User-friendly Search Possibilities for Early Latvian Texts: Challenges Posed by Automatic Conversion

Everita Andronova[1], Anna Frīdenberga[2], Lauma Pretkalniņa[1], Renāte Siliņa-Piņķe[2], Elga Skrūzmane[2], Anta Trumpa[2] and Pēteris Vanags[2]

[1] *Institute of Mathematics and Computer Science, University of Latvia, Raiņa bulv. 29, Rīga LV-1459, Latvia*
[2] *The Latvian Language Institute, University of Latvia, Kalpaka bulv. 4, Rīga LV-1050, Latvia*

Abstract

This paper deals with the Corpus of early written Latvian and explains the methodology for normalising historical spellings found in texts from the 16th–18th cc. It describes the types of replacements which will make searching early texts more convenient.

Keywords

Historical corpus, conversion of old spelling into modern, replacement algorithms

1. Introduction

Diachronic corpora are of high importance not only for linguistic research but also for those interested in other fields of humanities (literature, history, sociology, etc.). Historical spelling is a considerable obstacle for broader use of the Corpus of early written Latvian texts (henceforth the Corpus) among non-linguists. Work is currently underway on providing user-friendly search possibilities in the corpus.

2. The Corpus of early written Latvian texts: some remarks on its history and scope

The Corpus of early written Latvian texts was launched in 2003 after a short one-year project, but the origins of this initiative date back to the 1990s, when some texts from the 17th c. were manually typed in at the Institute of Mathematics and Computer Science, University of Latvia (henceforth IMCS, UL). A great deal of work in the digitalisation of Latvian texts covering different time periods has been actively carried out, but the main emphasis has of course been on modern texts, as they were crucial for Latvian language processing [1].

In 2002, the Corpus was developed with financial support from University of Latvia. This was a joint activity gathering together researchers from the IMCS, UL and the Department of the Baltic Languages, UL. It was one of the first projects in digital humanities in Latvia. The various stages and methodology of development of the corpus have been presented to the scholarly community elsewhere [2, 3].

The original sources were acquired from the National Library of Latvia, scanned and returned to the library. Both the Academic Library of the University of Latvia and the National Archive of Latvia, State Historical Archives of Latvia have become cooperation partners as well. One of the aims of the Corpus was to give researchers an opportunity to access these early Latvians texts in one repository, therefore

The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022.
EMAIL: everita.andronova@lumii.lv (A. 1); anna.fridenberga@lu.lv (A. 2); lauma.pretkalnina@lumii.lv (A. 3); renate.silina-pinke@lu.lv (A. 4); elga.skruzmane@lu.lv (A. 5); anta.trumpa@lu.lv (A. 6); peteris.vanags@lu.lv (A. 7)
ORCID: 0000-0003-1865-4611 (A. 1); 0000-0002-6444-5581 (A. 3); 0000-0002-5553-2165 (A. 4); 0000-0001-6022-0433 (A. 6); 0000-0003-0718-364X (A. 7)

not only word and frequency indices and a concordancer, but also facsimiles are available on the corpus platform (http://senie.korpuss.lv/). For a long time, this was the only public resource providing access to Latvian early texts, and it was highly appreciated by scholars and students. At the moment Latvian sources are available not only at the Latvian National Digital Library (http://gramatas.lndb.lv/), the largest resource of Latvian books, periodicals, maps and recordings, but also scattered across European libraries where intensive digitalisation is taking place. For instance, the digital copy of G. Dreszell's Catechism 'Swähta Bährno=Mahziba' (1682) is housed at the Royal Danish Library (https://www.kb.dk/e-mat/dod/12089000708F-color.pdf), G. Elger's 'Geistliche Catholische Gesänge' (1621) is kept at Vilnius University library and they have kindly passed the scan to developers of the Corpus. The development of the Corpus is still in progress and is still being supplemented with new sources (cf. [4] on adding short texts to the Corpus, mostly the occasional poetry of the 18th c.).

