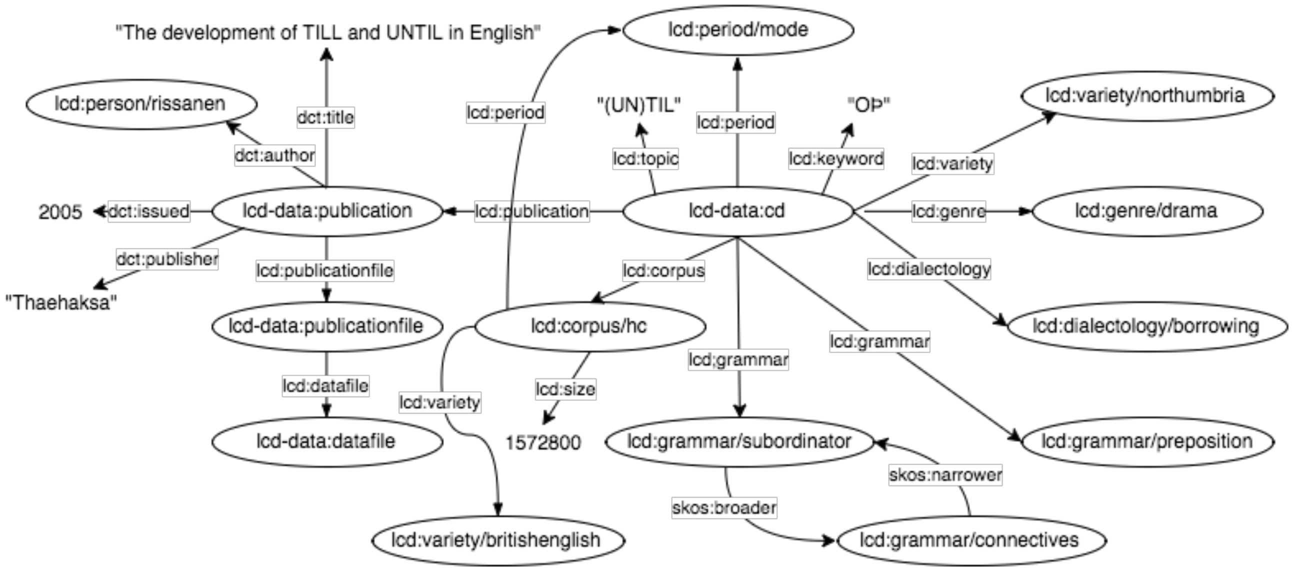


Figure 7. Part of the graph stored for our sample study. This is the version of the data that is created for machines to understand. See Figure 1 for a more human-friendly version of the same data.

2.4. LCD as linked linguistics data

Using linked data for creating, sharing and publishing linguistic data has many benefits that have been outlined in Chiarcos et al. (2013). In this paper, we use the term *linked data* for data that adheres to the Linked Data Principles first proposed by Tim Berners-Lee (e.g. 2006):

1. Use URIs [Uniform Resource Identifiers] as names for things.
2. Use HTTP URIs (URLs) so that people can look up those names.
3. When someone looks up a URI, provide useful information, using web standards.
4. Include links to other URIs, so that they can discover more things.

Linked data can be used to represent any type of linguistic resource in a way that is structurally and conceptually interoperable, i.e. allows for straightforward integration of data and schema resources through the use of URIs, provides tools for data fusion (Bizer et al. 2009), and is expressive enough to be used to denote not only the data but also the metadata, provenance and ontological constraints related to the data.

Applying RDF based graph technologies to describe, create and distribute research data is not a novelty amongst the linguistics community. The first RDF version of WordNet (Van Assem et al. 2004), a large lexical resource for the English language, was created using the web ontology language OWL in 2004, and a more light-weight, linked data version of the resource was made available by McCrae et al. in 2014.

Chiarcos et al. (2011) coined the term “Linguistic Linked Open Data” (LLOD) to refer to a set of linguistically relevant resources which are maintained and/or distributed using the best practices of linked data management (see section 2.2), take advantage of semantic web technologies, and are available under permissive, open licenses. These resources include any language-related data that can be used for the purposes of linguistic research or natural language processing (McCrae et al. 2016). The Open Linguistics

Working Group (OLWG) has maintained a dataset of LLOD resources since 2012 and provides a visualization of the data in the form of the LLOD cloud.[11]

The LCD is not currently connected to the linguistic linked open data cloud, but the model has potential for external linking. The bibliographic data in the LCD can be linked to external resources using pre-existing or generated identifiers related to the publications, such as URNs, DOIs and ISBNs. The LCD could then be used to add more in-depth information about the research results of the publication via ContentDetails descriptions. Furthermore, the corpus resources of the LCD can be easily implemented for external linking. Corpora regularly contain URL links to other resources, such as corpus documentation or URL access to the web user interface of the corpus. A similar field could be used to store linked data URIs in order to provide access to descriptive metadata about the corpus. Finally, since all the LCD categories (genre, grammar, variety, etc.) are expressed as SKOS schemas, they are natural candidates for SKOS vocabulary based mappings onto other classification schemes related to linguistic research.

3. Extending LCD for meta-analysis

One of the main goals of the LCD is to provide researchers with the possibility to re-use existing data for different purposes, such as replication studies and meta-analyses, and the data model described in the previous section has been designed with these goals in mind. The first step in performing a meta-analysis includes the identification and selection of all the articles that are relevant to the research question at hand. This is followed by the laborious task of going through all the selected articles, as well as extracting and possibly harmonizing the related data tables in preparation for analysis. In this section, we present two extensions to the LCD data model that are intended to provide concrete help to anyone interested in using the LCD for meta-analysis. These extensions are designed to make it easier to identify and aggregate relevant subsets of values in tabular data across multiple tables and publications, and to normalize raw word frequency values either automatically or semi-automatically.

3.1. Annotated data files

In the original LCD file model, data files containing the tabular data manually extracted from the original article are linked to the publication (see 2.3). In the current version of the LCD, data files are stored as Excel spreadsheets, and all the tables in the referenced publication are combined into one file with multiple sheets. These files provide the research data in a structured, open and text-based format, which is easy to copy, transform and combine. However, some additional metadata is required in order to support the automated processing of individual table cells. This can be achieved through semantic annotations that add structured data describing type or instance information about the target entity.

Inspired by the TabLinker tool (Meroño-Peñuela et al. 2012), we have developed a similar annotation approach that uses custom styles and formatting conventions to add a semantic layer to plain Excel files. Our solution is designed specifically with the corpus linguistics domain in mind and can only be used to create annotations that are compatible with the LCD data model (see section 2.3), as opposed to the more generic approach taken by the TabLinker tool.

[11] http://linguistic-lod.org/

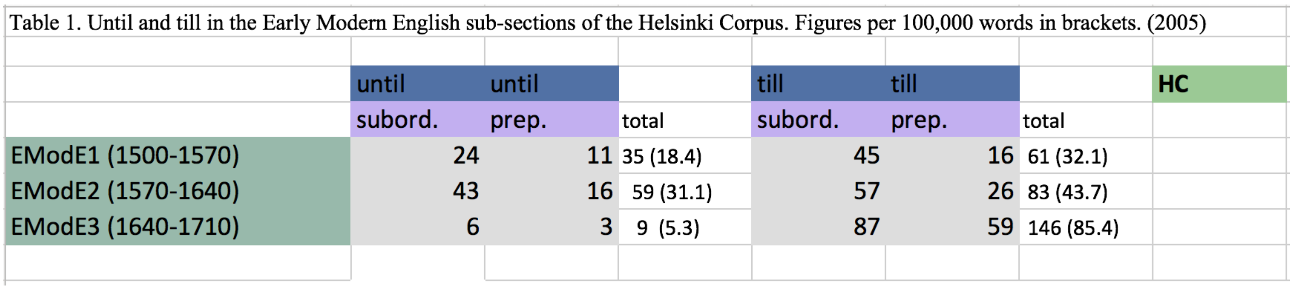


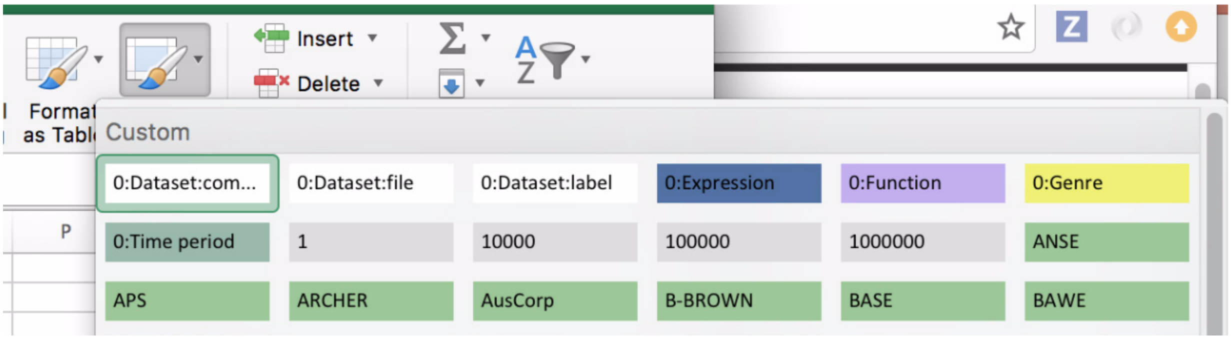
Table 1. Until and till in the Early Modern English sub-sections of the Helsinki Corpus. Figures per 100,000 words in brackets. (2005)

	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)	

Figure 8. Annotated table from Rissanen (2005).

In order to create an annotated data file, the LCD provides an automatically generated template with instructions and some prefilled annotations. Figure 8 shows an example where the corpus, time periods, word frequencies and the linguistic items and their functions are annotated with different colours. The program that processes the new annotated version can then associate the desired semantics and processing rules with specific colours. For example, values in the fields annotated as time periods might be parsed using different formats to indicate start and end years. Similar rules can also be associated with the formatting options, such as italics and boldface.

Figure 9 shows all the available dataset, word frequency value and dimension type annotation styles and part of the corpus annotations. Dataset annotations are used to link the annotated Excel file to the original file and to provide metadata (label and comment) for the table.



Custom

0:Dataset:com...	0:Dataset:file	0:Dataset:label	0:Expression	0:Function	0:Genre
0:Time period	1	10000	100000	1000000	ANSE
APS	ARCHER	AusCorp	B-BROWN	BASE	BAWE

Figure 9. Available LCD annotation styles. Cells can be annotated by Expression, Function, Genre, Time period, Value, and Corpus. The Value annotation (in grey) includes a normalization base, and the Corpus annotation (in green) links the cell to a specific corpus entry in the LCD.

Expression, function, genre and time period are dimension type annotations that classify the annotated cell values. For example, a cell that contains the value “Laws and documents” could be annotated as “Genre” to produce a compound genre. Expression annotation is normally applied to linguistic items, while function is used to contextualize their use. In Figure 8, cells containing the values “until” and “till” are annotated as expressions, while “subord.” and “prep.” are annotated as functions.

Word frequency value annotations (in grey) are used to markup cells that contain word frequencies. The annotation also provides information of whether the value represents an absolute or normalized frequency, and in the latter case, the normalization base of the value. Value cells are associated with dimension annotations that are either above or to the left of them. For example, in Figure 8, the value “24” in the second column is associated with the following three dimensions and their values: Expression = “until”, Function = “subord.”, and TimePeriod = “1500-1570”.

Only data files that are already added to the LCD should be annotated, because this makes it possible to generate a link back to the original data file (see Figure 10) automatically. It is, of course, also possible to add the link manually later, but that approach always comes with a risk of human error. When an annotated data file is added to the LCD, a link between the original and annotated versions of the `ExcelDataFile` is created. Figure 10 shows how the original LCD data model is extended with the annotated versions of `ExcelDataFile`s.

