Baltic J. Modern Computing, Vol. 12 (2024), No. 4, 570-584
https://doi.org/10.22364/bjmc.2024.12.4.18

The Chylinski Bible Database for Linguistic Research

Gina KAVALIŪNAITĖ[1], Wolf-Dieter SYRING[2], Felix THIES[3], Paweł BRUDZYŃSKI[4]

[1]Dept of Baltic Studies, Institute for the Languages and Cultures of the Baltic, Vilnius University, Lithuania, Universiteto g. 3, Vilnius
[2]TextDataSoft, Text Corpus Design and Database Development, Buxtehude, Germany
[3]Chair of Historical and Comparative Linguistics, Institute of German Language and Linguistics, Humboldt-University, Germany, Dorotheenstraße 24, Berlin.
[4]Dept of Baltic Studies, Faculty of Polish Studies, University of Warsaw, Poland, Krakowskie Przedmieście 26/28, Warsaw

gina.holvoet@flf.vu.lt; syring@textdatasoft.de;
felix.thies@hu-berlin.de; p.brudzynski@uw.edu.pl

ORCID 0009-0004-9167-8585, ORCID 0009-0008-9808-3647, ORCID 0009-0006-6328-0435, ORCID 0000-0002-1850-0818

Abstract. This article presents the Chylinski Bible website and its future extensions. After a brief outline of the publication history of the Chylinski Bible and the character of the texts to be digitized, we discuss the different approaches to the manuscript and printed parts. We further present the structure of the website and the methods chosen, focusing on data structures, compliance with TEI, annotation of structural data, annotation of manuscript corrections, editorial emendations, principles of grammatical annotation, database and interface design, search functionalities and guidelines for future expansion.
Keywords: digitized historical texts; Chylinski Bible; morphological annotation

1. Introduction

In recent years, multilingual parallel Bible corpora have gained increasing prominence as a valuable resource supporting various domains of linguistic research. Historical Bible corpora in particular have come to be recognized as important tools for the diachronic study of language. The domains to which Bible corpora are relevant include translation studies, linguistics, Biblical exegesis, computational linguistics, and natural language processing.

Bible study tools such as *Bible Hub*, and *Scripture 4 All* offer a wide range of tools such as glossaries, maps, interlinear Bible versions, and multilingual databases, some of them also used by linguists for academic research. However, not every collection of Bible texts accessible in digital space deserves the name of a Biblical corpus suitable for linguistic research. Annotation according to the standards accepted

for the individual languages and translations, as well as interlingual alignment, are among the defining features of what we could call a parallel Bible corpus (Resnik et al., 1999: 132). Alongside the many Bible websites containing a choice of translation texts, there are Bible corpora specifically intended for academic research[1]. Among the important diachronic corpora, we must mention the grammatically annotated diachronic corpus of 36 Bible translations, covering English, German, Dutch, and Swedish. It was compiled for the purposes of a diachronic research project into complex verb constructions in Germanic (a wealth of useful information can be found in (Bouma et al., 2020: 5232–5339)). Warsaw University offers a website covering 16th-century gospels: ewangelie.uw.edu.pl. It contains ten 16th-century Polish gospel translations, aligned verse by verse. Apart from parallel Bible corpora, there are databases covering individual Bible translations, such as the Dutch States Bible (www.statenvertaling.net), the Clementine Vulgate project (vulsearch.sourceforge.net) etc.

No database of Bible translations has been compiled in Lithuania to date, though such a database would account for a prominent part of a historical Lithuanian text corpus as each Western Christian church of Lithuania has produced several Bible translations. These are valuable specimina of Old Lithuanian writing. Many of them can be found on the Internet in different formats. A website created at the Institute for the Lithuanian Language[2] offers three New Testament translations made between the 17th and 19th centuries, the first edition of the Quandt Bible as well as several books from the first Lithuanian Bible translation, Bretke's manuscript Old Testament, in Word format, with concordances. There is a search functionality enabling retrieval of Bible verses across parallel translations, but it covers only a few printed versions. However, in the absence of a historical Lithuanian corpus, an urgent task for the future is to create an annotated diachronic Lithuanian Bible corpus allowing scholars to compare translations, establish their mutual relationships and search for data on the development of grammatical and lexical features. The Chylinski Bible website,[3] created in 2019, is intended to become part of a future Lithuanian Bible corpus. It now covers Chylinski's MS translation of the New Testament (henceforth: ChNT); at a further stage the extant text of his Old Testament (henceforth: ChOT) will be added, and search functionalities will be improved and expanded. The whole Bible text will be aligned verse by verse with the main and subsidiary translation sources – the Dutch Estates Bible (1657) and the Polish Gdansk Bible (1632). In this article we discuss the essentials: sources underlying the database, data structures of the electronic edition, source manuscript corrections, editorial remarks, annotation of lemmata for part of speech, and morphological properties. We discuss the general guidelines followed in lemmatization and morphological annotation, while a detailed analysis of problematic aspects will be given in a separate publication.

