doi:10.5128/ERYa19.11

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 19, 175–194

STRATIFIED HISTORICAL CORPUS OF ESTONIAN 1800–1940

Peeter Tinits

Abstract. The article introduces a stratified historical corpus of Estonian 1800–1940. A stratified corpus will allow for sociolinguistic comparisons of language use between past authors, considering their background and biographical details (e.g. native dialect area, age cohort, attained education) or the publication details (e.g. genre of publication or publisher). The corpus assembles texts from a number of different public archives and combines it with metadata on their publication details and the author's background. The corpus at the moment of publication consists of 4,412 works from 1,188 author names, constituting 11% of the works registered in the Estonian National Bibliography from 1800–1940. The author names are associated with biographical information where possible. Three use cases on studying orthographic variation are introduced as examples where the corpus can help study past language communities. The corpus is published online to allow updates as data is improved and more texts are digitized.

Keywords: corpus, language resource, metadata, historical sociolinguistics, sociolinguistic variable, written Estonian

1. Introduction

Historical sociolinguistics has expanded the research done in historical linguistics on past language communities and the linguistic norms present in them. The approach has brought forth previously marginalized communities in linguistic research (e.g. language of the lower classes, Vandenbussche, Elspaß 2007), hidden survival of minority languages (Havinga, Langer 2015), or hidden genres of discourse that impacted the spread and survival of linguistic norms (e.g. letter-writing, Van der Wal, Rutten 2013). The novel findings of historical sociolinguistics for language history have very often been enabled by the creation of new text corpora that bring forward a less-studied variety in the language or under-represented genre in the language. Historical sociolinguistics has considered it a good goal to study corpora where a broad spectrum of the society (e.g. by social class, education, geographical

background) can be represented (see e.g. Raumolin-Brunberg, Nevalainen 2007). It is then up to the researcher to use the corpus in a way that answers their research question.

One particular need, established by recent research in the field, has been the generation of corpora that would capture the variation of language use not just between authors, but also within authors over their lifetime and in different contexts of use (Ulbrich, Werth 2021). For example, it has been shown how letters written in a mental asylum from 1850–1936 show systematic variation based on their intended audience, the intended image of the speaker or other contextual factors (Schiegg 2022). Individuals may also change their writing during their lifetime as a reflection of diachronic changes or changes in social roles – like in the documented changes in Queen Elizabeth's letters from 1544–1603 (Evans 2013). Recently, systematic efforts have been made to create corpora that would allow questions on intra-individual variation to be approached: for example, the Early Modern Multiloquent Authors (EMMA) corpus, which contains samples of the 50 most prolific English writers born in the 17th century to allow a study of language change in their writing across their lifespans (Petré et al. 2019). Tracking intra-individual variation across a writer's lifetime can provide insight into the role that, for example, language contact, linguistic prescription, education, or generational change may have played in the history of a language.

The chosen period is based on the coverage of existing corpora, on the availability of materials to be included, and a few particular research questions on the history of Estonian that acted as an impetus for it. The two main corpora that have so far offered an insight into this time period are the Corpus of Old Written Estonian (VAKK) and the Corpus of Estonian Literary Language of the 20th century (Hennoste et al. 2001). The first of these primarily covers the earliest writings from the 15th to the 17th century in its core corpus, but also includes some writings from the 18th and the 19th century. The second one aims to cover the period from 1890–1990, providing a relatively balanced sample for each decade. It provides a few pages of text from each work, with a few hundred works in each decade. As it was compiled some time ago, some metadata is not fully preserved, and due to some unsolved considerations during the compilation, the years 1920–1935 have been not included in the same format. The materials of both of these corpora are also included here, adding up to a total of 27 complete texts and 214 text snippets from 1800–1940. There are a total of 111 authors represented, with 19 of them having more than 2 works included, and 5 authors with more than 5 works, were we to be interested in changes over the course of an author's lifetime.

The corpus presented here is built in an attempt to collect all digitally available printed texts from the period into one place and collection. Texts from newspapers and other periodicals have been excluded here, as they follow a different pattern of publication. This corpus follows other collections with broad goals, often built by non-linguists. For example, Gutenberg Books, which is often used as a corpus, is a collection of texts with a license allowing easy republication which has been made into a standardized collection providing more than 50,000 books from 1800–2019 in multiple languages (e.g. Gerlach 2020). Eighteenth Century Collections Online (ECCO) is a collection of English texts from 1701–1800 containing around 184,386 publications, which is around 54% of the titles listed in a bibliography of the era

(Tolonen et al. 2022). The digitization project HathiTrust contains more than 17 million digitized publications where text mining is permitted, allowing a corpus of 210,266 fiction texts to be created (Underwood et al. 2020). These text collections are not balanced – the representativeness of these contents for any research question is in itself an open research issue (see e.g. Tolonen et al. 2022) – which leaves the selection of relevant works up to the researcher. By contrast, while traditional linguistic corpora have been specifically designed to be balanced, there is increasingly discussion of using more complex collections for linguistic or historical study by a more thorough understanding of the representativity of each collection used (e.g. Mäkelä et al. 2020). The current corpus builds on the same thread – the goal is not to initially balance the selection of texts, but provide the metadata so that the user can do it themselves.

2. Sources

The collection of texts was aimed to be maximally comprehensive – to find any texts that could be used for this purpose. For the final stages, the metadata at the ENB (as of Jan 4 2022) was relied on for guides, while throughout the project text collections were approached and collected also separately from that. As a result, the final collection includes some texts that were not included in ENB metadata, with some of them likely to be added there soon. The texts added are based on the availability of full-text access within these collections, and as new sources emerge the collection is hoped to be updated.

The main sources for the texts are given short designations to simplify reference to them in this article as the following.

DIGAR (5878 texts) – Digital Archive of the National Library of Estonia (https://www.digar.ee) collects a number of digitized texts that have been collected for public use. Their availability depends on the licensing of particular texts; some of them are available from a public access point, some for use in the library premises, and some are merely archival copies not available to the public. The DIGAR collection has in many cases used OCR to retrieve the text from digitized images, while some of the publications remain in simple image format. In the collection, the publications that were available from a public access point and had a text layer added by OCR were used.

KIVIKE (1861 texts) – The "virtual cellar of the Literary Museum" (http://kivike.kirmus.ee) is the electronic repository of the Estonian Literary Museum, which contains a variety of digitized materials and links to their archival items. The database also includes a number of digitized print publications, some of which have been manually transcribed by in-house researchers. The items that were linked to by ENB or had a transcribed component (as of July 2 2018) were included in this collection.

DSPACE_UT (255 texts) – Dspace (https://dspace.ut.ee) is the repository for digital materials at the University of Tartu, which among other things also includes digitized copies of older publications. Some of these publications were available only as images, some had a layer of OCR texts attached to them. The OCR layer was accessed directly for the texts linked to by ENB or the texts given in the local search for 1800–1940 publications.

