Baltic J. Modern Computing, Vol. 12 (2024), No. 4, 548-559
https://doi.org/10.22364/bjmc.2024.12.4.16

New Possibilities for Exploring Early Latvian Texts: Switching to the *NoSketchEngine*

Everita ANDRONOVA[1], Anna FRĪDENBERGA[2],
Lauma PRETKALNIŅA[1], Renāte SILIŅA-PIŅĶE[2],
Elga SKRŪZMANE[2], Anta TRUMPA[2], Pēteris VANAGS[2]

[1] Institute of Mathematics and Computer Science, University of Latvia (IMCS UL)
[2] The Latvian Language Institute, Faculty of Humanities of the University of Latvia

everita.andronova@lumii.lv, anna.fridenberga@lu.lv,
lauma.pretkalnina@lumii.lv, renate.silina-pinke@lu.lv,
elga.skruzmane@lu.lv, anta.trumpa@lu.lv, peteris.vanags@lu.lv

ORCID 0000-0003-1865-4611, ORCID 0009-0007-0445-6127, ORCID 0000-0002-6444-5581,
ORCID 0000-0002-5553-2165, ORCID 0000-0003-2237-134X, ORCID 0000-0001-6022-0433,
ORCID 0000-0003-0718-364X

Abstract. The variable writing system in early Latvian texts is a bottleneck for non-linguists wishing to explore SENIE, the Corpus of early written Latvian texts. The writing system also poses many challenges for linguists. The Unicode version of SENIE, launched on the *NoSketchEngine* platform (https://nosketch.korpuss.lv/#dashboard?corpname=senie_unicode) in 2022, offers significant new possibilities. After the process of normalization of historical spelling the access to the Corpus has become more user-friendly. Queries made in the Latvian National Corpus Collection (LNCC) (https://korpuss.lv/search) display search results in the early texts as well.

Keywords: diachronic corpora, historical writing, normalization, Latvian, the Corpus of early written Latvian, digitization, *NoSketchEngine*

1. Introduction

The development of diachronic corpora and analysis of historical language data meets several issues, starting from the availability of sources scattered around different places and even different countries and raising the question of representativity of such a corpus in general, and ending with historical orthography with a high level of spelling variants, posing barriers to the application of modern language tools to earlier texts. A lot of effort is put into tackling this and finding a comprehensive solution (cf. Piotrowski, 2022 for a detailed state–of–arts description).

In Latvia, the biggest resource developers and holders of early written texts are the National Library of Latvia (further in the text — NLL), the Institute of Literature, Folklore, and Art, University of Latvia (further in the text — ILFA, UL), the Latvian

Language Institute, UL (further in the text — LLI, UL) and the Institute of Mathematics and Computer Science, UL (further in the text — IMCS, UL).

The NLL, together with the National Archives of Latvia, the National Cultural Heritage Board, and the Cultural Information Systems Centre implemented a couple of activities concerning the digitization of cultural heritage of Latvia; this resulted in the Digital Library publicly available[1]. Here, *periodika.lv* is one of the most important resources dealing with the major challenges of processing historical texts (Zogla and Skilters, 2010). It should be emphasized that early Latvian texts here are available scanned, their OCR is done automatically, but they lack any post-editing; thus, the search results done in these early texts are quite noisy (see e.g. the Latvian song book (1796) which has facsimiles and raw text[2]). The ILFA, UL is creating a Digital Archive of Latvian folklore[3], which now has been incorporated in the digital platform *humma.lv* (Laime and, Reinsone, 2024). The IMCS, UL together with partners from LLI, UL has been involved in the long-term development of the Corpus of Early Latvian texts 'SENIE'[4] (further in the text — the Corpus).

