Open Science for English Historical Corpus Linguistics: Introducing the Language Change Database

Joonas Kesäniemi[1][0000−0002−3770−0006], Turo Vartiainen[2][0000−0002−4760−750X], Tanja Säily[3][0000−0003−4407−8929], and Terttu Nevalainen[2][0000−0003−3088−4903]

[1] Helsinki University Library, Helsinki, Finland,
joonas.kesaniemi@helsinki.fi
[2] University of Helsinki, Department of Languages, Helsinki, Finland
[3] University of Helsinki, Department of Digital Humanities, Helsinki, Finland

Abstract. This paper discusses the development of an open-access resource that can be used as a baseline for new corpus-linguistic research into the history of English: the Language Change Database (LCD). The LCD draws together information extracted from hundreds of corpus-based articles that investigate the ways in which English has changed in the course of history. The database includes annotated summaries of the articles, as well as numerical data extracted from the articles and transformed into machine-readable form, thus providing scholars of English with the opportunity to study fundamental questions about the nature, rate and direction of language change. It will also make the work done in the field more cumulative by ensuring that the research community will have continuous access to existing results and research data.
We will also introduce a tool that takes advantage of this new source of structured research data. The LCD Aggregated Data Analysis workbench (LADA) makes use of annotated versions of the numerical data available from the LCD and provides a workflow for performing meta-analytical experimentations with an aggregated set of data tables from multiple publications. Combined with the LCD as the source of collaborative, trusted and curated linked research data, the LADA meta-analysis tool demonstrates how open data can be used in innovative ways to support new research through data-driven aggregation of empirical findings in the context of historical linguistics.

Keywords: historical corpus linguistics, re-use of research data, linked data, meta-analysis, software tools

1 Introduction

1.1 English Corpus Linguistics

English corpus linguistics provides an early example of open science in the field of linguistic research. The idea behind the creation of the first English digital

corpora was for researchers to be able to share data with each other by distributing these resources free of charge. The history of English corpus linguistics extends back to 1964, when the first structured electronic corpus of the English language, the *Brown Corpus*, was published (Francis & Kučera 1964). The Brown Corpus was a synchronic corpus of American English, and although it did attract the attention of some empirically-minded linguists, and inspired the compilation of many other corpora in the decades to follow, the time was not yet ripe for corpus-based research to become truly mainstream.

Things started to change in the late 1970s and the 1980s thanks to the rapid development of personal computers. The first synchronic corpus of British English, the *Lancaster-Oslo/Bergen Corpus*, was published in 1978 (Johansson, Leech & Goodluck 1978), and the first diachronic corpus of English, the *Helsinki Corpus of English Texts*, in 1991. The *Helsinki Corpus* (HC) covered the entire history of the English language from Old English to the end of the Late Modern period, and it has since been used as a source of hundreds of academic studies describing variation and change in English.

Thanks to the rapid technologization of society, the success of diachronic corpus linguistics has continued to the present day, and the study of language change has benefited from the compilation of many large-scale corpora, such as the *Corpus of Historical American English* (COHA, Davies 2012), and the *Corpus of Late Modern English Texts* (CLMET, De Smet 2005). Corpus-based research articles on the history of English number in the thousands, and they shed light on various questions concerning the ways in which the English language has changed during its history, which spans some 1,300 years into the past. Importantly, the quantitative evidence offered by corpora has allowed linguists to study language change as a phenomenon that is typically manifested through gradual changes in the frequency of the studied items rather than categorical splits.

1.2 LCD: Making New Use of Old Research

Now, almost thirty years after the publication of the *Helsinki Corpus*, we feel that the field of English historical corpus linguistics has come of age. We therefore propose that in addition to the more detailed questions about specific areas of grammar and lexis, it is time to start asking new kinds of questions that can offer us insight into some of the most profound processes of language change. These concern the nature, rate and direction of change, and the following questions are just some examples of the more general topics of inquiry that would be worthwhile to study from a corpus-linguistic perspective.