The scope of the Corpus is Latvian texts from the beginnings of the written tradition in the early 16th century until 1800. These are mostly printed Latvian monolingual sources (with some supplementary texts in German or Latin). A couple of bilingual dictionaries (German-Latvian and Latvian-German) have been added to the Corpus. Although the major sources are printed texts, some transcripts of the manuscripts have been also included (see [5] on the issues of decrypting the Statutes of Linen weavers (1625) housed in the National Archive of Latvia, State Historical Archives of Latvia). Typically for the time, the texts represented in the Corpus are mostly religious ones (hymnals, texts of catechisms, holy scriptures, the Lord's Prayer, etc.) and mostly translations from different German sources (but also from Latin and Polish). Therefore, we can trace a number features of German and Latin origin in the language of early printings (for more on the linguistic characteristics of early texts, see [6, 7, 8]). Nevertheless, original texts have also been produced; one of the most remarkable is the 1,200-page 'Sermon book' by G. Mancelius published in three parts in 1654 and comprising historical and ethnographic facts along with nice rhetorical figures of speech.

Due to the fact that the original sources are scattered across Europe, the developers opted for a full-text corpus in order to facilitate access to them. This explains the choice of interactive word indices for almost every single source (large sources lack this), with the possibility of navigating from the index within the entire text, see Fig. 1.

Figure 1: Interactive word index and a window with full text

Some non-standard metadata were added, thus selection of texts by author, century and text type is offered (see Fig. 2.).

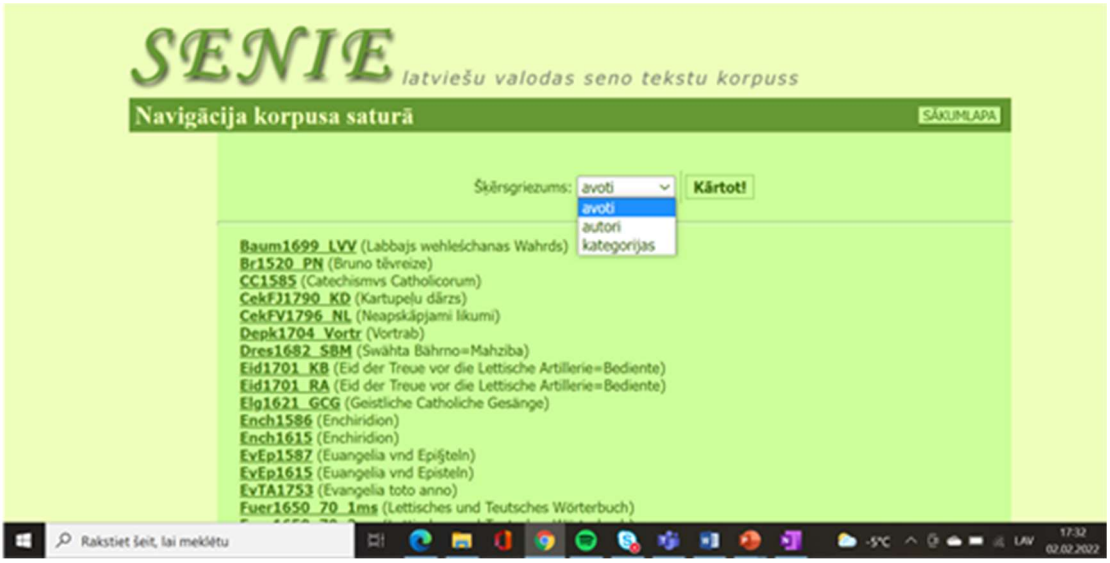

Figure 2: Navigation in the Corpus

Originally the Corpus was supported by in-house mark-up and a Java-based concordancer. At the end of 2021, it had a volume of 1.1 million running words.