Every institution has a different experience and practice in developing their resources: the intensive process of large scale scanning and digitization (NLL), crowdsourcing (ILFA, UL), and tiny text processing with manual post-editing of the Corpus (IMCS, UL). The common feature of all the resources is an inconsistent writing system in flux which means a high number of spelling variants. Although the main emphasis usually lies on the spelling of separate words, in terms of text processing, punctuation and hyphenation is also important. Thus, e. g., *periodika.lv* covers a three-century time span, the earliest source in their databases dates back to 1768–1769 when 'Latviešu Ārste', the first periodical in Latvian, was published. The Corpus of Early written Latvian texts includes sources from the 16–18th cc. One of the crucial issues common to all the above-mentioned sources is how to provide user-friendly search possibilities and to assure that your query returns accurate data.

This article will shed light on the experience with the development of the Corpus of Early written Latvian and switch to the *NoSketchEngine* (further in the text — *NoSkE*) platform.

The development of the Corpus has undergone several stages. First, a database of the first printed Latvian texts was initiated at the IMCS, UL in the mid-90s. The tasks comprised text collection, manual input, and manual crosscheck against the originals of the 16th–17th texts available in Latvia. Second, the development of the Corpus was carried out in 2002, and it was launched in close cooperation between IMCS, UL and LLI, UL, as well as the Faculty of Humanities, UL. For the needs of the Corpus, manual structural mark-up was applied, and a tailor-made corpus platform offering wordlists and frequency lists of each single source and the whole corpus were created at the IMCS, UL. In addition, all the scanned facsimiles were made publicly available to facilitate the studies of language history. The size of the Corpus in 2002 was 800 828 occurrences

[1] https://www.digitalabiblioteka.lv/?set_lang=en

[2] https://gramatas.lndb.lv/periodika2-wiewer/?lang=fr#panel:pp|issue:651104|article:DIVL17

[3] https://garamantas.lv/?lang=en

[4] https://korpuss.lv/id/Senie

(Andronova, 2007). Later on, new sources were added, expanding the scope of genres and timespans, and adding transliterations of handwritten sources. All the texts were added in the ASCII format to the Corpus using single and combined symbols. In 2017, the switch to the *Unicode* system was completed (Andronova, 2020) which allowed the move to more sophisticated software. Recently, the Corpus moved to the *NoSkE* platform and was included in the Latvian National Corpora Collection[5]. Thus, queries made in *korpuss.lv* return results from all the corpora, and data from 'SENIE' is listed among data from other corpora.

2. Corpus data characteristics

2.1. General Characteristics

On the new website[6], the Corpus SENIE[7] includes 172 documents, 2,087,165 words, and 2,827,101 tokens, of which 2,461,791 tokens, or 93.1% refer to the text in Latvian. This means that the number of documents in the corpus has almost doubled in the last two and a half years, while the number of Latvian words has increased by more than 40% (Andronova et al., 2022a).

In addition to this basic data, the website also provides a wide range of information on the metadata of the corpus, which at the same time provides an opportunity to perform various narrowed or refined queries on the corpus. The corpus includes the following metadata:

1) text author (42): 41 known authors, but more than 30 texts without authorship are listed under the heading 'unknown author';
2) genre (3): spiritual texts (128 documents), secular texts (39), dictionaries (5);
3) sub-genre (17): Bible (93 documents), Old Testament (43), New Testament (31), Apocrypha (17), occasional poetry (16), business texts (14), catechisms (10), oaths (9), the Lord's prayer (8), songs (7), etc;
4) a marker indicating whether the document is a printed text (161) or a manuscript (11);
5) document ID, title, year, century, etc.

Three collections have been created, covering all the primary texts of the Bible: the Old Testament, the New Testament and Apocrypha. This is done to allow queries within these document sets (each book of the Bible is a separate collection), as the category 'sub-genre' under these headings includes parts of the Bible from different editions at different times, not just the text of one edition of the Bible (this is important for the needs of the Historical dictionary of the Latvian language where the precise address is added to every single lexeme).

[5] https://korpuss.lv/
[6] https://nosketch.korpuss.lv/#dashboard?corpname=senie_unicode
[7] Data from September 2024.

