DE GRUYTER OPEN

ICAME Journal, Volume 40, 2016, DOI: 10.1515/icame-2016-0006

Language Change Database: A new online resource

Terttu Nevalainen, Turo Vartiainen, Tanja Säily, Joonas Kesäniemi, Agata Dominowska and Emily Öhman, University of Helsinki

Abstract

We introduce the Language Change Database (LCD), which provides access to the results of previous corpus-based research dealing with change in the English language. The LCD will be published on an open-access linked data platform that will allow users to enter information about their own publications into the database and to conduct searches based on linguistic and extralinguistic parameters. Both metadata and numerical data from the original publications will be available for download, enabling systematic reviews, meta-analyses, replication studies and statistical modelling of language change. The LCD will be of interest to scholars, teachers and students of English.

1 Introduction

One of the major challenges in linguistics is unravelling the process of language change. Sociolinguists have made great strides in analysing language variation and change synchronically, comparing successive generations at a certain point in time. Information on real-time change is sparser, but thanks to the digital turn in the humanities, there is now much more material, corpora and databases, that enable the diachronic approach. However, empirical work on change over time is still fragmented, and some of it is not easily available to scholars. In our four-year project (2014–18), we aim to make empirical research on linguistic change more sustainable and cumulative by compiling an online open-access database that provides real-time baseline data on language change from Old English to the present day (Language Change Database or LCD).

The database, which is currently being compiled by a research team at the VARIENG research unit and the Helsinki University Library, documents the findings of hundreds of corpus-based studies on the history of English that have originally been published in books or as journal articles. The LCD will be part of VARIENG's open-access e-resources, and will have a user-friendly interface

that allows rapid data retrieval across a number of parameters. The database will also provide numerical information from the studies in a downloadable format as well as links to the original papers and research reports (for copyright reasons, the papers themselves will not be distributed through the LCD). The LCD will be made available on a wiki-style platform with a graphical search interface, and scholars around the world will be able to collaboratively update it with their own work.

The justification for compiling the LCD is threefold. First, we believe that access to an extensive and cumulative resource on empirical studies will highlight the value of basic research by contributing to a better understanding of the various ways in which linguistic changes arise and diffuse across communities, varieties, registers and language systems over time. Secondly, the database will help bridge the gap between specializations within English linguistics: those scholars who know the historical data first hand and, for example, those who need historical baseline data to be able to analyse and interpret present-day variation. This annotated resource will bring together in one repository a wealth of research that is currently scattered and fragmented, and often only available in print publications. Thirdly, we believe that the LCD will improve the visibility of the individual studies carried out over the years, now made accessible to future work of various kinds, including systematic reviews and replication with other data sets. Our paper introduces the project and the solutions that have so far been adopted and implemented regarding the structure of the database.[1]

This article is organized as follows. Section 2 provides an overview of the current state of the LCD and discusses how all the articles included in the database are described and annotated. Section 3 takes a closer look at how grammatical information is organized in the LCD. Section 4 focuses on the technical details of the platform on which the LCD is run, and Section 5 provides concrete examples of the data submitter and end user interfaces of the database. Section 6 concludes the paper with a brief summary and some prospects for future work.

2 What is included in the LCD

All studies included in the LCD must meet two criteria. They should i) increase our understanding of how English has changed in the course of history (i.e. they are either diachronic studies or synchronic historical studies of English), and ii) include numerical data from corpora or other databases, such as dictionaries. New data is continuously added into the database, which currently includes detailed information of nearly 200 studies. In order to maximize the usefulness of the LCD even at this early stage of development, we decided to start compil-

ing the database with slightly older corpus-based studies from the 1990s. In our opinion, these studies are still highly relevant today, but because many of the articles were originally published in edited volumes which are not available online (or even as printed copies), they do not necessarily receive the attention they deserve. What is even worse, researchers may actually spend a great deal of time and effort in redoing older research simply because they are unaware of the existence of these hard-to-find articles.