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

While these topics have been raised before (e.g. Labov 1994, 2001, 2010; Trudgill 2011; see Tagliamonte 2012 for a summary), their systematic investigation from a corpus-linguistic perspective has been severely constrained by lack of data. What we need, therefore, is unrestricted access to the large body of past work on diachronic corpus-based research. The problem is, of course, that this research has not been collected into a single repository; rather, it has been published in dozens of journals, edited volumes and working papers, and some of these publications, especially the older ones, may be extremely hard to locate these days. Furthermore, even if the relevant piece of research is accessible to the research community, it is usually in a form that does not allow for its easy reuse: the data are typically expressed numerically as tables or visually as graphs in PDF files, and converting this information into a form that can later be used as a basis for new research is a very labour-intensive, and therefore expensive, process. Finally, even if an individual researcher or a research group made the effort to transform some data into a machine-readable form, it is very unlikely that the modified data would be shared with the research community, and so the overall situation would remain largely unchanged.

What is needed, then, is a focused effort to produce a resource – a database – that would facilitate the reuse of existing data and enable the study of the kinds of fundamental questions outlined above. In this paper, we introduce a resource that is intended to meet this precise challenge: a Language Change Database (LCD) that draws together information about hundreds of articles on the history of English, summarising their results and providing a rich annotation scheme to ensure accurate data retrieval. Importantly, the LCD will include numerical data extracted from the articles in an annotated form, which the end users of the database can download and use in their own research. We will also introduce a tool with which the users of the LCD can choose and sort relevant data and carry out preliminary experimentations and analyses based on them. The functionalities designed for this purpose are included in a tool called LADA (LCD Aggregated Data Analysis workbench). LADA transforms the annotated tables extracted from the LCD into RDF Data Cubes, which can be easily queried, manipulated and aggregated using semantic web and linked data tools (Meroño-Peñuela et al. 2012).

An important aspect of our project is to follow the principles of open science. Although both the LCD and LADA are still under development, we hope to publish them completely free of charge as open-access web applications. The primary users of the LCD and LADA will probably be scholars working on the history of English, but we envision that these resources will also be useful for university students and teachers, providing them with detailed information on the history of English in an easily accessible form, and offering them suggestions for coursework and thesis topics.

We will continue this paper by providing a more detailed overview of the Language Change Database in section 2. Section 3 focuses on the LADA tool, and section 4 concludes the paper with a discussion of some open questions and future prospects.

2 The Language Change Database (LCD)

The LCD is an online resource which draws together earlier corpus-based research on English historical linguistics, with the goal of making the work done in the field more accessible and cumulative by providing comparative baseline data from earlier studies (Nevalainen et al. 2016). The information included in the LCD goes far beyond basic bibliographies in that it includes detailed descriptions of the results of the studies along with quantitative data in tabular format (Fig. 1). The LCD is currently in beta stage, and it includes c. 270 entries that have been selected and prepared by the research team and student assistants. Later the database will be opened to the wider research community, and researchers will be able to add their own studies into the database directly. For copyright reasons, the articles themselves will not be distributed through the LCD (although they may be stored for restricted administrative use), but the metadata describing both the publications and the related data files will be openly available and citable through URI identifiers.

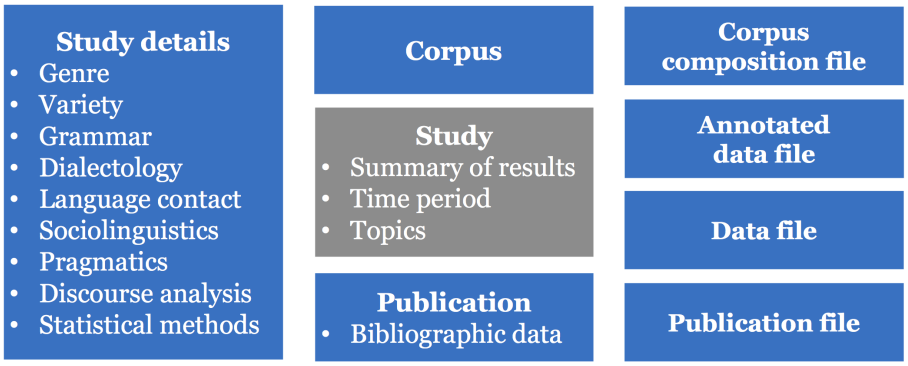

Fig. 1. A schematic view of the information included in the LCD.