2.2. Switch to Unicode

In 2002, when the development of the Corpus was started, the Gothic letters relevant to Latvian early texts were replaced by certain ASCII symbols or their combinations, which do not always correspond exactly to the original. For example, the Gothic *ſ* was replaced by § (corresponding to *ſ* in the Unicode encoding), the Gothic by ś (corresponding to in Unicode encoding), the *a* with a circumflex over it was replaced by the symbol combination *a^* (corresponding to *â* in the Unicode encoding), and the *m* with a tilde over it by the symbol combination *m~* (corresponding to *m̃* in the Unicode encoding).

As mentioned before, in order to make the Corpus more user-friendly and the texts more graphically similar to the original, the corpus was converted to the Unicode standard in 2017. Prior to this, the possibilities of Unicode[8] were explored in search of symbols as close as possible to the original. For the most part, ready-made Unicode Latin blocks of letters such as *Latin-1 Supplement*, *Latin Extended-A*, and *Latin Extended-B* were used, the ASCII characters mentioned above were replaced by the corresponding *Unicode* symbols *ſ*, and *â*, but in some cases combinations of symbols such as *m̃* were also used.

For every single corpus source, an individual symbol conversion table was created, according to which conversion rules were then carried out. This process also eliminated some pre-existing inconsistencies in the representation of characters in the corpus, where the same Gothic character could be represented by a different combination of ASCII symbols in two or more different sources, e.g., *a* with a dot above was represented by both *a&* and *a'*. In the Unicode tables, this symbol was represented by *ȧ* in all sources. After the conversion of the texts, post-editing was carried out, and the code-matching table was refined.

In 2017, 73 sources available in the Corpus at that time were converted to Unicode; the new version of the Corpus consists of 172 Unicode sources (see also Andronova et al, 2022a). Thus, we offer reliable, double-checked sources with metadata for the user community.

The Corpus in the Unicode version is also made publicly available in TEI format[9] via CLARIN-LV[10]. Currently we don't have a fixed release schedule, but we aim to provide a new data version on the end of each Corpus-related project.

2.3. Ambiguity of graphemes

Religious texts written in Gothic script in the late 16th and early 17th centuries are characterised by a high level of inconsistency in spelling and the ambiguity of graphemes and grapheme combinations (Andronova et al., 2022a), which sometimes cannot be predicted beforehand even if one is familiar with the old writing system (cf. *dſelige* as *žēlīgs* 'mercifully' in the Catholic catechism from 1585). Most of the phonemes in the first books printed in Latvian do not yet have stable graphic designations (Bergmane, 1986). This is illustrated by the example of the representation of the phoneme /ʒ/ in the first two printed books in Latvian (Tab. 1).

---

[8] https://unicode.org/charts/

[9] The TEI Guidlines: https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

[10] Latest version: http://hdl.handle.net/20.500.12574/90.

Table 1. Representation of the phoneme /ʒ/ in the first two Latvian catechism books (1585, 1586)

CC1585				Ench1586			
	Original spelling	Modern standard spelling	Meaning		Original spelling	Modern standard spelling	Meaning
ſ	ab**ſ**elo	ap**ž**ēlo	‘have mercy’	*ſz*	mu**ſz**ige	mū**ž**īga	‘eternal’
ſc	**ſc**eleſtib	**ž**ēlastība	‘mercy’	*Sz*	**Sz**eelyx	**ž**ēlīgs	‘merciful’
ſſ	da**ſſ**ekaert	da**ž**(e)kārt	‘sometimes’	*ſſ*	mu**ſſ**ige	mū**ž**īgi	‘for ever’
ſſch	da**ſſch**ekaert	da**ž**(e)kārt	‘sometimes’	*ſſch*	Gaſpa**ſſch**e	gaspa**ž**a	‘Mrs’
dſc	**dſc**eleſtibe	**ž**ēlastība	‘mercy’	*ß*	alla**ß**in	alla**ž**iņ	‘always’
dſ	**dſ**elige	**ž**ēlīgi	‘mercifully’	*ſch*	Muy**ſch**e	muiža	‘manor’

Therefore, the source conversion tables for this group contain a relatively large number of rules[11], see table 2.