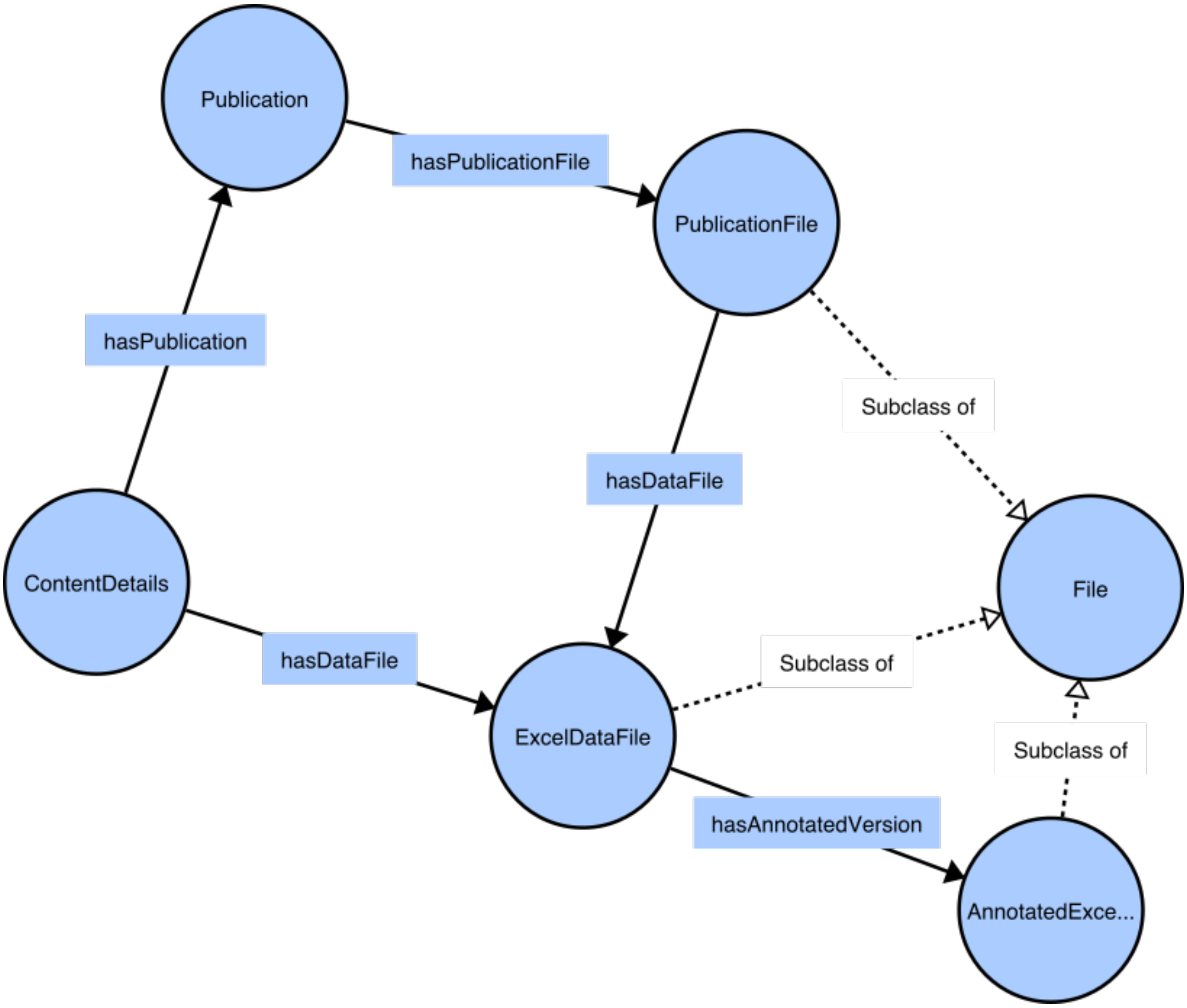


Figure 10. Original LCD data model extended to support the annotated versions of files.

Although the `ExcelDataFile` can have multiple annotated versions, an `AnnotatedExcelFile` is always linked to exactly one `ExcelDataFile`. Versioning can be used, for example, to fix annotation errors or to extend previously published annotations. Another case for versioning, albeit more subtle, is to provide different interpretations of tabular data with respect to the LCD's model. Every `AnnotatedExcelFile` has a unique URI identifier, and the researchers re-using the data should be able to select the version best suited for their purposes based on descriptive metadata. From the perspective of data management, versioning of files is also preferred to deletion, since there is no way of knowing whether someone has already re-used the files in their work.

3.2. Corpus compositions

3.2.1. Introduction

Some of the articles stored in the LCD report only absolute values for word frequencies in their findings. In order to efficiently combine data across publications, we need to be able to normalize those values in an automated manner, and for that we need information about the distribution of corpus word counts in a machine-processable format. To this end, we have created a vocabulary called RDF Data Cube for Corpus Compositions (QB4CC), which allows us to describe the required structural corpus data in linguistic terms in a way that is compatible with the LCD data model and the RDF Data Cube vocabulary. The RDF Data Cube vocabulary (QB)[12] is a W3C recommendation specification for publishing multi-dimensional data, such as statistics, on the web.

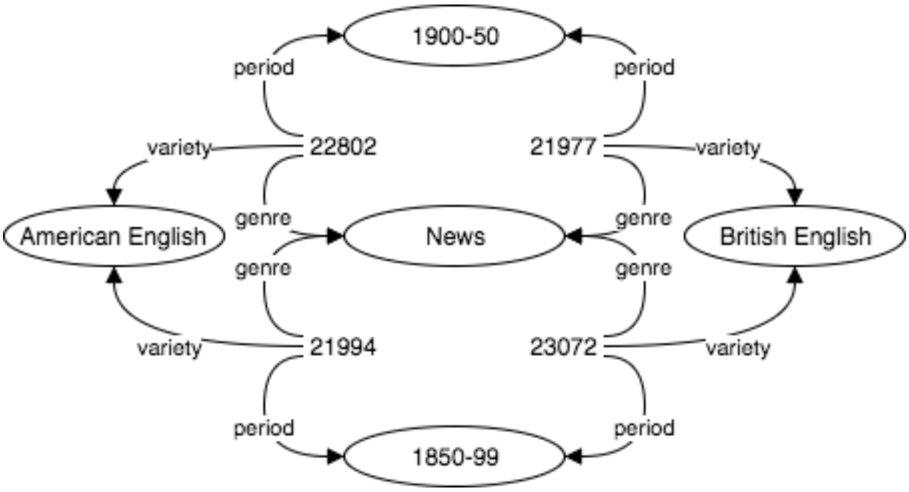


Figure 11. Example of corpus composition data describing the word counts in the ARCHER corpus.

The QB dataset consists of observations that adhere to a certain data structure definition (DSD). The observations can be likened to values in a data table, and they are contextualized, categorized and interpreted using dimensions, measures and attributes, respectively. A set of dimensions identifies a single observation by describing what the observation applies to, i.e. the context, such as time, geographic area or text genre, in which it was observed. What exactly was observed (e.g. absolute or normalized word frequency), is described by the measure property, which in turn can be described in more detail using attributes, such as unit, scaling factor or any other piece of metadata that can be helpful when utilizing the measured value. Put together, dimension, measure and attribute descriptions form a component part of the data structure definition. Since the DSD can have its own unique identifier, it can be re-used by multiple datasets with compatible content (i.e. observations). Figure 11 shows an example of corpus composition data extracted from the CoRD website[13] that contains four word count measures, each described using "variety", "time period" and "genre" as their dimensions.

[12] https://www.w3.org/TR/vocab-data-cube/

[13] http://www.helsinki.fi/varieng/CoRD/corpora/ARCHER/updated%20version/archer%203_2_structure.html

3.2.2. QB4CC

We define *corpus composition* as a partition of the corpus word count into distinct, i.e. non-overlapping, corpus parts, where each part is identified by one or more dimensions from the QB4CC vocabulary as per the RDF Data Cube specification. A single corpus can have multiple compositions as long as the sum of word counts over all the corpus parts belonging to a specific composition is equal to the total word count of the corpus. For example, the Helsinki Corpus[14] could be represented using three different compositions: by genre (e.g. history or fiction), prototypical text category (e.g. secular or religious instruction) and time period (e.g. Middle English 1 or Middle English 2).

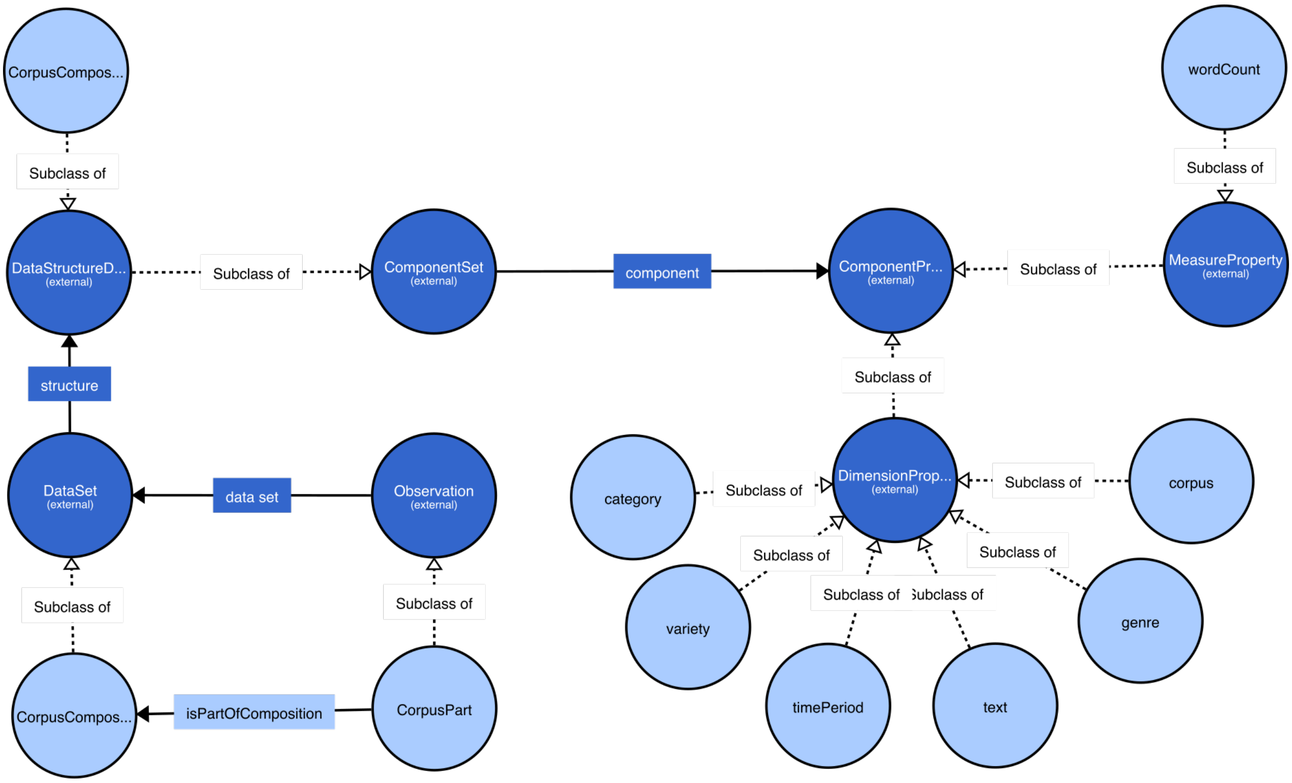


Figure 12. VOWL visualization of QB4CC and its relationship to the RDF Data Cube vocabulary concepts (external). Adapted from https://www.w3.org/TR/vocab-data-cube/.