The information included in the LCD will be published with permissive licensing and distributed via open interfaces, which makes it possible to integrate it into existing services or build new applications on top of it. For example, our search application prototype (Fig. 2) takes advantage of the LCD's classification schemes and hierarchical grammar concepts (for details see Nevalainen et al. 2016), allowing the user to zoom in on specific kinds of content in the database.

In addition to historical linguistics in general, the data in the LCD is tailored to the needs of researchers interested in statistical modelling, systematic reviews, replication of earlier research and sociolinguistic typologies (e.g. Trudgill 2011). The most recent additions to the LCD data model include corpus composition data and annotated data tables. Corpus compositions are data structures that describe the distribution of word frequencies in a corpus along the dimensions

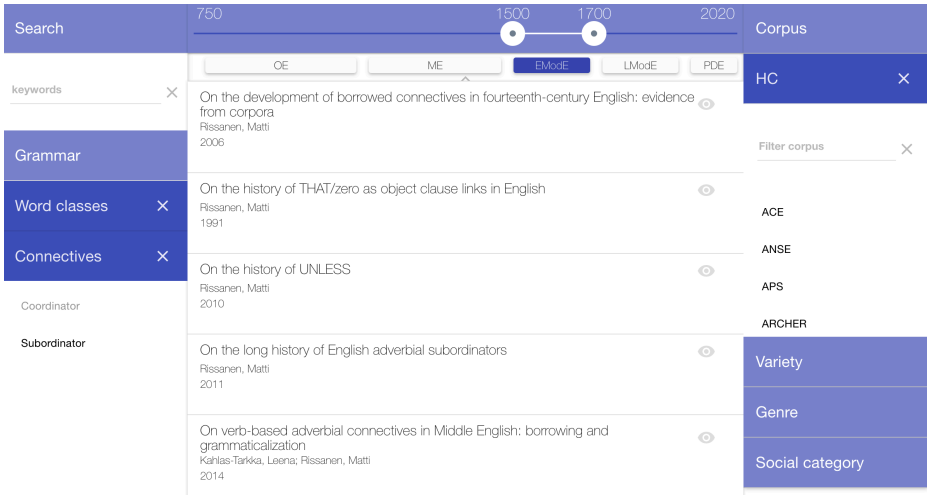

Fig. 2. Search interface prototype. The user has searched for studies of connectives (Grammar > Word classes > Connectives) in Early Modern English (EModE, 1500–1700) based on the *Helsinki Corpus* (HC).

specified in the LCD data model, such as time period, genre and language variety (Kesäniemi et al. in preparation). The annotated data tables are linked to the data model in a way that makes the original research data semantically compatible with the LCD (Fig. 3). Together, the corpus composition files and the annotated data tables turn the LCD into a valuable source of structured data for the purposes of meta-analysis (e.g. Ellis 2006), as described in the next section.

3 The LCD Aggregated Data Analysis Workbench (LADA)

LADA is an application which provides corpus linguists with a systematic workflow to perform exploratory meta-analyses based on earlier research results. The LADA workflow re-uses data from the LCD and provides the user with the tools to filter, review and normalize the existing data in order to create a new aggregated dataset, which can then be visualized and exported for further analysis. In general, the three main stages of meta-analysis are the sampling of relevant studies, the coding of data pertinent to the research question, and the analysis of the new aggregated data (cf. Lipsey & Wilson 2001). The LADA workflow provides its own take on these steps based on the annotated research data available from the LCD.

