Short Texts in the Corpus of Early Written Latvian (www.korpuss.lv/senie)

Everita Andronova

Institute of Mathematics and Computer Science, University of Latvia
everita.andronova@lumii.lv

Abstract. The paper introduces the Corpus of early written Latvian 'SENIE' (16th-18th cc.). This corpus provides access to texts which are scattered across different libraries and archives and which therefore have not been much investigated. The corpus is an important resource for linguists (whom is the main user community at the moment), but it can also provide some knowledge in culture and social studies.
The corpus-building methodology is presented, giving insight into the enrichment of the variety of text registers by adding shorter texts: individual texts such as occasional poetry, Latvian texts found in sources written in other foreign languages, and shorter texts in Latvian appended to some individual Latvian sources.

Keywords: Corpus of Early Written Latvian, Historical Dictionary of Latvian.

1   Introduction

Early written Latvian texts are important sources not only for linguistics, but also for culture and social studies. Latvian texts (and indeed Latvian culture as a whole) show consistent Germen influence (from clergy, educated men, artists, craftsmen, etc.). These texts were produced in a cross-cultural context of Catholicism and Protestantism and also display elements from local folklore. Especially the earliest of the texts can be considered a mirror both of ideas originating in Europe and of elements of the Baltic religion.

2   Early Written Latvian Texts

The very early texts of the 16th c. are mostly religious (hymnals, Lord's prayer, sermons, translations of parts of the Holy Bible), since the 17th c. the scope of genres of written sources widens: lexicographical[1], legal[2], and other secular[3] texts are pub-

---

[1] There are four dictionaries included in the Corpus at the moment of writing this paper: the first German-Latvian dictionary 'Lettus' (1638) by G. Mancelius, its supplement 'Phraseologia Lettica'(1638) and two transcripts of an unpublished Latvian-German dictionary by C.

lished, and the number of sources reaches 100[4]. During the 18th c. the spectrum of genres widens considerably: among the texts published at that time we find localized and original fiction, calendars, primers, cooking books and popular science texts[5]. The rise of new genres is closely connected with the growth of the literacy level of Latvians (it is estimated that around 1800 2/3 of peasants in Vidzeme and 1/3 of peasants in Courland could read[6]). The texts cover information on historical events, habits, values and virtues of the local inhabitants and provide important ethnographical data. Thus, in the Latvian sermon book by G. Mancelius (1654[1]) we can find information on plague in Livonia (Manc1654_LP1[7]$_{184}$). Both the above-mentioned Latvian sermon book and a wedding song by an anonymous author (GD_1698) present descriptions of the dress and shoes worn in the 17th c.

From Latvian sermons we learn the notion *zeltnesis* (literary 'one carrying the gold') in the meaning 'nobleman, landlord'. Nowadays the word has the meaning 'ring finger'; the earlier meaning is fixed in the dictionary of Mühlenbach-Endzelin (1923–1932), the widest dictionary providing etymology of Latvian lexemes supplied with samples from early texts, folklore, and fiction of the 19th–20th c. The dictionary entry is based on evidence from this exact sermon book (Manc1654_LP2$_{19}$). More examples of the earlier meanings and their change over time have been discovered during the ongoing project of the Corpus-based electronic historical Latvian dictionary [1] and have been discussed in the recent publications of the dictionary compilers [see 8, 9, 10].

In spite of at least century long tradition of research, early monuments of the Latvian language still require attention of linguists and certain international team of scholars is working with them. Some new sources have been discovered recently. One should mention a manuscript of a Latvian-Polish word-list found in the National Library of Russia in 2017. Before this it had been unknown in Latvia (at least the most detailed description of Latvian lexicography till 1900 published in 1961 [12] does not mention this word-list). According to preliminary observations, at least part of this list is excerpted from G. Elger's published Polish-Latin-Latvian dictionary (1683). The

---

Fürecker (appr. 1650–1670). G. Elger's 'Polish-Latin-Latvian' dictionary (1683) is in process to be included in the corpus.