[1] The creation of the first linguistically informed parallel annotated Bible corpora goes back to the late 20th century. It was followed by others, among which a synchronic corpus of 100 parallel New Testament translations, aligned at sentence level, is worth mentioning. It is the only one including a Lithuanian Bible translation, that of Kavaliauskas and Rubšys. The texts were collected from various freely accessible Bible websites. The corpus targets researchers in the domains of linguistics and natural language processing (Christodouloupoulos, Steedman 2015: 375–395).

[2] https://seniejirastai.lki.lt

[3] https://www.chylinskibible.flf.vu.lt

2. Chylinski, his Bible, and its publication

Samuel Boguslaus Chylinski's Bible is the second complete Bible translation into Lithuanian and the first to be (partly) printed. The printing started in 1660 with the support of prominent English men of letters and King Charles II himself, but was discontinued and never resumed. The printed part comprises ChOT up to Job 6. A report of the Lithuanian Synod's delegate states that 3000 copies were printed (Kavaliūnaitė, 2015: doc. 51). Only three printed copies are known, of which only one is now certainly extant; they are referred to as the Vilnius, Berlin, and London copies respectively, based on their former location. The lost Vilnius copy was the longest (416 pages). The Berlin copy, lost since World War II, was somewhat shorter (384 pages). The only extant copy, held by the British Library in London (shelfmark C 51.b.13), is also the shortest (176 pages). For publication on the website, the London copy was used, complemented by prewar photographs of the Berlin copy discovered in 2008 as well as photographs of fragments from the Vilnius copy reproduced in publications by Jerzy Broel-Plater (Jurgutis and Žukas, 1963: 193–203), Adam Jocher (1842: 109–111), Eduard Wolter (1887: 71–102), Maurycy Stankiewicz (1889: 55–57) as well as in the preface to the Quandt Bible (1735).

In 1926, the British Museum acquired the ChNT manuscript (shelfmark MS 41301). The Polish scholar Stanisław Kot researched the MS and its history (Kot 1958: XLIII–LXVII). In 1958, the Polish Academy of Sciences published a transcription of the MS, prepared for publication by Profs. Jan Otrębski and Czesław Kudzinowski. Their edition consists of three volumes. In addition to the transcriptions (Otrębski and Kudzinowski, 1958) it comprises an index of word forms (Kudzinowski, 1964) and photographs of the MS (Kudzinowski, 1984).

Research on Chylinski and his Bible translation was resumed in the early 21st century. A facsimile edition of ChOT (Kavaliūnaitė, 2008) was followed in 2019 by a new high-quality facsimile of ChNT, published together with a study of the MS (Kavaliūnaitė, 2019; Čapaitė, 2019). In parallel with the volume dedicated to ChNT, the website www.chylinskibible.flf.vu.lt was launched. Its goal is to provide access to Chylinski's work. Currently, it allows the user to read a transcription of the ChNT text and compare it with photographs of the original. The entire transcription is provided with philological tools to assist in research. The website offers word indexes, search engines for various word forms, and tracking of the text's editorial layers, along with proposals for the reconstruction of some fragments. The website is constantly being improved and expanded, which makes it a comprehensive tool for examining texts with a complex editorial history. In addition to the text itself, the website contains information on the author and his work. Ultimately, the Chylinski Bible project seeks to be a repository presenting Chylinski's work and information about the author, but also an overview of research.

3. The data sources and the task

The Chylinski Bible is extant in two shapes: the ChOT as a printed text and the ChNT in the manuscript. The two constitute the backbone of the website. At the next stage, the ChB text will be aligned verse by verse with the basic source, the Dutch Estates Bible (1657), and the subsidiary source, the Polish Gdansk Bible (1632). An editor working on

the diplomatic transcription and annotation of such a miscellaneous source for publication on the internet must adopt a dual approach.

The publication of the ChB on the Internet started with the ChNT manuscript. The manuscript text has gone through many correction rounds, and its opening and closing pages contain many jottings, glossary-like lists of words and phrases, and multilingual fragments. In preparing the text for publication, we decided to set the framework apart from the NT translation and to classify the entries into thematic groups. The framework entries are also provided with linguistic annotation showing in what circumstances the author of the MS switched between language codes. The manuscript is transcribed faithfully, with all characters and punctuation marks rendered. During transcription, common end-of-word abbreviations, Bible book titles, and other abbreviations are expanded. Many characters used by Chylinski, especially consonants, correspond to the modern letter inventory, but others differ. Some of the consonants had additional symbols absent from the modern language, based on the German-Polish spelling tradition; they are all rendered as in Chylinski's MS. The vowel system was probably complex and different from the modern system, as each sound has several representations, e.g., *ȧ, ȯ* is used in words where two vowels in separate syllables collide, as in *Izȧȯkas*.

Linguists being the main target group of the website, annotation focuses on the character and sequence of the corrections at the expense of formal features of the MS: whether corrections are written above or beneath a word or in the margin is not marked, as the user can see all this by opening a window showing the relevant fragment in facsimile.

A rare feature of the database – appealing to different users according to whether they are interested in text history or linguistic variation – is the possibility of retrieving lists of changes classified according to linguistic features (lexical, morphological, morphosyntactic, syntactic, other). For this classification see section *4.4 Source corrections*.