KRZW100 (250 texts) – Kreutzwald Century (http://krzwlive.kirmus.ee) is an interactive portal compiled by the Estonian Literary Museum to introduce Estonian cultural history from the early 19th century until 1918. This collection also includes a number of texts that have been manually transcribed by researchers at the museum from first editions of the publications, where possible.

WIKISOURCE (68 texts) – Wikisource (https://et.wikisource.org/wiki/Esileht) is a website maintained by the Wikimedia Foundation with the intention of making texts that have no copyright restrictions publicly available. These texts have been in most cases manually transcribed by the visitors to the page, and a workflow has been set in place that ensures some quality control to the public texts.

ESTLIT (119 texts) – the digitization programme Estonian Literature. A number of old publications have been republished as e-books based on the manuscripts made by Estonian Literary Museum and a few publishing houses. They vary in their quality and editing practices. Some of the books contain the statement, "language unaltered" (*keeleliselt muutmata*), while others have been transformed into modern language. Only the books which were marked as "language unaltered" or known to be unaltered were used for the collection.

CELL (218 texts) – The Corpus of Estonian Literary Language of the 20th century (https://www.cl.ut.ee/korpused/baaskorpus) was constructed by researchers at the University of Tartu (Hennoste et al. 2001), and has so far been the primary text source for corpus studies of the time. It contains snippets of texts from a selection of books chosen to characterize the language of the era. These texts have been manually transcribed, although at least some undocumented editing practices for spelling can be observed.

VAKK (31 texts) – The Corpus of Old Written Estonian (http://vakk.ut.ee) is a collection that has been built with a focus on 15th–17th century Estonian texts, where it contains an almost complete record. From the 18th and 19th centuries, a selection of texts has been added. For our purposes all texts from the 19th century have been included, as the texts were shared by the maintainers.

As of Jan 4 2022, there were a total of 39,442 Estonian language publications published from 1800–1940 entered into the Estonian National Bibliography. 8,062 of them had a linked digital copy. This registry also includes reprints of earlier works, which are not differentiated in the dataset. Thus, roughly 20.4% of all the print publications have been digitized, with somewhat better coverage in earlier works. The number of text files retrieved from each archive are given in the section above – many of them overlapped between the archives.

In total, it was possible to retrieve 8,680 text files from 1,744 author names, although not all these texts proved suitable for the corpus. Excluding texts of bad quality and duplicate texts, 4,412 works were included in the corpus, which amounts to 11.2% of the texts published. See more details in the corpus description section. The texts from CELL are snippets, but the rest of the corpus should be full-length publications. In some cases the publications are collections of texts that contain writings from several different authors – these have not been disaggregated in the corpus.

3. Processing and formats

In many cases, the digitized publications were available as pdf files with a text layer on them (e.g. DIGAR). To extract the text from the pdf, they were processed with the PDFminer 'pdf2txt' tool (De Smedt 2012). This tool separates the text layer, but follows the structure given in the file, which often does not follow the visual sequence. As a result, sentences may sometimes be mixed together across different paragraphs, making detailed text analytics complicated, even though all the words are there. In some cases, the OCR text layer was available for access separately from the pdf (DSPACE_UT). In this case the texts were downloaded; however, imprecision from the quality of digitization is also possible here. In KIVIKE the text could be accessed in a text field, which was used in full. This was usually manually transcribed. For KRZW100, the maintainers were contacted and they shared the manuscript versions, which are included here. VAKK texts were provided by corpus maintainers, and reprocessed to match the format. ESTLIT and WIKISOURCE files were available as epub files, which were converted to text files. CELL texts are available online in snippets. Texts from other corpora were linked to ENB IDs, or given new IDs if a match was not found. Metadata from original collections was processed and used to augment the ENB data.

All these texts were tokenized and lemmatized with the Estonian Natural Language Processing Toolkit for Python (EstNLTK, Orasmaa et al. 2016). EstNLTK uses the morphological analyser Vabamorf, based on an earlier C++ library to analyse words based on word-form and sentence context (Kaalep, Vaino 2001). This software is tuned to modern Estonian vocabulary and spelling and hence does not work very well with linguistic variants that have fallen out of use. Lemmatization was performed under two regimes:

1. To detect the word-forms that could be seen as non-standard by modern norms, the lemmatizer was run with no disambiguation or guessing. Thus when the lemmatizer failed, no lemma was offered.
2. To detect best possible matches, lemmatization was performed with guessing and disambiguation.

4. Metadata enrichment

The metadata on publications was enriched by information on the authors, translators and editors involved in publishing these books. Here, in dealing with linguistic analyses, we consider translators as the authors, as what we are interested in their language choices. If an author was missing, but the publication had an editor, then they were used as an author as they very often had done the writing themselves. This should be checked for individual cases if these details are important in a particular study. The original authors of translated works are also included in the metadata, in case content is the focus of a research topic.

4.1. Biographical information

The book authors dataset was enriched by biographical information from several databases:

ENB – The Estonian National Bibliography keeps an authority file of persons that have been associated with print publications. This database is linked to VIAF on the basis of the ID.

VIAF – Virtual International Authority File is a project aimed to link together authority files from many national libraries.

DNB – The German National Bibliography keeps information on individual biographies that has been linked through VIAF to the ENB person index. This includes information on the place of birth and occupation for some Estonian authors as well.

Wikidata – Wikidata is a collaboratively edited database that is used as a repository for Wikimedia projects and gathers information in structured form. There is an active effort to transport information from Wikipedia to database format in Wikidata, although not all of this information has been transmitted.

Wikipedia – Wikipedia is a collaboratively edited encyclopedia which contains articles on historically significant individuals. The lead section was used along with Wikidata materials to provide locations for the birthplaces.

ISIK – ISIK (http://www2.kirmus.ee/biblioserver/isik) is a biographical database compiled by the Archival Library of the Estonian Literary Museum based on the records of bibliographic archives. The information is given in semi-structured format, with some text processing needed to export the data.

VEPER – VEPER (http://isik.tlulib.ee) is a biographical database of the Estonian diaspora maintained by Tallinn University. The database includes structured and semi-structured information on the individual biographies. The website uses the same engine as ISIK.

4.2. Geographical information

The places associated with individual biographies (some already added via DNB_GEO) were enriched with geographical coordinates on the basis of GEONAMES and EKI_PLACENAMES. They were then placed on a map of Estonian traditional dialects published before (Uiboaed, Kyrolainen 2015) (EST_DIALECT_MAP).

DNB_GEO – The German National Bibliography has a list of placenames and coordinates that are linked to the individual biographies.

GEONAMES – GeoNames is a free downloadable database. Places in Estonia were extracted from this and linked with the placenames that did not have a location.

EKI_PLACENAMES – The Place Names Database (KNAB) at the Estonian Language Institute was used to augment the available placename information. This database includes also historical placenames that are no longer in use.