[2] The war law text 'Savādi kara tiesas likumi' (1696) for soldiers in Swedish Livonia (http://senie.korpuss.lv/source.jsp?codificator=SKL1696_KB) and 'Mūsu augsta un varena Ķēniņa sodu likums pret bērnu mušināšanu" (1684) – a text on punishment against the killing of new-born babies (http://senie.korpuss.lv/source.jsp?codificator=SL1684).

[3] Cf. "10 Gespräche" by G. Mancelius (1638) is a good example of spoken language, as there are 10 dialogues published.

[4] Catalogue 'Die älteren Drucke in lettischer Sprache 1525–1855' [7] lists 5 printed sources in the 16th c., 100 sources printed till 1700, and 572 sources published from 1701–1800. In addition to printings, numerous manuscripts and their transcripts have been in use.

[5] The first Latvian encyclopaedia 'Augstas gudrības grāmata no pasaules un dabas' ('The book of high wisdom of the world and nature') (1774) by the Baltic German Lutheran pastor Gotthard Friedrich Stender is in a process of inclusion in the corpus now.

[6] https://enciklopedija.lv/skirklis/22533, last accessed 2020/01/26.

[7] Source abbreviations are used the same as in the corpus (http://senie.korpuss.lv/abbrevs.jsp).

discovered source is not processed for the corpus at the moment, but it might be copied and presented for a larger audience in close future. Early written Latvian sources are very much treated isolated and in many cases are used for quite narrow purposes (mostly by linguists and historians).

There was a serious lack of general overviews introducing the sources and studies on them, and more important, even now there are still a few interdisciplinary studies carried out. Fortunately, the last two decades have seen a growth in popularization and dissemination of the early written sources, e.g. a number of 17th–18th c. manuscripts have been prepared for printing, supplemented with commentary of their historical and cultural context, as well as linguistic analysis (cf. T.G. Fennell's input in decrypting the lexicography sources of L. Depkin; stories of Livonia written by F. B. Blaufuss in 1753 (2015) and some others). In 2018 the edited facsimile of the first Catholic song book by G. Elger (1621[1]) was published [4]. This publication contains a detailed description of the beginnings of religious poetry in Latvia, some characteristics of the song's Latvian orthography and a description of six Polish songs included in the 1st ed. of the song book. The authors conclude that '[i]t offers us a glimpse of the earliest Latvian poetry, and constitutes a peculiar reflection of the late Renaissance – early Baroque world, with its predilection for playing with form, means of expression, and topics situated at the crossroads of the European Christian tradition and the social culture of Latvian peasants' [4: CXII]. The Library of Vilnius University keeps the only known copy of the first edition. The corpus compilers benefit from the good co-operation with the Library, the defective facsimile (the pages after p. 206 are missing) is available online[8] from the corpus.

The 21st c. brought new chances for lesser-used and lesser-studied languages, namely, the era of digitalization has resulted in the development of different general and special corpora. In 2007 the National Library of Latvia initiated and in 2009–2012 continued a large-scale project of digitalizing newspapers, now more than 2.5 million pages from ca 3000 periodicals have been processed and are available for searches at www.periodika.lv [13]. Apart from this, the valuable project 'The Latvian National Digital library' has been launched, offering not only books, but also maps and audio recordings and different special collections[9].

3   The Corpus of Early Written Latvian

The diachronic corpus of early written Latvian was created in 2003 as a result of initiatives started already in the beginning of 1990s, when the Institute of Mathematics and Computer Science at the University of Latvia started to deal with language technologies. Then the most significant Latvian printed sources of the 17th century were typed in (eg. The New Testament (1685), the first Lutheran song book printed in Riga (1615), as well as the Latvian Sermon book of G. Mancelius (1654)). In the mid-1990s these texts were included in a database and word indices were created, but this project soon stopped due to the lack of funding, and only individual research has been

---

[8] http://senie.korpuss.lv/faksimili/17gs/Elg1621_GCG/, last accessed 2020/01/26.
[9] https://www.lnb.lv/en/resources/digital-library, last accessed 2020/01/26.

carried out using the data. In 2002 a joint project of the IMCS and the Department of Baltic languages, Faculty of Philology, UL was started with the goal of developing an electronic dictionary of the 17th c. One of the project applications was the Corpus of early written Latvian texts [6, 2]. The most important aim was to make early Latvian texts open for the public in one place in order to stimulate more research on them. Looking at the experience of others today, there are languages with more than 4 diachronic corpora available (e.g. English, Russian, Polish[10]), but a more common practice is to have one diachronic corpus. The diachronic corpus might be a sub-corpus of the National corpus (like Czech Diakorp[11], Russian[12]). However, in Latvia there is as yet still no Latvian National Corpus.