Figure 12 depicts the main classes and properties of QB4CC and how they are linked to the RDF cube concepts. An OWL representation of the whole vocabulary is available from Github.[15] We have employed subclassing to create an LCD specific view of the RDF Data Cube in order to connect it to the linguistics domain. The QB4CC model provides the following, fixed set of dimension properties: corpus, genre, variety and time period. The QB4CC dimensions are all linked to the LCD data model, which makes the data modelled according to the QB4CC vocabulary compatible with the data in the LCD. Through such explicit domain-specific properties, it is easy to build tools that utilize corpus compositions in a consistent manner. With the RDF Data Cube aligned QB4CC ontology, corpus compositions can be automatically processed into valid data cubes and combined with any other RDF Data Cube dataset. This is exemplified through the LADA tool in section 4.

[14] http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/middleenglish.html
[15] https://github.com/jkesanie/LCD/blob/master/lcd-ontology.owl

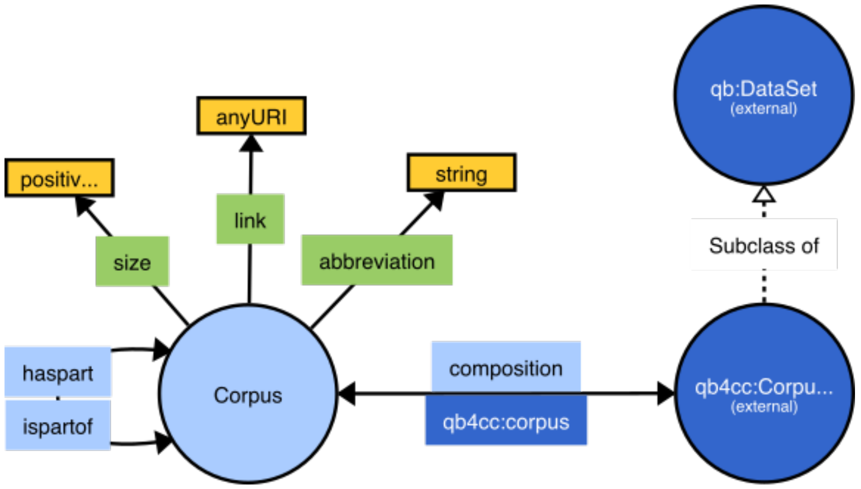


Figure 13. Corpus composition and corpus are linked with a bidirectional relationship.

In the LCD data model, a corpus composition can be regarded as a corpus version where the corpus clusters together all related compositions (see Figure 13). Every composition also has a unique URI identifier, which can be used to refer to a specific composition when documenting the re-use of corpus composition data. The URI identifier of the corpus composition acts as a link back to the LCD database, which can be used to retrieve the content of the composition.

4. LADA

LCD Aggregated Data Analysis workbench (LADA) is an application which provides researchers with a systematic workflow to perform exploratory meta-analyses based on earlier research results. LADA makes extensive use of the LCD data, especially the annotated data tables (see section 3.1) and corpus compositions (see section 3.2).

LADA takes as its input a set of LCD publications that are deemed relevant to the research question at hand. This initial set of data is then filtered, reviewed and normalized in order to create a new aggregated dataset, which can then be visualized or exported as raw data. The whole experiment can be exported as a LADA Experiment Exchange Package (LEEP), which contains all the original data, any intermediate datasets generated in the different stages of the workflow, as well as the final aggregated dataset and its related visualizations. Finally, LEEP comes with comprehensive provenance information that links together the inputs, outputs and parameters of the experiment, allowing other researchers to re-run it in order to validate, extend or comment on the results.

The LADA implementation consists of a Python-based back-end component and a Javascript React front-end component. All the data processing activities are done in the back-end against a set of RDF graphs. Although LADA has both server and client parts, it is designed to be run on a personal computer. The technical details of LADA are beyond the scope of this paper, but the source code is available online.[16]

4.1. Processing annotations – from Excel tables to RDF data cubes

Annotated data tables are the main source of initial data for LADA. The internal processing of LADA works on top of a set of graphs, so the tabular data in Excel files must first be transformed into RDF. Similarly to the corpus compositions discussed in section 3.2, LADA uses the RDF Data Cube vocabulary as the target model for the data. The specification provides a standard way of moving data tables extracted from linguistics papers to the world of Linked Data. The resulting data, with interlinked and machine-processable dimensions, measures and observations (see section 3.2.2), allow for flexible queries across datasets, a requirement for conducting meta-analysis.

As part of the transformation process, the URI identifiers generated for the created resources are all put under a configurable, experiment-specific namespace (see section 2.2) as shown in Listing 1 (`experiment1`), and the name of the identifier is based on the values and structure of the annotated data file (e.g. `pub1-Table6-o-2_4_1`). This name stays the same if the data is regenerated from the same input file, but the namespacing can be used to distinguish between different experiments using the same data.

Every sheet in the Excel file is transformed into a single RDF Data Cube Dataset (`qb:DataSet`) using the cell data annotated as dataset metadata (see section 3.1). A resource of type `qb:Observation` is created for every word frequency value annotated cell in a table. Dimensional data for each `qb:Observation` is then added according to the available annotations (see section 3.1) with some additional LADA specific properties, such as the row and column of the value, which are used for internal processing. LADA applies some simple rules for dealing with complex cell values for different types of

[16] https://github.com/jkesanie/lada

annotations. For example, for cells annotated with the time period dimension, LADA can parse start and end years from cell values such as 1800-, 1800-1900, 1960s. Listing 1 shows an abbreviated example of transformed data with one dataset and one observation.

```
  <experiment1:otables_2005_rissanen-Table1-o-2_5_1>
      rdf:type qb:Observation ;
      lada:frequency 24 ;
      lada:per 1 ;
      lada:corpus <lcd:corpus/hc> ;
      lada:expression "until" ;
      lada:function "subord." ;
      lada:endYear 1570 ;
      lada:startYear 1500 ;
      lada:col 2 ;
      lada:row 5 ;
      lada:sheet "Table1" ;
      ns4:dataSet <experiment1:otables_2005_rissanen-Table1-o--dataset-Table1> .

  <experiment1:otables_2005_rissanen-Table1-o--dataset-Table1>
      rdf:type qb:DataSet ;
      rdfs:label "Table1" ;
      rdfs:comment "Table 1. Until and till in the Early Modern English sub-
sections of the Helsinki Corpus. Figures per 100,000 words in brackets. (2005)"
;
      lada:file <lcd:annotated-file/otables_2005_rissanen> .
```

Listing 1. Example observation expressed in Turtle[17] *syntax where the frequency value 24 is an absolute value since the normalization base (lada:per) is 1. The corpus (lada:corpus) is linked to the LCD description of the* Helsinki Corpus*.*

LADA does not perform any clean up of data, beyond the rules discussed above, or provide any additional linking as part of the initial transformation process, but relies solely on the data available from the annotated Excel file. This means that some mapping is always required in order to make use of the data in the context of the specific research question.

4.2. LADA workflow

The UI of the LADA workbench is divided into five steps (Figure 14) that cover specific tasks related to the LADA meta-analysis workflow. The steps are linked together through input and output datasets, which all have their own unique identifiers. The following sections give a technical overview of the different stages, their functionality, and associated inputs and outputs. For a quick overview of the process, please see the video at https://doi.org/10.5281/zenodo.1202371.

[17] https://www.w3.org/TR/turtle/

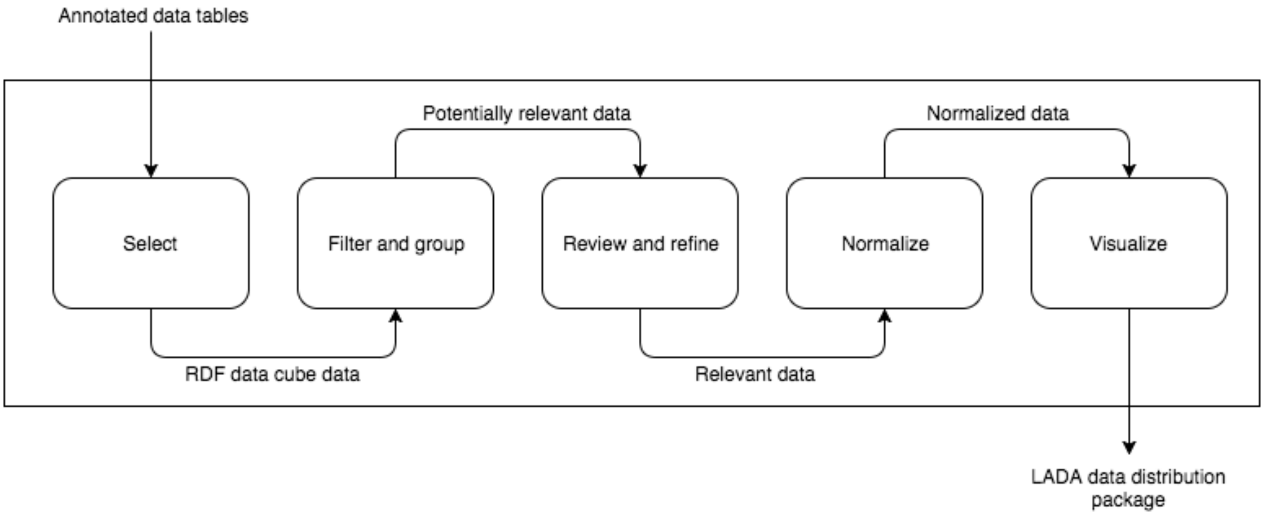


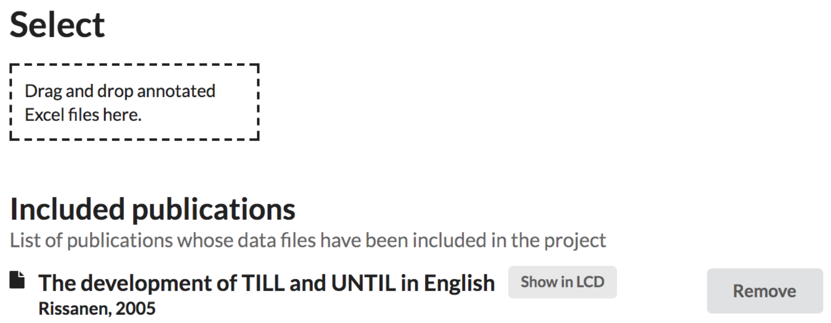
Figure 14. The LADA meta-analysis workflow with stage specific inputs and outputs.

4.2.1. Select

- Input: Annotated data tables from the LCD
- Output: RDF Data Cube version of the annotated data

The “Select” step is used to manage the source data available for the meta-analysis workflow. Annotated data tables are provided by the LCD, and they can either be added individually or as a set using the distribution package provided by the LCD’s end-user search application (see section 2.1). LADA performs the initial transformation described in section 4.1 upon upload, and shows if there are other versions of the annotated file available in the LCD. For example, it is possible that two versions of the annotated Excel file are available, one of which includes more comprehensive annotations. LADA also provides a direct link to the LCD entry, allowing the user to view more detailed information about the publication (Figure 15).

This stage also allows the user to either remove or exclude publications from the dataset. The former option deletes all initially generated data from LADA’s internal storage, whereas the latter simply marks it as excluded from the active dataset, without the need to upload and transform the file again if the publication is later decided to be included.


Select

Drag and drop annotated Excel files here.

Included publications

List of publications whose data files have been included in the project

The development of TILL and UNTIL in English
Rissanen, 2005

Figure 15. The “Select” step with one added publication.

