The Corpus of Early Written Latvian: Current State and Future Tasks

Everita Andronova[1]

1. Introduction

The history of written Latvian dates back to the late 16th century, when both Protestantism and Catholicism reigned. Although the first physically available book in Latvian – Catechismvs Catholicorum – was published not in Latvia, but in Vilnius, the capital of neighbouring Catholic Lithuanian, in 1585, texts in Latvian and copies thereof were distributed in Riga much earlier. Martin Luther's ideas on preaching in the native language became very popular here and there is written evidence of the first book in Latvian published in 1525, but it has not survived. Research on the history of written Latvian has been carried out rather fragmentary. The delayed development of this branch of the Baltic philology might be explained by the view expressed by Jānis Endzelīns, one of the most influential and well-known Latvian linguists, that the earliest texts were "written incorrectly (by Germans!)" and that the language in the texts is "full of mistakes" (Endzelīns, 1951: 22, 20). Another prominent linguist, Artūrs Ozols, stated that early written Latvian "is a distortion of the people's language, a grouping of the words of this language according to the model of the German language" (Ozols, 1965: 8). These statements influenced the study of the Latvian language in the first written texts until the early 1980s. This also resulted in a situation where the research on the Early Latvian texts for a long time focused on describing mistakes in separate sources, and in only a few cases some attempts were made to see the reflection of the language system of the time through the mistakes and erroneous and sometimes obscure spelling.

The Corpus of Early Written Latvian named *SENIE* (www.ailab.lv/SENIE) is an effort to change the existing statements and to support a completely new view on the language as a system in these texts. The Corpus was first launched in January 2003, but its development is still in progress (the approximate size of the corpus is now about one million running words). The aim of the Corpus is to facilitate diachronic studies of Latvian, to support variant and language standardization studies, to serve a basis for a historical dictionary of Latvian, as well as to popularise early written sources and to support their re-evaluation.

2. A historical background of the idea of collection of data from early Latvian

Several times in the past the necessity to collect as much data as possible about the Latvian language together in one repository has been emphasised. In the 1930s attention was paid to the *Thesaurus linguae letticae*, where all Latvian words both from spoken language and written texts could be collected (Endzelīns, 1933: 818). Up until now the only Latvian dictionary giving a deeper insight into the lexis of early written Latvian is Mühlenbachs' *Lettisch–deutsches Wörterbuch* (Mühlenbachs,

[1] Institute of Mathematics and Computer Science, University of Latvia
e-mail: everita@ailab.mii.lu.lv

1923–1932; Endzelin and Hausenberg, 1934–1946). Apart from this, the only work where some data can be found are the two volumes of the Latvian Etymological Dictionary (Karulis, 1992). On this background, the need to create a new lexicographical source where lexis from the early written sources is present has been voiced several times. The Latvian linguist Rūķe-Draviņa, who after World War II continued her linguistic activities in Sweden, voiced the need for compiling the complete material of the Latvian language and proposed to develop a dictionary of the early written texts that should open new perspectives in the study of the history of Latvian (Rūķe, 1961). Unfortunately, this idea did not get to be realised. Also during the 1990s the idea to initiate a historical dictionary of the Latvian language was proposed (Baldunčiks, 1994). But no dictionary can be compiled without the appropriate data – including primary lexicographical sources (written texts) and secondary sources (previous dictionaries). Today, the prerequisite of any lexicographer is data in electronic form.

The development of a database of the first printed Latvian texts was initiated in the 90s, when the most significant printed sources were manually typed in. The Institute of Mathematics and Computer Science (henceforth IMCS) at the University of Latvia (henceforth UL) initiated the work on the digitalisation of early written Latvian texts. In 1992–1994 the text of the first translation of the Holy Bible was digitalised at the Laboratory of Artificial Intelligence, IMCS (Spektors and Baltiņa, 1994). Shortly after – in 1995–1996 – the work on creating a database of the early written Latvian texts was initiated and several early printed sources (from the 17th century) were prepared in an electronic form (Ozoliņa, 1997; 1998). In order to ensure the possibility of printing early Latvian texts, the Latvian software company Tilde was engaged in order to design the fonts FrakturaSpecial in 1996.

Unfortunately, the data collection as well as the processing of the database and its supplementing with new material at the Artificial Intelligence Laboratory has been interrupted for a longer time, but parts of the database have been used in studies of the history of the Latvian language and to study the language of particular authors.