The LADA workflow starts with a selection stage, where the user picks and imports a set of potentially relevant publications from the LCD. LADA then retrieves the annotated data files linked to these publications and converts them

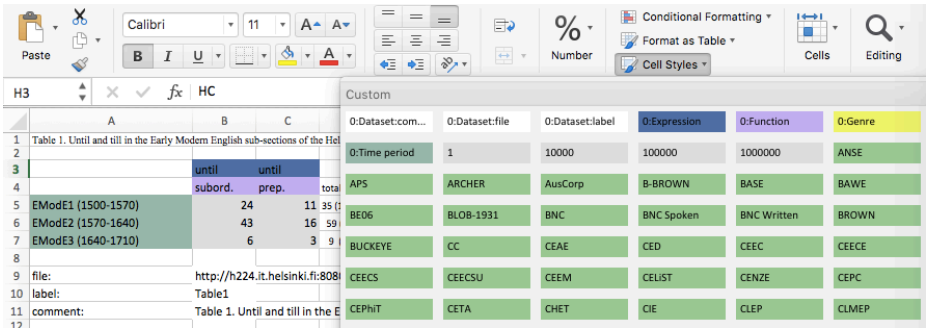

Fig. 3. Annotating data tables with Excel styles. Cells can be annotated by Expression, Function, Genre, Time period, Value, and Corpus. Word frequency values are annotated in grey (e.g. 1 for absolute values), and the Corpus annotation in green links the cell to the corpus entry in the LCD.

into the Resource Description Framework (RDF) format using Data Cube vocabulary. Recommended by the World Wide Web Consortium, the RDF Data Cube vocabulary (2014) is designed for expressing multidimensional data (e.g. statistics) on the web as linked data.

RDF is an especially suitable format for implementing a meta-analysis workflow by virtue of the following characteristics (Bizer et al. 2009):

1. RDF allows for effortless integration of data from multiple datasets, or in our case numerical data from multiple publications.
2. RDF data can be easily mapped to any other RDF data, which facilitates the coding of relevant information for meta-analytic purposes.
3. RDF tools provide a powerful way to query over multiple datasets.
4. RDF can be used to express not only the data, but also its metadata, provenance and ontological constraints.

The following figures illustrate the key stages of the LADA workflow after the initial data transformation: Filter and group (Fig. 4), Review (Fig. 5), and Visualize (Fig. 6). The linguistic case study here is the diachronic development of the connectives *till* and *until*, based on information extracted from several tables in Rissanen (2005). For details on using LADA for meta-analysis, see Kesäniemi et al. (in preparation).

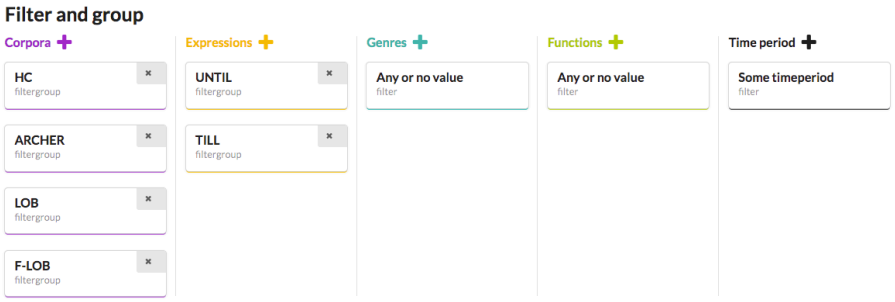

Fig. 4. Data exploration begins by filtering and grouping the initial dataset according to user-specified parameters for e.g. corpus, linguistic expression and genre. LADA shows the number of matching publications, tables and values, which are updated in real time.