A large proportion of the research included in the current version of the database is based on the Helsinki Corpus and done by the researchers at the University of Helsinki. The reason for this is largely practical: first, as we are very familiar with the structure and categorization of the Helsinki Corpus, we have been able to use this information to our advantage when making decisions about the structure of the LCD. Second, by piloting the database with studies by our close colleagues, we have been able to benefit from their feedback when inserting the information from their articles into the database. The most recent LCD entries also include data from articles that have been published in top-ranking journals, such as *Language Variation and Change* and *Journal of English Linguistics*. Although at this point we still prioritize older articles that are not easily accessible, the users of the database will also be able to submit their most recent research on the history of English regardless of availability.

The information used to describe each article in the LCD is divided into *publication details* and *content details*. Publication details include basic bibliographical information, such as the source publication in which the article appeared (either a journal or a book), publisher and place of publication, but they also include an abstract of the article and numerical data which have been extracted from the tables published in the original paper. If the publication is available online, a link to it will also be provided.

Content details, on the other hand, cover a broad range of hierarchically organized categories, including the major category of grammar (discussed further in the next section) and such usage-based fields as dialectology, sociolinguistics, pragmatics and discourse analysis. Some of these categories have been designed in relatively fine detail in order to facilitate the feeding of the data and to make user queries more accurate and efficient. Other categories, on the other hand, only include a short list of terms at present. The idea is that the users of the LCD will be able to add new keywords into these categories as they enter the information from their own articles into the database. Indeed, our primary aim at this point has been to provide the users with a relatively simple architecture that will be further elaborated as the database becomes more representative of the whole spectrum of English historical linguistics.

In addition to the categories and fields already mentioned, each article will be marked for the following content-related information:

- the topic of the study (e.g. the development of connectives in OE and ME)
- corpora and databases used in the study
- genre and regional variety studied
- time periods studied
- summary of the results.

The summary of the results comprises the main research findings of the article, and it is by far the most extensive field included in the content details. We have chosen to represent these summaries as bullet points in order to allow the users to gain a quick idea of the contents of the article (see Appendix 1 for an example). The summaries are based on our close reading of the article, making particular use of the discussion and/or the conclusions of the study.

3 Grammar in the LCD

The most detailed category that we provide to the end users of the LCD is 'grammar'. We fully acknowledge the challenge of drafting even a simple grammatical description of English that would satisfy the demands of researchers with different interests and backgrounds. When designing 'grammar', we have consulted both a large number of diachronic studies of English and the widely-read synchronic reference grammars of Present-day English (Quirk *et al*. 1985; Huddleston and Pullum 2002). Ideally, of course, we would have liked to follow both the organization and the terminology of a single reference grammar, but this was not feasible: the organization of the large reference grammars of English does not correspond sufficiently well with the actual research questions studied in English historical linguistics. Therefore, the organization of 'grammar' in the LCD is more reminiscent of a linguistics textbook than a grammar of English, including major categories like 'syntax', 'word class', 'morphology' and 'phonology', all of which are divided into several sub-categories. Whenever possible, we have tried to avoid adopting terminology that would be specific to a single grammatical framework.

Table 1 gives an example of how *nouns* are organized in the 'grammar' component. The users can choose to feed and retrieve data at different degrees of granularity, and new subordinate categories can always be added to complement the existing ones. For example, it would be possible to add 'proper name' under 'proper noun'.

Table 1: How nouns are organized in LCD's grammar component

Level 1	**Level 2**	**Level 3**
Word class	Noun	Proper noun
		Common noun
		Count noun
		Non-count noun/Mass noun
		Collective noun
		Gerund

Table 1 also shows that a term may be given an alternate label: in our model, 'non-count noun' and 'mass noun' can be used interchangeably when performing searches or feeding new information into the database. Alternate labels are also given to terms with variant spellings and/or forms (e.g. 'grammaticalization'; 'grammaticalisation'; 'grammaticization'). Again, we acknowledge that some linguists might prefer to make a distinction between some of the terms that we have categorized as alternatives. However, we are primarily concerned with the discoverability of the articles and data that are included in the database, not with fine-grained theoretical distinctions that may be subject to debate within the linguistic community in any case.