When the user transitions from the “Select” step, LADA performs another round of transformations targeting all the expression, genre and function dimension values of the observations generated from the selected publications. This second transformation is used to turn the cell values interpreted as strings into `skos:Concept` resources with unique URI identifiers. For example, the expression “until” becomes a resource with a URI such as `<experiment/pubX-TableY-o-1_2_3/expression/until>`. Listing 2 shows how the example from the previous section has been extended with the `lada:expression` property, which points to a resource that has a unique label. These new resources cluster distinct expression, function and genre values together, in the context of a single data table, and provide URIs to use for mapping in the next stage.

```
<experiment1:otables_2005_rissanen-Table1-o-2_5_1>
  rdf:type qb:Observation ;
  lada:frequency 24 ;
  lada:per 1 ;
  lada:corpus <lcd:data/corpus/hc> ;
  lada:expression <:otables_2005_rissanen-Table1-o-2_5_1/expression/until> ;
  lada:expression <:otables_2005_rissanen-Table1-o-2_5_1/function/subord.> ;
  lada:endYear 1570 ;
  lada:startYear 1500 ;
  lada:col 2 ;
  lada:row 5 ;
  lada:sheet "Table1" ;
  lada:dataSet <experiment1:otables_2005_rissanen-Table1-o--dataset-Table1> .

<:otables_2005_rissanen-Table1-o-2_5_1/expression/until>
  rdf:type skos:Concept ;
  skos:prefLabel "until"  ;

<:otables_2005_rissanen-Table1-o-2_5_1/function/subord.>
  rdf:type skos:Concept ;
  skos:prefLabel "subord."  ;
```

Listing 2. Example output of the second step transformation of the RDF Data Cube data.

4.2.2. Filter and group

- Input: All the observations generated from the annotated data tables
- Output: Observations relevant to the research question with appropriate groupings

The “Filter and group” step is used to formulate the filters and groupings that satisfy the desired framing of the research question. The user can create filters, groups and filter groups using the same five dimensions that are available for data table annotations (see section 3.1): corpora, expressions, genres, functions and time periods (Figure 16).

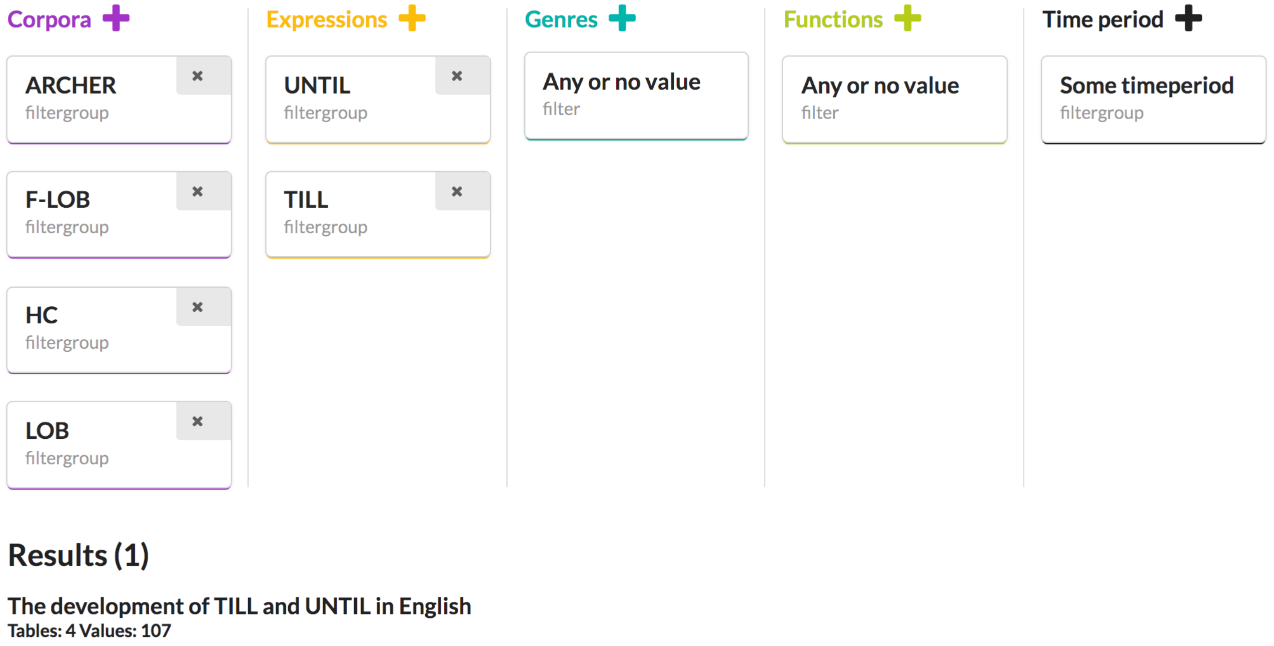
Groups are the mapping targets for data table specific dimension values. For example, one could first create a genre group called “Statutory” in accordance with one of the prototypical text categories in the *Helsinki Corpus*[18], and then map all the related, data table specific genres, such as “Legal” and “Law”, to this new group. After that, all the observations linked to the new genre group can be treated as a single group

[18] http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/genres.html

in the later stages of the workflow. The values available for mapping are restricted to the distinct values found for the given dimension in the data cubes generated from the input data.

Filtering is used to decrease the number of observations generated as an output. There are three generic filters that can be used to make broad selections: “No value”, “Any value” and “Any or no value”. The first one uses negation to only include observations that do not contain a value for the selected dimension, for example observations that are not linked to any genre. Conversely, “Any value” includes observations that contain a value for the dimension in question, and the “Any or no value” option removes all restrictions related to a given dimension. Value filters work in the same way as groups and allow the user to cluster values from different datasets. For example, a filter called “preposition” can be created to include different abbreviated versions of the term found in the data, such as “pre.”, “prep” and “prep.”.

By default, LADA creates a filter group that, as the name suggests, combines features of filters and groups and is suitable for most use cases, e.g. when the user wants to include observations only from certain corpora and also group them based on the corpus dimensions.



Corpora ✚
ARCHER filtergroup ×
F-LOB filtergroup ×
HC filtergroup ×
LOB filtergroup ×

Expressions ✚
UNTIL filtergroup ×
TILL filtergroup ×

Genres ✚
Any or no value filter

Functions ✚
Any or no value filter

Time period ✚
Some timeperiod filtergroup

Results (1)

The development of TILL and UNTIL in English
Tables: 4 Values: 107

Figure 16. The “Filter and group” step with three active dimensions. Our sample study combines results related to the selected linguistic variable from four corpora. The “Results” preview shows that the source data has four relevant tables which contain 107 cells that are annotated with dimensions that match at least one of the combinations from the added corpus and expression filter groups.

The effect of filtering is immediately visible from the “Results” list, which shows the number of tables and observations matching the current filters grouped by publication. The filtered values list shows how the observations potentially related to the research question are distributed across the source data and gives a rough estimate of how much work is needed in the next step, “Review”.

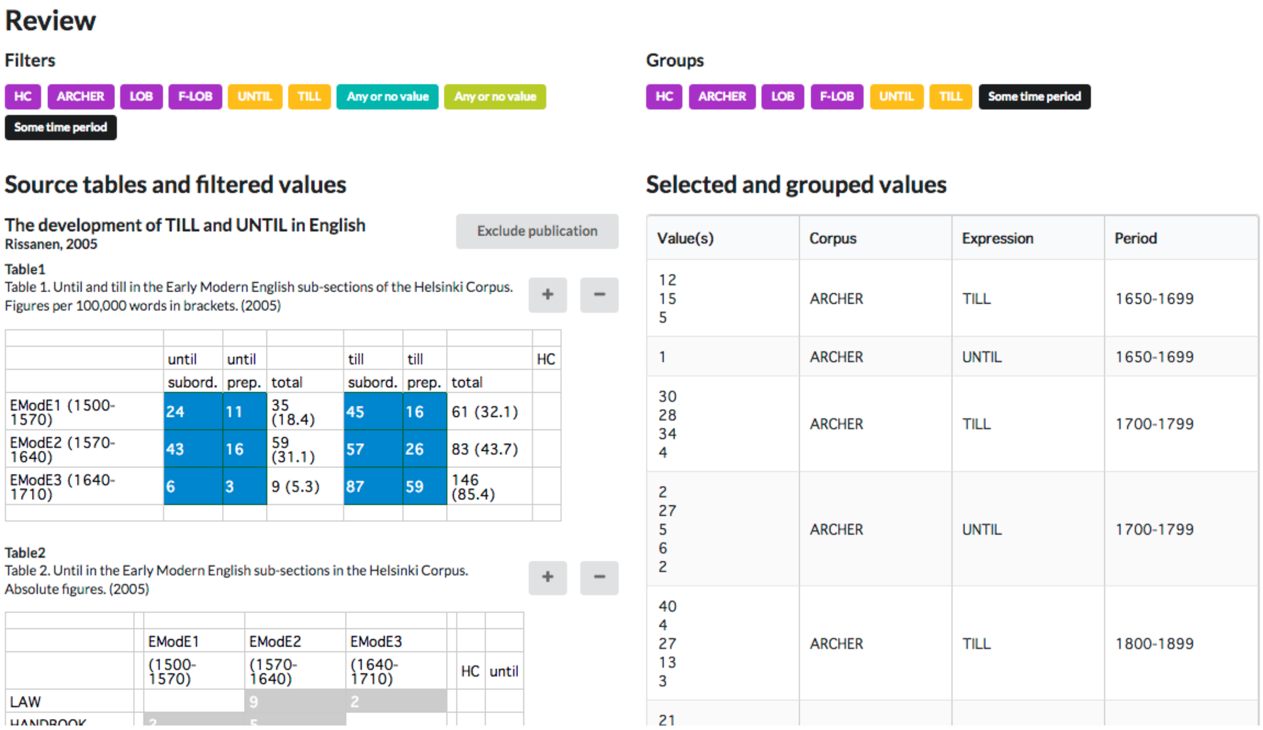
4.2.3. Review and refine

- Input: Filtered and grouped observations

- Output: Final set of observations

The purpose of the "Review" step is to finalize the observations included in the result dataset by reviewing individual observations included in the active dataset based on selections made in the previous stages of the workflow, and by refining that dataset by excluding specific observations or entire tables.

For example, the same data can appear in multiple tables across the source dataset, as a result of previous research being referenced or summarized in the included articles. This causes duplicate observation values to appear in the result dataset, since the observations share the same dimensional properties, so they must be explicitly excluded. An individual observation can be excluded and included by clicking a value cell in the table preview (Figure 17, left column), or the user can exclude/include an entire table or publication with a single click of a button. Included observations are shown with a blue background, which is changed to grey for excluded observations.



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)	(1640-1640)	(1640-1710)		HC	until
LAW			9	2			
HANDBOOK		2	5				

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
21			

Figure 17. The "Review" step. 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. Table2 has been excluded because it contains the same data as Table1 only broken down by genre.

The result table (Figure 17, right column) adds a column for every dimension that has at least one group or filter group. Figure 17 shows an example where the Corpus, Expression and Period dimensions have groupings. The Values column is then populated with word frequencies found in observations that match every distinct combination of group values, hence the number of rows in the result table is the number of available combinations of the group values. The frequency and, for normalized values, the normalization base is shown for every matched value. Currently, all absolute values are summed together in the normalization step, so the user must pay close attention to rows that contain multiple absolute values in order to make sure that the values can indeed be aggregated without sacrificing the statistical soundness of the resulting data set.

LADA users will most likely spend most of their time switching between the “Filter and group” and “Review” steps when fine-tuning the available data values and their groupings. When the result data includes only the relevant values, it is time to move forward to the “Normalize” step.

4.2.4. Normalize

- Input: Final set of relevant raw observations
- Output: Result set of observations with frequency values normalized to a common base

The input dataset for the “Normalize” step can include both absolute and normalized values. The goal of the normalize step is to make those values comparable with each other by normalizing them all to the same base.