EST_DIALECT_MAP – A dialect map composed to assist in visualizing Estonian dialects. It does not have a high level of detail, but is enough for us to connect the placenames with dialect areas.

4.3. The process

The bibliographical entries on the publications in ENB contain information on the authors of these works which is designed to link to a database on people associated with these publications. The associated persons have been linked to the VIAF database, which also contains links to other datasets. From the VIAF database, the links connecting to DNB and Wikidata were collected. In addition to the entries that were directly connected, also the individuals associated with Estonia were collected from DNB and the individuals who had an article on Estonian Wikipedia were collected. These were checked for links manually. The biographical collections in ISIK and VEPER were collected from their websites and transformed into a structured dataset that contained biographical information on each of the individuals.

5. Description of the corpus

5.1. Overview of texts

A total of 8,065 text files were collected, consisting of 6,658 unique texts (some publications had several digital copies). Some files were not suitable for research use – they had either an empty layer of text attached to them or very poor character recognition. Thus, files with less than 50 tokens of text and files where less than 1% of the tokens could be recognized by the EstNLTK lemmatizer were excluded. Were we to (again) perform OCR to extract text from images, we would be able to expand the corpus here. This left 4,608 text files and 4,412 unique publications. From these, 142 publications which had the exact same titles and authors as an earlier publication in the corpus were also excluded, as the text is likely to have been a reprint of an earlier edition with the text largely overlapping, leaving 4,270 publications in the set.

The texts were combined with ENB metadata. All publications had information of the publisher's location, while 2,593 had information on a publisher's name. 2,486 had an author associated with them. Of the latter, 1,791 works had information on the birthyear of the author, 1,367 on the birthplace of the author, of which 1,153 could be associated with a dialect area. 658 of the works had information on the author's chosen domicile later in life, 779 had information on the author's attained level of schooling, 510 publications had information on each of these parameters. 395 authors had more than 2 works included, and 151 authors with more than 5 works, were we to be interested in changes over the course of an author's life. Additionally, 1,672 had information on the number of copies that were initially printed.

ENB also contained information on the genre of the publication. The list of genres was harmonized into a few basic types, where one publication could belong to several genres. Looking at the metadata availability by genres shows an uneven distribution. Figure 1 shows the data availability for each genre. While the majority of fiction works had available both the name of the author and some biographical details on them, the bureaucratic publications and letters very rarely had the author's name included. For publications that could not be associated with a genre, more than a third had distinguishable authors.

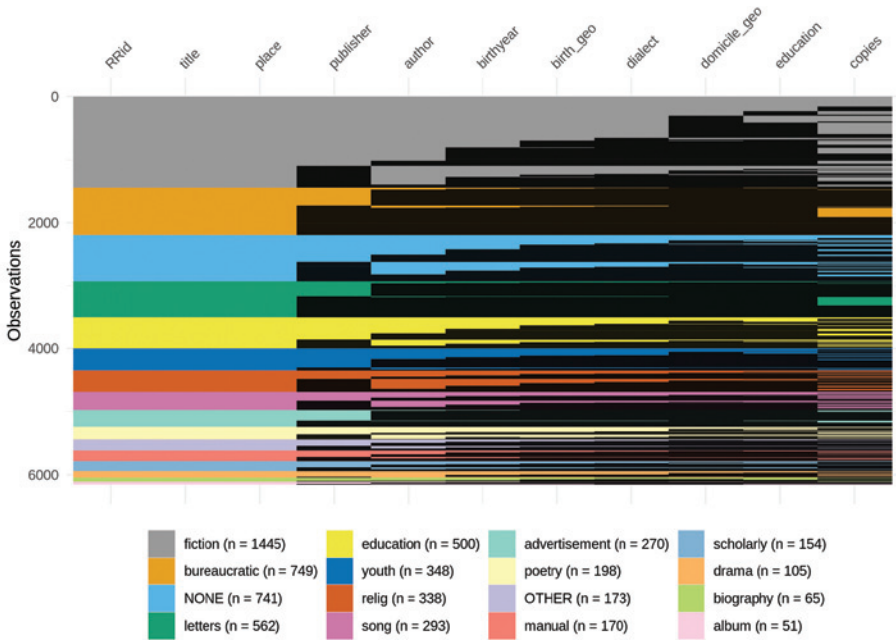

Figure 1. The availability of metadata by genre, excluding genres with less than 50 publications. The works are listed as observations and ordered by the availability of the information within genres. Genres are marked in different colors, black shows missing data. Note that because some works belong to several genres, the number of rows in the figure is 6,162

5.2. Language and text quality

In order to assess the text quality, we took the EstNLTK annotation layer and checked for the words that had not been recognized. EstNLTK relies on models of modern Estonian for this process, and thus is unable to understand archaic word-forms, obsolete spelling variants or OCR errors. We can use the results as a first proxy to the quality of the individual text files, Figure 2 plots the results. However, since this is a long time period where major changes in orthography took place, we can most clearly see the two transitions towards modern spelling. One, from the old German-based spelling tradition to the new Finnish-based tradition, was proposed by Ahrens in 1830 and was widely adopted in the 1870s – texts in older orthography had less than 50% of words recognized. Another major transition visible from lemmatization is the adoption of V instead of W for the same sound, focussed around the 1920s – this entails a move from 25% words not recognized, as many words contained a W, to around 5% words not recognized, likely due to OCR errors and a few more archaic forms. In addition to these major transitions, it is possible to use this information to find writings that rely on dialectal forms that were not adopted by the written standard later on – for example in looking more closely at the texts that had less than 50% words recognized in the 1900s.

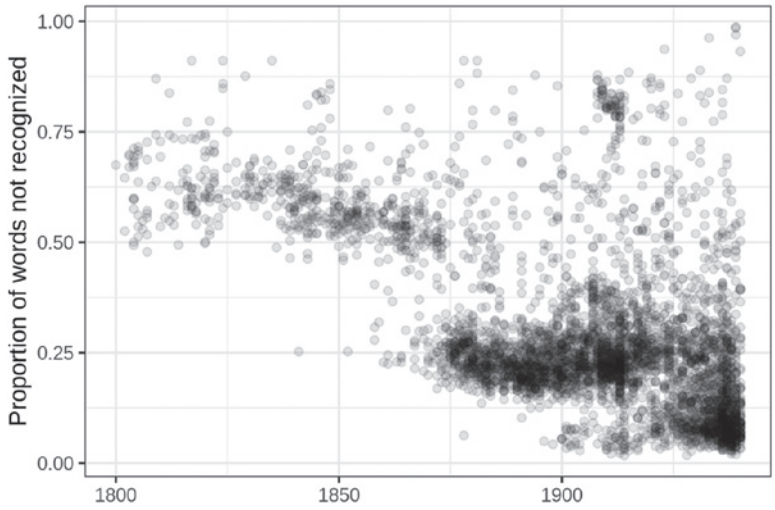

Figure 2. The proportion of tokens not recognized by EstNLTK

5.3. Representativeness and metadata availability

The acquired texts in the collection can be compared with the Estonian National Bibliography to understand how well it represents the contemporary landscape of written texts. Figure 3 shows the number of texts in the corpus on the left, and the proportion of the bibliography covered on the right. The corpus represents the bibliography better in the earlier times, with above 20% coverage before 1900. As the number of publications in the bibliography grew, a smaller proportion of them has been digitized amounting to a coverage of around 5–7% from 1920–1940.