In our model of grammar, we make a rough division between formal/structural categories (e.g. 'word class', 'phrase structure', 'word order') and semantic/functional categories (e.g. 'animacy', 'definiteness', 'person', 'number', 'modality'). This decision is based on the observation that many semantic/functional categories are relevant to more than one formal category (e.g. both nouns and verbs are marked for person and number; animacy is relevant to both nouns and pronouns; modality is expressed through verbs, adjectives and adverbs etc.). The linked-data architecture of the LCD allows users to combine different categories freely, which not only helps keep the number of grammatical terms more manageable but also improves the accuracy of user queries (see Sections 4 and 5 below). For instance, a user may only be interested in finding articles that discuss the development of modal adjectives in the history of English. By linking 'adjectives' with 'modality', the user can exclude all articles that discuss the development of modal verbs and adverbs from the search results. Similar links can be made, for example, between 'third person' and 'pronouns', 'anaphoric reference' and 'demonstrative determiners', and the 'agent role' and 'syntactic subjects'.

The end user can perform searches based on the grammatical categories (e.g. 'noun', 'third person'), the usage-based categories (e.g. sociolinguistic and pragmatic information) – or any combination of these. An example of how this information can be used in a complex search will be given in Section 5 below.

Finally, we would like to point out that both the grammatical and the usage-based categories in the LCD have been designed with the end user in mind; the categorization is not intended to reflect our stance on some fundamental questions concerning the relationship between grammar and usage, the modularity of grammar, and so on. We would also encourage the users of the LCD to adopt a pragmatic attitude to their own data and to consider the annotation scheme as a general guide that will help other users find their research in the database.

4 Linked data platform for research data collection, curation and dissemination

4.1 From Excel to web, tables to graphs

The draft version of the LCD was initially stored in Microsoft Excel .xlsx file format, which offered us more flexibility in terms of editing, sharing, organizing, and exporting the data than a database tool (MS Access) would have. We also found that the increasing complexity of the categories and their relationships could not be handled sufficiently well in the Access environment. In the early stages of the project, work was simultaneously focused both on gathering and inserting data from the articles to the Excel database and on finding a suitable platform to host the final version of the database.[2]

Excel and Access are good tools in their own right, but as file-based solutions, they are inappropriate for shared and collaborative data collection. Moreover, the relational data model with a rigid database schema is not an ideal solution for data that does not have, or follow, standardized or de facto structure and semantics. There is hardly any prior work available on how to describe and store research results on language change, and the existing resources that we know of are general data repositories, while the LCD is a knowledge management system.

To create a data model for this particular type of research data, we opted to represent the relationships between categories as a mathematical graph (or a network) stored in a graph database. Such databases are well suited for creating a data model in an iterative manner, since they are occurrence-oriented as opposed to schema-oriented relational databases. A graph-based approach allowed us to start entering minimal but meaningful data quickly, and then let the data model evolve over time without having to worry about changing the database schema.

The graph model that we chose for data description is the resource description framework (RDF). RDF models data as statements in the form of subject-predicate-object expressions, also known as 'triples', which form a labelled directed multi-graph. An RDF database also supports the generation of new data with ontological inference and rules. Using standards such as Web Ontology Language (OWL), we can, for example, create class and property hierarchies, add transitive or symmetric characteristics to properties, or define inverse relations between properties.

RDF, being one of the World Wide Web Consortium's specifications, is designed for the web, and it is the specification used to distribute and consume Linked Data. Linked Data refers to the following three principles for creating and publishing structured data, first outlined by Tim Berners-Lee (2006): 1) All conceptual things are identified by Uniform Resource Locators (URLs), 2) URLs can be 'dereferenced' to look up the description of the identified thing, and 3) the description contains URLs that are links to other related things identified by URLs. The LCD allows the creation of research data that adheres to these principles.