In 2022, a new release of the Corpus is in progress. It will be moved to the new corpus platform (more on this below) and a number of new sources have been added to the corpus. The recent characteristics of the Corpus are as follows: the size is ca 1.75 million tokens; there are more than 100 different sources included. Every single source has kept its unique identifier which was assigned at the beginnings of the Corpus and shows some metadata about the source. It consists of an abbreviation of the author, source and year of publishing, thus ensuring sustainable development and not confusing users.[2] For example, the identifier *Manc1638_PhL* stands for the German-Latvian phrase book 'Phraseologia Lettica' published by Georg Mancelius in 1638. 35 known authors and a number of unknown authors are represented in the Corpus.

3. Historical spelling in early prints: experience of others

A good overview and comparison of different methods for normalising historical spellings is presented in Bollmann's study [9]. Here, we would like to describe the experience of researchers in neighbouring countries sharing a similar history of the development of early printed texts.

Estonian colleagues working on the Corpus of old written Estonian (https://vakk.ut.ee/) have developed the converter *Vakker* [10], which also uses conversion rules and later consults a dictionary to deal with early sources. Polish researchers developing the KorBa corpus (17th–18th cc. texts with morphosyntactic annotation, https://korba.edu.pl) offer their users transliterated and transcribed (normalised) texts. They deal with spelling normalisation only, keeping the original inflectional endings and lexis unchanged. The aim is to have the spelling of the texts as similar to modern Polish as possible in order to facilitate search in the corpus [11].

Lithuanian researchers also aim to create a universal search engine dealing with different spellings in their Database of old writings (http://seniejirastai.lki.lt/home.php). Thus, historical spellings are rewritten in the modern Lithuanian alphabet, unifying graphemes and ignoring orthography, but taking into account normalisation of phonetics (eliminating dialectal features and solving assimilation issues). The pre-processing handles forms where rules cannot be applied. Although the morphology is not changed, some of the rewriting rules apply to the morpheme level. The number of rules applied varies from 74 to 495 [12].

Researchers dealing with Early New High German texts have presented their semi-automatic normalisation tool *Norma* [13]. The normalisation is performed in two stages. A distinction is made

[2] The list of source abbreviations is available at http://senie.korpuss.lv/abbrevs.jsp.

between *normalisation* (preferring forms close to original) and *modernisation* (preferring forms close to modern language).

To sum up, several approaches can be applied in the normalisation of old writings (like wordlist substitutions, the rule-based method). The frequently used terms *transliteration* [cf. 14] and *transcription* in Latvian studies are from time to time used with different meanings. Our working group uses the term *conversion* to denote the process and result of such a change, as it covers several steps: transliteration, transcription, and adaptation [15]. Rules are set for every single source in the Corpus, and subsequently the software engineer converts the text into normalised spelling.

4. Coding in the Corpus: From *Windows-1257* to *Unicode*

The original texts were mostly printed using Fraktur and other blackletter typefaces, but in the Corpus they are presented in Latin transliteration. The ASCII code set – single and combined characters – was used in text processing. These combined symbols represented a number of diacritic marks typical for early writings. Thus, we have 7 letters a with different diacritic marks encountered in the Corpus:

1. à 00E0 C1
2. á 00E1 C1
3. ã 00E3 C1
4. â 00E2 C1
5. ȁ 0227 Latin Extended-B
6. ä 00E4 C1
7. ⱥ 2C65 Latin Extended-C

In 2017, the conversion into *Unicode* format[3] was carried out to ensure more precise visualisation of the original text and to facilitate its comprehension. Linguists created 73 tables for every single source. Unicode files were added to the Corpus as separate items for downloading, and all used symbols were merged in a single table (see http://senie.korpuss.lv/unicode/tabula.pdf). The conversion of new sources added to the Corpus continues, and the number of tables has thus far reached 168.

Since the Corpus was created over a long period of time and sources were added gradually, different symbols were introduced for the same grapheme, e.g., the grapheme *ë* in VLH1685_Sal was represented as *e#*, but the same letter *ë* in the manuscript source Fuer1650_70_1ms was presented as *e″*. During the conversion to *Unicode* such cases were unified, and this grapheme is in all cases represented with *Unicode* symbol *ë* (00EB), which visually is the closest version to the original graphemes.