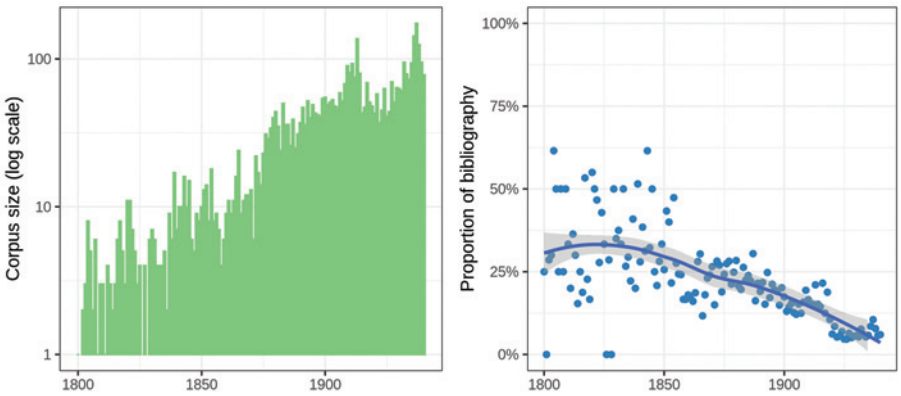

Figure 3. The number of publications in the corpus (left) (n = 4,412) and the proportion of the Estonian National Bibliography (right) per year with the modelled means of the data (using a loess smooth fit). The number of publications is shown on a log scale

We can also have a look at texts by genre and the texts of the most prolific authors. On Figure 4, we can see the representation is quite similar across genres at around 15% of the works published. On Figure 5, we can see the most prolific authors and their lifespans. On average they have 33% coverage (sd = 17%) in the corpus. Some areas and authors are covered quite well, others less so. The representativeness of the corpus should be considered by the researcher on a case-by-case basis.

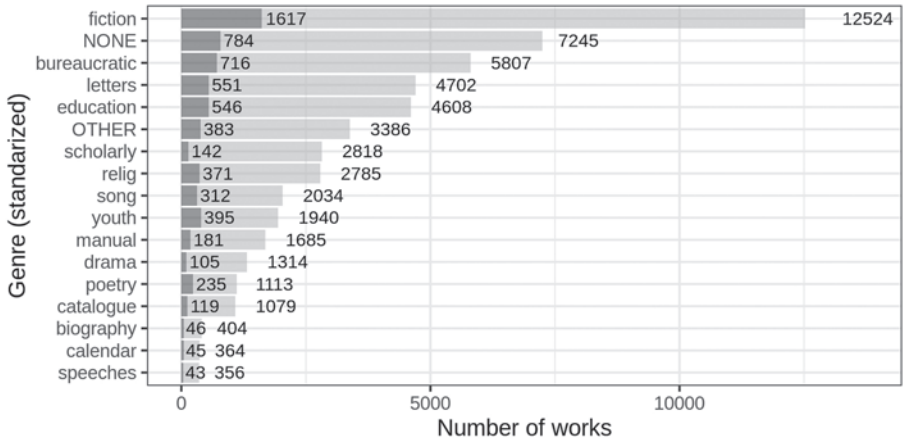

Figure 4. The number of works in the corpus vs ENB by genre. Genres are based on the ENB genre information that has been combined into aggregated types. OTHER refers to genres that did not fit the main categories, NONE refers to works that had no genre assigned to them. The full list of the genres included is given among the corpus files

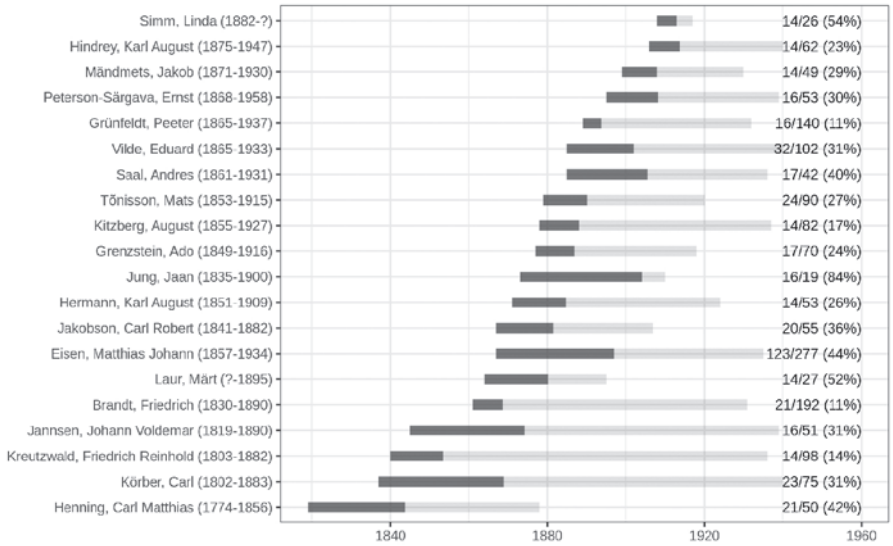

Figure 5. The number of works published shown over the range of publication years for the 20 authors with the most works in the corpus

6. Use cases

The corpus users need to keep in mind that the corpus is not balanced in terms of genre, authors, or social background – it is a reflection of what has been made available as digital copies. Many of the texts have been digitized automatically, so there are likely to be OCR errors that transform individual words or sentences – not all the words can be reliable – for example in looking for hapax legomena, this corpus is likely to provide many digitization errors instead of true contemporary linguistic variation. On the other hand – some texts originate from digitization frameworks that also edited and modernised the language use, which can interfere in studying some linguistic questions, but help in content analysis. Whichever the researcher's interests, it is up to the user to check the corpus contents in any study.

One linguistic research topic that can be approached with data with the aforementioned problems is linguistic variation. This is generally understood in terms of alternating variants that are used by people with a different educational, social, or dialectal background to mean mostly the same thing (e.g. color vs colour in American vs British English) (Labov 1972). However, because these variants can be used performatively (e.g. when representing dialect speech, or trying to appear educated), this understanding is usually accompanied with some caveats (e.g. Silverstein 2003). Since studying linguistic variables works with the use of one variant over another, then, even if some words have been corrupted by OCR errors, their impact on the general trend is likely to be marginal. When working with single orthographic words, also scrambled sentences or OCR errors in nearby words would play little role. Looking at linguistic variants can be pretty robust thus even for low or medium quality digitization if we expect writers to be mostly consistent with their choice of variants.