4.2 Architecture

The architecture of the LCD platform is divided into two parts: back-end and front-end. The main responsibilities of the back-end are data management and storage. All the data is primarily stored within an RDF database through a web-based user interface. The back-end is used by the database administrators and data submitters. The back-end also includes an Application Programming Interface (API) with restricted access that provides data for the front-end.

The front-end is comprised of an API and an end-user interface. The API is responsible for providing scalable, fast and read-only access to the subset of the back-end data that has been made publicly available. It in turn uses the search index and document storage. Documents are snapshots of the graph that have been anchored to a certain identified concept such as 'grammar', 'corpus' or 'publication'. The content of the documents is determined by a mapping, where it is decided which features can be retrieved and are searchable for different domain types. For example, each entry in the database is marked for 'status' (i.e. stage of completion; 'draft' or 'final'). We could choose not to map 'status' information, simply because all the publicly available data should have the status 'final'. This would also mean that the end user would not be able to query documents by their 'status' information. So, the entire private data graph is not traversable by the end user, who must work in the context of publicly available documents. Finally, the public API works as the data source for the web inter-

face that allows users to search, browse and download research data. Figure 1 shows the logical components of the LCD platform and their relationships.

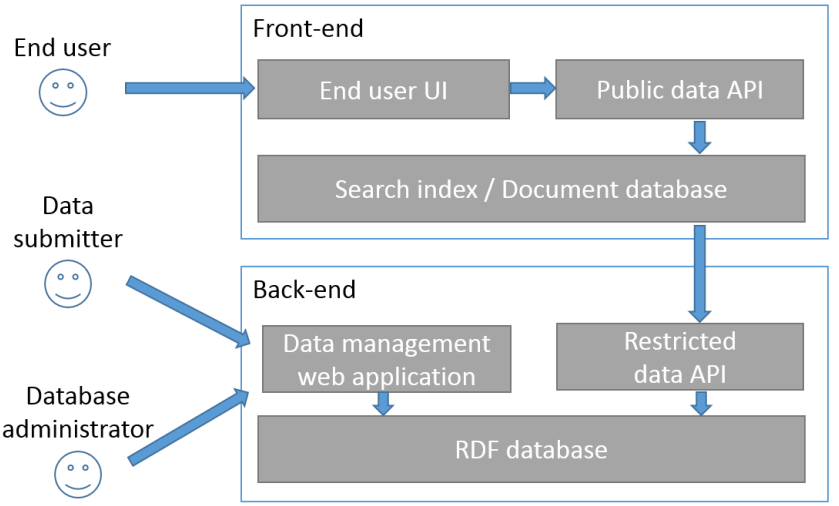

Figure 1: Conceptual architecture and user roles of the LCD platform

By keeping the back-end and front-end data separate it is possible to generate multiple different versions of the same underlying graph for public consumption. These versions can differ both in structure and content. We are currently looking into the possibility of implementing versioning of the publicly available data using parallel timestamped document storages. Implementation details depend on whether we want to publish small pieces of data continuously or focus on more comprehensive, scheduled data releases. The current architecture allows for maximum flexibility both internally and externally without sacrificing scalability or performance. The downside of this arrangement is that internal data flows become much more complicated, and more work is required to keep the front-end data up to date with the back-end data.

Semantic MediaWiki (SMW) and Callimachus were the two technologies reviewed for the implementation of the back-end data management functionality. SMW is based on the MediaWiki software that also runs, for example, the Wikipedia. Semantic functionality is added to the basic MediaWiki with a bundle of plugins that allow one to work with structured data in addition to wiki

text. SMW is a fairly popular tool for different kinds of knowledge management scenarios. Callimachus, on the other hand, is more of a niche framework aimed at building linked data applications. It is easy to install and take into use. It is also built from the ground up to support web standards and linked data principles. Both solutions are active open source projects with regular new releases, and both of them allow the building of web applications using nothing but the browser.