Due to different matters and obstacles, the idea of creating a historical dictionary of Latvian as mentioned before has not yet been realised. Thus again, in 2001 Trevor Fennell from Flinders University of South Australia invited and encouraged scholars in Latvia to start work on the dictionary of Old Latvian. He put forward a number of questions to be solved and issues concerning such a dictionary (Fennels, 2002).

In response to Fennell's call, a pilot project on the development of an electronic dictionary of 17th century Latvian was initiated early in 2002, funded by the Latvian Culture Foundation. This was a joint project of the Department of Baltic Languages at the Faculty of Philology, UL, and the IMCS, UL. The project was headed by Pēteris Vanags and carried out by a team comprising both linguists and software engineers. The task of this pilot project was to reach awareness of the data necessary for such a dictionary and to draw up a methodology for compiling the dictionary, as well as to test the possibilities of modern technologies in dealing with Old Latvian texts. A pilot project on the development of an electronic dictionary of the early printed texts using XML technologies (Milčonoka, 2002) served as a test bed and contributed towards building up the first publicly available Latvian corpus.

In 2002 the first stage of the development of the Corpus of early written Latvian texts was finished and the Corpus was made accessible on-line.

3. The corpus of early written Latvian

3.1. Corpus design

The aim of the Corpus is to facilitate diachronic studies of the language, as well as to popularise early written texts to a wider audience. Nevertheless, the main purpose of creating this Corpus was to ensure the necessary data for a historical dictionary of Latvian. We may speak about three stages in the development of the corpus (1st stage – until 2002, 2nd stage – 2004; 3rd stage has started in 2005 and is still on-going). Due to different reasons (including lack of human resources) the work on the project has been carried out with varying intensity during its different stages, but the main point is that work is carried out on an ongoing basis.

The main focus lies on the 17th-century texts. The first task one had to cope with was the selection of printed sources representing this time period. The union catalogue of ancient prints in Latvian (Seniespiedumi latviešu valodā 1525–1855, 1999), published by the National Library of Latvia, lists 101 entries referring to the printed sources of the 17th century. This catalogue includes all the editions of sources printed in the 17th century.

One of the tasks to be solved was to make a selection of text types to be included in the Corpus and to make a decision about the proportions of each type of text. The 17th century is the time when the first secular texts and lexicographical sources were created and a translation of the Holy Bible was carried out. After examining the union catalogue a solution was found. At the outset it was decided to include only the first editions of any source, leaving the inclusion of repeated editions as a task for the future (they might serve as a basis for comparative analysis). It turned out that the dominant types of printed texts representing this period are ecclesiastical:

- scripture;
- religious prose;
- Church hymns.

Apart from these, we find some other text types which are represented only by a few titles:

- bilingual, trilingual and quadralingual dictionaries;
- narrative prose;
- grammar texts in German (and Latin) with Latvian examples and paradigms;
- ABC books;
- congratulatory poems;
- theory of poetics in German with Latvian examples;
- legal texts.

The texts that were manually typed in at the IMCS in the mid-90s were mainly religious texts (catechisms, hymnals, a sermon book, the Holy Bible) with great cultural and linguistic value. Therefore it was decided to proofread these texts, unify the character encoding (transforming them from Yamaha MSX coding to Windows ASCII coding) and add some text mark-up. Ecclesiastical texts made the core of the Corpus, covering hymnals and different types of prose (e.g. sermons consisting of a

fragment of scripture followed by some historical narrative, didactical prose and narrative prose). Apart from this, some new texts were scanned and added to the Corpus. Here, the availability of the particular source was taken into consideration – only those texts which were kept in Riga's libraries were scanned. The compilers of the Corpus benefited from the co-operation with the Department of Rare Books and Manuscripts at the National Library of Latvia, which kindly offered valuable sources. Scanned facsimiles were handed over to the Library for its internal use.

Another solution that had to be made concerned the size of the sample. Due to the fact that the compilers' aim was to introduce the texts to a wider audience (some 17th-century sources are of rare availability), a decision was made to include all the texts *in toto*. We did not tackle the issue of how large a sample should be in order to be representative, but texts of different length were included, e.g. sixty-six running words of Our Lord and more than 270 000 running words of a sermon book written by Georgius Mancelius. This, of course, causes some problems concerning the influence of one particular author's language in the general corpus, which should be prevented in the future by adding new sources.