The normalization step utilizes the corpus composition data (see section 3.2) available from the LCD to calculate normalized values for observations with absolute frequencies. Mappings are currently done at the table level, so that every distinct corpus resource within the observations from one specific table can be mapped to exactly one composition. The approach is a compromise between usability and maximum mapping granularity, which limits the amount of manually created mappings, but also makes it impossible to correctly handle situations where the same table contains values from different versions of the same corpus. We did not encounter such situations in any of the data tables related to our case study.

Corpus to composition mapping is a manual process, because the annotation does not contain information about the specific version of the corpora, but just a reference to the parent corpus, such as “Helsinki Corpus”. However, a list of available mapping targets, i.e. corpus versions, for each corpus is populated with only the kind of corpus compositions that match the dimensions found in the input dataset. Figure 18 shows an example of two tables, one with observations from only one corpus (ARCHER) and the other with observations coming from two corpora (LOB and F-LOB).

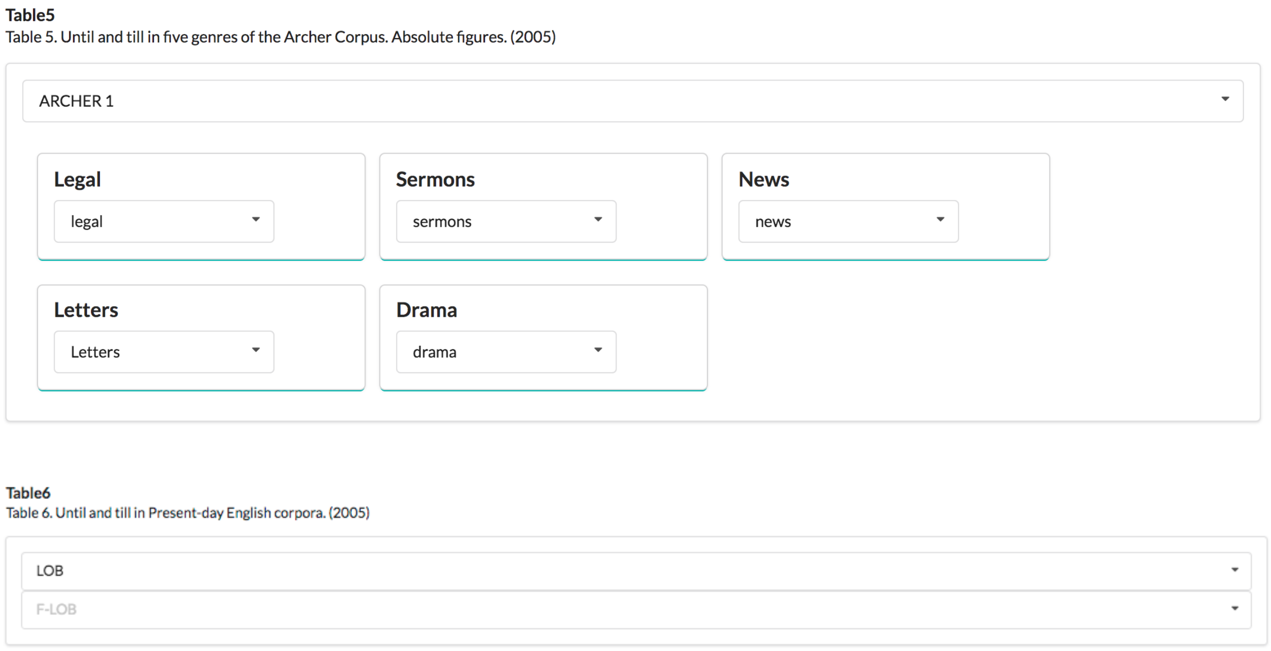


Table5
Table 5. Until and till in five genres of the Archer Corpus. Absolute figures. (2005)

ARCHER 1

Legal: legal
Sermons: sermons
News: news
Letters: Letters
Drama: drama

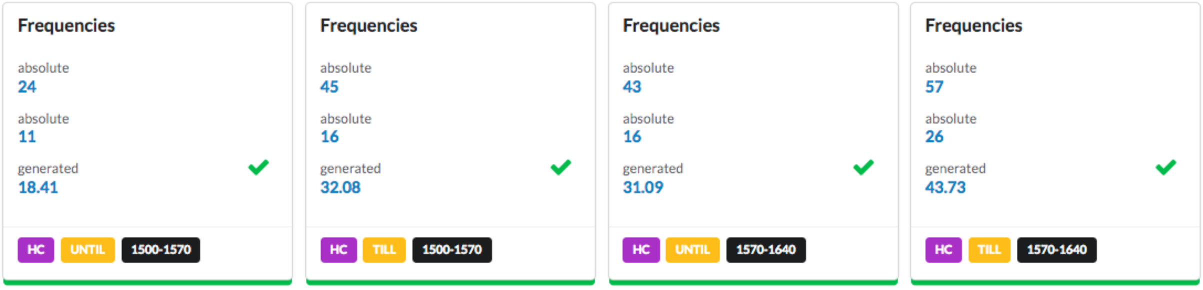
Table6
Table 6. Until and till in Present-day English corpora. (2005)

LOB
F-LOB

Figure 18. Part of the "Normalize" step: selecting corpus compositions for tables and mapping the genres found in the corpus composition to the genres found in the observations.

If the input observations contain a Genre dimension, it is also necessary to create a mapping between observation specific and corpus composition genres (Table5 in Figure 18). For example, the observation can have an ad hoc genre called "L&D", which the author of the publication has created by combining texts from two or more corpus genres, such as "Letters" and "Documents".

Whenever a corpus to composition mapping is modified, the output dataset is updated with the newly calculated normalized frequency values for observations linked to the corpus in question. The results of the mapping process are visualized as part of the normalization stage using a card-like UI (Figure 19), which shows the absolute and normalized frequency values coming from the input data and the generated, normalized frequencies for every observation in the input dataset. The card is marked with a green border if the observation contains a normalized value that uses the common normalization base. Otherwise the observation is decorated with a red border, and will not be included in the output dataset. If the observation contains a normalized value coming from the source data as well as the generated value, the user must choose which one is selected (checkmark in Figure 19) for the output dataset.



Frequencies
absolute
24
absolute
11
generated
18.41 ✓
HC UNTIL 1500-1570

Frequencies
absolute
45
absolute
16
generated
32.08 ✓
HC TILL 1500-1570

Frequencies
absolute
43
absolute
16
generated
31.09 ✓
HC UNTIL 1570-1640

Frequencies
absolute
57
absolute
26
generated
43.73 ✓
HC TILL 1570-1640

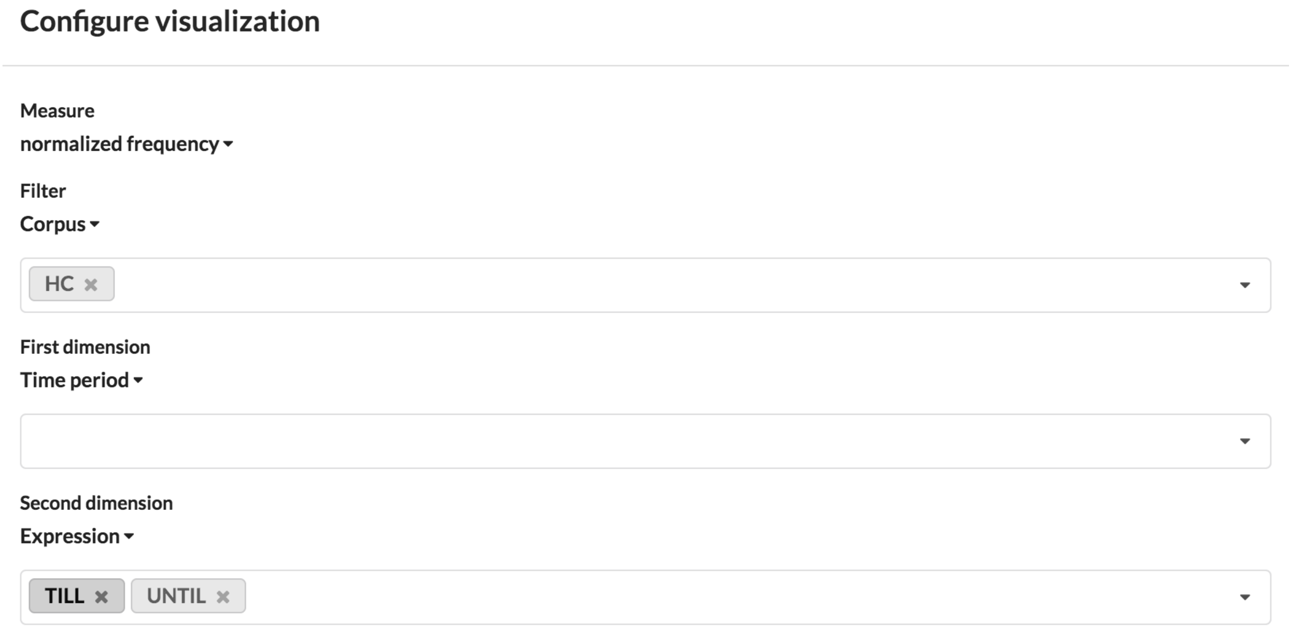
Figure 19. Part of the "Normalize" step, with normalized values generated based on absolute values and corpus composition data. The figure shows the aggregated and normalized frequencies for the values shown in Figures 8 and 17. As the generated frequency values match the normalized values of the sums reported in the earlier figure from the original data, we can see that the selected corpus to composition mapping is correct.

4.2.5. Visualize

- Input: Normalized observations
- Output: Set of visualizations based on the result dataset

The "Visualize" step has a preprocessing task, which takes the filter and group configuration from the previous stage as an input, and generates a valid RDF Data Cube data structure definition (DSD) for the input dataset. The DSD defines a set of components that describe the dimensions, measures and attributes related to the observations (see section 3.2). LADA creates two measures: a normalized word frequency with a specific normalization base, and a percentage of occurrences within a dimension group. The first one is simply the asserted or generated normalized frequency selected in the previous step. The second measure could be used, for example, to calculate the proportion of the expression "until" in the ARCHER corpus between 1700–1799 in the genres "Legal", "Sermons", "News", "Letters" and "Drama". The normalized frequency measure is useful for analysing change in the overall frequency of the expression over time, whereas percentage values are better suited for showing variation in the use of the expression across e.g. genres or grammatical functions.

The visualization UI consists of one or more individually configurable visualization blocks (see Figure 21). Each block is an instance of a generic web-based RDF cube visualization component that uses the data structure definition together with the input dataset, to create a data-based configuration form for modifying the query parameters that drive the generated visualization (Figure 20).