Ultimately, the choice between the two technologies came down to the decision between wiki conventions and web standards. The learning curve with SMW is quite steep both in terms of syntax and development philosophy. Callimachus was chosen as the implementation because with knowledge about web technologies, such as HTML and JavaScript, one can start building complex linked data applications immediately. Callimachus uses RDFa (Resource Description Framework in Attributes) to embed RDF data in the HTML code as extended element attributes. This means that the process of changing or evolving one's data model corresponds to changing the code that produces the form that is used to edit the actual data.

Elasticsearch was chosen as a basis for all front-end data layer implementation because of its distributed and scalable basic functionality, ease of use and extended support for hierarchical documents through the Siren plugin. Elasticsearch is used both as a search index and as document storage. The public API is also implemented as an Elasticsearch plugin. The user interface is implemented as an AngularJS JavaScript application. The concrete implementation components are summarized in Figure 2:

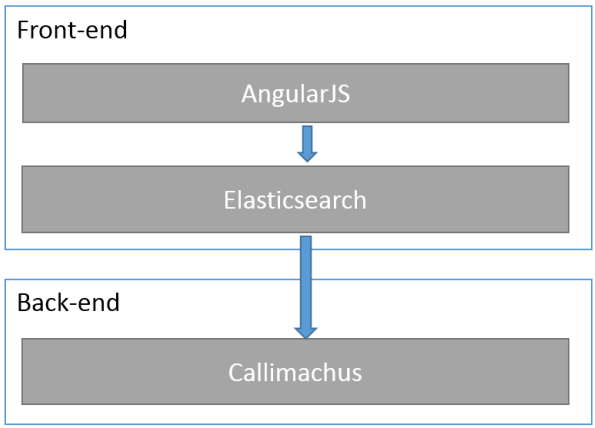

Figure 2: Components used to implement the LCD platform

5 Navigating the LCD

The administrative side of the LCD website is organized into wiki-style pages. The main page offers summary information on the different data types: studies (separated into publication and content details), authors, grammar concepts, corpora, genres and time periods, with links to recently changed entries. Each of the data types has a page of its own for browsing the entries within it, and each entry constitutes a page of its own. The pages incorporate the following wiki-style tabs: 'view', 'edit', 'discussion', 'describe' and 'history'. Thus, every page is easy to edit and describe, the editing history is easily accessible and any problems related to the page can be conveniently discussed with other users.

For adding a new entry, there is an 'add' button below each of the data types on the main page. Figure 3 shows the forms for adding a new study to the database. The idea is to first add the publication details, i.e. the standard bibliographical information, and then the related content details. Tables and other files related to the publication can be uploaded at the bottom of the publication details form. Many of the fields in the forms provide an auto-complete feature: as the user begins to type in the field, it automatically offers matching content from the field in the database. If no suitable content exists, the user can add new content by selecting 'add' in the drop-down menu. In Figure 3, this functionality is shown for the 'author' field. Note that linguistic items are capitalized in the content details.