Table 2. Proportion of rules applied only once in the conversion process for the texts of the very early period

Source ID[12]	Number of tokens	Number of conversion rules	Number of rules applied only once	Rules applied only once (%)
Br1520_PN	58	37	26	70%
Ench1586	8691	635	300	47%
EvEp1587	39833	1262	565	45%
Ps1615	37864	1649	714	43%
Manc1654_LP1	149326	937	309	33%

[11] More on conversion, see Chapter 3.
[12] The source ID contains information about the publishing year and author (if known), and an abbreviation of the source

The number of conversion rules is determined by the time, size and quality of the source, which is directly related to the author's knowledge of Latvian. Thus, in order to carry out the conversion of 58-word Lord's prayer by Bruno (1520), 37 rules have been created, and 26 rules are used only once (70% of the total). This shows that it would be quicker to convert (such) short texts manually. In the sources of the 90s of the 16th century, the number of rules applied just once has decreased (Ench1586 – 47% and EvEp1587 – 45%), showing a common improvement of the quality of the texts. But in general, a problematically uneven quality of the texts is still observed in later sources as well. For example, the psalms of 1615 are similar to the ones already mentioned: although they have the largest amount of conversion rules in total, they still keep the proportion of single-application rules (43%). This in turn confirms the efforts made to improve the written language of that time. From the point of view of conversion, the overall quality of the text remains at the previous level. Positive trends are observed over time, as the proportion of single-application rules decreases, e.g., for Manc1654_LP1 it is 33%, i.e. 309 rules out of more than 149 thousands tokens. For comparison, the Lutheran catechism Ench1586 has a similar amount of rules in the text which is just 5,82% the amount of tokens compared to the Manc1654_LP1 text. This allows us to conclude that the quality of both the early printed texts and the conversion is improving at the same time. A simple manual conversion would be of good use only for small sources and only if the sole purpose of the activity is to prepare a text for the modern reader. In the larger sources, the most commonly used rules are applied in thousands of cases (e.g. in the book of sermons (Manc1654) *Jef* > *Jēz* 1301x, *töw* > *tev* 1022x, *ß* > *s* 9921x, etc.). However, our goal is to carry out a rule-based conversion of all early texts in the Corpus, regardless of their size. Applying correctly selected criteria for rule-based conversion is crucial for text processing, regardless of the quality and quantity of early prints. Each table is a research system that can be combined or compared with other tables created according to the same criteria. The conversion process, both in terms of time and effort, is determined by the quality of the texts, characterized by the number of conversion rules: the lower the number of the rules applied once, the higher the quality of the text.

The works of Georg Mancelius (mid–17th century) have an improved and more systematic orthography compared to the texts of the previous period in Latvian, but the principle of 'one phoneme — one designation' has not yet been implemented; there are several variants for almost all phonemes, both positional and optional (Vanags et al., 2023). The translation of the Bible by Ernst Glück, published at the end of the 17th century, and several other Latvian texts of this period are characterised by greater consistency in orthography; letters and ligatures are no longer as ambiguous as in earlier texts. Consequently, the conversion tables are shorter, unambiguous correspondences are dominating (Andronova et al., 2022b), and thus the conversion rules become more universal.

3. The rule-based conversion

As mentioned before, handling spelling variation is one of the most challenging tasks for NLP. Terminology to address the issue might be slightly different, thus Piotrowski mentions *spelling difference* (which in the long term may be different *historical spellings*), *spelling variance* (different ways in a single text); detecting *diachronic* and *synchronic variation* (Piotrowski, 2022). Different methods are applied to solve problems related to normalization and standardization of historical spelling (see Bollmann, 2019 for an overview). Some corpora offer modernized spelling next to the historical one; e.g., the platform *HistCorp* available from Uppsala University contains texts from the Gender and Work project (GaW)[13] with both historical and normalized forms (see (Pettersson and Megyesi, 2018)).

The normalization of early texts for the corpus search engine has also been developed for the Lithuanian corpus of early texts *Senieji raštai*[14]; see (Šinkūnas, 2018) for a method for 'automatically generating modern writing from historical writing forms based on empirical rules while preserving the characteristics of the original writing'.