Configure visualization

Measure
normalized frequency ▾

Filter
Corpus ▾

HC ×

First dimension
Time period ▾

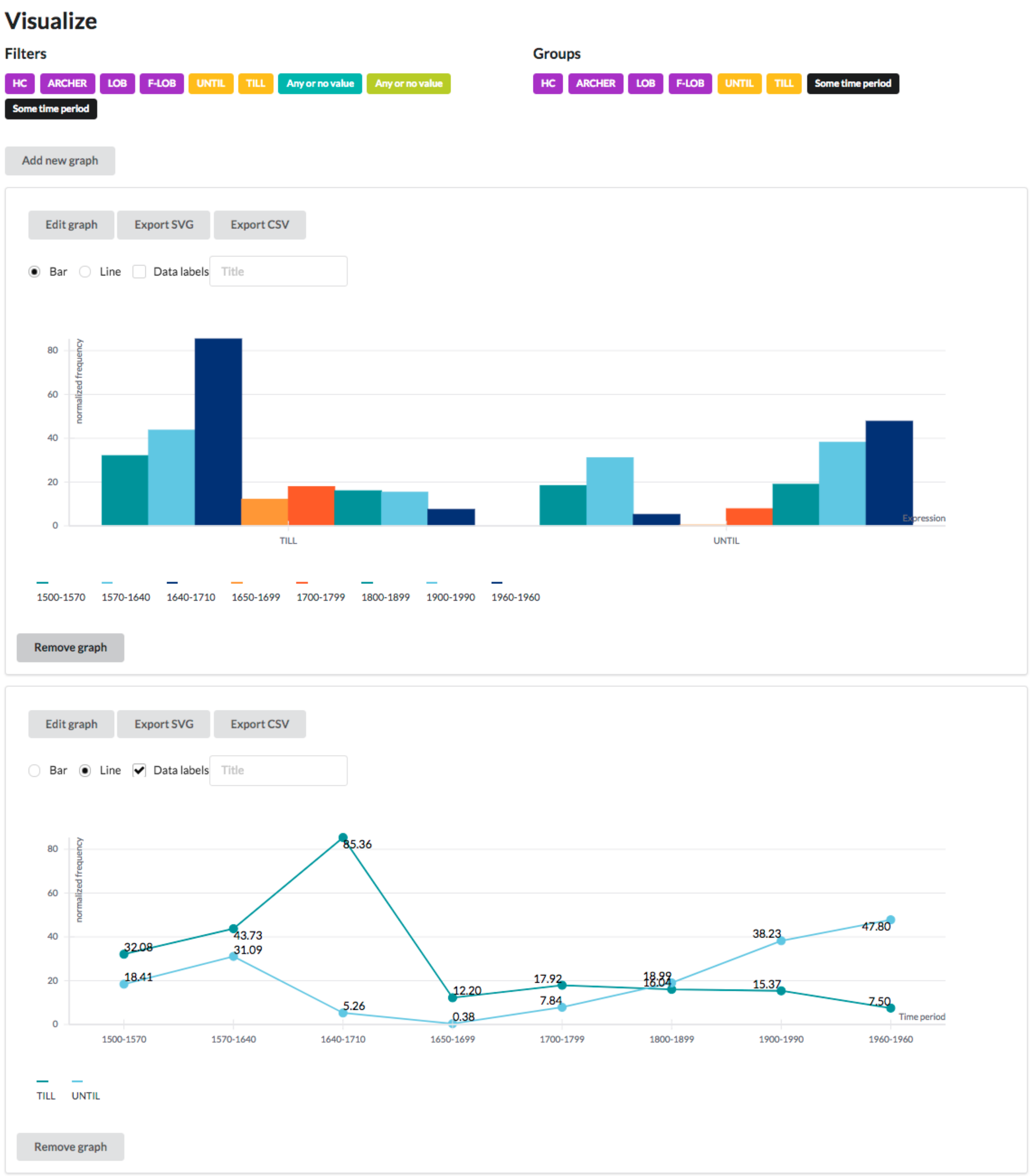
Second dimension
Expression ▾

TILL × UNTIL ×

Figure 20. Visualization configuration for a single graph. Selection values are based on the generated data structure definition of the final aggregated dataset.

The user must first select one of the two generated measures to be projected as the values on the Y axis. The values on the X axis are determined by the first dimension, whereas the second dimension selection can be used to add additional groupings, i.e. measure values, to each X axis value. By default all of the dimension values are displayed, but the user can also filter them, as shown in Figure 20 for expressions. Finally, the user can add a global filter using one of the available dimensions and its values. For example, if the input dataset contains observations from multiple corpora, global filtering can be used to create separate visualizations for each corpus (see Figure 20).

After configuring the query parameters, the user is able modify the visualization output by selecting between line and bar charts, toggling data point labels and adding a title for the figure. The goal is that visualizations generated with the LADA could be used “as is” as part of publications related to LADA experimentation. Finally, the user can export individual visualizations as PNG or SVG images and store them for later use. It is also possible to export the data behind every image in a CSV format for further analysis.



● Bar ○ Line ☐ Data labels

○ Bar ● Line ☑ Data labels

Figure 21. The “Visualize” step with two visualization components. The diachronic development shown in the figure is discussed in section 5.

4.3. LADA Experiment Exchange Package (LEEP)

LADA implements export and import functionality through a custom format called LADA Experiment Exchange Package (LEEP). This feature allows the user to create and restore the full state of the LADA workflow in a single file. LEEP contains all the inputs and outputs of each LADA workflow stage as well as any data, such as mappings, the imported annotated data files, and any visualizations exported as image files. In addition to this, LEEP also includes provenance metadata related to the experiment expressed using the PROV-O[19] ontology, which is a graph that connects all the different parts of the experiment together.

PROV-O is a W3C recommended specification for describing the lineage of any type of data. The main concepts of PROV-O are Agent, Activity and Entity, and basic provenance is modelled as a set of activities performed by agents that are connected with entities that are used and generated by the activities (Figure 22).

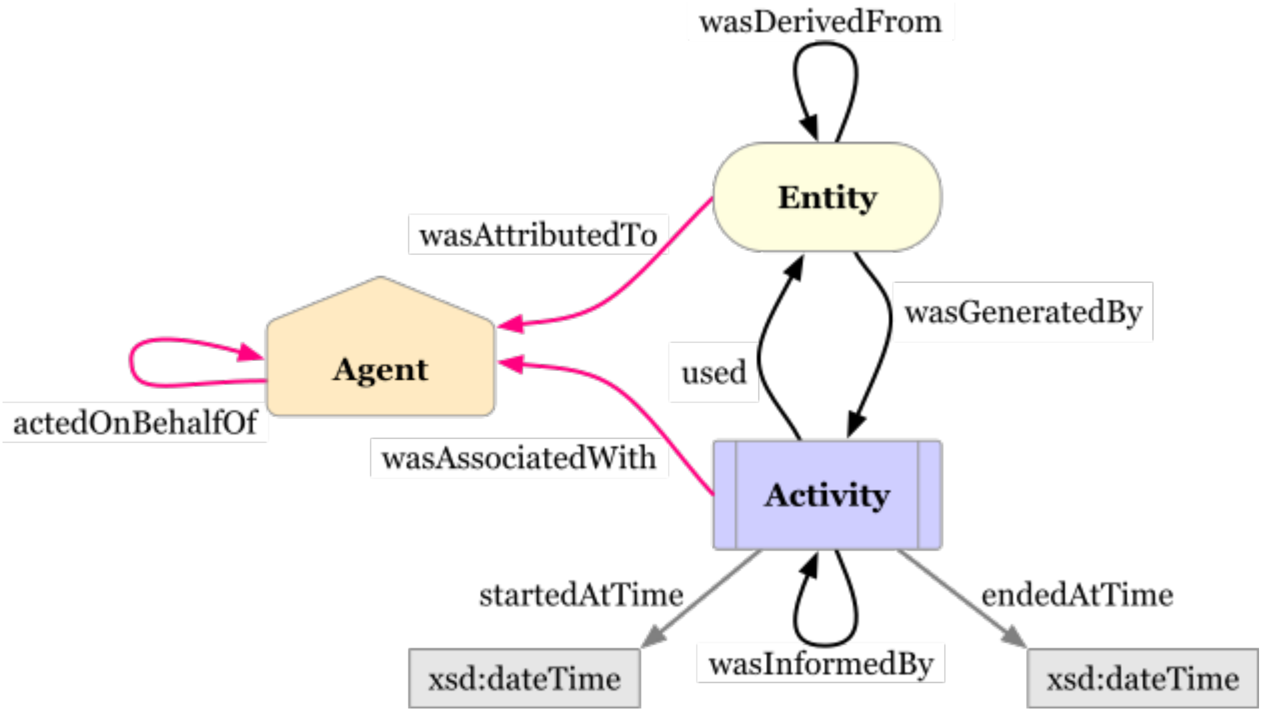


Figure 22. Core concepts of the PROV-O model (quoted from the PROV-O specification available at https://www.w3.org/TR/prov-o/).

Figure 23 shows a provenance graph example that is generated based on the actions performed by the user in the "Select" stage of the LADA workflow. Every sheet is transformed into a distinct dataset, but the example only shows the dataset that represents the data from the first sheet. Ovals represent RDF resources identified with URIs (e.g. `lcd-data:publication`). For example, `lcd-data:publicationfile` is a resource that describes the actual PDF file, which might or might not be stored in the system.

The input dataset, i.e. the annotated data file, was transformed through an activity performed by the LADA software agent with version number 0.1, into an RDF graph identified by the URI `lada:exp1_pub1_table1_ds`. Using the data from the LCD, we can also assert that the annotated data file was derived from another file with an identifier that points back to the LCD, and that the original file was extracted from a certain publication that is also described in the LCD, which contains bibliographic information about the publication and, in the best-case scenario, the original published PDF.

[19] https://www.w3.org/TR/prov-o/

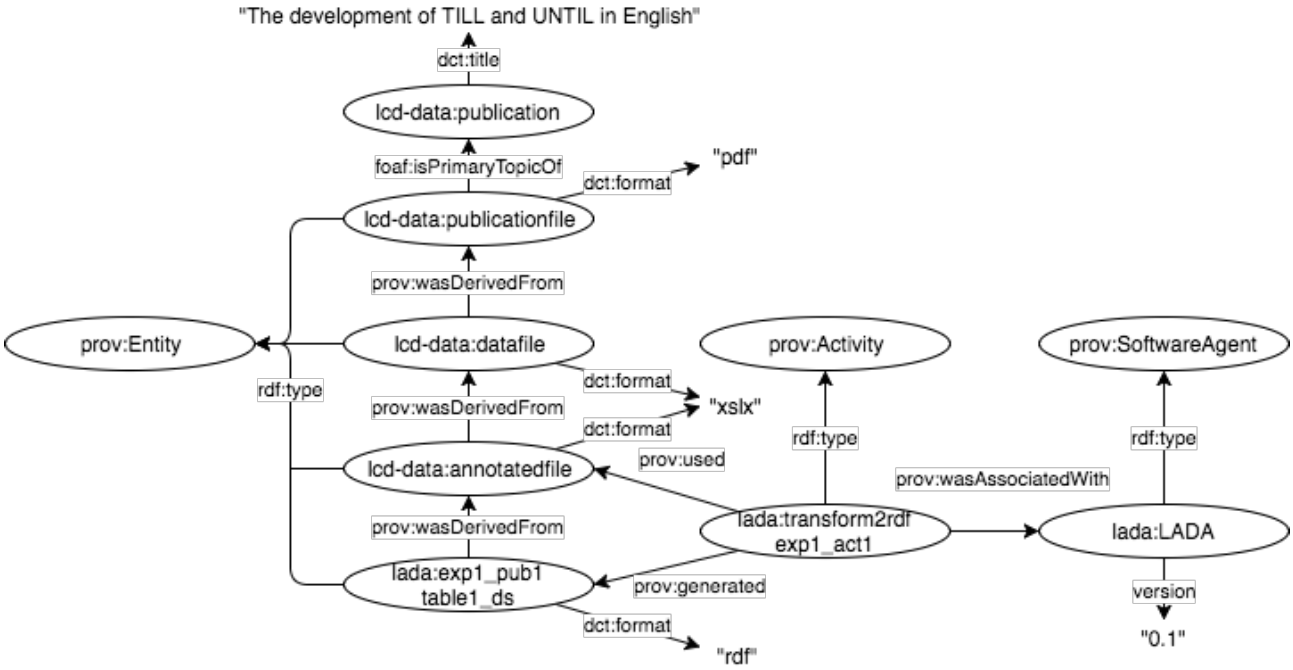


Figure 23. Example provenance information recorded based on the actions performed in the first stage of the LADA workflow, where an annotated Excel data file was converted with the LADA software version 0.1 into an RDF representation.

This kind of information is extremely valuable when trying to solve and remedy data errors that are related to a human error in either the extraction or the annotation part of the data management process. The explicit link between datasets and publications also provides researchers with a convenient way to gather all the material needed to validate, scrutinize and reproduce the results of the experiment.