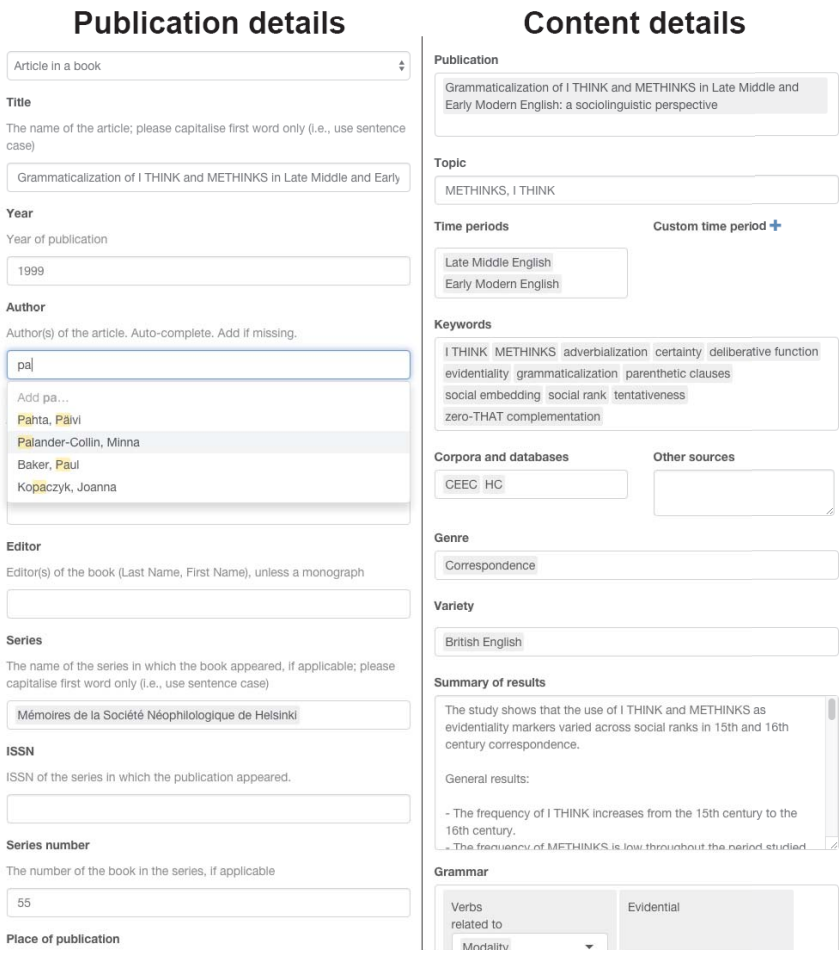

Figure 3: Adding a new study to the LCD: screenshots of the publication and content details pages (see Appendix 1 for the full entry)

Figure 4 is a screenshot of a working prototype of the LCD search interface. The time period of interest can be selected in two ways – by adjusting the start and end sliders on the timeline at the top, or by clicking on the buttons representing the standard periods of the English language: Old English (OE), Middle English

(ME), Early Modern English (EModE), Late Modern English (LModE) and Present-Day English (PDE). In Figure 4, we have selected Early Modern English, 1500–1700. On the left, we have a keyword search targeting all of the keyword-type fields in the database. Below it is the grammatical hierarchy: here, we have selected 'word classes' to search for studies dealing with any word class. We could drill down by selecting e.g. 'adjectives' below 'word classes'. On the right, we may filter the search by corpus, variety, genre or social category; we have selected the Helsinki Corpus (HC). The middle of the page is reserved for the search results – in this case, a list of all studies dealing with Early Modern English, any word class and the Helsinki Corpus. From here, we may proceed to view any of the entries by clicking on them. We will also be able to download the entries and the tables associated with them.