Here we perform a few such example queries to the corpus. The analysis tries to understand the use of some orthographic variants that were used in the late 19th century, and thus could be approached by this corpus. In this case, each variable is understood in terms of a lexicon of fixed alternating variants, one of which is a standard form in modern Estonian. This standard form is designated with a score of 1 on the visual graphs, all other forms are designated with 0 (although usually there was just one common alternative). The lists and the workflow to repeat the steps are provided in the supplementary appendix.

6.1. Studying patterns in long-term linguistic change

The corpus presents a sufficiently long time span that it can allow the study of language changes that took longer than a generation. It is possible to look at long term patterns of change and try to understand the mechanisms that took us there. Here, we can look at four orthographic variables with some variation in the corpus: 1) *naene* (e.g. /naene/ vs /naine/), 2) *nnud* (e.g. /annud/ vs /andnud/), 3) *te* (e.g. /õiete/ vs /õieti/), 4) *üi* (e.g. /nüid/ vs /nüüd/). In each of the example pairs, the first variant was used as a common option in earlier texts, but has by now mostly fallen out of use (though still used sometimes as a stylistic variant or dialectal speech). Figure 6 presents the variation in the corpus of texts. The points show the usage

proportion for each text, ranging from 1 (100% modern usage) to 0 (0% modern usage). The points on all figures are colored ranging from red to black (light to dark in print), with x marking texts with no suitable variants. Noticeably, some works use both variants within them. This can be an interesting question in its own right – sometimes it shows free variation, sometimes contributions of several authors within one publication, sometimes possible stylistic uses (e.g. showing dialectal speech in a fiction text) or distinctions that had not been made in the methodology. Right now, we can use these patterns as an example.

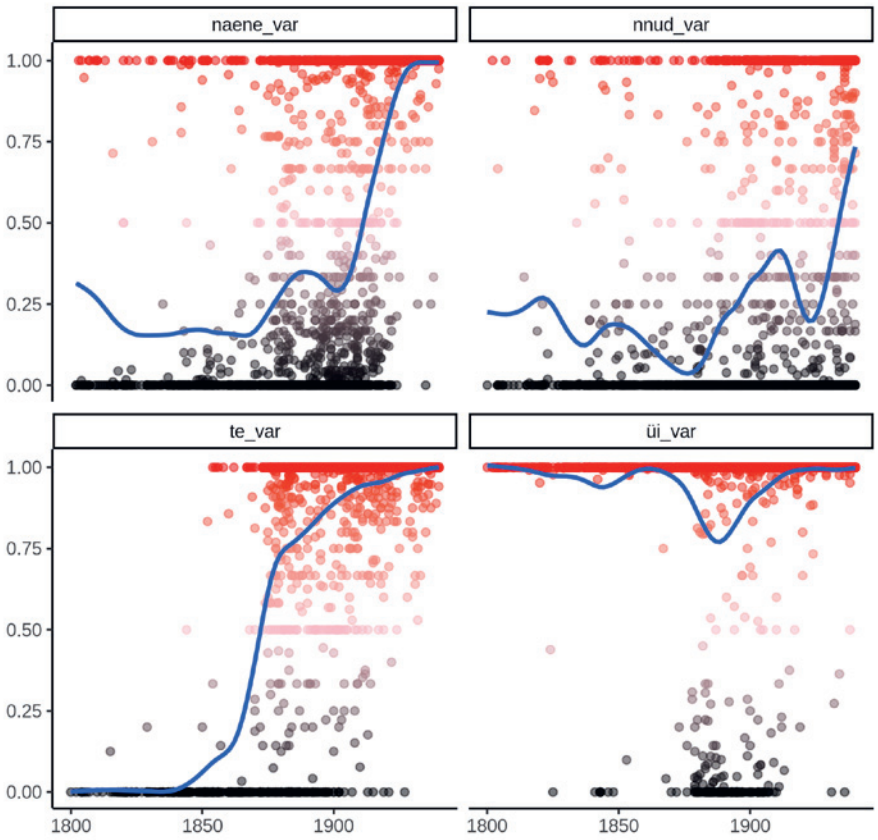

Figure 6. The use of four linguistic variables across the time span of the corpus 1800–1940

We can see that the variants show quite different overall trends. Three of the four (*naene*, *te*, *üi*) became established as dominant by the end of the period, while the fourth (*nnud*) is increasing in frequency and reaches dominance after the corpus concludes. However, the patterns of growth are quite different between the variables – *te* seems to go up quite abruptly and decisively, even though some variation remains. *Naene* and *nnud* show much slower process of change, with some noticeable bumps on the way. *Üi* follows the modern standard early on, and new variation is introduced around the 1880s, which again disappears around 1910. While all the variables became the standard eventually, likely different social processes were behind them.

One thing that we can look at with the corpus for some authors is their dialectal background. It is possible to take their birth location from the abovementioned databases and check where on the Estonian dialect map it was. This is not an exact measurement, as some speakers may have had parents with other backgrounds or moved away to another location when they were young (though moves to outside of dialect areas were not common at the time), we can use this as a proxy. On Figure 7, we can see the use of the *üi* variable across different dialectal areas. It can be seen that the bump we saw in Figure 6 is mostly driven by the Central dialect and the Western dialect which have *üi* as a spoken language feature. It can be supposed that the introduction of the *üi* variant in the late 1880s was connected with more writers coming to participate in the written language community and experimenting with using their own dialects in writing. Showing this connection would require a somewhat more detailed study into the writers and the publications.