5. On-going modernisation of the Corpus (2020–2022)

In 2020 the project 'Digital Resources for Humanities: Integration and Development' was initiated to support development and a wider access of digital resources. The modernisation of the Corpus will be developed further within this project: the conversion from the old spelling into the modern one is being carried out and a switch to the elaborated corpus management system *NoSketch Engine* is in progress.

6. Conversion into Modern Latvian orthography

Unicode files now serve as input data for conversion of the texts into Modern Latvian orthography, which will provide easier search and comprehension of the corpus material. This task presupposes not only transliteration, but also morphological adaptation of Old Latvian spelling to the modern one. Both procedures can be facilitated and accelerated by elaboration of certain rules of automatic conversion. This paper presents problems that occur when performing automatic transliteration.

[3] The project was funded by University of Latvia within the project of academic development 'Switch of the Corpus of Early Written Latvian to *Unicode*' (LU, No AAP2017/63).

6.1. Methodology

As this research deals with the very first Latvian sources, the number of spelling and morphology versions is very high and differs from source to source. The Corpus comprises both printed and handwritten texts of different length. There is no stable Latvian orthographic system in this period yet; we can observe attempts of different authors to offer their own writing systems. Taking into account the facts mentioned above, we conclude that each source or at least each author requires an individual conversion approach. In order to provide the best possible results, we opted for creating hand-crafted conversion rule tables for every single source. These rules do not use any Latvian lexicon or language processing tools, because to our best knowledge there are none for early Latvian. The historical dictionary of Latvian (16–17th cc.) (www.tezaurs.lv/lvvv) is still too small (only ca 2000 entries) to be of significant help for large-scale transliteration.

The accepted conversion process consists of the following steps:

1. Development of tables of conversion rules for every source. Each rule is deterministic, i. e, is applied for every token it matches and rules can stack on each other, namely, each token can undergo multiple rule applications to reach its final converted form.
2. Implementing of tables in the software algorithm and automatic conversion.
3. Post-editing: rereading of the converted texts (all or part of it, if the text is huge) and detecting errors.
4. Error analysis and supplementation/correction of the tables, evaluating the usefulness of correction if possible.
5. Repeated automated conversion.
6. Quality assessment.
   It should be noted that converted text will not be the same as modern standard Latvian (the main emphasis lies in the recognisable root of the word, length of vowel in suffixes is ignored at this stage).

6.2. Characteristics of early Latvian sources and spelling conversion applied

Our recent experience is based mostly on the texts of the 1st (from the beginning of the 16th c. until the 1620s) and the 2nd (1631–1680s) period of Old Latvian. These texts are characterized by the greatest amount of spelling variation, and thus they hopefully cover most of the potential issues.

A high level of inconsistency in spelling and ambiguity of graphemes and grapheme combinations is typical for the first period sources, which consist of mostly anonymous translations of religious texts of various length. This can be illustrated by the large variety of spellings of the word 'heart' within one single source, *Szyrdtcz*, *Szirdes*, *ßirde*, *ßirdtcz*, *ßyrdtz*, *ßyrdtcz* (UP1587). A comparison of several sources reveals even greater diversity: *Szirdees*, *Sczyrdtcz*, *czirdtcz*, *ßirde* (Ench1615). Of course, the conversion tables for these sources include rules converting a letter to another letter (*ä>ā*), a grapheme combination to a letter (*ſch>š*), or one grapheme combination to another (like *dcz>dz*), but due to the high orthographic inconsistency, the source tables in this group have a disproportionate number of so-called individual correspondences when the root is replaced by the root or a whole word for a whole word (*czedaatz>dziedāts* 'sung'). As a result, the number of conversion laws in this group of sources is relatively high, for example, the 'Vndeudsche Psalmen' (UP1587) has 1024 laws.