The developers have faced two issues: first, do only printed texts represent the language of that time; second, how to get good data for the forthcoming dictionary. Texts representing the 17th century are mainly ecclesiastic ones: hymns, Evangelists' books and Epistles, a translation of the Holy Bible, a sermon book etc. In order to vary the content of the Corpus and to get a more precise picture of the language of that time, a decision was made to add some of available transcripts of manuscript dictionaries. During the second stage of the corpus development in 2004 the corpus was supplemented with data from two manuscript dictionary transcripts (the deciphering and publishing of these data were carried out by Trevor G. Fennell (Fennell 1997; 1998)). While the Corpus until then had been described as a corpus of "early printed" Latvian texts, this was now changed to "early written" Latvian texts.

No	Text ID	Title	Type	Running words
1	JT1685	Tas Jauns Testaments	ecclesiastical (scripture)	161 359
2	Manc1654_LP1	Lang=gewünschte Lettische Postill I	ecclesiastical (sermon)	127 534
3	Manc1654_LP2	Lang=gewünschte Lettische Postill II	ecclesiastical (sermon)	99 646
4	LGL1685_K1	Lettische geistliche Lieder vnd Collecten	ecclesiastical (hymns)	75 064
5	LGL1685_V5	Lettische geistliche Lieder vnd Collecten	ecclesiastical (hymns)	72 302
6	Manc1654_LP3	Lettische Lang=gewünschte Postill III	ecclesiastical (sermon)	49 538
7	Manc1631_LVM	Lettisch Vade mecum	ecclesiastical (scripture)	38 179
8	Manc1631_LGL	Lettische geistliche Lieder vnd Psalmen	ecclesiastical (hymns)	36 024
9	EvEp1615	Euangelia vnd Episteln	ecclesiastical (scripture)	32 444
10	Ps1615	Psalmen vnd geistliche Lieder	ecclesiastical (hymns)	30 478
11	Manc1631_Syr	Das Haus=, Zucht= vnd Lehrbuch Jesu Syrachs	ecclesiastical (scripture)	24 986
12	Elg1621_GCG	Geistliche Catholiche Gesänge	ecclesiastical (hymns)	17 283

13	Fuer1650_70_2ms	Lettisches und Teutsches Wörterbuch	dictionary	16 073
14	Fuer1650_70_1ms	Lettisches und Teutsches Wörterbuch	dictionary	14 524
15	Manc1637_Sal	Die Sprüche Salomonis	ecclesiastical (scripture)	13 280
16	Ench1615	Enchiridion	ecclesiastical (scripture)	8 260
17	Manc1631_Cat	Der kleine Catechismus	ecclesiastical (scripture)	7 971
18	Reit1675_UeP	Eine Übersetzungsprobe	ecclesiastical (scripture)	3 074
19	SKL1696_KB	Sawadi Karra=Teesas Likkumi	secular (legal texts)	1 403
20	SKL1696_RA	Sawadi Karra=Teesas Likkumi	secular (legal texts)	1 395
21	SL1684	Sohdu=Likkums prett to Behrno=Muschinaschanu	secular (legal texts)	542
22	Reit1675_OD	Oratio Dominica XL Linguarum	ecclesiastical (scripture)	66

Table 1: List of 17th century sources included in the Corpus.

As a result, the diversity of the Corpus data turned out as follows: ecclesiastical texts – 96 %; dictionaries – 3.6 % and secular texts – 0.4 %.