In transcribing the ChNT text, Kudzinowski and Otrębski's edition (1958) was often consulted, and it was of invaluable help in clarifying many readings. In many places, however, the new transcription diverges from that of the Polish scholars. Having no access to the translation source, they read some words incorrectly and misunderstood some entries. For instance, Dutch *mande* 'basket' is read *maude* (ChNT 220v: 15), with *maude* apparently interpreted as a Lithuanian word.

The ChNT database is already operational. Once expanded to comprise the whole Chylinski Bible, it will offer new search facilities and allow users to download derived lists. The search criteria will be word form, word class, frequency, and – for the ChNT – type of editorial change. Users will be offered easy navigation by MS and book page, and every verse will be linked to a facsimile fragment.

The surviving part of ChOT is printed, and the characters are not fundamentally different from those of ChNT, so that a diplomatic transcription is unproblematic. It will be based on the London copy and, for the parts missing from it, on the Berlin and Vilnius copies. The surviving ChOT text from the London and Berlin copies is expected to be digitized, morphologically annotated, and uploaded to the database in the near future. A trial run of ChOT annotation is now under way. Based on Genesis, it comprises c. 34000 words on 42 pages. For a trial run, an optical character recognition (OCR) of Genesis was conducted, using *Abbyy Finereader 14*. The results were corrected manually, yielding a diplomatic transcription. Due to the wide extent of variation and

lack of standardization in the language and spelling of Lithuanian texts, automatic annotation is impossible. The texts are annotated using *Field Linguist's Toolbox*, a program for dictionary-based semi-automatic annotation. As no specialized dictionaries for Old Lithuanian are available, dictionary entries are created during the work process, i.e., on every first occurrence. The annotation process will be different for the two text parts. It will start with ChOT; annotating ChNT will be technically more challenging due to the many editorial layers and the lack of a final editorial touch.

The website home page will contain background information on Chylinski and his translation. The website will also include smaller printed texts published by Chylinski and the metadata of the texts building the backbone of the website. The user will finally find an instruction on how to use the search facilities and a discussion of the criteria underlying the classification of editorial changes in the MS.

4. Data structures of the electronic edition

In this part, we describe the data structures of the electronic edition. The preliminary decision was to create a TEI-compliant XML text as a basis for all database structures, providing

- structural information regarding page layout and biblical subdivisions,
- textual corrections applied to the original manuscript and assigned to layers,
- editorial emendations by the publishers,
- linguistic annotation of lemmata, part of speech, and morphological properties.

We discuss the general guidelines employed for lemmatization and morphological annotation, whereas the more minute problems can only be touched upon. Finally, we outline the database structure and the interface design.

Data arising from scientific and scholarly research should be published observing the FAIR principles (www.go-fair.org/fair-principles) ensuring that the data are Findable, Accessible, Interoperable, and Reusable. Funding institutions often require these or similar principles of data handling for project applications. The Chylinski project uses two basic types of data structures, viz. (a) the well-established XML-based TEI format to provide interoperable and reusable data, and (b) a well-defined relational database schema to provide findable and freely accessible data in a web interface (see 5). The TEI-compatible text will be made accessible to the international scholarly community.

4.1. Structural elements

In the domain of Bible texts, manuscripts, and printed versions are always structured by their physical appearance and by their content-related divisions. Most pages consist of one or more columns of Biblical text, often with margins providing references and notes, in many cases enriched by introductions and summaries. Additionally, a headline may show the page number and/or the book title as well as a bottom line with catchwords and/or foliation marks.

The **page layout** is represented by the TEI tags *pb* (page break), *cb* (column break), and *lb* (line break). These are so-called milestone tags indicating the beginning of a physical or layout unit respectively, without embedded tags.

The tag *fw* (form work) allows several *type* attributes like ‘head’, ‘pageNum’, or ‘catch’. It always follows or precedes a page break, containing headline text, page numeration, or catchword indicating the beginning of the next page.

The content of the **Biblical units** is represented by typed division tags *div*, using the types ‘book’, ‘chapter’, and ‘verse’, which are hierarchically organized.

Additional meta-content like captions, introductions, or summaries are placed at the beginning of a unit and realized as *head* tags, using the same types as the encasing *div* tags. Many Bible translations also indicate the end of a book by a subscript, marked by the *trailer* tag.

```
<div type="book" n="2Chr">
   <pb n="328"/><fw type="pageNum">328</fw>
   <head type="book"><lb/>ANTRA KNIGA <lb/>KRAYNIKU.</head>
   <head type="introduction">...</head>
   <cb/>
   <div type="chapter" n="13">
      <head type="chapter"><lb/>PAGUL. XIII.</head>
      <head type="summary">...</head>
      <div type="verse" n="1">
         <lb/><head type="verse">[1]</head> ASzmoſe liekoſe metoſe Karalaus Jero-
         <lb/>beamo, Abia tapo Karalumi and Judos.
      </div> <!-- verse -->
      <div type="verse" n="2">
         <lb/><head type="verse">2</head> Karalawo per treis metus Jeruzaley: ó
         <lb/>wardas motynos jo buwo Michaja, dukte Urielo
         <lb/>jiz Gibeos: ir buwo wayna terp Abios, ir terp
         <lb/>Jerobeama.
      </div> <!-- verse -->
      <div type="verse" n="3">
         <lb/><head type="verse">3</head> Ir ſuryſzo Abia kowę, kareys kareywingu
         <lb/>wiru, kiatwertu ſzymtu tukſtanciu iſzrynktu
         <lb/>wiru; Jerobeam wel ſuſtate prieſz ghi kowes
         <lb/>redą iſz aſztoniu ſzymtu tukſtanciu iſzrynktu
         <lb/>wiru, budru galunu.
      </div> <!-- verse -->
   </div> <!-- chapter -->
   <trailer><lb/>Efezump raßyta iß Rima <lb/>(&o_; nusiusta per &a;. ...</trailer>
</div> <!-- book -->
```