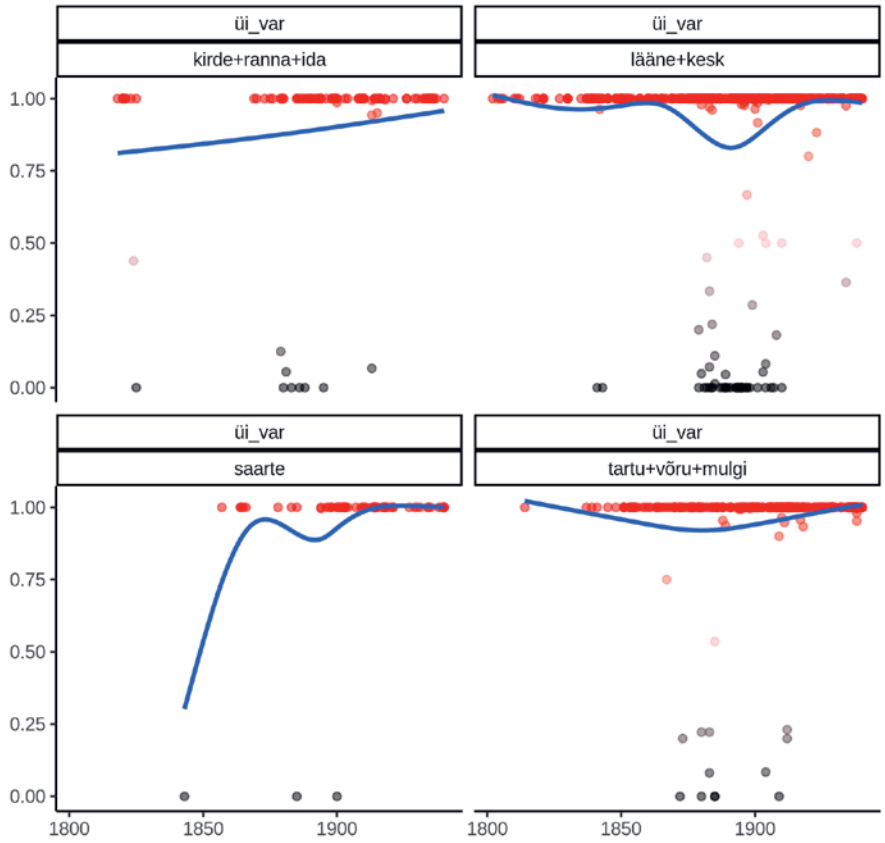

Figure 7. The variety of use of *üi* by dialect area with some neighboring dialect areas combined

6.2. Studying intra-individual variation

As mentioned above, since the corpus includes multiple works from a number of authors, it is possible to use this to study the changes in language use over their lifetime and possible societal trends that they may reflect. In Figure 8, we can see the writers with the most works in the corpus and their use of the *üi* variable that we discussed above. Looking at these writers, we can see that almost only Mats Tõnisson uses the archaic variant a lot, while a few of the writers use it once or twice. Given that Mats Tõnisson is such a prolific writer in the corpus, it should be checked whether the bump in the *üi* use may be just due to one or two authors. On a closer look, we find that of the writers active in 1880–1910, when the bump in the data occurs, 16 out of 60 (27%) Western/Central dialectal writers relied mainly on the *üi* form in at least one work, while only 13 out 64 (20%) writers from a different dialectal background did the same. A subtle difference is apparent, but individual linguistic histories may be quite complex. Based on this basic overview it is difficult to say if the forms emerged through dialectal influence or some other social trend.

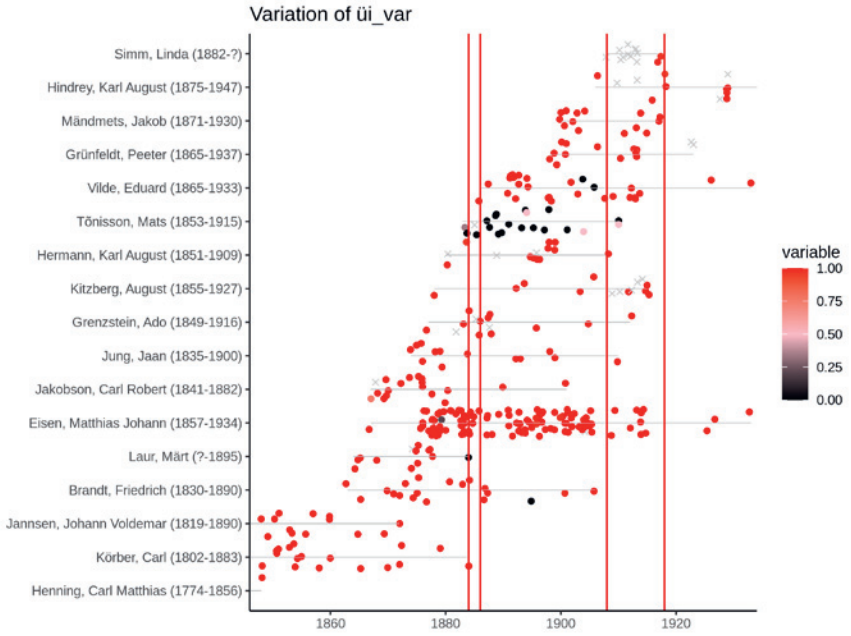

Figure 8. The top 20 most prolific writers in the corpus and their use of the *üi* variable. The vertical lines are points in time when prescriptive advice was given to use the variant of the same color

Figure 9 gives us an overview of another variable, *naene*, which shows a bit more variation. Some writers use their preferred forms throughout their career, while others change during their lifetime. Here, we can turn to another possible layer of analysis – external influences on language. On both these figures, vertical lines have been drawn to mark an outside event, here a prescriptive recommendation made by language activists in some publication – either a newspaper or a dictionary

(collected from Raag 2008). For example, Karl August Hermann recommended in his 1884 Grammar of Estonian Language for Schools and Self-learning that the form *naene* should be preferred to *naine* from the then competing forms. *Naene* is currently considered an archaic form and thus marked with black on the figure, while *naine* and recommendations to use this are marked with red. This seems to have had little influence on the writers: among the most prolific writers, very few switch from *naine* to *naene* even temporarily. On the other hand, in the 1909 language conference and with Johann Voldermar Veski's prescriptive orthographic dictionary of 1918, a recommendation was given to use *naine* which does seem to have had some impact – for example, Jakob Mändmets and Peeter Grünfeldt switch to *naine* around that time. These connections can be found with other variables and general patterns as well, although they require a more in depth study. Studying the role of external events in language change, such as prescriptive works, literary circles that authors joined or cities they moved to, can be an interesting avenue of research here.

Figure 9. The top 20 most prolific writers in the corpus and their use of the naene variable. The vertical lines are points in time when prescriptive advice was given to use the variant of the same color

6.3. Studying the background of the writers

There is a variety of background knowledge on the writers presented in this corpus, collected from a few biographical databases – their education, their profession, whether they were married, etc. It is possible to use that information to study the language use in different parts of society from a historical sociolinguistic vantage

point. Here, we consider simple metadata that most authors in the dataset have – their year of birth. There are a total of 1188 unique author names represented in this corpus and 810 of them have this information. The year of birth can be used to study the cohort effects in the mechanisms of language change. Linguistic variants can spread horizontally within a community, or they may show cohort effects where older authors are likely to stay conservative when new forms spread.

In Figure 10 we can see the birthyears of authors on the vertical axis and the year of publication on the horizontal axis. The variants are again shown with distinct colors. In addition to the point data, we can look at the density graphs on the side of the figure. They are higher where there are more points of the given type, so their patterns show both – the number of works as well as the forms chosen within these works. Keeping this caveat in mind, that absolute peaks may be due to more writers or writings on some years, we can still look at the relative variation in particular variants. The *te* variant seems to be quite common up until the people born in the 1840s and onward and up until around 1875. Afterwards, some older writers continue using their old form, but new writers that start using the *te* variant quickly become quite rare. Comparing the density graphs, the modern *ti* variant abruptly becomes more common around the birthyear of 1850. If we compare this pattern to the *nnud* variant, then around 1850 both variants were quite common, and both variants show a peak there, following mostly very similar distributions.