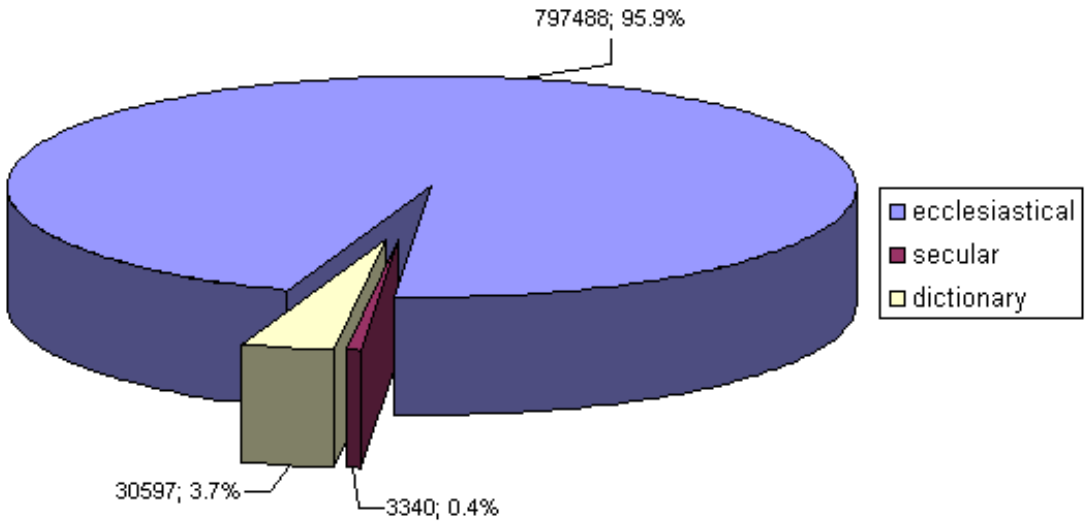

Figure 1: Corpus data (running words and proportion) according to the text type.

Concerning the content of the Corpus, the question of representativeness is usually discussed. The developers adhere to the opinion that all the texts written in the period under consideration are representative of that time. In the future we would like to include all available texts from the 16th–17th century in our Corpus.

In 2004, three (out of four available) sources from the 16th century were added to the Corpus, all of them religious texts: a catechism, Evangelists' books and Epistles, as well as church hymns.

No	Text ID	Title	Type	Running words
1	Ench1586	Enchiridion	ecclesiastical	7068
2	EvEp1587	Euangelia vnd Epiſteln	ecclesiastical	32519
3	UP1587	Vndeudsche Psalmen	ecclesiastical	13055

Table 2: List of 16th century sources added to the Corpus.

In order to widen the scope of the corpus and to provide better data for a dictionary, a decision was made to enlarge the Corpus by adding some fiction texts and texts on popular science from the 18th century. As there are fairly many texts from this time, priority was given to texts covering the lexis of different subjects: medicine, agriculture, legal texts, as well as science fiction. Only first editions were included. New 18th-century texts are currently being added to the Corpus; our task is to prepare five new titles (covering texts on medicine, science, agriculture and a piece of drama) before the end of 2007.

No	Text ID	Title	Type	Running words
1	CekFJ1790_KD	Kartupeļu dārzs	secular (didactical prose, agriculture)	1 511
2	CekFV1796_NL	Neapskāpjami likumi	secular (didactical prose, agriculture)	1 372
3	Depk1704_Vortr	Vortrab	dictionary	1 777
4	Eid1701_KB	Eid der Treue vor die Lettische Artillerie=Bediente	secular (legal texts)	203
5	Eid1701_RA	Eid der Treue vor die Lettische Artillerie=Bediente	secular (legal texts)	201
6	EvTA1753	Evangelia toto anno	ecclesiastical	16 715
7	Hag1790_IM	Īsa mācība priekš latviešiem	secular (didactical prose, medicine)	2 200
8	Lod1775_SEAPP	Sprediķis pie iesvētīšanas	religious prose	3 057
9	Lod1778_WTMD	Vārdi tās mūžīgas dzīvošanas	religious prose	13 937
10	MD1788	Mīļi Draugi!	religious prose	2 613
11	Rav1767_SD	Svētas domas	religious prose	2 498
12	SL1789	Skolas likumi	secular (didactical prose, education)	5 649
13	StendGF1789_SL	Ziņģu lustes	fiction (poetry)	8 702
14	Sulc1764_ARMST	Aizkraukles muižas un Rīmaņmuižas likumi dzimtļaudīm	secular (legal texts)	2 228
15	TII1790	Tā īsa izstāstīšana	religious prose	1 572

Table 3: List of 18th century sources added to the Corpus.

The current size of the corpus is 958 077 running words covering forty sources:

- 16th century – three sources and 52 642 running words;
- 17th century – twenty-two sources and 829 876 running words;
- 18th century – fifteen sources and 75 559 running words.