With the transition to the new corpus platform *NoSkE* and the inclusion of the Corpus 'SENIE' in the *Latvian National Corpus Collection*[15], it was decided to develop a search engine with the possibility to enter a query in modern script. In order to do this, it was necessary to re-convert all corpus sources that had already been converted to the *Unicode* encoding into a determined modern spelling, preserving the early text and dialect features (Andronova et al., 2022b). Since the 16th–17th century Latvian texts are very heterogeneous, the methodology was as follows:

1) a unique table of conversion rules was developed for every single source,
2) the tables were then used in the programming algorithm,
3) the rules were used to automatically convert all the texts,
4) error analysis and correction of the tables were performed, followed by
5) re-conversion and
6) re-assessing the quality of the conversion (Andronova et al., 2022b).

For now, conversion tables have been prepared and 172 texts have been converted.

During the first project, the research team not only developed the conversion methodology but also carried out theoretical studies, such as proposing new terminology, classifying conversion rules into groups such as *unambiguous graphemic correspondences*; *positional* (graphemic and morphemic) *correspondences*; *individual* (lexical) *correspondences*, and subgroups (more on this in (Andronova et al., 2022a), (Andronova et al., 2022b)).

Initially, when creating conversion tables, researchers found it more convenient to first solve orthographic problem cases with many exceptions, namely individual (lexical) correspondences (e.g., the highly inconsistent use of *ſ*), and include the unambiguous explicit rules at the end of the table, e.g., *ah>ā*, *w>v*. But later it turned out that different approaches can be applied: it is just as successful to start the table with the unambiguous rules, the number of rules does not change significantly. In any case, the relationship

[13] https://www2.lingfil.uu.se/person/pettersson/histcorp/
[14] https://seniejirastai.lki.lt/accept.php
[15] https://korpuss.lv/

Table 3. An extract from ‘Die Sprüche Salomonis’ (1685) (VLH1685_Sal, 14B)

Text in the Unicode	**Rule-based converted text**	**Modern Latvian**
5. Kas Wa□□ara=Laikâ krahj/ tas irr Gudrs/ bet kas Pļaujamâ Laikâ gull/ tas tohp Kaunâ.	5. kas vasara=laikā krāj/tas ir gudrs/ bet kas pļaujamā laikā gul/ tas top kaunā.	5. Kas vasarā iekrāj, tas prāta vīrs, kas ražu noguļ, tam jākaunas.
6. Us ta Tai□na Galwas irr ta □wehtiba bet us to Besdeewigo Mutt wiņņu Wal□chķiba uskrittihs.	6. uz ta taisna galvas ir ta svētiba bet uz to bezdievigo mut viņu valšķiba uzkritīs.	6. Svētība rotā taisnā galvu, ļaundaru mute slēpj pārestību.
7. Ta Peeminne□chana to Tai□no paleek eek□ch □wehtibas/ bet to Besdeewigo Wahrds isnihks.	7. ta pieminešana to taisno paliek iekš svētibas/ bet to bezdievigo vārds iznīks.	7. Taisno piemin ar svētību, bet ļaundaru vārds satrūd.
8. Kas no □irds gudrs irr/ peejemmahs tohs Bau□lus/ bet kam Ģeķķa Mutte irr tohp kults.	8. kas no sirds gudrs ir/ piejemās tos bauslus/ bet kam ģeķa mute ir top kults.	8. Sirdsgudrais klausa, ko viņam liek, bet balamute taps nogāzts.
9. Kas nenoſeedſigi dſihwo/ tas dſihwo droh□chi/ bet kas □amaita □awu Zeļļu/ taps ſinnams.	9. kas nenoziedzigi dzīvo/ tas dzīvo droši/ bet kas samaita savu ceļu/ taps zinams.	9. Kas krietnumā staigā, tas staigā droši, kas iet līkus ceļus, tiks piemeklēts!
10. Kas ar Azzim mirķ□chķina/ Behdas darrihs/ in kam Ģeķķa Mutte irr/ tohp □akults.	10. kas ar acim mirķšķina/ bēdas darīs/ in kam ģeķa mute ir/ top sakults.	10. Kas miedz aci, tas aizvaino, bet balamute taps nogāzts.