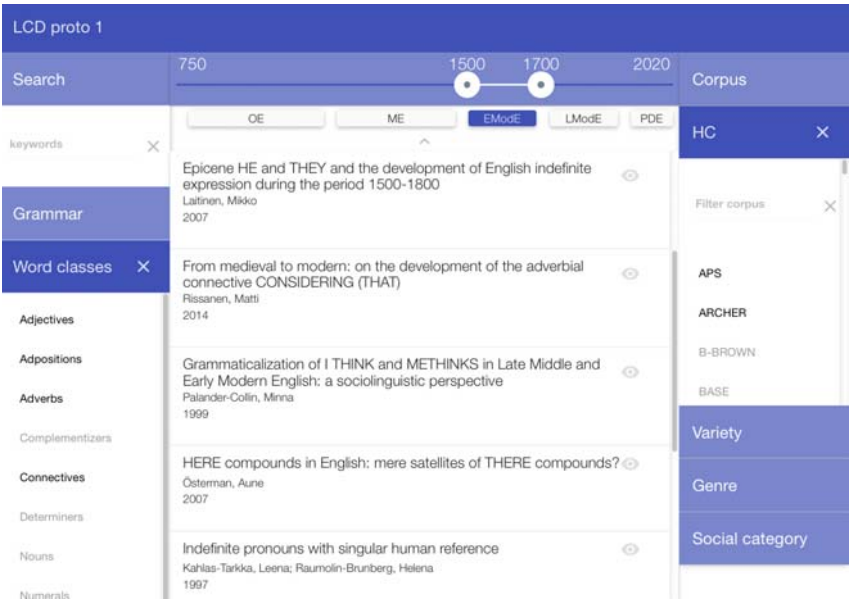

Figure 4: Prototype of the LCD search interface

6 Concluding remarks

In this paper we have discussed the development, design and aims of the Language Change Database (LCD). We have argued that the LCD will be useful to the research community for a number of reasons. First, the end users of the data-

base can gain a quick understanding of earlier corpus-based historical research of English by performing simple searches through the LCD's graphical user interface. Second, scholars of English can make their research better known to the research community by entering the descriptions and data of their own articles into the LCD. Third, the data included in the LCD can be used for new research, such as meta-analyses and replication studies. By publishing the LCD as an open-access online resource, we wish to be part of the growing trend in the humanities where both research results and the data on which the research is based are made freely available to the scholarly community.

In the short term, the project's work will concentrate on making the first version of the LCD available to other scholars. As the database grows in size, version control will be implemented to document all the changes made to the architecture and contents of the database. Our long-term plans include designing and adding new functions to the LCD. These include an export function with which the end users can extract the publication details of the articles in standard bibliographical formats, such as BibTeX and RIS, and new search functions, such as interactive maps. We are also very interested in acquiring data from the great body of research on the history of English written in other languages (e.g. German, Japanese); to achieve this aim, the help of our international colleagues will be especially valuable. Finally, although the database has only been designed with the history of English in mind, we see no reason why the LCD could not function as a platform (or a model) for the mapping of diachronic research of other languages.

Notes

1. The Language Change Database is funded by the Academy of Finland for the period 2014–18 as part of the project "Reassessing language change: the challenge of real time" (PI: Terttu Nevalainen). Joonas Kesäniemi is responsible for developing and implementing the linked data platform for data collection, curation and dissemination, and the project's research assistants have keyed in the publications currently included in the database in close collaboration with and under the supervision of the senior project members.
2. Special thanks are due to Professor Emeritus Matti Rissanen, who kindly provided us with the very first draft entries for the database. We should also like to thank Anne Kingma (University of Groningen) and Kimmo Koskinen (Helsinki University Library) for assisting us during those early stages of the project.

3. This category is for general keywords related to any aspect of the study. It may duplicate some of the keywords in other categories, and it may also contain keywords that do not easily fall within the other categories or that are not common enough to merit a keyword under e.g. 'grammar'.

References

AngularJS. https://angularjs.org.

Berners-Lee, Tim. 2006. Linked Data. http://www.w3.org/DesignIssues/LinkedData.html.

Callimachus. http://callimachusproject.org.

CoRD = Corpus Resource Database. http://www.helsinki.fi/varieng/CoRD/.

Elasticsearch. http://github.com/elastic/elasticsearch.

HC = *The Helsinki Corpus of English Texts*. 1991. Compiled by Matti Rissanen (Project leader), Merja Kytö (Project secretary); Leena Kahlas-Tarkka, Matti Kilpiö (Old English); Saara Nevanlinna, Irma Taavitsainen (Middle English); Terttu Nevalainen, Helena Raumolin-Brunberg (Early Modern English). Department of Modern Languages, University of Helsinki. http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/.

Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

LCD = Language Change Database. http://www.helsinki.fi/lcd/.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

RDFa = Resource Description Framework in Attributes. http://www.w3.org/TR/rdfa-syntax/.

Siren. http://siren.solutions/siren/overview/.

SMW = Semantic MediaWiki. http://semantic-mediawiki.org.

Appendix 1. Sample draft entry illustrating the publication and content detail categories in the LCD.