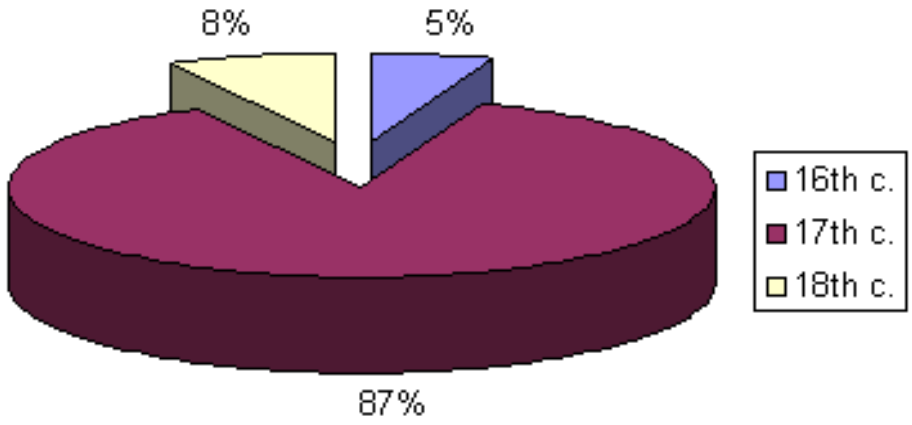

Figure 2: Corpus data according to chronological diversity.

As a result, our Corpus, at first intended as a synchronic one (to get the picture of the 17th century), turned into a diachronic one, showing a continuum of three centuries. This allows scholars to carry out several studies: to examine the history of spelling and morphological variants, to trace the beginnings of standardization of the Latvian language, to investigate the morphological system and lexical changes in the vocabulary, to pay attention to loan words in Latvian, their entry through the mediation of German and the adaptation process, etc. We strongly hope that the Corpus will encourage more studies, opening more possibilities to look at the early written Latvian language as a system, not only a collection of mistakes.

3.2 Corpus creation

When the digitalisation of Old Latvian texts started in the 90s, all the texts were manually keyboarded and Gothic script was transliterated into Latin script with some additional special characters. The collaboration with The National Library of Latvia established in 2002 made it possible to scan early printed texts, returning the scanned texts to the Library afterwards. The software program ABBYY Fine Reader 6.0 was trained to recognise the scanned texts with Gothic script; due to the established methodology, a precision of 80–90 percent has been achieved. Mistakes in scanned texts were much more predictable in comparison to human typing errors. The most typical mistakes left after scanning are: a match of some letters and digits, e.g. *l* (letter) and *1* (number); some combinations of two letters might be recognised as one letter and vice versa (e.g. *ni* vs. *m*); capitalization of some letters (usually the capital letters in old texts are decorated with special ornaments). Scanned facsimile pictures (600 dpi, colour and 200 dpi, black and white) are also available on-line.

A readily understandable abbreviation has been assigned to each text. If the author (or editor) of the particular text is known, the abbreviation consists of a shortened form of the name of the author, the year when the source was published and the first letters of the title, e.g. the abbreviation *Manc1631_LVM* stands for the text "Lettisch Vade mecum", which was edited by Georgius Mancelius and published in 1631. If there is no information about the author (or editor) of the book, the abbreviation consists of the first letters of the title and the year of its publishing, e.g.

the abbreviation *UP1587* stands for the church hymnals "Vndeudsche Psalmen", which were published in 1587. For the Holy Bible text we use abbreviations suggested by the Latvian Bible Society, e.g. *Mk* stands for the "Book of Mark".

Every single source has its own passport – it covers bibliographical information taken from the union catalogue mentioned before, an interactive index and the text itself, frequency indexes and word lists (taking into account case sensitivity), as well as a facsimile (if available).

Figure 3: The passport of the text "Lettisch Vade mecum".

The creation of the Corpus has undergone several stages: typing or scanning → proofreading → introducing some text conventions → adding some structural mark-up → automated verifying and processing.