between the origin of the text and the number of conversion rules has remained constant: the older the text, the less developed the orthographic system, and the higher the number of conversion rules. While sources from the 16th and the first half of the 17th century

may need almost a thousand rules, sources from the 18th century, if the text is small, may need only 30 rules. See Tab.3 with an extract from 'The Proverbs of Salomon' (1685, 10:5–10) with the converted text, and in Modern Latvian spelling next to it.

At the moment, the development of the pilot converter for the 18th c. texts is in progress. Afterwards, the experiments with the pilot converter will be carried out with different sample text fragments to evaluate the quality.

4. Switch to *NoSketchEngine*

When the work on the corpus started in the nineties, it was made searchable via a custom in-house indexing solution. However, nowadays better solutions are available for this task — tools like *SketchEngine/NoSketchEngine* or *Korp* (spraakbanken.gu.se) are made specifically for work with corpora, and they provide a wide variety of functionality for querying and overviewing data. Furthermore, using a ready-made corpus querying tool instead of making one's own reduces the human effort needed for maintaining the corpus infrastructure. The decision to use *NoSkE* was made mostly thanks to the fact that it is already used for hosting modern Latvian corpora. Thus, it is more familiar to Latvian users and the IMCS already had a *NoSkE* server running.

To load the Corpus in *NoSkE*, two most important tasks are to calculate a precise address for each token and to tokenize correctly, including connecting together hyphenated words.

Addresses are obtained by slightly different means depending on the type of source. All addresses contain source ID as the first part. If the source is a collection of several books (this is true for the Old Testament, the New Testament and Apocrypha), the next address part is the book ID. For laws and rules, the address then contains the verse number. For Bible verses, the address then contains chapter number and verse number. For other sources, the address then contains page number and line number. Finally, all token addresses contain the index of the token in the line of verse. In this way, both tokens, lines, and verses can be addressed uniquely. These addresses are included in the TEI export published in CLARIN-LV as well — to facilitate interoperability between various tools, and projects and researchers using the Corpus.

Tokenization is generally done on spaces and punctuation, however, the equals sign`=` is an exception — this mark is used in compounds between compound parts and is not considered to be a place where tokens must be separated. Another thing that complicates the tokenization process is that source data files contain text separated in the original lines, and occasionally a word is split between two lines, adding the hyphen `-` in the first line. Automatic de-hyphenation itself is simple enough; however, if combined with corrections marked in text and places where before page break a fragment of the next line is added in the previous page, occasional errors can happen if the corrections are marked inconsistently. When we identify such errors, we strive to enhance the source data; however, we have not yet identified all such mistakes. Currently, the TEI export published in CLARIN-LV features both the original and de-hyphenated version.

5. Search

Typing *vīrs* 'husband' into the Basic search field results in a concordance page with several forms of 'husband': <*wirs, wyrs, Wyrs, wiers, Wiers, wihrs, Wihrs*>. This answer allows us to further investigate in which sources the forms are used. By choosing to type one of the original forms in the search field, it is possible to further explore the usage of each group, for example, the use of capital and lowercase for nouns. In addition, it is also possible to analyse the use of forms in a single source or author's text, which will allow us to understand the distribution of variants and trends in spelling.

In order to find out how to write the pronoun form in *Dat.Sg. fem. tai* 'for her' and *Loc. Sg. masc., fem. tai* 'in that', you can type the word in the simple search field. The result will be several forms: <*tai, taei, thaei, tay, thay, taî*>. In addition to the differences in the different sources, a functional distribution can also be seen. The spelling <*taî*> was used specifically for the *Loc.Sg.* form in sources since the end of the 17th century. This distribution and its regularity in the relevant sources can be analyzed separately. Thus, by searching for forms in modern writing, it is possible to find the maximum number and variety of spellings, while further work on the original spellings allows a more in-depth study of the distribution and development of the spelling techniques concerned.