Figure 1. An example taken from ChOT

The strict distinction between milestones and content tags as well as the separation of Bible texts from meta-information on the tag level enables a quick addressing of physical and Biblical units and allows the definition of specific search domains and a parallel presentation of different translations.

4.2. Annotation of manuscript corrections and layers

The text of the ChNT is heavily edited, with several layers of corrections by the translator himself and three other editors (Čapaitė 2019: clxiij–clxxj). In the annotation engine, each text change was first evaluated as a complete change or a synonymous variant: A **complete change** consists of a word or text fragment being crossed out or otherwise erased (e.g., overwritten) and replaced with a new one. A **synonymous variant** consists in that a word or text fragment is not deleted (though underlined in

some cases), but a synonymous variant is given above it or in the margin. In this case, the translator was not quite sure of the final rendering; the latest written version is included in the final version of the text by decision of the publisher. A **restored variant** consists of a complete change being made, e.g., by crossing out and writing another word above it, and a previous variant being then restored by placing a dotted line or dots below the deleted word. The restored version is included in the last version, and all editing steps are reflected in the text version with all editorial layers (all versions).

In the ChNT corpus, the text is presented so that the reader can choose from three text versions: **the initial text,** without any changes, is called the *first version*. It consists of the basic MS text without immediate (*Sofortkorrekturen*) or later-stage corrections. Unedited text is shown in black in the transcription, edited text in green. The corrected places contain language errors that have been corrected by the publishers (see 4.5 *Editorial remarks: annotation of emendations*). **The final text version** with obvious errors corrected by the publishers, words restored where empty spaces were left, and the order of words changed according to the numbers written above the words by the translator, is called *last version*). The final version includes the latest correction in the manuscript. In some places, if the word controlling the inflected words was edited or in similar cases, the change was made only in one segment of the construction, therefore the text is not smooth in the final version. The last version is a text with all editorial layers. It reflects all editing steps and is referred to as *all versions*.

In corpus linguistics, the most common types of linguistic annotation of written texts are lemmatization, part-of-speech tagging focusing on syntactic and morphological annotation, syntactic parsing, and semantic annotation (Gries, Berez 2017: 383–387). In the nearest future, we plan to carry out the morphological annotation (see 4.4 *Annotation of lemmata and morphological properties*). The ChNT manuscript database also contains less common types of annotations relevant to researchers of text history, editorial process, and linguistic variation. All changes in the ChNT text were classified according to grammatical features. The following types are distinguished:

Lexical changes (tagged *lex*) comprise changes where a word is replaced with a synonym, or a word with the same lexical stem but a different derivational affix. This label also applies to cases where the lexical substitution (in prepositions) further involves a change in governed case; where the change involves only a verbal prefix but not the verbal stem; when enclitic discourse particles are added to a lexeme or removed; when a noun is moved to the opposite gender class; and when changes affect the presence or absence as well as the exponency of verbal reflexivity. Long and short forms of indeclinables are counted as different lexemes, and so are variant forms of Biblical names.

Morphological changes (tagged *morph*) involve changes in the morphological exponency of a grammatical feature, while the morphosyntactic category is not affected. This includes alternations between long and short variants of inflectional affixes; shifts to other inflectional classes; alternating stem forms; addition or removal of connecting vowels; and changes in the placement of clitics and their reduplication.

Morphosyntactic changes (tagged *morph-synt*) comprise changes from one morphosyntactic category to another. This may be the replacement of one case with another; replacement of a propositional phrase with a bare case form (this could also be tagged as syntactic); changes from singular to plural or from indefinite to a definite form of the adjective; changes in tense form or voice (active to passive), etc.

Spelling changes (tagged *spell*) comprise addition or removal of palatalization marks; changes in the rendering of inflectional affixes or stems; addition or removal of dots on vowel or consonant characters, nasal signs, or the stroke marking unpalatalized *l*; changes from lower-case to capital letters and vice versa; introduction or removal of consonantal gemination; and replacement with another letter that could have the same phonetic value, like *w* alongside *v*, etc.

```
<div type="verse" n="1">
  <lb/><head type="verse">1.</head>Kad tada Jeʒus gime
  <choice>
    <del>Betheehme</del>
    <add type="writ">Bethleheme</add>
  </choice>
  <lb/>(mieſte gulinciame)
  <choice>
    <del rend="ul">Judeoÿ</del>
    <add type="lex synt">Zydu
      ziamey</add>
  </choice>,
  dienoſe
  <lb/>Karalaus Heroda. Sztey
  (<choice>
    <del> nekurieÿ</del>
    <add type="morf synt">nekurie</add>
  </choice>)
  <lb/>Jßmintingi nog
  <choice>
    <del rend="ul">Uztekieima Saułes</del>
    <add type="lex synt">Saułe-tekia</add>
  </choice>
  <lb/>atajo Jeruʒaleń.
</div>
```