From the very beginning this project has been more of a in-house activity, using a corpus platform developed at the IMCS in 2002. As its size grows, the user community widens, and technologies develop, it should obviously be moved to another, more elaborated corpus management system.

The corpus is now intended to cover the history of written Latvian of the 16th–18th cc., and its aim is to facilitate studies of early Latvian in general and to serve as the basis for the Historical dictionary of the Latvian language. This is a good example of successful co-operation between linguists and software engineers in creating a new kind of dictionary in Latvian lexicography. 1200 pilot entries covering all types of part-of-speech were published on the web in 2016: www.korpuss.lv/lvvv [1], and the dictionary project dealing with lemmas beginning with the letter A continues in 2018–2020. The development of the corpus and the work to extend its coverage is an ongoing collaborative research task. There has never been a project specifically for corpus development, usually this is just one task among many in some other project dealing with lexicography or the study of Latvian in early texts. Corpus compilation has gone through several phases.

3.1 Acquisition of Sources

Corpus sources are examined based on the information in the united catalogue of Latvian prints [7]. Early written Latvian texts have been acquired thanks to close co-operation with Latvian and Lithuanian libraries, as well as with researchers across Europe interested in the history of early Latvian texts. This is a good example of networking. Undergraduate students at the University of Latvia and St. Petersburg State University (Russia) have also been involved in the process of transliterating some shorter texts during the compilation of the corpus. This has served to raise the interest of the history of the Latvian language, and subsequently some bachelors and master level, as well as PhD theses have been defended on the basis of these texts.

The first digitized text copies were handed over to the National Library of Latvia in 2002. It is worth mentioning that some new sources have been discovered since the

---

[10] E.g. http://martinweisser.org/corpora_site/historical_corpora.html lists English diachronic corpora. See an overview of other diachronic corpora on https://www.clarin.eu/resource-families/historical-corpora, last accessed 2020/01/27.

[11] See Diakorp at https://wiki.korpus.cz/doku.php/en:cnk:diakorp, last accessed 2020/01/27.

[12] http://ruscorpora.ru/new/en/search-mid_rus.html, last accessed 2020/01/27.

catalogue was published: thus e.g. a unique copy of *Agenda Parva* (1622), earlier reported unknown, has recently been published on the website of the Warmia-Mazury Digital Library[13]. We are presently processing Latvian fragments in this Agenda that will be added to the corpus. Manuscript of the Latgalian grammar by J. Rymkiewicz (1810) which had been considered lost during the WWII was found by S. Kalvāne in 2013 in the Archive of the Latvian Roman Catholic Curia church.

3.2   The Compilation of the Corpus

This project took place in 2002. As it was a one-year project, the emphasis was placed on the 17th c. data. Texts that had already been typed in were checked twice and structural markup was introduced; new sources acquired from the National Library of Latvia were scanned and Gothic script was OCR-read and transliterated into Latin script. As it turned out, the Gothic script used in Latvian texts slightly differs from German Gothic script, as in Latvian there might be additional diacritical marks used above the letter (see the characters used in the transliterated corpus at http://senie.korpuss.lv/notations.htm). The ASCII code set was used in the corpus. The bulk of diachronic corpora comprises transliterated texts, although some corpora are diplomatic transcribed (like the reference corpus of Middle Low German and Low Rhenish texts – ReN[14]), partly normalised to modern spelling (Old Hungarian[15]) or transcribed (like Diakorp).