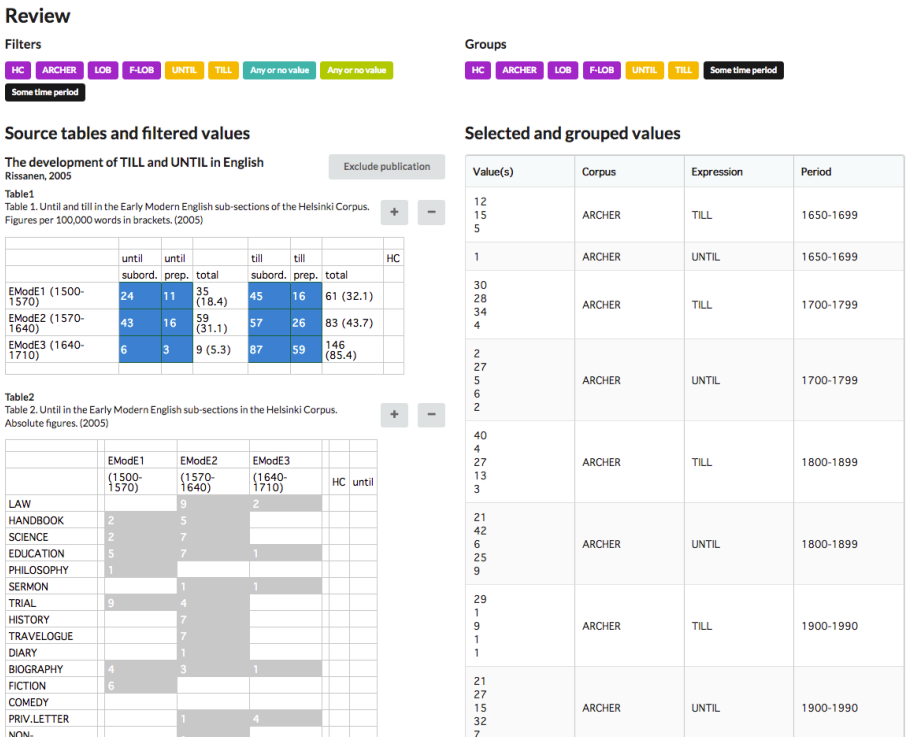

Fig. 5. The Review stage provides an overview of the filtered data and allows the user to remove any irrelevant values. The left-hand column displays the filtered tables and individual cells, while the right-hand column shows the new aggregated data table with user-defined groupings and dimensions.

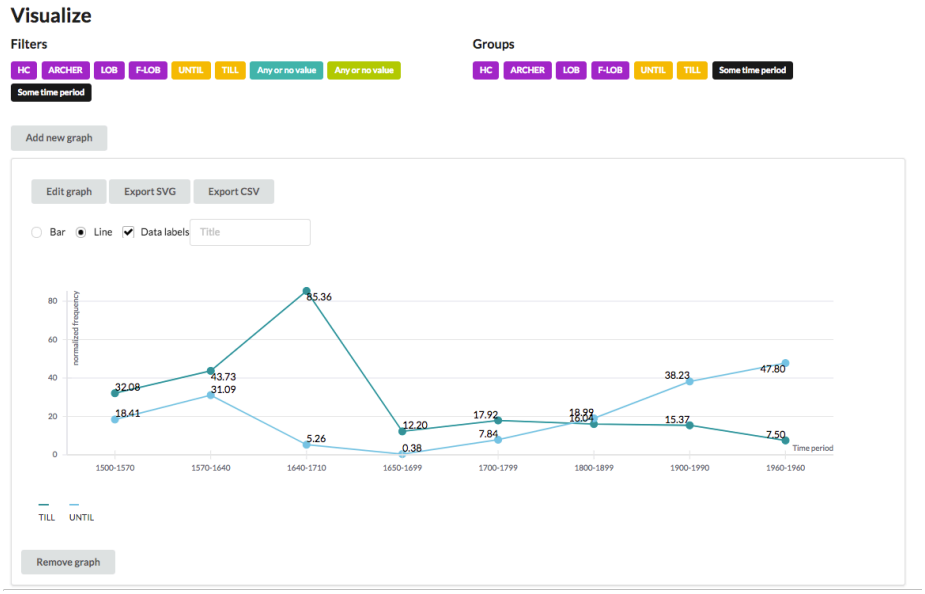

Fig. 6. In the final stage, the new aggregated dataset is visualized. The researcher is now able to see the development of till and until across multiple corpora over time. This dataset can be exported in CSV format for further analysis.

4 Discussion and Conclusions

English corpus linguists were early adopters of the latest digital technology in the era of big mainframe computers. Both corpus linguistics and digital technology have since come a long way. The LCD summarizes the results obtained in English historical corpus linguistics in the past decades to make this work based on shared data sources more accessible to the growing but dispersed research community. A database like the LCD cannot, and is not intended to, replace the reading of the published articles because it only provides a synopsis of their contents. However, by providing access to the numerical information reported in the studies, the LCD aims to make the work in the field more cumulative, always attributing the information to the original researchers. As noted above, we hope to involve the research community in this joint effort to improve the accessibility and sustainability of the field by inputting their findings in the LCD. By contributing to the project, scholars will also gain visibility for their own research.