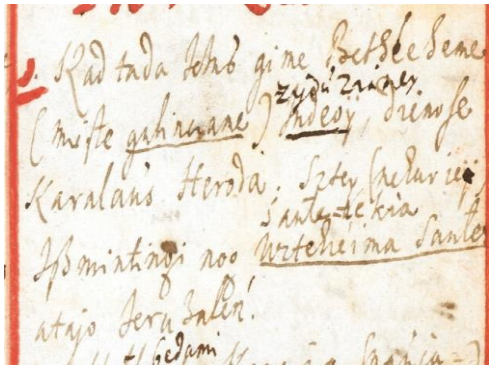

Figure 2. Annotating manuscript corrections

4.3. Editorial remarks: annotation of emendations

The transcribed text is published in two versions: 1) an exact transcription without interventions from the publisher (referred to as **original)**; 2) a text with scribal errors corrected and illegible sequences restored by the publishers (referred to as **editorial**). Restored text is marked in annotation as *emend*. Part of page 7r of the MS (Mt 5:13–17) was covered with ink and illegible; it is presented based on Otrębski and Kudzinowski's reconstruction. In other places, the publishers' corrections were limited to obvious errors – missing letters were restored, mistakes in verse numbering were corrected, and punctuation was unified. Publishers' interventions occur where a non-Lithuanian word is written in the text, an open space is left for the word to be inserted later, or a comparison with the original shows that a word was omitted. If there was a Dutch, Polish, or Latin word left in the translation of the NT, the missing Lithuanian equivalent was restored based on analogous places in Chylinski or, these lacking, in Samuel Bythner's NT (Königsberg, 1701). The restored version can then be found in the version with the publishers' corrections (editorial) and the last version, while the other versions contain a non-Lithuanian word written by Chylinski, or a mark that the word is missing (gap), III) **Both** the original transcript and the version corrected by the publisher are shown (on the webpage – *both*).

```
<div type="verse" n="13" note="reconstructed by Otrębski and Kudzinowski (1958)">
  <cb/>
  <lb/><head type="verse">13.</head>Jus eſte druska ʒiames. Kad druſka
  <lb/>pateroja ſuruma ſawo kuo ją apſudyſi?
  <lb/>
  <choice>
    <del>Niekam</del>
    <add type="m">
    Niekop</add>
  </choice>
  daugiaus nedera tykt oran
  <choice>
    <sic>est</sic>
    <corr    note="illegible">
    […]</corr>
  </choice>
  <lb/>ißmeſta    ir    kojomis
  ʒmoniu paminta.
</div>
```

Figure 3. An example of a reconstructed text

4.4. Annotation of lemmata and morphological properties

The main obstacle in automatically annotating Old Lithuanian text lies in the ambiguities of its spelling and different writing traditions. The grapheme *e* can represent the phonemes *e*, *ė*, *ę* or *ie*; the phoneme *ž* can be spelled *ß* or *sch*, depending on the region where the text is written, etc. Though there have been attempts to automatically standardize Old Lithuanian spelling,[4] no operational results have been achieved. Therefore, before using a lemmatizer developed for Standard Lithuanian like *Lemuoklis* or *Semantika.lt*, the text would have to be transposed into modern spelling. Using *Toolbox*, the first step of the annotation is also a standardization of the historical form into modern spelling. However, only the first occurrence of each form is entered into a dictionary file manually, each later occurrence is retrieved from there. Although this method has some disadvantages and sometimes creates artificial or ahistoric forms, it has proven itself to be the most time-saving approach and the advantages outweigh the disadvantages compared to directly annotating the historical form; besides, the standardized forms are only a tool within the project, not visible in the resulting data. This data is stored as a separate dictionary, as due to the different spelling traditions of the Old Lithuanian authors, reusing it from a different text creates difficulties. A second dictionary contains information on the morphological features and the lemma of the standardized form, and a third dictionary gives information on the part of speech.[5]

One disadvantage is that the infinitive is normalized as *-ti*, thus coinciding with the nominative plural of the past passive participle (e.g. *abdęgti* $6b_8$ ‘covered’), even though Chylinskis always uses the ‘short’ form of the infinitive in *-t* (*abdęgt* $265a_{52}$ ‘to

[4] https://drive.google.com/file/d/1cU1OajUFxOp_W2q8uPUId3NFFH_0phzI/view

[5] Due to a similar structure, these dictionaries could be provided by the projects *Altlitauische Kleintexte (ALKT)* and *Emergence of register*, both implemented by Berlin Humboldt University, where they have been trained on Johannes Bretke’s 1591 postil and several minor OLith texts. As for the dictionary file with standardized spelling, experience has shown it is more efficient to keep them separate for each author due to different spelling traditions and idiosyncrasies of the authors.

cover'), thus keeping both forms separate. On the other hand, the standardization allows for differentiation, e.g., between instr. sg. and gen. pl. of the *a*-stems (e.g., instr. *vaiku* 'with a child', gen. *vaikų* 'of the children', both spelled *wayku*) or 3rd pres. and the infinitive of *eiti* 'to go' and its derivates (OLith. *eiti*, StdLith. inf. *eiti*, 3rd pres. *eina*).