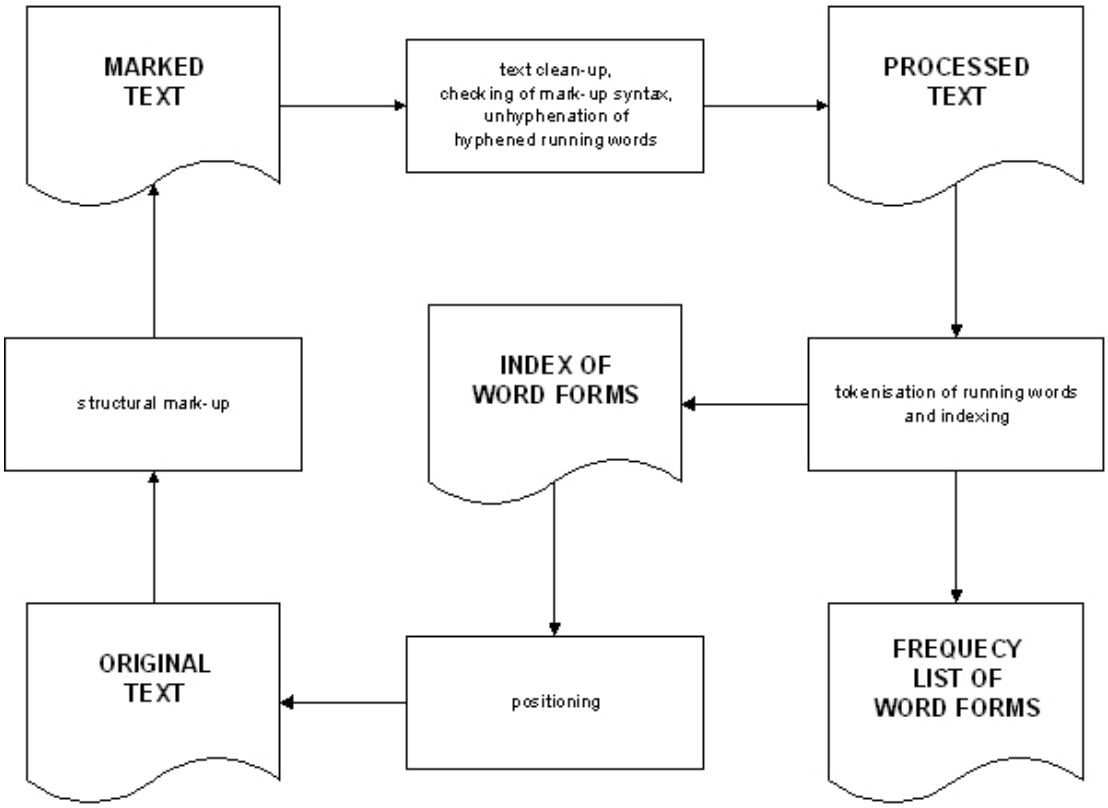

Figure 4: Process of corpus development.

The process of corpus development is still in progress.

3.3 Corpus annotation

The aim of the developers was to preserve the original text (except the character set) and its layout, thus we tried to reduce the amount of changes as much as possible. All Gothic script texts are converted into the Windows-1257 code page and compound symbols are introduced for accented letters and special characters. General format (lines, paragraph, word boundaries, punctuation, hyphens and dashes) is preserved. Concerning the font style, different size of letters as well as bold and italic type, these have not been marked; this information is kept and seen only for some texts which were scanned. A facsimile of the text is linked with a recognised *.doc* file where most formatting is kept. Some minor changes have been introduced in order to facilitate the indexing process, e.g. in early texts '=' is used both to separate parts of compound names and prefixes and to hyphenate a word; where used for hyphenation, '-' has been introduced. Different accents have been coded by placing the accent mark after the accented letter, e.g. the character *â* in the Corpus appears as a combination of two symbols: *a^*. Some new symbols were introduced to find a similar letter to Gothic script, which for the end-user might seem a bit strange (e.g. the use of *ſ* to mark the Gothic letter *ƒ*).

A number of conventions have been introduced to code obvious errors in a text (sometimes they are listed in an errata list at the end of the book). Words with a corrected form in curly brackets appear together as a separated item in word list, e.g:

auxtahx{autahx}, meaning ‘higher’. In the text (Manc1631_LVM 151.lpp. 13.r.) the same unit is presented in the following way:

```
12: Acklis Acklam Czeļļu rahdiet? Negg kriet tee abbidiwi
13: Beddreh ? Tas Mahzeklis nhe gir auxtahx{autahx} ka Mei§ters /
```