The categorizations in the LCD have been designed for the study of the English language, but they can be easily extended to accommodate other languages as well. The same applies to the data model itself, thanks to the flexibility afforded by the linked data approach. Moreover, the transformation process implemented by the LADA tool is of course only one way of utilizing the annotated data tables. New tools can implement transformations to other output formats

or even link some of the data to resources in other fields within the digital humanities.

The LADA tool presents the first concrete case of re-using LCD data for meta-analysis and shows a way forward for data-driven exploration. To make the most of linguistic applications such as LADA, it is important to be able to contextualize the numerical findings of historical developments by linking them to particular periods and genres. This has been possible in the case of small historical corpora such as the *Helsinki Corpus*. However, as noted in the Introduction, one of the major developments in historical corpora and databases has been their growth in size. While large corpora enable the study of low-frequency linguistic phenomena, they typically only include limited metadata, especially in terms of language-external information. Further contextualization is therefore needed, and this is indeed a growth area in the field of corpus linguistics, as shown by e.g. the *Old Bailey Corpus* (Huber et al. 2012) and the *Hansard Corpus* (Alexander & Davies 2015). To maximize its potential for re-use, contextual information should be implemented and distributed in the form of interoperable metadata (e.g. TEI; see TEI Consortium 2017).

In an ideal world, researchers should be able to share not only the published tables but their original spreadsheet files showing the analytic decisions they made in the course of the research process, but this has not been much in evidence so far. The fact that languages change across time makes data analysis in historical linguistics less straightforward than in many other fields of science and scholarship. Moreover, as Edward Sapir (1921: 39) famously put it: "Unfortunately, or luckily, no language is tyrannically consistent. All grammars leak." He implied that linguistic categories do not necessarily come with clear-cut boundaries at any time. We hope that access to systematic studies, multiple sources and interpretations will make the labour-intensive work carried out in the field more sustainable, paving the way for even more openness and transparency in the future (Flanagan 2017).

Acknowledgments. We would like to thank the many people who have helped us prepare the information for the LCD: the participants of our project course in 2015 as well as our students Melissa Haug, Marjut Kontinen, Noora Kumpulainen, Tina Lin, Matti Myllynen, Peppi Santaniemi and Emmi Seppälä, and our current and former research assistants Agata Dominowska, Aatu Liimatta and Emily Öhman. Agata and Aatu have also been involved in developing LADA, and Agata has had the main responsibility for the day-to-day running of the project from its inception. We owe a special debt of gratitude to Professor Emeritus Matti Rissanen, who gave the project a head start by providing us with detailed information on his extensive research into the history of English connectives. We also acknowledge the Academy of Finland funding for the project ("Reassessing language change: the challenge of real time", grant number 276349), which has enabled us to carry out our research plan in the first place.