For the lemmatization, the first and most authoritative source is the Lithuanian Academic Dictionary, *Lietuvių kalbos žodynas* (LKŽ). However, not all historical variants are attested there; additional lexicographical sources include J. Palionis' historical glossary (2004), P. Skardžius' work on Slavic loanwords (1931), and the etymological-historical dictionary of Old Lithuanian, *Altlitauisches Wörterbuch* (ALEW). If a word is not attested in any of these, a modern form is transposed based on the patterns of modern Lithuanian word formation. During the trial run, we operated with the following parts of speech: *sm*. (masculine substantive), *sf*. (feminine substantive), *adj*. (adjective), *adv*. (adverb), *vb*. (verb), *prn*. (pronoun), *card*. (cardinal numeral), *ord*. (ordinal numeral), *num*. (other class of numeral), *part*. (particle).

As can be seen from the examples, different types of numerals are distinguished, while all pronouns are subsumed under one type. This is because types of numerals can easily be recognized by their form, while for the function of pronouns (demonstrative, interrogative, indefinite, relative), information is provided only by the wider context. A more detailed analysis can thus more easily be done in the edition and correction process.

In the final digital presentation, the abbreviations of the parts of speech and the morphological properties can be displayed in different ways, e.g., *d.* or *dat.* for 'dative' or even *naud*. for Lith. *naudininkas* 'dative'; *sf., subst. fem.* or *dkt. mot.* for 'feminine substantive.'

A problem with annotating historical stages of languages like Old Lithuanian is that in certain cases it is unclear whether a specific form is still a case form or has already been lexicalized as, e.g., an adverb or preposition. Cf., e.g., *tiesa*, historically the instrumental singular of *tiesa* 'truth', which can be used as an adverb in the sense of 'indeed': *Uźdawe jam tieſa ſzawejey kartumą* (41b$_{10}$ = Gen 49,23) 'Indeed the archers have done him bitterness'. Similarly, *slėpčioje* in *walgis nes jos słepcioy* (158b$_{68}$ = Dtn 28,57 'she will eat them secretly' and *slėpčiomis* in *Kodel iźbegey słepciomis* (25b$_{30}$ = Gen 31,27) 'why did you run away secretly?' are historically feminine forms of either a *u*-stem adjective *slėptus* or a *ja*-stem adjective *slėpčias* (loc.sg.f., and instr.pl.f. respectively). In Chylinski's text, however, they have become lexicalized in adverbial use. Such forms can be marked both as lexical adverbs and as case forms of the corresponding substantive or adjective.

Another problem lies within the orthographic peculiarities of the Old Lithuanian authors. As can be seen, e.g., from *buwo łabay didy* (49b$_{62}$ = Ex 9,24) 'they were very big' alongside *buwo łabay galintyngeys* (43a$_{38}$ = Ex 1,7) 'they were very powerful', Chylinski uses both the nominative (StdLith. nom.pl.m. *didi* 'big') and the instrumental (StdLith. instr.pl.m. *galintingais* 'powerful, mighty') case in predicative position. However, the nom.sg. and instr.sg. of *o*- and *ė*-stem substantives and adjectives are identical in his spelling, both ending in *-a* resp. *-e*. Thus, telling these cases apart is not possible in constructions like *Źiame buwo puſta ir tuſzcza* (1b$_{57}$ = Gen 1,2) 'the earth was empty and barren'. After completing both the New and the Old Testament, the data can be reviewed; if a conditioning factor is found, the morphological annotation will be corrected accordingly.

Word	ne	Iszeytu	te	Kalba	anogimet	
Std	ne	Iseitu	te	Kalba	anuomet	Gi
Morf	-	3.cnd.	-	3.prs.	-	-
Lex	ne	Iseiti	te	kalbeti	anuomet	Gi
PoSp	neg.	vb.	part.	vb.	adv.	ptcl.

Figure 4. An example of lemmatization and morphological annotation

4.5. The annotation process

Previous work on annotating the text of Chylinski's New Testament had shown that an iterative process performed on manageable parts of the text is the most efficient way of working on large historical texts.

We started with the overall annotation of the structural elements, which could largely be done automatically. Then portions of c. 20000 words (e.g., a gospel or a set of epistles) were chosen and each team member worked with well-defined guidelines decribing the categories of manuscript corrections and editorial emendations. Unclear phenomena where marked and discussed during periodic team meetings. Annotation of lemmata and morphological properties was not part of this project, but a partly machine-aided way of creating these data was tested, now again on the whole text. The results were promising, but in line with other projects involving Old Lithuanian texts it was decided to adopt their experience with Toolbox (see 4.4).

The annotation of early modern texts shows the limitations of automatic procedures, as spelling variation and divergences in grammatical features between historical and modern language cannot be handled efficiently. One way to do so would be to use a kind of pre-processor assigning a modern form to each word. Instead, a 'learning' database was used providing previously analysed words as suggestions during the annotation process. Recent attempts at applying AI tools might open up new horizons for analysing historical text corpora.