Due to the short terms only the first editions of the printed sources were chosen. At that time ecclesiastic texts made up the dominant part of the corpus (17 different sources – The New Testament, Sermon book, collections of hymnals, parts of the Bible, catechisms), and only 3 small law texts represented other genres. The size of the corpus in 2003 was 800,828 running words, and the influence of one particular author was clearly observed (texts by G. Mancelius were dominating). Keeping in mind the idea of availability for public, it was decided that the corpus will not use sample selections (like e.g. historical sermon corpus on the Nachfeld HIPKON[16]), but whole texts exclusively.

In 2004 two manuscript dictionary transcripts were added, thus the former concept of the corpus of early printed texts turned into the concept of early written Latvian texts. The new data (30,597 running words) had little impact on the contents of the corpus: ecclesiastic texts constitued 96% of the total, lexicographical sources – 3.6 % and secular texts 0.4%. In order to widen the time span, new sources were processed and added to the corpus: three 16th c. sources were added in 2004 and 15 sources from the 18th c. added in 2007 [2], see Table 1:

---

[13] http://wmbc.olsztyn.pl/dlibra/doccontent?id=926, last accessed 2020/01/28.

[14] https://www.slm.uni-hamburg.de/en/ren/ueber-ren.html, last accessed 2020/01/28.

[15] http://omagyarkorpusz.nytud.hu/en-descr.html, last accessed 2020/01/28.

[16] Historisches Predigtenkorpus zum Nachfeld (HIPKON) available at https://corpora.uni-hamburg.de/sfb632/b4hipkon/, last accessed 2020/01/28.

Table 1. The size of the corpus in 2007.

Period	Sources	Size (running words)
16th c.	3	52,642
17th c.	22	829,876
18th c.	15	75,559
Total:	40	958,077

In 2007 we set ourselves two tasks: (1) keeping in mind the idea of the historical dictionary of Latvian we should pay more attention to all text types present at that time in the Latvian written language; (2) the corpus management system should be improved.

Due to different objective circumstances, the corpus retained the same size and design up until 2013. In 2013 the project 'A Corpus-based Electronic Historical dictionary of Latvian' (covering only the 16th and 17th c.) received a 4-year grant from the Latvian Council of Science (No. 212/2012), and the Institute of Latvian Language, UL joined the project to start the dictionary writing in earnest and to facilitate the development of the corpus. As a result, in 2016 the size of the corpus was extended not so much in size as in registers (among the texts processed were dedication poems, oath texts and a transcript of handwritten Linen statutes (for more see [1] and [11]), see Table 2.

Table 2. The size of the corpus in 2016.

Period	Sources	Size (running words)
16th c.	10	57,484
17th c.	46	974,203
18th c.	17	79,855
Total:	73	1,111,542

One of the challenges in this work is the crucial need of comparison between different editions of the same source and an analysis showing the tradition of circulation of different parts of religious texts from one source to another.

After a year's break the dictionary project received its next funding for 2018–2020. The addition of (mainly) 17th c. sources continues with the goal of sometime in the future covering all the 100 printed sources listed in the Union catalogue (if they are available in any library). At the moment we cannot predict when this will happen.

4 Inclusion of Short Texts in the Corpus

Quite a wide range of short texts have either been added to the corpus recently or are in the process of being included now; these texts can be divided into 3 groups:

1. individual short texts, e.g., occasional poetry, oath texts;

2. Latvian texts found in sources written in other foreign languages, e.g., the prayer *Pater Noster* published in the 16th c.; sentences in Latvian in several editions of 'Stratagema oeconomicum oder Akker-Student' written in German by S. Gubert in the 17th c. or Latvian text in Agenda Parva (1622 and later editions);
3. shorter texts in Latvian appended to some individual Latvian sources.

4.1 Individual Short Texts

These include both poetry and certain legal texts (different oaths, laws of war court). The bulk of the sources in this group is occasional poetry, written in the 17th and 18th cc.

The beginnings of Latvian occasional poetry have recently been the object of in-depth studies. A broad inspection of the 16th and 17th c. poetry in the cultural context has been carried out by Māra Grudule [5]. The book shows the long way of evolution of this type of texts: they were profoundly influenced by German culture but later little by little turned into Latvian poetry. Three early dedication poems were already added to the corpus in 2016. In 2019 almost 70 poems from 10 sources (6143 running words) were collected in different libraries and are now in the process of being included in the corpus. One of these new poems is a unicum kept at the Russian National Library – 'Mūsu visu upurs tai priecas dienā' (1791). These new poems are of wide thematic range, covering different occasions – birthday congratulations, wedding songs, popular New Year's wishes, which can be printed on cards or written in letters, funeral songs and others (a bachelor's thesis has already been defended at the St. Petersburg State university on this topic [3]).