Problems will arise if a phonemic feature, such as long vowels or palatal consonants, is not marked in the original text. The search engine will only find such forms if they are systematic rules or sporadically manually corrected during text conversion.

Advanced search in *NoSkE* allows us to combine different metadata (year, author, genre, sub-genre, print or manuscript, language).

6. Applications

The main user community of the Corpus is humanitarians: first of all, linguists, but it serves as a good source for studies in literature, stylometry, the development and influence of the Reformation in Livonia, history and ethnography, and some other fields of social sciences and arts. At the moment, the main beneficiaries are the compilers of the Historical dictionary of Latvian[16] (Andronova et al., 2016). First of all, *NoSkE* concordancer is used. With the conversion to modern spelling, a faster search for lexemes and forms, as well as better results, including earlier unnoticed exceptions, are received.

With the inclusion of the Corpus 'SENIE' in the LNCC on *korpuss.lv*, researchers have a great opportunity to examine the usage of lexemes of early texts diachronically as well. E.g., it is possible to find out that the lexeme *āļot, āļoties* 'get confused' is not uniquely a phenomenon of 17th c. lexicography, one can trace its usage both in the 19th c. Latvian newspapers, as well as in the press of the Soviet era and in modern short prose as well. So, if you type *veselīb** or *veselib** into a search engine, you will find that the word *veselība* 'health' is already used in texts from the late 16th century.

[16] https://lvvv.tezaurs.lv/

7. Conclusion

The *NoSketchEngine* platform allows users to operate with elaborated queries, combing and refining metadata. The concordancer assists in the writing of new entries in the Historical Dictionary of Latvian (16th–17th c.). Results give a deeper insight into the history of Latvian orthography, providing more precise data. A pilot converter for the 18th c. texts is in progress. It is intended to be an open-source tool and will be publicly available to deal with historical texts in the Unicode format.

Finally, this will open a new opportunity in Early Latvian data processing using NLP tools for Modern Latvian.

Acknowledgements

Extensive work on the conversion of corpus texts has been carried out within the framework of National Research Programme 'Digital Resources of the Humanities (DH VPP)' (No. VPP-IZM-DH-2020/1-0001) (2020–2022). This work has been conducted within the DHELI (Towards Development of Open and FAIR Digital Humanities Ecosystem in Latvia; No.VPP-IZM-DH-2022/1-0002) (2022–2025) framework in 2023–2025 by the Latvian Language Institute, Faculty of Humanities of the University of Latvia and the Institute of Mathematics and Computer Science, University of Latvia.

References

Andronova E. (2007). The Corpus of Early Written Latvian: current state and future tasks, *in* Davies M., Rayson P., Hunston S., Daniellson P. (eds.), *Proceedings of Corpus Linguistics, 2007 (University of Birmingham, UK, 27-30 July, 2007)*, available at http://ucrel.lancs.ac.uk/publications/CL2007/paper/245_Paper.pdf.

Andronova E. (2020). Short Texts in the Corpus of Early Written Latvian, *in* Reinsone S., Skadiņa I., Baklāne A., Daugavietis J. (eds.), *Proceedings of the 5th Conference on Digital Humanities in the Nordic Countries* (*DHN* 2020) (Riga, Latvia, October 21–23, 2020) (CEUR Workshop Proceedings; Vol. **2612**), pp. 173–183, available at http://ceur-ws.org/Vol-2612/short1.pdf.

Andronova E., Frīdenberga A., Pretkalniņa L., Siliņa-Piņķe R., Skrūzmane E., Trumpa A., Vanags P. (2022a). User-friendly Search Possibilities for Early Latvian Texts: Challenges Posed by Automatic Conversion, *in* Berglund K., La Mela M., Zwart I. (eds.), *Proceedings of Conference 'Digital Humanities in the Nordic and Baltic Countries': 6th Conference* (Uppsala, Sweden, 15–18 March, 2022) (CEUR Workshop Proceedings; Vol. **3158**), pp. 168–176, available at http://ceur-ws.org/Vol-3232/paper13.pdf.