References

1. Alexander, M., Davies, M.: The Hansard Corpus 1803–2005. (2015). http://www.hansard-corpus.org, last accessed 2017/10/31.
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked data – the story so far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009). doi:10.4018/jswis.2009081901
3. Davies, M.: Expanding horizons in historical linguistics with the 400 million word Corpus of Historical American English. Corpora 7(2), 121–157 (2012). doi:10.3366/cor.2012.0024
4. De Smet, H.: A corpus of Late Modern English texts. ICAME Journal 29, 69–82 (2005). http://clu.uni.no/icame/ij29/, last accessed 2017/10/31.
5. Ellis, N.C.: Meta-analysis, human cognition, and language learning. In: Norris, J.M., Ortega, L. (eds.) Synthesizing research on language learning and teaching, Language Learning & Language Teaching, vol. 13, pp. 301–322. John Benjamins, Amsterdam (2006). doi:10.1075/lllt.13.16ell
6. Flanagan, J.: Reproducible research: Strategies, tools, and workflows. In: Hiltunen, T., McVeigh, J., Säily, T. (eds.) Big and rich data in English corpus linguistics: Methods and explorations, Studies in Variation, Contacts and Change in English, vol. 19. VARIENG, Helsinki (2017). http://www.helsinki.fi/varieng/series/volumes/19/flanagan/, last accessed 2018/01/22.
7. Francis, W.N., Kučera, H.: Manual to accompany a standard sample of present-day edited American English, for use with digital computers. Original edn. 1964, revised 1971, revised and augmented 1979. Department of Linguistics, Brown University, Providence, RI (1979). http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM, last accessed 2017/10/31.
8. HC = The Helsinki Corpus of English Texts. Compiled by Rissanen, M. (Project leader), Kytö, M. (Project secretary); Kahlas-Tarkka, L., Kilpiö, M. (Old English); Nevanlinna, S., Taavitsainen, I. (Middle English); Nevalainen, T., Raumolin-Brunberg, H. (Early Modern English). Department of Modern Languages, University of Helsinki, Helsinki (1991). http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/, last accessed 2017/10/31.
9. Huber, M., Nissel, M., Maiwald, P., Widlitzki, B.: The Old Bailey Corpus. Spoken English in the 18th and 19th centuries. (2012). http://www1.uni-giessen.de/oldbaileycorpus/, last accessed 2017/10/31.
10. Johansson, S., Leech, G., Goodluck, H.: Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers. Department of English, University of Oslo, Oslo (1978). http://clu.uni.no/icame/manuals/LOB/INDEX.HTM, last accessed 2017/10/31.
11. Kesäniemi, J., Vartiainen, T., Säily, T., Dominowska, A., Nevalainen, T.: Exploring meta-analysis for historical corpus linguistics based on linked data. (In preparation).
12. Labov, W.: Principles of linguistic change, volume 1: Internal factors, Language in Society, vol. 20. Wiley-Blackwell, Malden, MA (1994).
13. Labov, W.: Principles of linguistic change, volume 2: Social factors, Language in Society, vol. 29. Wiley-Blackwell, Malden, MA (2001).
14. Labov, W.: Principles of linguistic change, volume 3: Cognitive and cultural factors, Language in Society, vol. 39. Wiley-Blackwell, Malden, MA (2010).

15. Lipsey, M.W., Wilson, D.B.: Practical meta-analysis, Applied Social Research Methods Series, vol. 49. Sage, Thousand Oaks, CA (2001).
16. Meroño-Peñuela, A., Ashkpour, A., Rietveld, L., Hoekstra, R., Schlobach, S.: Linked humanities data: The next frontier? A case-study in historical census data. In: Kauppinen, T., Pouchard, L.C., Keßler, C. (eds.) Proceedings of the 2nd International Workshop on Linked Science 2012 (LISC 2012) – tackling big data. CEUR Workshop Proceedings, Aachen (2012). urn:nbn:de:0074-951-6
17. Nevalainen, T., Vartiainen, T., Säily, T., Kesäniemi, J., Dominowska, A., Öhman, E.: Language Change Database: A new online resource. ICAME Journal 40, 77–94 (2016). doi:10.1515/icame-2016-0006
18. RDF Data Cube vocabulary. (2014). http://www.w3.org/TR/vocab-data-cube/, last accessed 2017/10/31.
19. Rissanen, M.: The development of till and until in English. In: Fisiak, J., Kang, H.-K. (eds.) Recent trends in medieval English language and literature in honour of Young-Bae Park, vol. I, pp. 75–92. Thaehaksa, Seoul (2005).
20. Sapir, E.: Language: An introduction to the study of speech. Harcourt & Brace, New York (1921). https://archive.org/details/languageanintrod00sapi, last accessed 2017/10/31.
21. Tagliamonte, S.A.: Variationist sociolinguistics: Change, observation, interpretation. Wiley-Blackwell, Malden, MA (2012).
22. TEI Consortium (eds.): Guidelines for electronic text encoding and interchange. (2017). http://www.tei-c.org/P5/, last accessed 2017/10/31.
23. Trudgill, P.: Sociolinguistic typology: Social determinants of linguistic complexity. Oxford University Press, Oxford (2011).