5. Database and interface design

The challenges posed by structuring a text database and presenting the search possibilities and the results are interrelated design tasks, intended to create and to operate queries efficiently.

5.1. Database structure

The XML data are transformed into a relational database because of the necessity of additional tables in order to create dictionaries and indexes, both for performance reasons and for the sake of research flexibility.

The main table contains all **words** taken from the XML data. A second table represents all **lexemes** and their properties, created by condensing the data of the word table. Both tables with their indexes provide the basis for all research questions on word level, looking for morphological properties or semantic fields. Further tables are added to extract syntactic features derived from linear word order, intended to allow the study of phrase structure as a basis for sentence formation.

Each **word** entry provides the word and its standardized spelling, the base form with its affixes, the lemma, the editorial level, the position within the physical and the biblical structure, and looks as in the following abridged example:

Id	**Phys**	**Bibl**	**Word**	**Pref**	**Stem**	**Suff**	**Lemma**	**Edit**	**...**
21643	35_02_02_31	02_40_04_09	Kaypo	-	Kaypo	-	kaipo	-	...
21644	35_02_02_31	02_40_04_10	neturytetu	ne	turyte	gu	turęti	origsp	...
21645	35_02_02_31	02_40_04_11	neturitegu	ne	turite	gu	turęti	corrsp	...

Figure 5. Word table entries

A **lexeme** entry contains the lemma, part of speech with a subcategory, and further inherent properties like gender and (lexical) number; again, we give an abridged example:

Id	**Lemma**	**PoSp**	**Sub**	**Gen**	**Num**	**...**	
1004	durys	Noun	comm	f	p	...	*plurale tantum*
1623	kaipo	Advb	-	-	-	...	
4054	Steponas	Noun	prop	m	-	...	*proper noun*
4487	turėti	Verb	-	-	-	...	

Figure 6. Lexeme table entries

Apart from the indexes on these tables some additional tables are created from these data in order to accelerate more complex searches like linear sentence structures and word-field investigation.

5.2 Interface design

The interface design follows the needs of the user, presented either as a simple search interface looking for a word or a lemma, or as a more sophisticated search combining sets of elements and properties.

Textual databases collect different sources, in most cases various individual texts selected according to the focus of a database project, based on criteria related to language, age, form, topic, and other. In some cases the main research interest focuses on the comparison of different versions of the same text, translated and reproduced in different languages and re-edited and revised for several centuries. The ChB project belongs to the latter type and is intended to be the basis of a Lithuanian corpus providing Bible texts and Bible-related sources like lectionaries, postils, and other texts. The texts are set alongside their sources and/or parallel texts, in this case the Dutch *Statenvertaling* and the Polish *Biblia Gdańska*.

The **basic search** interface works like traditional search utilities. The occurrences of a lemma or a word form are listed and can be presented in their context and with their parallel texts taken from translations or other versions.

The **extended search** offers the user the possibility to select sets and to define sequences.

- A search covers, by default, all texts in the database, but it can also be restricted to a **set of texts**, which could be a list of selected texts or a self-defined set of text parts, for example in order to understand the differences between prose and poetry.
- Another feature will be the definition of **lexical sets**. Users are enabled to define and edit sets of lexemes or word forms, e.g. in order to carve out semantic fields.
- This procedure can also be applied to define **sets of grammatical properties** which provide the possibility to investigate case usage or translation principles.
- The study of collocations is extended to **sequences of elements** defined by lexical and/or grammatical properties enabling the user to look for the position of adjectives in the phrase or of an adverb in a clause.

The main idea behind these more sophisticated search options is to allow scholars to extract the lexical or grammatical data they are interested in, without the necessity to deploy query languages like SQL (en.wikipedia.org/wiki/SQL_syntax) or CQL (www.loc.gov/standards/sru/cql). This idea corresponds to the aim of keeping the accessibility on a level which affords non-professional users an in-depth but also straightforward look into texts belonging to the Lithuanian cultural heritage.

6. Concluding remarks and an outlook toward a Lithuanian Bible Corpus

In creating the Chylinski Bible database, international standards are observed, and the creators hope that it will one day become part of a diachronic Lithuanian Bible corpus. The annotation system created for the database is suitable for the publication of both historical manuscripts and printed texts. With some improvements it could also be used for the publication of the first Lithuanian Bible translation, the manuscript Bretke Bible, as well as of more recent Lithuanian Bible translations extant in print. The Lithuanian Bible database could comprise not only complete Bible translations but also pericopes printed as parts of Lithuanian postils as well as gospel and epistle fragments used in the liturgy over the centuries. Such a corpus should comprise both facsimiles of Bible books and diplomatic transcriptions, it should be morphologically annotated and offer lists of word forms arranged according to different criteria. It should enable the user to compare Bible texts among themselves and ideally also with the translation sources.