Andronova E., Frīdenberga A., Pretkalniņa L., Siliņa-Piņķe R., Skrūzmane E., Trumpa A., Vanags P. (2022b). Latviešu valodas senāko rakstu pieminekļu konvertācija mūsdienu rakstībā: iepriekšējā pieredze un automatizācijas mēģinājumi = The Conversion of Early Written Latvian Texts into Modern Spelling: Previous Experience and Automatization Attempts, *in* Helviga A. (ed.), *Aktuālas problēmas literatūras un kultūras pētniecībā*, **27,** LiePA, Liepāja, pp. 346–358, available at https://dom.lndb.lv/data/obj/1035006.html.

Andronova E., Frīdenberga A., Siliņa-Piņķe R., Skrūzmane E., Trumpa A., Vanags P. (2022). Variantums kā konvertācijas izaicinājums: Georga Manceļa tekstu atveide mūsdienu rakstībā

= Variation as a Challenge for Conversion: Presenting Texts by Georg Mancelius in Modern Spelling, *Letonica* **47**, pp. 188–207, available at http://lulfmi.lv/files/letonica/Letonica_47.pdf.

Andronova E., Siliņa-Piņķe R., Trumpa A., Vanags P. (2016). The Electronic Historical Latvian Dictionary Based on the Corpus of Early Written Latvian Texts, *Acta-Baltico Slavica* **40**, pp. 1–37, available at https://ispan.waw.pl/journals/index.php/abs/article/view/abs.2016.018/2481

Bergmane A. (1986). Latviešu grafētikas izveide = The development of Latvian graphetics, *in* Bergmane A., Blinkena A., *Latviešu rakstības attīstība. Latviešu literārās valodas vēstures pētījumi*, Zinātne, Rīga, pp. 18–79.

Bollmann M. A. (2019). Large-Scale Comparison of Historical Text Normalization Systems, *in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume **1** (Long and Short Papers), Association for Computational Linguistics, pp. 3885–3898, available at: https://liu.diva-portal.org/smash/get/diva2:1798251/FULLTEXT01.pdf.

Laime, S., Reinsone, S. (2024) HUMMA.LV: Towards a Collaborative Digital Platform for Humanities and Arts in Latvia, *Baltic Journal of Modern Computing*, **12**(4), 487–492.

Pettersson E., Megyesi B. (2018). The HistCorp Collection of Historical Corpora and Resources, *in* Mäkelä E., Tolonen M., Tuominen J. (eds.), *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference* (Helsinki, University of Helsinki), pp. 306–320, available at https://www.diva-portal.org/smash/get/diva2:1205788/FULLTEXT01.pdf.

Piotrowski M. (2022). *Natural Language Processing for Historical Texts. Springer Nature*, Switzerland.

Šinkūnas M. (2018). Senųjų raštų rašybos keitimas paieškos sistemai = The Normalization of Old Lithuanian Orthography for Usage in a Search Engine, *in* Judžentytė-Šinkūnienė G., Zubaitienė V. (eds.) *Baltų kalbų tekstų ir žodžių reikšmės*, Vilniaus universiteto leidykla, Vilnius, pp. 389–407, available at https://www.zurnalai.vu.lt/open-series/article/view/12999/11825.

Vanags P., Frīdenberga A., Trumpa A. (2023). Georga Manceļa rakstības principi: darbības pirmais posms (līdz 1631. gadam) = Georg Mancelius' Principles of Orthography: the First Period (Until 1631), *Baltistica* ***LVIII(2)***, pp. 329–353, available at https://www.baltistica.lt/index.php/baltistica/article/view/2529/2428.

Zogla A, Skilters J. (2010). Digitization of Historical Texts at the National Library of Latvia, *in* Skadiņa, I., Vasiļjevs, A. (eds.), *Human Language Technologies – the Baltic Perspective. Proceedings of the Fourth International Conference Baltic HLT*, pp. 177–184.

Received November 29, 2024, accepted December 5, 2024