In order to identify foreign language material and to exclude it from subsequent word indexes, word lists and concordances, a mark-up for Latin, German, Polish and Greek words has been introduced.

```
Abbejads @v{der es mit bey"den halt,} @l{neuter}
      Abbejadi warr śazziht, @v{man kan}
      @v{bey"des, bey"derley" sagen.}
Ahbole. @v{der Apfelbaum}, Abols
      @v{der Apfel}. Ahbolinsch @l{dimin.}
      [Mesch] Ahboli @v{busch holtz Äpfel.}
      [Wilk=Ahbole]. @v{[Hagedorn]} @v{wilde}
      @v{[Rosen, Engel=Thier.]}
      [Semmes Ahboli] @v{dieses landes}
            @v{[einheimische Äpffel]}
      [Wahz=Ahboli] @v{[Teutsche Garten=Äpffel]}
```

Figure 5: An example of a text with mark-up for foreign language text.

In addition, an annotation is introduced for cross-notes to the related parts of the text (this is usually done in the text of the Holy Bible, in the New Testament there are some references to the text of the Old Testament), structural containers and other elements. Thus, the introduced mark-up refers to the text itself and its representation. No grammatical tagging is applied. Manual mark-up without appropriate software requires huge human resources. Unfortunately, we cannot apply software developed for Modern Latvian due to the fact that the early texts are rich in morphological and orthographical variants which are hard to foreseen and elaborate in an automated mark-up software. Several automated test procedures and methods have been introduced to check the mark-up and char set consistence.

3.4 Corpus exploration tools

In order to ensure the maintenance of our corpus and the successful exploration of corpus data, a corpus platform has been developed. This platform makes it possible to get general statistical information both about the whole corpus and about a particular text in the form of the number of individual word forms (case sensitive and lower case) and the total size of running words. The average number of instances per word form is 10.6, but the dictionary data show a completely different result: 14 524 running words and 8 772 different word forms are met in Fürecker’s dictionary. Here, the instance score is 1.7. Navigation through the Corpus content is ensured, giving the user the opportunity to choose different authors and text types (three types are offered: ecclesiastical, secular and dictionaries).

Figure 6: Navigation in the Corpus: selecting a particular author and receiving information about all the texts which that person has authored or edited.

On-line search in word lists and frequency lists is ensured (Grūzītis, 2003; Milčonoka, 2003). Three types of queries are available:

- a single word or word form (e.g. *Kungs* 'Lord') – the result is instances with the word form *Kungs*;
- a part of a word or word form followed by any one letter (e.g. *Kung_*) – the result is instances with *Kung'*, *Kunga*, *Kunge*, *Kungi*, *Kungo*, *Kungs*, *Kungy*, *Kungu*; '_' might stand in any position;
- a part of a word or word form followed by any chain of letters (e.g. *Kung%*) – the result is instances with *Kung*, *Kung'*, *Kunga*, *Kungam*, *Kungam{Kunam}*, *Kungam{Kungan}*, *Kunga~*, *Kunge*, *Kungeem*, *Kungeems*, *Kungen*, *Kunges*, *Kunge{de}*, *Kungh'*, *Kungha*, *Kungha=Preeka`*, *Kungham*, *Kungha{Knngha}*, *Kungha{Kungha}*, *Kunheem*, *Kungheems*, *Kunghee~*, *Kunghi*, *Kungho*, *Kunghs*, *Kunghu*, *Kunghus*, *Kunghu{Kungho}*, *Kunghu{Kuughu}*, *Kungi*, *Kungim*, *Kungims*, *Kungo*, *Kungs*, *Kungs=ißkai§śa*, *Kungsteht*, *Kungs{Knngs}*, *Kungs{Kunss}*, *Kungß*, *Kungu*, *Kungus*, *Kungy*; '*%*' might stand in any position.

If a particular source is relatively small, an interactive search in a list of word forms is possible as well, and the context of the word form can be rendered. The context is either a verse in the Bible text or a sentence (and a corresponding page) in all other texts. Apart from this, a reverse dictionary has been created for every single source. All data are integrated in a corpus database together with source texts and processing results are available also for downloading.

A *kwic*-concordance program which deals with early Latvian texts has been developed at the IMCS. The types of available queries are the same as in word

indexes, several criteria and bounding of scope can be applied. A possibility to get to the extended context is available.

Figure 7: Concordance lines with the query *Kri§tus*.