References

ALEW – *Altlitauisches etymologisches Wörterbuch* [Old Lithuanian Etymological Dictionary], www.alew.hu-berlin.de

Bouma, G., Coussé, E., Dijkstra, T., Van der Sijs, N. (2020). The EDGeS Diachronic Bible Corpus, *in Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 5232–5239, Marseille, 11–16 May 2020. www.neerlandistiek.nl/2021/06/nieuw-edges-een-diachroon-bijbelcorpus

Christodouloupoulos, C., Steedman, M. (2015). A massively parallel corpus: the Bible in 100 languages, *Language Resources and Evaluation* **49**(2), 375–395. DOI 10.1007/s10579-014-9287-y).

Čapaitė, R. (2019). Chylinski's New Testament in the context of the Latin cursive, *in* (Kavaliūnaitė, 2019), clxiij–clxxj.

Gries, S., Berez, A. (2017). Linguistic Annotation in/for Corpus Linguistics, *in* Ide, N., Pustejovsky, J. (eds.). *Handbook of Linguistic Annotation*, 379–409. DOI: 10.1007/978-94-024-0881-2_15.

Jocher, A. (1840–1857). *Obraz Bibliograficzno-historyczny Literatury i nauk w Polsce* [Bibliographical and historical survey of literature and the sciences in Poland], Zawadzki, Wilno.

Jurgutis, V., Žukas, V. (1963). J. Plioterio darbas „Trumpa žinia apei tą iszdawima lietuwiszkos biblijos Londone" [J. Plater's "Brief report on the publication of a Lithuanian Bible in London"], in *Lietuvos TSR Aukštųjų mokyklų mokslo darbai, Bibliotekininkystės ir bibliografijos klausimai* **3**, 193–203.

Kavaliūnaitė, G. (ed.) (2008). *Samuelio Boguslavo Chylinskio Biblija. Senasis Testamentas* **1**: *Lietuviško vertimo ir olandiško originalo faksimilės = Biblia Lithuanica Samueli Boguslai Chylinski* **1***: Vetus Testamentum Lithuanicâ Lingvâ donatum a Samuelo Boguslao Chylinski. Unâ cum texto belgico*, LKI, Vilnius.

Kavaliūnaitė, G. (ed.) (2019). *Samuelio Boguslavo Chylinskio Biblija* **2**: *Naujasis Testamentas Viešpaties mūsų Jėzaus Kristaus lietuvių kalba duotas Samuelio Boguslavo Chylinskio = Biblia Lithuanica Samueli Boguslai Chylinski. Tomus* **2***: Novum Testamentum Domini Nostri Jesu Christi Lithvanicâ Linguâ donatum a Samuelo Boguslao Chylinski*, Vilnius: Vilniaus universitetas, p. 572.

Kot, S. (1958). Chylinski's Lithuanian Bible; Origin and Historical Background, *in* Kudzinowski, Cz., Otrębski, J. (eds.) (1958) *Biblia litewska Chylińskiego. Nowy Testament* **2**, Tekst, Poznań: Zakład Narodowy im. Ossolińskich.

Kudzinowski, Cz. (1964). *Biblia litewska Chylińskiego. Nowy Testament* **3**, Indeks [Chylinski's Lithuanian Bible. The New Testament 3. Index], Państwowe Wydawnictwo Naukowe, Poznań.

Kudzinowski, Cz. (1984) *Biblia litewska Chylińskiego. Nowy Testament* **1**, Fotokopie [Chylinski's Lithuanian Bible. The New Testament 1. Photographs], Wydawnictwo Naukowe UAM, Poznań.

Kudzinowski, Cz., Otrębski, J. (eds.) (1958) *Biblia litewska Chylińskiego. Nowy Testament* **2**, Tekst, [Chylinski's Lithuanian Bible. The New Testament 2. Text]. Ossolineum, Poznań.

Palionis, J. (2004). *XVI-XVII a. Lietuviškų raštų atrankinis žodynas* [A Selective Glossary of 16th and 17th-c. Lithuanian Texts], Mokslų ir enciklopedijų leidybos institutas, Vilnius.

Quandt, J. (1735). Vorrede [Preface], in: *Biblia, Tai esti: Wissas Szentas Rasztas* [...], Hartung, Königsberg.

Resnik, P., Olsen, M. B., Diab, M. (1999). The Bible as a parallel corpus: Annotating the 'Book of 2000 tongues', *Computers and the Humanities*, **33(1)**, 129–153. doi:10.1023/A:1001798929185.

Skardžius, P. (1931). *Die slavischen Lehnwörter im Altlitauischen* [Slavic Loanwords in Old Lithuanian], in *Tauta ir žodis* **7**, „Spindulio" spaustuvė, Kaunas, 4–252.

Stankiewicz, M. (1889). *Bibliografia litewska od 1547 do 1701 r.* [Lithuanian Bibliography for the Years 1547–1701], Gebethnera i Ska, Kraków.

Wolter, E. (1887). Ob ètnografičeskoj poězdkě po Litvě i Žmudi lětom 1887 goda. Privat"-docenta È. Vol'tera [Report on an ethnographical journey to Lithuania and Samogitia in the summer of 1887 by Privatdozent E. Wolter]. Supplement to Vol. 56 of the Proceedings of the Imperial Academy of Sciences, No 5, St Petersburg, 71–102.

Received November 30, 2024, accepted December 5, 2924