5. Case study: history of the English connectives

To further illustrate the potential of LADA for meta-analysis, we take a bird's-eye view of the history of English connectives, of which our earlier example of *till* and *until* forms a part. Connectives represent an active area of English grammar that has acquired new members at various periods throughout history, while some other members have faded into obsolescence. In particular, a number of new connectives were introduced to English in the Late Middle English period, which led to increased competition between different forms (Rissanen 2003). We will now analyse the overall situation in the *Helsinki Corpus* from Late Middle to Early Modern English, combining data from several articles written by Matti Rissanen (see the list of primary sources).[20] After conducting a search in the LCD, we import the data tables into LADA. Next, we filter the tables by corpus, time period and expression, selecting all of the available connectives. In the review step, we remove some duplicate tables that have been included in more than one article. After normalizing the frequencies, we are able to visualize the results. The entire process is shown in detail in our video tutorial,[21] and the resulting visualization is displayed in Figure 24.

[20] We are very grateful to the late Professor Emeritus Matti Rissanen for providing us with provisional LCD entries based on his extensive work in the field, and our research assistants, particularly Agata Dominowska, for refining the entries and entering the data in the LCD.

[21] https://doi.org/10.5281/zenodo.1202371

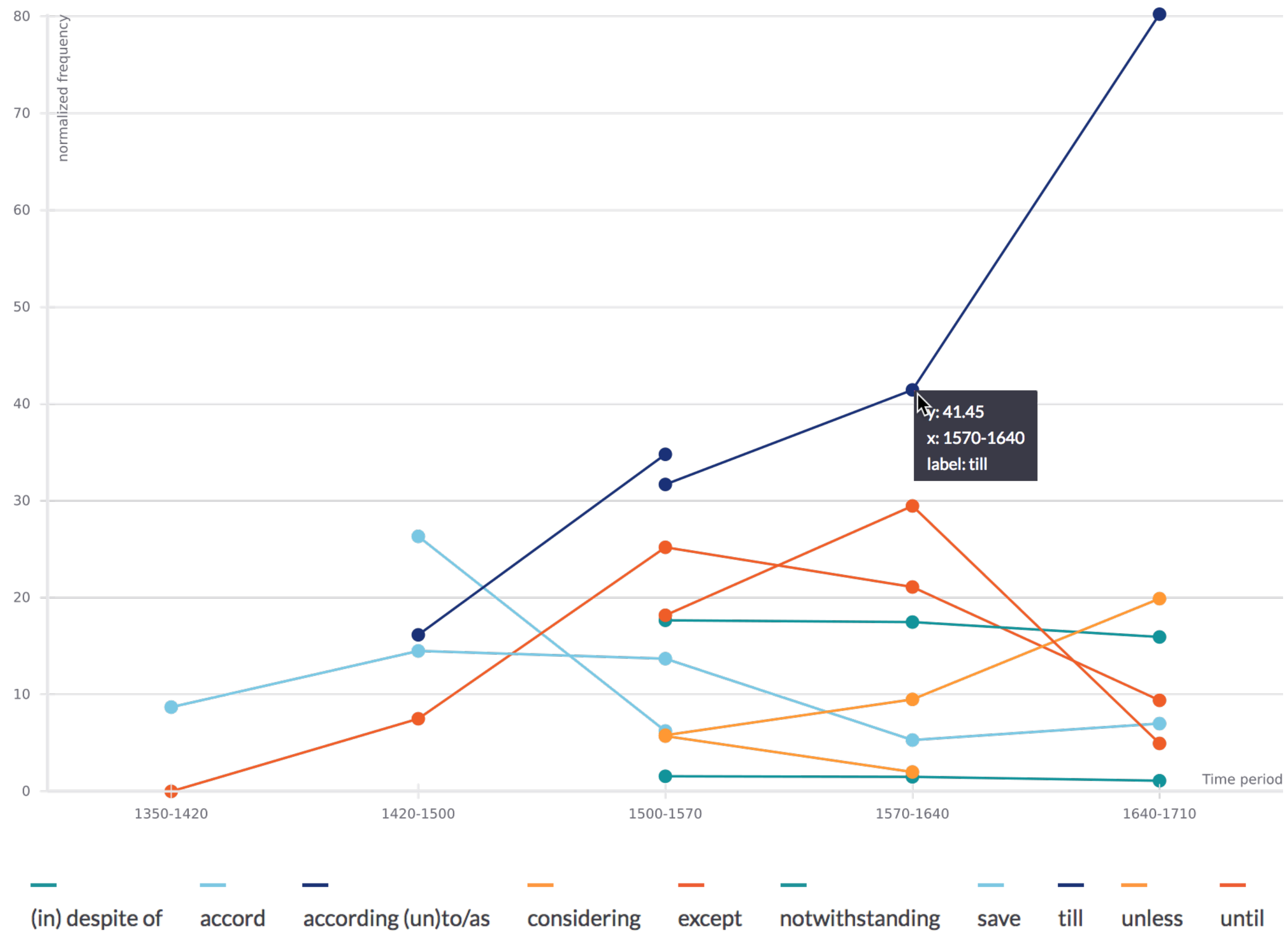


Figure 24. Some English connectives in the Helsinki Corpus, 1350–1710 (figure generated using LADA). The interactive visualization provides more information on mouse hover.

We are immediately able to perceive several changes of interest in Figure 24. The most dramatic of these seems to be the increase in the frequency of *till* at the end of the Early Modern English period, matched by a corresponding decrease in the frequency of *until*. Another competing pair of connectives that undergoes change during this period is *except* and *unless*, the latter of which gains ground at the expense of the former. What could be the reason for this heightened activity at the end of the Early Modern English period? Earlier research in other areas of English has posited a so-called Civil War effect (Raumolin-Brunberg 1998, Lijffijt et al. 2012), which refers to an increase in the rate of language change during the English Civil War (1642–1651) owing to an increased number of weak social ties. We might hypothesize that the Civil War could have influenced the developments in connectives as well.

However, we are also interested in later developments. Thanks to Matti Rissanen's pioneering work, we have data on frequency changes in *till* and *until* that extend beyond Early Modern English, all the way to Present-day English, with the help of several different corpora. While all of this data is presented in a single article (Rissanen 2005), they are spread out over a number of tables, which only provide absolute values rather than normalized frequencies. Using the LCD and LADA, we can easily combine the data from different corpora, normalize the absolute values and explore the results in a single visualization (Figure 21 above).

Figure 21 reveals that the dramatic change at the end of the Early Modern English period in the *Helsinki Corpus* (1640–1710) is reversed in the Late Modern English data from ARCHER. In fact, *until* becomes more frequent than *till* by the period 1800–1899, and the frequencies continue to diverge, so that by the 1990s in the FLOB corpus, *till* has been marginalized, while the frequency of *until* has more or less stabilized. Rissanen (2005) shows that these developments can be connected with changes in the genres and functions in which the two connectives are used: in ARCHER, *until* occurs most often in legal texts at first and is later generalized to other registers, while *till* starts out as more of a feature of news texts and is relegated to the more informal registers of drama and letters by the twentieth century. In the Present-day English *British National Corpus*, *till* is almost exclusively a spoken feature and is used more often as a preposition than as a subordinator, whereas *until* is more frequent in written than spoken language and functions more often as a subordinator than as a preposition.

This case study illustrates the ease with which data from different sources can be combined and explored in LADA. By focusing on a single topic, the history of connectives, and drawing data from an extensive body of work of a single scholar, we were able to overcome some potential issues related to the comparability of data: in his research, Matti Rissanen had made every effort to ensure that the data was comparable across time periods, which enables us to form a general overview of the history of this particular domain of English grammar. It is true that data-related issues may turn out to be more severe with other research topics and datasets, especially when data is taken from various corpora and databases. It is not our purpose to trivialize these matters, and we would like to emphasize that while these resources facilitate the re-use of existing data, it is crucially important to know both what the original data is like and what principles the researchers had followed in data collection.

Our case study is admittedly based on a rather small number of research articles, and its main purpose is to illustrate the functionality of the LCD and LADA. However, as the number of articles in the LCD grows, researchers will be able to conduct large-scale meta-analyses of multiple changes in the English language, which could result in significant advances in theoretical accounts of language change as well as descriptions of the history of English (see e.g. Nevalainen et al. in preparation).

6. Discussion

Meta-analyses have a great potential to enrich our view of linguistic variation and change. In this paper, we have given one example of meta-analysis, providing an overview of the research done by an individual scholar over a long period of time. In the future, we can start asking more fundamental questions about the sociolinguistic reality of language change. Some related questions on which meta-analyses can potentially shed more light include the following (e.g. Trudgill 2011; Kesäniemi et al. 2018): i) which population groups are eager to embrace new grammatical forms and which are more conservative in their usage, ii) how is language change affected by contact situations between speakers of different dialects and languages, and iii) how do cataclysmic social events, such as wars, famines or pandemics, contribute to language change? As we have argued, the investigation of such questions requires a database in which earlier research results are freely available to the research community, and our solution, the Language Change Database, is designed to meet this challenge. In addition to providing baseline data for new research, the LCD will also facilitate the accumulation of knowledge within the field of historical corpus linguistics, and it will thus be of service to anyone interested in specific linguistic developments from Old English to Present-day English.

Extracting and annotating historical research data for the LCD is a time-consuming task and requires precision and knowledge of corpus linguistics to produce something than can be utilized for further research activities. However, through collaborative services like the LCD, the workload can be distributed and the results shared in a way that promotes re-use through APIs and a common self-describing data format (RDF). We envision that the LCD can also become part of the evolving ways of scholarly communication as a service where researchers can contribute their own results and data in a way that links it to the other LCD data. The LCD allows researchers to publish their data tables separately, eliminating the need for an additional extraction task. While our approach to annotating data tables is geared towards files created using Microsoft Excel, it is possible to add support for other spreadsheet formats as well. However, since the specification for Excel's current data format is openly available, there are many open source tools that can be utilized for reading and processing the annotated files.

The dimensions currently available for annotations were selected based on the evaluation of the data tables in the LCD. The annotation scheme could be extended to include e.g. the social categories already present in the LCD data model as well as text level information, which would provide the user with a more granular view to the data. There are also many cases where a cell in a data table refers to the ambiguous value "other". For example, in Rissanen (2009), Table 2 reports frequencies in the genres "official" and "other". A similar example can be found in Rissanen (2014b), Table 4, where the linguistic variables under investigation are reported as "According (un)to/as" and "Accord (other forms)". The first case is something that might be possible to handle with the existing data, since it refers to a finite and known set of values (genres) in a specific corpus. In the latter case, only the first expression is re-usable in most situations, as the "other forms" of *accord* have not been specified.

The workflow implemented in LADA is a simplified view to the meta-analysis process and it still requires a knowledgeable user in order to create valid new datasets. For example, when working with data from multiple corpora, the user needs to know and understand the structure and compilation principles behind each corpus in order to make sound decisions as to whether or not the results are suitable for integration. We acknowledge that making straightforward comparisons of observed word frequencies can be a source of statistical error, but taking into consideration the exploratory focus of the tool, we argue that the approach is viable for discovering possibly interesting phenomena over an aggregated dataset.

When normalizing absolute frequency values found in the data tables, it is important to use the right version of the corpus. These versions can range from corpora with planned and scheduled update cycles to custom-built corpora. In any case, there can be significant differences in size between different versions of the same corpus. This affects the LADA normalization step where the user is required to select the corpus composition to be used for word frequency normalization. If the article does not explicitly mention the specific corpus version, the user must deduce it from his or her previous knowledge or use the contextual information available, such as the publication year of the article, to narrow down the possible alternatives.

Once the data is available in the RDF Data Cube format, it can be transformed and queried using standard semantic web tools. There are also approaches that utilize RDF data directly. Kalampokis et al. (2014) demonstrate how RDF data cubes can be integrated with the popular R statistics software through SPARQL queries with the OpenCube Toolkit. This toolkit also supports exporting data in the RDF Data Cube format, which makes it straightforward to integrate the analysis data with the source data. LADA does not currently support the analysis of produced datasets beyond basic visualizations, but the generated dataset can be easily exported in CSV format for further analysis in statistical software, such as R (cf. Flanagan 2017).