Nine mid-17th century sources by Georg Mancelius make up the largest group of the second period. G. Mancelius has an improved and more systematic spelling in comparison to texts of the previous period; therefore, it was assumed that letter-to-letter replacement or letter-to-grapheme combination correspondences would predominate in the conversion of Mancelius' works. However, this assumption was not completely borne out [16]. For example, in 'Ten conversations' by Mancelius (Manc1638_Run) it is possible to replace part of letters or grapheme combinations with a particular letter or grapheme combination in modern writing. So, *w>v* (*pļawas>pļavas* 'meadows'), *v>u* (*vs>uz* 'to', *Vppe>Upe*

‘river’), *ñ>n* (*mañ>man* ‘for me’), *ä>e* (*rättais>retais* ‘the seldom’, *wätz>vecs* ‘old’, *Bährni>Bērni* ‘children’), *à>ā* (*Zeemà>Ciemā* ‘guest’), *ee>ie* (*Deena>Diena* ‘day’); in turn *ie>ī* (*brienums>brīnums* ‘miracle’), *gh>g* (*ghann>gan* ‘enough; ever’), *tſch>č* (*tſchettrus>četrus* ‘four’) etc. Double consonants in most cases can be replaced by one consonant, e. g., *bb>b* (*drebb>dreb* ‘shiver’, *labba>laba* ‘good’), *ļļ>ļ* (*zeļļu>ceļu* ‘I pick up’, *packaļļ>pakaļ* ‘after’), *nn>n* (*mann>man* ‘to me’), *tt >t* (*Ratti>Rati* ‘carriage’), *rr>r* (*turr>tur* ‘keeps’, *Barribu>Barību* ‘food’), *ŗŗ>ŗ* (*kuŗŗam>kuŗam* ‘to whom’) etc., a short vowel and the following letter *h* can be replaced by a long vowel, e. g., *äh>ē* (*Dähls>Dēls* ‘son’), *uh>ū* (*truhx>trūks* ‘will lack’), *eh>ē* (*Drehbes>Drēbes* ‘clothes’), *ih>ī* (*dſihrehß>dzīrēs* ‘was going to’); the only exception is *oh>o* (*Ohrmans>Ormans* ‘a coachman’), denoting a diphthong.

The order of conversion rules is also crucial in many cases. For instance, the replacement *iß>iz* and *Jß>Iz* should be completed before all other changes involving *ß>s*. However, the number of exemptions for several letters or grapheme combinations is still very high. During the process of conversion, it was observed that, e. g., the usage of long *ſ* without a stroke corresponds to modern s and z in 50/50 cases; the grapheme combination *ſch* corresponds to modern *š* and *ž* equally. A decision was made to make conversion laws for separate grapheme combinations, namely, *dſ>dz*, *ſi>zi*, *ſt>st*, *ſp>sp*, *ſm>zm*, *ſl>sl*. If necessary, lexical substitution of root to root or lexeme to lexeme was carried out, eg., *Meſch>Mež* ‘forest’. The same issues concern the conversion of long s with a stroke *ſ*, as well as *z*, *y*, *x*, *ß* and the grapheme combination *tz*. After different attempts there are 190 rules set to be applied in certain order.

6.3. Description of conversion rules

On the basis of the conversion rule tables for the 16th and 17th cc. Latvian sources developed so far, we may identify three main conversion rule groups, each with subgroups:

1. Unambiguous graphemic correspondences:

1) grapheme-to-grapheme conversion, e. g., *à>ā* (*Dahrſ**à**>dārz**ā*** ‘in a garden’);

2) grapheme combination to letter, e. g., *tſch>č* (*Lah**tſch**us>lāč**us*** ‘bears’ Pl.Acc.);

3) letter to grapheme combination, e. g, *x>ks* (*attma**x**aht>atma**ks**āt* ‘to repay’);

4) grapheme combination to grapheme combination, e. g., *ee > ie* (*p**ee**därr>p**ie**der* ‘belongs’).