These songs may be interesting not only for literature and linguistic studies, but also in order to examine the culture, history and ethnography in Livonia at that time. Thus, we can find reference, e.g. to the clothes of wedding participants and description of food served at wedding (GD_1698); commonly used music instruments (GD_1698, ZP1685, Baum1699_LVV, Bruehn1756_DLWS); typical birds and animals killed at hunting (StendAJ (1793) etc. Thus, we would like to encourage not only linguistic, but all other kinds of studies by means of the corpus. These texts will be included in the corpus as individual sources.

4.2 Latvian Inscriptions in Texts Written in Other Languages

This group covers single words, phrases, sentences and longer passages in Latvian in books printed in other languages. Latvian proper names – personal names and places names – have been found in several sources dated to the 15th century (e.g. chronicles). The lists of craftsmen guilds from the 16th c. should be examined and excerpted for the purposes of the corpus. The history of written Latvian rises with the period of Reformation and the claim of Martin Luther to use native language. There are already a number of prayers *Pater Noster* from the 16th c. in the corpus, before including them a linguistic analysis was performed in order to define which prayer to include [11].

At the moment 2 new sources are being processed for inclusion in the corpus:

1. Agenda Parva (1622) with its texts written in Polish, German, Estonian and Latvian. For the needs of the corpus only the Latvian sentences are excerpted and processed, and a Latvian word-list will be created on the basis of this material.

2. The popular 17th c. book by S. Gubert, 'Stratagema oeconomicum oder Akker-Student' (1st ed. 1645 and later editions in the 17th c.), is a good example of so-called Hausväterliteratur and is a valuable source for ethnographical studies among others (e.g. the description of instruments and agriculture cultures known in Livonia at that time; 'Bauer=Prognosticon' for weather forecast is often mentioned, later included in the volumes of Latvian beliefs compiled by P. Šmits (1940–1941). In this book we can find Latvian phrases and hymnals at the end (the last edition printed in 1757 excludes hymnals). Single words and phrases are encountered within the German sentences, commonly introduced by the phrase 'die Bauern nennen'. In this case the whole sentence will be copied and marked as German, but only the Latvian phrase will be included in the word list. There are some hymnals added at the end of the book both in German and Latvian. All the songs will be included in the corpus in order to facilitate the analysis of the source text in German and its translation into Latvian.

4.3 Texts in Latvian Added (later) to Some Individual Latvian Sources

At the moment we have only one such source – a letter written by the peasant Anšs to the priest Loder dated June 1771 and added to the transcript of the 'Lettisches und Teutsches Wörterbuch' by Ch. Fürecker. This letter has already been included in the Corpus (http://senie.korpuss.lv/static/V1771_SZA.html) as a separate item. At the moment it is the only example of the correspondence in the corpus.

5 Corpus annotation

Apart from (rich) structural markup many diachronic corpora are POS tagged, some of them are annotated using modern language processing tools (Diakorp, Corpus of old written Estonian[17]). In our corpus only structural markup is applied to the texts to keep the layout of the text and to exclude words in foreign languages (e.g. German, Latin, Polish, Greek and some others) from the Latvian word-list and frequency list. If an obvious spelling mistake or similar error is observed, it is also marked, but the compilers have tried not to interfere with the original text. All the source abbreviations used in the corpus contain the name of authors, date of publishing and name of source. One of the advantages of this corpus is that it provides the exact location of a word-form (usually the abbreviation of the source, page and line number of the text or the Bible Book, chapter and verse, see Fig.1.). This makes it easy to accurately cite

---

[17] Prillop, K.: Vana kirjakeele korpus. Center of Estonian Language Resources. (2013) https://doi.org/10.15155/1-00-0000-0000-0000-00075L. http://vakk.ut.ee/, last accessed 2020/01/28.

the corpus data. There is a possibility to look at facsimiles of the sources as well, which gives an extra added-value to this resource.