7. Conclusion

In this paper, we have argued that historical corpus linguistics has come of age: existing data and technologies allow us to ask ever more challenging and fundamental questions concerning language change. For this purpose, we have introduced a database that can be used as a baseline for a variety of research questions on the history of English: the Language Change Database (LCD). At present, the LCD includes data from c. 300 studies, and it can already be used to carry out small-scale meta-analyses, as was discussed in section 5. When compiling the LCD, we have paid special attention to the so-called FAIR principles: findability, accessibility, interoperability and reusability (Wilkinson et al. 2016), with the aim of making the data freely available to the research community in as accessible a form as possible.

We have demonstrated one possible of way of re-using the data in the LCD by creating a tool called LADA, which allows researchers to conduct experiments with data tables extracted from LCD publications. LADA provides a workflow that supports the steps that are typically taken in meta-analysis: defining the criteria for eligible data, transforming existing data into a format that can be used for automated processing, harmonizing and finally combining the selected data into a new dataset. The LCD and LADA make use of Linked Data and semantic web technologies, thus anticipating a future where both research results and research data are regularly presented in a way that allows for their verification, validation and re-use – not just in English historical corpus linguistics, but in any language, or any field in the digital humanities.

Acknowledgments

This research was funded by the Academy of Finland, grant number 276349. We are grateful to the two anonymous reviewers for their valuable comments and the conference audiences at ISLE 4 and 5, ICAME 38, and HELDIG Digital Humanities Summit 2017.

References

Primary sources

Rissanen, M. (2002) *Despite* or *notwithstanding*? On the development of concessive prepositions in English. In A. Fischer, G. Tottie and H. M. Lehmann (eds.) *Text Types and Corpora: Studies in Honour of Udo Fries* 191–203. Tübingen: Gunter Narr.

Rissanen, M. (2005) The development of *till* and *until* in English. In J. Fisiak and H.-K. Kang (eds.) *Recent Trends in Medieval English Language and Literature in Honour of Young-Bae Park, Vol. I* 75–92. Seoul: Thaehaksa.

Rissanen, M. (2009) Grammaticalisation, contact and adverbial connectives: the rise and decline of *save*. In S. Watanabe and Y. Hosoya (eds.) *English Philology and Corpus Studies: a Festschrift in Honour of Mitsunori Imai to Celebrate His Seventieth Birthday* 135–152. Tokyo: Shohakusha.

Rissanen, M. (2011) On the long history of English adverbial subordinators. In A. Meurman-Solin and U. Lenker (eds.) *Connectives in Synchrony and Diachrony in European Languages* (Studies in Variation, Contacts and Change in English 8). Helsinki: VARIENG. Retrieved on 12 April 2018 from http://www.helsinki.fi/varieng/series/volumes/08/rissanen/

Rissanen, M. (2012) Grammaticalisation, contact and corpora: on the development of adverbial connectives in English. In I. Hegedűs and A. Fodor (eds.) *English Historical Linguistics 2010: Selected Papers from the Sixteenth International Conference on English Historical Linguistics (ICEHL 16), Pécs, 23–27 August 2010* 131–152. Amsterdam: John Benjamins.

Rissanen, M. (2014a) From medieval to modern: on the development of the adverbial connective *considering (that)*. In I. Taavitsainen, M. Kytö, C. Claridge and J. Smith (eds.) *Developments in English: Expanding Electronic Evidence* 98–115. Cambridge: Cambridge University Press.

Rissanen, M. (2014b) On English historical corpora, with notes on the development of adverbial connectives. In A. Sintes and S. Hernández (eds.) *Diachrony and Synchrony in English Corpus Linguistics* 109–140. Bern: Peter Lang.

Secondary sources

Battle, S., Wood, D., Leigh, J. and Ruth, L. (2012) The Callimachus Project: RDF as a web template language. In J. F. Sequeda, A. Harth and O. Hartig (eds.) *Proceedings of the Third International Conference on Consuming Linked Data* 1–14. CEUR-WS.org. urn:nbn:de:0074-905-3

Berners-Lee, T. (2006) Linked data. In *Design Issues: Architectural and Philosophical Points*. Retrieved on 12 April 2018 from https://www.w3.org/DesignIssues/LinkedData.html

Bizer, C., Heath, T. and Berners-Lee, T. (2009) Linked data – the story so far. *International Journal on Semantic Web and Information Systems* 5(3): 1–22.

Blythe, R. A. and Croft, W. (2012) S-curves and the mechanisms of propagation in language change. *Language* 88(2): 269–304.

Chaudron, C. (2006) Some reflections on the development of (meta-analytic) synthesis in second language research. In J. M. Norris and L. Ortega (eds.) *Synthesizing Research on Language Learning and Teaching* 323–339. Amsterdam: John Benjamins.

Chiarcos, C., Hellmann, S. and Nordhoff, S. (2011) Towards a Linguistic Linked Open Data cloud: the Open Linguistics Working Group. *Traitement Automatique des Langues* 52(3): 245–275.

Chiarcos, C., McCrae, J., Cimiano, P. and Fellbaum, C. (2013) Towards open data for linguistics: linguistic linked data. In A. Oltramari, P. Vossen, L. Qin and E. Hovy (eds.) *New Trends of Research in Ontologies and Lexical Resources* 7–25. Berlin: Springer.

Durrant, P. (2014) Corpus frequency and second language learners' knowledge of collocations. *International Journal of Corpus Linguistics* 19(4): 443–477.

Flanagan, J. (2017) Reproducible research: strategies, tools, and workflows. In T. Hiltunen, J. McVeigh and T. Säily (eds.) *Big and Rich Data in English Corpus Linguistics: Methods and Explorations* (Studies in Variation, Contacts and Change in English 19). Helsinki: VARIENG. Retrieved on 12 April 2018 from http://www.helsinki.fi/varieng/series/volumes/19/flanagan/

Francis, W. and Kučera, F. (1979) *Manual of Information to Accompany A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*. Providence, RI: Department of Linguistics, Brown University.

HC = *Helsinki Corpus of English Texts* (1991) Compiled by M. Rissanen (Project leader), M. Kytö (Project secretary); L. Kahlas-Tarkka, M. Kilpiö (Old English); S. Nevanlinna, I. Taavitsainen (Middle English); T. Nevalainen, H. Raumolin-Brunberg (Early Modern English). Helsinki: Department of Modern Languages, University of Helsinki.

Johansson, S., Leech, G. and Goodluck, H. (1978) *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*. Oslo: Department of English, University of Oslo.

Kalampokis, E., Nikolov, A., Haase, P., Cyganiak, R., Stasiewicz, A., Karamanou, A., Zotou, M., Zeginis, D., Tambouris, E. and Tarabanis, K. A. (2014) Exploiting linked data cubes with OpenCube Toolkit. In M. Horridge, M. Rospocher and J. van Ossenbruggen (eds.) *International Semantic Web Conference (Posters & Demos)* 137–140. CEUR-WS.org. urn:nbn:de:0074-1272-7

Kesäniemi, J., Vartiainen, T., Säily, T. and Nevalainen, T. (2018) Open science for English historical corpus linguistics: introducing the Language Change Database. In E. Mäkelä, M. Tolonen and J. Tuominen (eds.) *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference, Helsinki, Finland, March 7–9, 2018* (CEUR Workshop Proceedings 2084) 51–62. CEUR-WS.org. Retrieved on 16 November 2018 from http://ceur-ws.org/Vol-2084/paper4.pdf

Lijffijt, J., Säily, T. and Nevalainen, T. (2012) CEECing the baseline: lexical stability and significant change in a historical corpus. In J. Tyrkkö, M. Kilpiö, T. Nevalainen and M. Rissanen (eds.) *Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources* (Studies in Variation, Contacts and Change in English 10). Helsinki: VARIENG. Retrieved on 12 April 2018 from http://www.helsinki.fi/varieng/series/volumes/10/lijffijt_saily_nevalainen/

Lohmann, S., Negru, S., Haag, F. and Ertl, T. (2016) Visualizing ontologies with VOWL. *Semantic Web* 7(4): 399–419.

McCrae, J., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., de Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S. and Osenova, P. (2016) The Open Linguistics Working Group: developing the Linguistic Linked Open Data cloud. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk and S. Piperidis (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* 2435–2441. Paris: ELRA.

McCrae, J., Fellbaum, C. and Cimiano, P. (2014) Publishing and linking WordNet using Lemon and RDF. In C. Chiarcos, J. McCrae, P. Osenova and C. Vertan (eds.) *Proceedings of the 3rd Workshop on Linked Data in Linguistics*. urn:nbn:de:0070-pub-27327797

Meroño-Peñuela, A., Ashkpour, A., Rietveld, L., Hoekstra, R. and Schlobach, S. (2012) Linked humanities data: the next frontier? A case-study in historical census data. In T. Kauppinen, L. C. Pouchard and C. Keßler (eds.) *Proceedings of the 2nd International Workshop on Linked Science 2012 (LISC2012)*. CEUR-WS.org. urn:nbn:de:0074-951-6

Nevalainen, T., Säily, T., Vartiainen, T., Liimatta. A. and Lijffijt, J. (in preparation) History of English as punctuated equilibria? *Journal of Historical Sociolinguistics*.

Nevalainen, T., Vartiainen, T., Säily, T., Kesäniemi, J., Dominowska, A. and Öhman, E. (2016) Language Change Database: a new online resource. *ICAME Journal* 40: 77–94. doi:10.1515/icame-2016-0006

Newberry, M. G., Ahern, C. A., Clark, R. and Plotkin, J. B. (2017) Detecting evolutionary forces in language change. *Nature* 551: 223–226.

Norris, J. and Ortega, L. (2000) Effectiveness of L2 instruction: a research synthesis and quantitative meta-analysis. *Language Learning* 50(3): 417–528.

Raumolin-Brunberg, H. (1998) Social factors and pronominal change in the seventeenth century: the Civil-War effect? In J. Fisiak and M. Krygier (eds.) *Advances in English Historical Linguistics* 361–388. Berlin: Mouton de Gruyter.

Renehan, A., Tyson, M., Egger, M., Heller, R. and Zwahlen, M. (2008) Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. *Lancet* 371(9612): 569–578.

Reynolds, K., Lewis, B., Nolen, J., Kinney, G., Sathya, B. and He, J. (2003) Alcohol consumption and risk of stroke: a meta-analysis. *JAMA* 289(5): 579–588.

Rissanen, M. (2003) On the development of English adverbial connectives. In M. Ukaji, M. Ike-Uchi and Y. Nishimura (eds.) *Current Issues in English Linguistics* (Special Publications of the English Linguistic Society of Japan 2) 229–247. Tokyo: Kaitakusha.

Trudgill, P. (2011) *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

van Assem, M., Menken, M. R., Schreiber, G., Wielemaker, J. and Wielinga, B. (2004) A method for converting thesauri to RDF/OWL. In S. A. McIlraith, D. Plexousakis and F. van Harmelen (eds.) *The Semantic Web – ISWC 2004* 17–31. Berlin: Springer.

Wang, A., Wang, S., Zhu, C., Huang, H., Wu, L., Wan, X., Yang, X., Zhang, H., Miao, R., He, L., Sang, X. and Zhao, H. (2016) Coffee and cancer risk: a meta-analysis of prospective observational studies. *Scientific Reports* 6: 33711.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., ’t Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, B. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3: 160018.

Wood, D. (2015) Semantic composition of disparate data in GeoHealthUS for navigation, display and analysis. Poster, *International Semantic Web Conference (ISWC) 2015*. doi:10.13140/RG.2.2.33948.59528