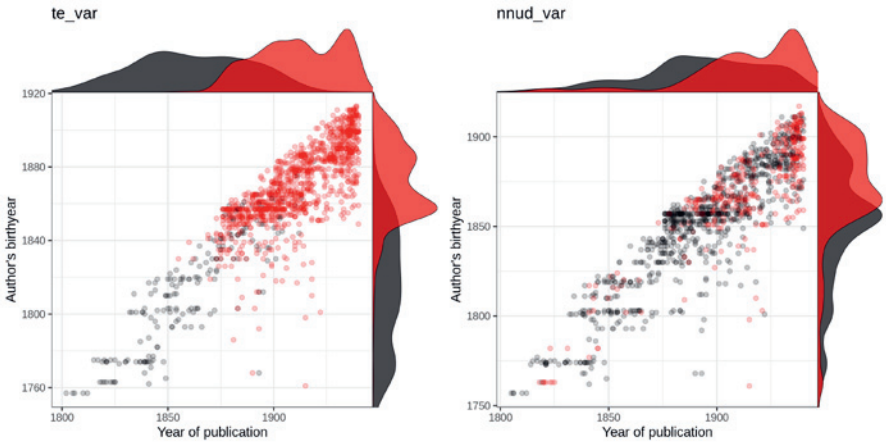

Figure 10. A scatterplot with density of the author's birthyear and the year of publication for te and nnud variables. The variables were here simplified by rounding them to either 1 or 0 to distribute them between two colors

For both variables, the later period shows a dominance towards the more common form, but for *nnud* both variants remain quite frequent, while for *te*, the modern variant has almost completely replaced the older variant. This coincides with the adoption of the new standard in spelling in the 1870s. What we can see though is that the change for the *te* variable is very abrupt, where the new generation of writers predominantly turn to the *ti* variant, with some older authors also catching up slowly. The changes in the *nnud* variable are much more gradual. Here, the modern variant *nud* emerges as an almost equal alternative to the older *nnud* for most of the

corpus time span. This can show that the language community found the old form still quite natural, or, judging by the rows of different colors on the graph, that writers were quite conservative in their choice in one or the other, and other background features besides the birth year may have played a role. While there are likely to be elements of generational change present in both cases, the changes for *te* are much more abrupt and the changes for *nnud* much more gradual. A more detailed study here could look at how conservative the individual writers were, and whether the choice in forms also correspond to the author's background or life events in some way. The visualization shown here can only provide a first exploratory look into it.

7. Conclusion

This paper presents a stratified corpus of written Estonian for 1800–1940. It is designed with historical sociolinguistic research in mind, but can be used for a variety of research questions. It makes the case that looking at a combination of different digitized materials along with their associated metadata allows researchers to address questions that would not be possible when considering just one or two sources. The corpus was built with the help of public databases on both the texts and the author metadata and relies on the focussed work of the maintainers of these archives; as such, it is more an enterprise in integration of databases than compiling one anew. Having the data in one place with an easy access allows questions to be addressed that previously have been difficult. The paper provides three aspects as examples: 1) studying patterns across many texts on long term language change, possibly linked with metainformation on the writers, 2) studying intra-individual variation and lifetime changes of the writers, possibly linked with external or lifetime events that may introduce them, 3) studying the mechanisms behind particular language changes based on the metainformation in the dataset. These are simple use cases provided to show the potential of the dataset as a research material, and for each of these questions more thorough analysis would be needed to understand the mechanisms that shaped the language use of the authors. Naturally, the corpus could be used for other types of analyses in linguistics or digital humanities.

8. A note on open science

This corpus publication is aimed to facilitate open science. The codes to analyse it are available with the data. The corpus is stored online and can be easily updated also by guest contributors or developed into bigger or smaller versions of it. The upside of having an unbalanced corpus is that it can be updated and improved in small increments as it is used in further studies. Much of the metadata provided with the database is based on automatic workflows, and has been verified for accuracy only in parts of it. The author has added a list of future steps to improve the corpus in the repositories. Researchers who use the corpus are encouraged to familiarize themselves with the content and consider any issues of representativeness of its contents for a particular research question themselves.

9. Data and code availability

The corpus along with basic processing scripts is shared here https://osf.io/zbup2. The supplementary materials to reproduce the examples and figures in this paper are stored here https://osf.io/7t5pk.

References

De Smedt, Tom; Daelemans, Walter 2012. Pattern for python. – The Journal of Machine Learning Research, 13 (1), 2063–2067.

Evans, Mel 2013. The Language of Queen Elizabeth I: A Sociolinguistic Perspective on Royal Style and Identity. Oxford: Wiley Blackwell.

Gerlach, Martin; Font-Clos, Francesc 2020. A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. – Entropy, 22 (1), 126. https://doi.org/10.3390/e22010126

Havinga, Anna; Langer, Nils (Eds.) 2015. Invisible Languages in the Nineteenth Century. Historical Sociolinguistics 2. Oxford: Peter Lang Verlag. https://doi.org/10.3726/978-3-0353-0760-3

Hennoste, Tiit; Kaalep, Heiki-Jaan; Muischnek, Kadri; Paldre, Leho; Vaino, Tarmo 2001. The Tartu University Corpus of Estonian Literary Language. Congressus Nonus Fenno–Ugristarum Pars V. Tartu, 337–344.

Kaalep, Heiki-Jaan; Vaino, Tarmo 2001. Complete morphological analysis in the linguist's toolbox. – Congressus Nonus Internationalis Fenno-Ugristarum Pars V. Dissertationes sectionum: linguistica. II. Tartu, 9–16.

Labov, William 1972. Some principles of linguistic methodology. – Language in Society, 1 (1), 97–120. https://doi.org/10.1017/S0047404500006576

Mäkelä, Eetu; Lagus, Krista; Lahti, Leo; Säily, Tanja; Tolonen, Mikko; Hämäläinen, Mika; Kaislaniemi, Samuli; Nevalainen, Terttu 2020. Wrangling with non-standard data. – Sanita Reinsone, Inguna Skadiņa, Anda Baklāne, Jānis Daugavietis (Eds.), Proceedings of the Digital Humanities in the Nordic Countries 5th Conference Riga, Latvia, October 21–23. CEUR Workshop Proceedings, 2612, 81–96. https://urn.fi/URN:NBN:fi-fe2021042826375

Orasmaa, Siim; Petmanson, Timo; Tkachenko, Alexander; Laur, Sven; Kaalep, Heiki-Jaan 2016. EstNLTK – NLP Toolkit for Estonian. – Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2460–2466.

Petré, Peter; Anthonissen, Lynn; Budts, Sara; Manjavacas, Enrique; Silva, Emma-Louise; Standing, William; Strik, Odile A. 2019. Early Modern Multiloquent Authors (EMMA): Designing a large-scale corpus of individuals' languages. – ICAME Journal: Computers in English Linguistics, 43 (1), 83–122. https://doi.org/10.2478/icame-2019-0004

Raag, Raimo 2008. Talurahva keelest riigikeeleks. Tartu: Atlex.