Fig. 1. Screenshot with the word index and the full text of the Gospel of Mathew from the corpus (http://senie.korpuss.lv/static/Mt.html).

6   Corpus management system

As mentioned earlier, the management system was initially developed in 2002 by N. Grūzītis and since then only minor changes have been done. The system offers word lists, frequency lists, inverse vocabularies and a concordancer. It supports navigation within the corpus (search for author or text genre – only 3 categories are offered: 1) religious, 2) ecclesiatic texts and 3) dictionaries). The online search in word-lists and in the concordancer offers search for exact match or with _ and *.

In 2017 seventy-three corpus sources were converted to the Unicode system (tables of convertion available at http://senie.korpuss.lv/unicode/tabula.pdf, see Fig.2.).

Fig. 2. Screenshot showing letter "e" with different combined simbols encountered in the corpus and their Unicode counterparts.

All corpus sources have been checked, a list of 60 different symbols created and converting rules into Unicode system made.

This is a step towards modernization of the corpus platform. Our future plans are to apply TEI guidelines and to consider switching to the NoSketch Engine platform to improve performance and allow for more sophisticated corpus queries statistical metrics, and to make it more user-friendly. We would like to make our corpus machine-readable in the standard TEI/XML format, instead of the legacy in-house format, in order to offer this data to other partners dealing with the same topic (there are some projects initiated dealing with Latvian early texts in Germany). But the original site of the Corpus could be left as a sort of electronic library – just for reading texts and obtaining facsimiles.

Acknowledgments

References

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

2. Andronova, E.: The Corpus of Early Written Latvian: current state and future tasks. In: Proceedings of Corpus Linguistics, 2007, Birmingham, UK. http://ucrel.lancs.ac.uk/publications/CL2007/paper/245_Paper.pdf

3. Bal'chunas, A.: Latyshskije hvalebnyje pesni konca XVII–XVIII vv.: lingvisticheskij i literuvodecheskij analiz [Russian: Latvian dedicative poems of the end of the 17th–18th cc.: a linguistic and literary analysis]. BA thesis. St. Petersburg State University (2019).

4. Elger, G.: Geistliche Catholische Gesänge... (Braunsberg 1621). Ed. facs. Grudule, M., Prusinowska, J., Solarz, M. (2018).

5. Grudule, M. Latviešu dzejas sākotne 16. un 17. gadsimtā kultūrvēsturiskos kontekstos. Rīga, 2017.

6. Milčonoka, E.: Latviešu valodas 17. gadsimta teksti internetā. In: Baltu filoloģija XII (1), pp. 139–150 (2003).

7. Seniespiedumi latviešu valodā, 1525–1855: kopkatalogs. Latvijas Nacionālā bibliotēka. Rīga (1999).

8. Siliņa-Piņķe, R. Trumpa, A.: Das Historische Wörterbuch der Lettischen Sprache (16.–18. Jahrhundert): Vorstellung und eine Fallstudie. In: Bock, B.; Kozianka, M. (Hrsg./eds.). Weiland Wörter-Welten – Akten der 6. Internationalen Konferenz zur Historischen Lexikographie und Lexikologie (Jena, 25.–27. Juli 2012.), pp. 261–272 (2014).

9. Trumpa, A.: Ar XVII a. latvių kalbos žodynai –patikimi senųjų reikšmių aiškinimo kelrodžiai? In: Baltų kalbų tekstų ir žodžių reikšmės 2018, pp. 44–60 (2018).

10. Trumpa, A.: Senās nozīmes 16. un 17. gadsimta latviešu tekstos. In: Linguistica Lettica 24, pp. 202–216 (2016).

11. Vanags, P.: Latviešu valodas vēsturiskās vārdnīcas projekts. In: Valodas prakse: Vērojumi un ieteikumi, pp. 97–109 (2014).
12. Zemzare, D.: Latviešu vārdnīcas (līdz 1900. gadam). Rīga 1961.
13. Zogla, A., Skilters, J.: 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 (2010).