2. Positional (graphemic and morphemic) correspondences:

1) depending on the position in a word, e. g., in the beginning or in the middle of the word: *tz>c* (***Tz**ilwäkī>**c**ilvēki* ‘men’), in the middle or at the end of the word: *tz>c* (*tape**tz**>tāpē**c*** ‘therefore’, *Swe**tz**e>sve**c**e* ‘candle’); at the end also *tz>ts* (*fäll**tz**>zel**ts*** ‘gold’) or *tz>ds* (*Ghall**tz**>gal**ds*** ‘table’);

2) depending on neighboring letters, e. g., *aya>āja* (*ißgh**aya**>izg**āja*** ‘went out'); but *ty>tī* (*nack**ty**>nak**tī*** ‘at night’).

3. Individual (lexical) correspondences:

1) word roots, e. g., *ſwäht>svēt* (***ſwäht**itam>**svēt**ītam* ‘blessed’), here we also deal with position in the word, e. g., beginning of the word *tytcz>tic*, (***tytcz**am>**tic**am* ‘we believe’), but at the end of word *tytcz>tīts* (*rax**tytcz** >raks**tīts*** ‘written’);

2) separate lexemes, e. g., *föv>sev* ‘for oneself’.

Undeniably, the older the source, the more inconsistency is observed in the representation of different graphemes and phonemes. This is the reason why positional and individual correspondences are prevalent in the process of conversion of the texts from the earliest period (before 1631), which in turn increases the number of rules applied. Of course, setting individual correspondences is a time-consuming task, but this is the only way to recognise a part of the instances where graphemes are ambiguous.

Taking into account the development of Latvian writing, the number of conversion rules gradually decreases as fewer individual rules are needed and as writing becomes more homogeneous. The newer a source is and the more consistent the spelling it displays, the smaller number of positional and individual correspondences and the fewer conversion rules needed.

The sequence of conversion rules is crucial, e. g, only after the implementation of the law *ie>ī*, can the rule *ee>ie* be applied. In general, the sequence of correspondence rules is as follows: lexical – morphemic, graphemic.

The number of conversion rules also depends on the size of text. The number of rules varies from 37 rules in the Lord’s Prayer to 1024 correspondences in ‘Vndeudsche Psalmen’ (1587). As the

orthography of the 18th c. texts is similar to the spelling predominating at the end of the 17th c., it could be possible to create a conversion template which might be used for the bulk of the sources, with some minor variations.

7. Switch from in-house platform to *NoSketch Engine*

In 2022 a new corpus version was released. The corpus was moved from an in-house platform to the *NoSketch Engine* platform (http://nosketch.korpuss.lv/#dashboard?corpname=senie_unicode), because this old platform is not maintained any more. The corpus is now available on a par with other Latvian language corpora.

During the migration to the new platform, we paid special attention to preserve the unique address of every token the same as it was in the old version. The address format makes it very convenient to cite the particular wordform in articles and in the corpus-based Historical dictionary of Latvian (https://tezaurs.lv/lvvv/; [5]). The address consists of source identifier, page, line or book of the Bible, chapter and verse, as in Fig. 3.

Figure 3: Part of word index of the token *un* 'and' with addresses (http://senie.korpuss.lv/index.jsp?wordform=un&source=SENIE&sort=asc&limit=50&cols=4)

In *NoSketch Engine* the address is presented in a separate window, showing text identifier, year of publication and page number. Simple metadata have been provided (author, century, year of publication, title, text genre, and type (printed/ handwritten)).

Different languages are encountered in the sources (mostly German, Latin, Polish, but also Greek, Hebrew), which have been appropriately marked in the corpus. Even though these languages are not of primary interest to this research, it is worth mentioning that *NoSketch Engine* will provide search possibilities in these parts of texts; they were excluded for analysis in the old system.

NoSketch Engine offers us a concordancer and wordlists of the original forms. At the moment, a search can be done either by original forms or regular expressions describing original forms, but after completion of conversion, searches will also be available by converted forms and regular expressions describing converted forms. However, the search results (concordancer and wordlists) will be presented in the original writing. At the moment we do not plan to publish conversions as whole texts, as we fear that converted but not standardised text may confuse a number of corpus users with no research background in early prints.