Raumolin-Brunberg, Helena; Nevalainen, Terttu 2007. Historical sociolinguistics: The corpus of early English correspondence. – Joan C. Beal, Karen P. Corrigan, Hermann L. Moisl (Eds.), Creating and digitizing language corpora. London: Palgrave Macmillan, 148–171. https://doi.org/10.1057/9780230223202_7

Schiegg, Markus 2022. Flexible Schreiber in der Sprachgeschichte. Intraindividuelle Variation in Patientenbriefen (1850–1936). Heidelberg: Winter. https://doi.org/10.33675/2022-82538575

Silverstein, Michael 2003. Indexical order and the dialectics of sociolinguistic life. – Language & Communication, 23 (3–4), 193–229. https://doi.org/10.1016/S0271-5309(03)00013-2

Tolonen, Mikko; Mäkelä, Eetu; Lahti, Leo 2022. The anatomy of Eighteenth Century Collections Online (ECCO). – Eighteenth-Century Studies, 56 (1), 95–123. https://doi.org/10.1353/ecs.2022.0060

Uiboaed, Kristel; Kyröläinen, Aki J. 2015. Keeleteaduslike andmete ruumilisi visualiseerimisvõimalusi. – Eesti Rakenduslingvistika Ühingu aastaraamat, 11, 281–295. https://doi.org/10.5128/ERYa11.17

Ulbrich, Christiane; Werth, Alexander. 2021 What Is Intra-individual Variation in Language? – Alexander Werth, Lars Bülow, Simone E. Pfenninger, Markus Schiegg (Eds.), Intra-individual Variation in Language. Trends in Linguistics: Studies and Monographs 363. Berlin–Boston: De Gruyter Mouton, 9–44. https://doi.org/10.1515/9783110743036-002

Underwood, Ted; Kimutis, Patrick; Witte, Jessica 2020. NovelTM Datasets for English-Language Fiction, 1700–2009. – Journal of Cultural Analytics, 5 (2), 13147. https://doi.org/10.22148/001c.13147

Vandenbussche, Wim; Elspaß, Stephan 2007. Introduction: Lower class language use in the 19th century. – Multilingua, 26 (2–3), 147–150. https://doi.org/10.1515/MULTI.2007.007

Van der Wal, Marijke J.; Rutten, Gijsberg J. 2013. Ego-documents in a historical-sociolinguistic perspective. – Marijke J. van der Wal, Gijsberg J. Rutten (Eds.), Touching the Past: Studies in the Historical Sociolinguistics of Ego-documents. Advances in Historical Sociolinguistics 1. Amsterdam: John Benjamins, 1–17. https://doi.org/10.1075/ahs.1.01wal

Databases and resources

ENB = Estonian National Bibliography. http://data.digar.ee

CELL = The Corpus of Estonian Literary Language of the 20th century. https://www.cl.ut.ee/korpused/baaskorpus

DIGAR = Digital Archive of the National Library. http://www.digar.ee

DNB = German National Bibliography. https://www.dnb.de/EN/Professionell/Metadatendienste/Metadaten/Nationalbibliografie/nationalbibliografie_node.html

DSPACE-UT = Electronic repository of University of Tartu. http://dspace.ut.ee

EKI_PLACENAMES = The Place Names Database (KNAB) at the Institute of the Estonian Language. https://www.eki.ee/knab

ESTLIT = The digitization programme Estonian Literature. https://et.wikipedia.org/wiki/Eesti_kirjandus_(toetusprogramm)

GEONAMES = GeoNames online database. https://www.geonames.org

ISIK = Biographical database at the Archival Library of the Estonian Literary Museum. http://www2.kirmus.ee/biblioserver/isik

KIVIKE = Database of items for Estonian Literary Museum. https://kivike.kirmus.ee

KRZW100 = The Century of Kreutzwald project. https://kreutzwald.kirmus.ee

VAKK = Corpus of Old Literary Estonian. https://doi.org/10.15155/TY.0005

VEPER = Biographical database VEPER maintained by Tallinn University. http://isik.tlulib.ee

VIAF = Virtual International Authority File. https://viaf.org

WIKIDATA = Wikidata database. https://www.wikidata.org

WIKIPEDIA = Wikipedia, the free encyclopedia. https://et.wikipedia.org

WIKISOURCE = Wikisource environment. https://et.wikisource.org

EESTI KIRJAKEELE KIHILINE KORPUS 1800–1940

Peeter Tinits
Tartu Ülikool, Tallinna Ülikool, Eesti Rahvusraamatukogu

Artikkel esitleb eesti kirjakeele kihilist korpust 1800–1940. Selle eesmärk on võimaldada sotsiolingvistilisi võrdlusi toonases keelekogukonnas, ühendades korpusandmeid autorite keelekasutusest nende kohta leiduva taustainformatsooniga (nt osadel autoritel sünniaeg, kodumurre, haridustee). Korpus on loodud koondades digiteeritud tekste avalikest allikatest ja digiarhiividest (nt Eesti Rahvusraamatukogu, Eesti Kirjandusmuuseum, Tartu Ülikool, Vikitekstid). Metainformatsiooni teoste kohta on kogutud Eesti rahvusbibliograafiast ja sellega ühendatud rahvusvahelistest andmebaasidest (VIAF, Wikidata, Saksa rahvusbibliograafia) ning isikuloolistest andmebaasidest ISIK ja VEPER. Metaandmeid on analüüsi hõlbustamiseks puhastatud, harmoniseeritud ja struktureeritud. Artikkel kirjeldab korpuse loomist ja selle sisu. Andmete koondamise tulemusena on kogus 4412 teksti 1188 erineva nimega autorilt, mis hõlmab umbes 11% sellel perioodil Eesti rahvusbibliograafias registreeritud teostest. Autoritega on seotud metainformatsiooni, kus võimalik. Artikkel esitab kolm näidisjuhtu, kus korpusest võib uurimistöös kasu olla:

1) pikaajaliste trendide analüüsiks keeles, kasutades ka kirjutajate taustainformatsioon;
2) autorite keelekasutuse muutumist eluea jooksul, sidudes neid väliste muutustega nagu näiteks avaldatud õigekeelsussoovitused;
3) keelemuutuste mehhanismide uurimine autorite taustainformatsiooni kaudu, näiteks tuues esile põlvkondlikke kihistusi keelekasutuses.

Toetudes suurele tekstikogule ja tekstidega seotud metaandmetele on võimalik detailsemalt uurida toonast keelekogukonda ning selle rolli eesti keele kujunemises. Korpus on avaldatud veebis, et võimaldada selle uuendamist metaandmete täiendamisel ja uute tekstide digiteerimisel.

Peeter Tinits (University of Tartu, Tallinn University, National Library of Estonia) has a background in linguistics and semiotics. His research interests encompass historical sociolinguistics, cultural evolution and digital humanities.
Ülikooli 18, 50090 Tartu, Estonia
peeter.tinits@ut.ee