4. Conclusion

Our tasks in the future are twofold. First, we will continue work on developing the contents of the corpus (in order to ensure more data for the future historical dictionary of Latvian) – more 18th century texts will be added, some manuscript dictionary transcripts should be processed. We would like to add more sources (repeated editions) from the 17th century as well, to carry out some comparative studies. Second, we will improve the corpus exploitation tools (detailed statistical analysis, sorting possibilities in search results and collocation analysis are required and foreseen). More distant future plans concern a selection of text fragments (the 16th–18th century) and a development of a representative Corpus of Early Written Latvian.

Acknowledgements

The Corpus of Early Written Latvian *SENIE* could be developed due to financial support from different institutions during several years. These are: Soros Foundation-Latvia (1994–1996), the Latvian Council of Sciences (1997), the Cultural Foundation of Latvia (2002 and 2003), the University of Latvia (2002 and 2005) and the Ministry of Education and Science of the Republic of Latvia (2003). It could be done by the joint efforts of many people – linguists, students of the Faculty of Philology, UL and software engineers. Among them Andrejs Spektors, Head of Artificial Intelligence Laboratory, IMCS, who has always been very enthusiastic about the idea of digitalisation of Latvian texts, Anita Ozoliņa, former software engineer of the

Laboratory, who carried out the development of the first database of the early written texts, Maija Baltiņa, who first came to the Laboratory in the late 1980s with an idea to analyse the Book of Job and was hoping to receive some support from software engineers, and Normunds Grūzītis, the present software engineer and the developer of corpus platform, should be mentioned.

References

Baldunčiks, J. (1994) Latviešu valodas vēsturiskā vārdnīca, in Valoda un tehnika Eiropā 2000. Baltijas perspektīva, p. 20. Rīga: LU MII.

Endzelīns, J. (1951) Latviešu valodas gramatika. Rīga: Latvijas valsts izdevniecība.

Fennell, T. G. (1997) Fürecker's Dictionary: The First Manuscript. Rīga: Latvijas Akadēmiskā bibliotēka.

Fennell, T. G. (1998) Fürecker's Dictionary: The Second Manuscript. Rīga: Latvijas Akadēmiskā bibliotēka.

Fennells, T. G. (2002) Veclatviešu valodas vārdnīca: domas par iespējamību un saturu, in Vārds un tā pētīšanas aspekti. Rakstu krājums 6, pp. 33–37. Liepāja: Liepājas Pedagoģiskā augstskola.

Grūzītis, N. (2003). Development of Text Corpus Using Java and XML. Unpublished Bachelor's thesis (in Latvian). Riga: Department of Computer Science, Faculty of Physics and Mathematics, UL.

Mühlenbachs, K. (1923–1932). Lettisch–deutsches Wörterbuch. Riga: Lettisch Kultufronds.

Karulis, K. (1992) Latviešu etimoloģijas vārdnīca. 2 sēj. Rīga: Avots.

Endzelin J., E. Hausenberg (1934–1946). Ergänzungen und Berichtigungen zu K. Mühlenbachs Lettisch–deutschem Wörterbuch. Riga: Lettisch Kultufronds.

Milčonoka, E. (2002) Корпусная лингвистика и историческая лексикография, in Материалы XXXI межвузовской научно-методической конференции преподавателей и аспирантов. Выпуск 1. Секция балтистики. Тезисы докладов, стр. 34. Санкт-Петербург: Санкт-Петербургский государственный университет.

Milčonoka, E. (2003) 'Latviešu valodas 17. gadsimta teksti internetā'. *Baltu filoloģija XII (1)*, 139–50.

Ozoliņa, A. (1997) '17. gs. tekstu datorfonda izveides programmlīdzekļi'. *Linguistica Lettica, 1*, 219–25.

Ozoliņa, A. (1998) 'Seno tekstu datorfonda izveides programmatūra'. *Lietuvių kalbos klausimai*, *XXXIX*, 211–18.

Ozols, A. (1965) Veclatviešu rakstu valoda. Rīga: Liesma.

Rūķe, V. (1961) 'Turpmākie uzdevumiem latviešu valodas pētīšanā'. *Ceļi X*, 5–16.

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

Spektors, A., M. Baltiņa (1994). Projekts "Latviešu valodas vēsturisko tekstu datu bāzes izveide, in Valoda un tehnika Eiropā 2000. Baltijas perspektīva, p. 30. Rīga: LU MII.