8. Results and issues

If any incorrectly recognised or typed wordforms are noticed, they are corrected in the *Unicode* file. It turns out that pre-editing is preferable, e. g. expanded spacing in a word in a header should be eliminated (like *J E S U>Jesu* 'Jesus'). Pre-editing concerns only formatting, but obvious original spelling mistakes are defined as separate replacement rules. Therefore, the number of rules grows, but we can re-use them and get a new version of the converted text.

There are replacement rules supplemented with a list of exceptions; the number of exceptions might reach ca. ten in some cases.

In some languages where normalisation of historical spelling is performed, pre-editing takes care of dialectal forms. However, we decided to leave them as in the original, e. g. *ūz-* (the prefix *uz-* in modern Latvian), *āz-* (the prefix *aiz-* in modern Latvian), the verb form *jir* (*ir*) 'is' etc. Our practice is not to intervene in the original text. In addition to this, we cannot solve highly complicated linguistic issues in historical texts (thus, *svēts* 'holy' and its variant *švēts* are left as two forms because there is no clear agreement on this yet).

In the result we got a converted text and a *Unicode* file which is as close as possible to the original. After conversion post-editing is performed and mistakes are evaluated, new replacement rules are written.

Although this process is very time-consuming, the results show that source-based rules give rather precise results. A major bonus of this approach is that differences between sources do not introduce new errors in other, differently written sources, which was major issue in [14]. Another major improvement compared to [14] is the elimination of multiple transliteration variants per single token - since all conversion rules in this project are mandatory and deterministic, only a single transliteration per token is generated.

Up to now, all replacement rules have been written by linguists; no machine learning method has been applied. Hopefully it would be possible to create a kind of template with base rules for the conversion of further texts of the 18th c. in which spelling is not so idiosyncratic.

9. Conclusions

In this paper, we have described the methods used for normalisation of early Latvian sources, identifying three main conversion rule groups with subgroups: 1) unambiguous graphemic correspondences; 2) positional (graphemic and morphemic) correspondences; 3) individual (lexical) correspondences. This will make texts more accessible to scholars in the humanities.

10.Acknowledgements

We would like to express our gratitude to the Department of Baltic Linguistics, University of Latvia and the Latvian language institute supporting the corpus development at different stages during 2002–2021. We would especially like to thank Andrejs Spektors and Normunds Grūzītis from the AILab, IMCS, UL for their long-term guidance.

The modernization of the Corpus of early written Latvian has been undertaken within the framework of the National Research Programme 'Digital Resources of the Humanities' (No: VPP-IZM-DH-2020/1-0001) funded by Latvian Council of Science of the Ministry of Education and Science. This article has been prepared within the same project.

11.References

[1] A. Spektors, Latviešu valodas datorfonda izveide, Latvijas Zinātņu Akadēmijas Vēstis A 2 (2001) 74–82. URL: http://ailab.mii.lu.lv/aspekt/dfond.htm.

[2] E. Andronova, The Corpus of Early Written Latvian: current state and future tasks, in: Proceedings of Corpus Linguistics, Birmingham, UK, 2007. URL:

https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2007/245Paper.pdf.
[3] E. Milčonoka, Latviešu valodas 17. gadsimta teksti internetā, Baltu filoloģija XII (1), (2003) 139–150.
[4] E. Andronova, Short texts in the Corpus of early written Latvian (www.korpuss.lv/senie), in: S. Reinsone, I. Skadiņa, A. Baklāne, J. Daugavietis (Eds.). Proceedings of the 5th Conference on Digital Humanities in the Nordic Countries (DHN). CEUR Workshop Proceedings, volume 26, 2020, pp. 173–183. http://ceur-ws.org/Vol-2612/short1.pdf.
[5] E. Andronova, R. Siliņa-Piņķe, A. Trumpa, P. Vanags, The Electronic Historical Latvian Dictionary Based on the Corpus of Early Written Latvian Texts, Acta-Baltico Slavica 40. Pogranicze bałtycko-słowiańskie w aspekcie leksykalnym i leksykograficznym (2016) 1–37. https://doi.org/10.11649/abs.2016.018.
[6] P. Vanags, Die möglichen Formen deutschen Einflusses auf die grammatische und syntaktische Struktur der ältesten lettischen Texte, Linguistica Baltica 2 (1993) 163–181.
[7] P. Vanags, Latvian texts in the 16th and 17th centuries: beginnings and development, in: K. Ross, P. Vanags (Eds.), Common Roots of the Latvian and Estonian Literary Languages, Peter Lang, Frankfurt am Main etc., 2008, pp. 172–197.
[8] P. Vanags, German Influence on the Christian Discourse of Early Written Latvian, in: M. Kaukko, M. Norro, K.-M. Nummila, T. Toropainen, T. Fonsén (Eds.), Languages in the Lutheran Reformation. Textual Networks and the Spread of Ideas, Amsterdam University Press, Amsterdam, 2019, pp. 273–301. doi-10.5117-9789462981553-ch12.
[9] M. Bollmann, A Large-Scale Comparison of Historical Text Normalization Systems, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1. Minneapolis, Minnesota, 2019, pp. 3885–3898. URL: https://aclanthology.org/N19-1389.pdf. doi:10.18653/v1/N19-1389.
[10] K. Prillop, Kuidas märksõnastada vanu eestikeelseid tekste?, Keel ja Kirjandus, 2, (2004) 90−99. URL: https://vakk.ut.ee/avaleht/Prillop-KK-2-2004.pdf.
[11] W. Gruszczyński, D. Adamiec, R. Bronikowska, W. Kieraś, E. Modrzejewski, A. Wieczorek, M. Woliński. The Electronic Corpus of 17th- and 18th-century Polish Texts. Lang Resources & Evaluation. 56 (2022) 309–332. https://doi.org/10.1007/s10579-021-09549-1.
[12] M. Šinkūnas, Senųjų raštų rašybos keitimas paieškos sistemai, in: G. Judžentytė-Šinkūnienė, V. Zubaitienė (Eds.), Baltų kalbų tekstų ir žodžių reikšmės, Vilniaus universiteto leidykla, Vilnius, 2018, pp. 389–407.
[13] M. Bollmann, S. Dipper, J. Krasselt, F. Petran, Manual and Semi-automatic Normalization of Historical Spelling – Case Studies from Early New High German, Proceedings of the KONVENS-Workshop on Language Technology for Historical Text(s) (LThist2012), Wien, Austria, 2012. URL: https://www.linguistics.rub.de/~dipper/pub/lthist12.pdf.
[14] L. Pretkalnina, P. Paikens, N. Gruzitis, L. Rituma and A. Spektors, Making historical Latvian texts more intelligible to contemporary readers. Proceedings of the LREC Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage Objects, LREC 2012, 29–35. URL: https://www.researchgate.net/publication/230800163.
[15] E. Andronova, A. Frīdenberga, L. Pretkalniņa, R. Siliņa-Piņķe, E. Skrūzmane, A. Trumpa, P. Vanags, Latviešu valodas senāko rakstu pieminekļu konvertācija mūsdienu rakstībā: iepriekšējā pieredze un automatizācijas mēģinājumi, Aktuālas problēmas literatūras un kultūras pētniecībā: rakstu krājums, atb. red. Anita Helviga. Liepāja, LiePA, 27, (2022) 346–358. URL: https://dom.lndb.lv/data/obj/1035006.html.
[16] E. Andronova, A. Frīdenberga, L. Pretkalniņa, R. Siliņa-Piņķe, E. Skrūzmane, A. Trumpa, P. Vanags, Variantums kā konvertācijas izaicinājums: Georga Manceļa tekstu atveide mūsdienu rakstībā, Letonika (2022). To apper.