Publication details

Type:
Article in a book

Title:
Grammaticalization of I THINK and METHINKS in Late Middle and Early Modern English: a sociolinguistic perspective

Year:
1999

Author:
Palander-Collin, Minna

Source:
Grammaticalization and Social Embedding: I THINK and METHINKS in Middle and Early Modern English

Editor:
–

Series:
Mémoires de la Société Néophilologique de Helsinki

ISSN:
–

Series number:
55

Place of publication:
Helsinki

Publisher:
Société Néophilologique

Start page:
193

End page:
227

Abstract:
The article deals with the evidential expressions I THINK and METHINKS. The approach is a sociolinguistic one and attention is paid to the role of social status in the use of these expressions in Late Middle and Early Modern English. The changes in frequencies of these expressions from the fifteenth to the sixteenth centuries are shown as well as their use as evidentiality markers across social ranks.

DOI:
–

Link:
–

Parallel publication:

Palander-Collin, Minna. 1998. Grammaticalization of I THINK and METHINKS in Late Middle and Early Modern English: a sociolinguistic perspective. *Neuphilologische Mitteilungen* 99(4): 419–442.

Linked files:
[Tables in the .xlsx format]

Status:
Draft

Content details

Publication:
Grammaticalization of I THINK and METHINKS in Late Middle and Early Modern English: a sociolinguistic perspective

Topic:
METHINKS, I THINK

Time periods:
Late Middle English, Early Modern English

Custom time periods:
–

Keywords:[3]
I THINK, METHINKS, adverbialization, certainty, deliberative function, evidentiality, grammaticalization, parenthetic clauses, social embedding, social rank, tentativeness, zero-THAT complementation

Corpora and databases:
Corpus of Early English Correspondence, Helsinki Corpus

Other sources:
–

Genre:
Correspondence

Variety:
British English

Summary of results:
The study shows that the use of I THINK and METHINKS as evidentiality markers varied across social ranks in 15th and 16th century correspondence.

General results:
- The frequency of I THINK increases from the 15th century to the 16th century.
- The frequency of METHINKS is low throughout the period studied, and its frequency decreases further in the 16th century.
- There is variation with regard to the structures with which METHINKS is used, but usually METHINKS is followed by a zero-THAT complement clause. After 1570, METHINKS is typically used with clause-initial zero-THAT clauses or in clause-medial and clause-final parenthetic structures. These structures are also argued to represent the most grammaticalized (adverbialized) uses of METHINKS.
- For I THINK, the frequency of THAT-complementation decreases from 1440-1500 to 1540-1600, while parenthetical uses become more common; zero-THAT complementation is common in both periods.

Sociolinguistic and pragmatic results:
- The gentry, including knights, esquires and gentlemen, used both I THINK and METHINKS more frequently than the nobility or the non-gentry (particularly the merchant rank) in the 15th century.
- In the 16th century, the differences in the frequency of use of I THINK were no longer statistically significant between the ranks studied. However, the data revealed statistically significant differences in the grammaticalized vs. non-grammaticalized uses of I THINK according to social rank throughout the period studied; for instance, in the 16th century merchants used both

zero-THAT and parenthetic constructions considerably more frequently than the gentry or social climbers.

- Both I THINK and METHINKS can be used as politeness markers, and they may signal different kinds of politeness. The author suggests that the gentry may have used these expressions in a deliberative function, i.e., to express certainty or reassurance. The non-gentry, on the other hand, may have used I THINK and METHINKS to express uncertainty and tentativeness. The author proposes that the adverbialized uses of I THINK and METHINKS may have developed in the tentative function, as evidenced by the high frequency of grammaticalized uses of I THINK by the merchant rank in the 16th century.

Grammar:
Verbs related to Modality, Evidential, Grammaticalization

Dialectology:
–

Language contact:
–

Sociolinguistics:
Social category

Social categories:
Social rank

Pragmatics:
Politeness

Discourse analysis:
–

Statistical methods:
Chi-square test

Status:
Draft