

An aerial photograph of a historic city, likely Vilnius, Lithuania. The image shows a dense urban landscape with numerous buildings, including a prominent white church with a tall, ornate tower in the foreground. The city is surrounded by greenery and hills in the background. The sky is overcast with soft, grey clouds. The text is overlaid on the upper half of the image, which is slightly faded to ensure readability.

Senujų baltiškų šaltinių skaitmeninė prieiga
Digitālās pieejas senajiem baltu valodu avotiem
Digital approaches to Old Baltic linguistic monuments

EVERITA ANDRONOVA

Stockholms universitet
everita.andronova@balt.su.se

Challenges and some observations about part-of-speech tagging for early Latvian texts

Part-of-speech (POS) tagging for historical texts presents a well-known challenge in natural language processing (NLP). Common difficulties include historical spelling, a high degree of morphological variability, and the scarcity of linguistically annotated data suitable for training POS tagger.

The current POS tagger for modern Latvian is based on the Stanford POS tagger and, as reported, the accuracy is higher than 97% (cf. Paikens, Rituma, Pretklaniņa 2013). Znotiņš and Bārzdīņš (2020) presents the LVBERT model and claims the accuracy of POS tagger reaching 98.1%. However, no tagger for early Latvian texts is available. There is at least one attempt known to provide morphological annotation for early Latvian texts using SIL Toolbox ('A grammatically annotated corpus of the old Latvian Postil of Georg Mancelius') (Goymann, 2023).

This presentation shares preliminary observations from applying two existing tools for modern Latvian: a POS tagger via <https://nlp.ailab.lv/> and Stanza, an open-source Python toolkit (Qi, Zhang et al 2020), to texts by G.Mancelius from the Corpus of Early Latvian texts (<https://senie.korpuss.lv/>).

Although preprocessing is designed to be language-independent, in practice it is often highly source-specific. Tokenization of text in original spelling with Stanza reveals significant difficulties. Segmentations such as *Tas gir tee fcham tees .* and *Tas gir tee fcham tees .* demonstrate inconsistent handling of hyphenation. Ligatures, too, present challenges, e.g., *Ko mahzais tu no teems Wahrdeems no to Ghräko pammef fchanu ?* is not consistently tokenised.

Both taggers produce quite similar results. However, in texts with original spelling all special Unicode characters and non-standard wordforms tend to be unrecognised (marked as 'X'). Additionally, capitalised common nouns in Stanza are treated as proper nouns (see *Parradu* 'debt' tagged as PROPN in Table 1).

Manc1631_Cat: Stanza	Manc1631_Cat: nlp.ailab.lv																																																
Muhrf<u>u</u> X deenif<u>chku</u> X Mais X dohd X mums PRON fchodeen NOUN / PUNCT Vnd X peedohd X mums PRON muhrf<u>u</u> VERB Parradu PROPON / PUNCT	<table><tr><td>1</td><td>Muhlu</td><td>Muhlu</td><td>NOUN</td></tr><tr><td>2</td><td>deenichku</td><td>deenichka</td><td>NOUN</td></tr><tr><td>3</td><td>Mais</td><td>Mais</td><td>X</td></tr><tr><td>4</td><td>dohd</td><td>dohd</td><td>X</td></tr><tr><td>5</td><td>mums</td><td>mums</td><td>X</td></tr><tr><td>6</td><td>fchodeen</td><td>fchodeen</td><td>X</td></tr><tr><td>7</td><td>/</td><td>/</td><td>PUNCT</td></tr><tr><td>8</td><td>Vnd</td><td>Vnd</td><td>X</td></tr><tr><td>9</td><td>peedohd</td><td>peedohd</td><td>X</td></tr><tr><td>10</td><td>mums</td><td>mās</td><td>PRON</td></tr><tr><td>11</td><td>muhlu</td><td>muhlu</td><td>NOUN</td></tr><tr><td>12</td><td>Parradu</td><td>paradāt</td><td>VERB</td></tr></table>	1	Muhlu	Muhlu	NOUN	2	deenichku	deenichka	NOUN	3	Mais	Mais	X	4	dohd	dohd	X	5	mums	mums	X	6	fchodeen	fchodeen	X	7	/	/	PUNCT	8	Vnd	Vnd	X	9	peedohd	peedohd	X	10	mums	mās	PRON	11	muhlu	muhlu	NOUN	12	Parradu	paradāt	VERB
1	Muhlu	Muhlu	NOUN																																														
2	deenichku	deenichka	NOUN																																														
3	Mais	Mais	X																																														
4	dohd	dohd	X																																														
5	mums	mums	X																																														
6	fchodeen	fchodeen	X																																														
7	/	/	PUNCT																																														
8	Vnd	Vnd	X																																														
9	peedohd	peedohd	X																																														
10	mums	mās	PRON																																														
11	muhlu	muhlu	NOUN																																														
12	Parradu	paradāt	VERB																																														

Table 1. Example from Catechism by Mancelius (1631) in original (historical) spelling tagged by Stanza and nlp.ailab.lv.

To improve tagging accuracy, it is common practice to apply spelling normalization prior to using a POS tagger trained on modern data (cf. Bollmann 2013). Rule-based normalisation of the historical spelling of early Latvian texts has recently been undertaken (e.g., Andronova, Frīdenberga et al. 2024). Consequently, both original and normalised spelling versions of G. Mancelius’ texts were tagged. Normalised spelling visibly improved the results; however, non-standard forms tend to be treated as foreign words in `nlp.ailab.lv` tagger:

Manc1637_Sal: Stanza		Manc1637_Sal: nlp.ailab.lv				
13.	ADJ	13.	13.	ADJ	xo	NumType=Ord
kas	PRON	kas	kas	PRON	pr000nn	Case=Nom Pro
labu	ADJ	labu	labs	ADJ	afmsanp	Case=Acc Deg
ar	ADP	ar	ar	ADP	spsa	—
ļāunu	ADJ	ļāunu	ļāuns	ADJ	afmsanp	Case=Acc Deg
atmaksa	NOUN	atmaksa	atmaksa	NOUN	ncfsn4	Case=Nom Ger
/	PUNCT	/	/	PUNCT	zx	—
no	ADP	no	no	ADP	sppd	—
tā	DET	tā	tas	PRON	pd3msgn	Case=Gen Ger
namu	NOUN	namu	nams	NOUN	ncmpg1	Case=Gen Ger
ļāun	NOUN	ļāun	ļāun	X	xf	Foreign=Yes
ne	CCONJ					
atstās	VERB					

Table 2. Example from *Die Sprüche Salomonis* by Mancelius (1637) in normalised historical spelling tagged by Stanza and nlp.ailab.lv.

In order to improve POS tagging for historical Latvian, future work will involve evaluating the applicability of the modern Latvian tagset, identifying unrecognised or recognised incorrectly worforms in early texts, incorporating them into the tagset, and using the results to develop new training data. At present, the necessary amount of annotated historical data for Latvian remains unknown. By comparison, Bollmann reports that in German, ‘[u]sing only 250 manually normalized tokens as training data, the tagging accuracy of a manuscript from the 15th century can be raised from 28.65% to 74.89%’ (Bollmann 2013:11).

References

- Andronova E., Frīdenberga A., Pretkalniņa L., Siliņa-Piņķe R., Skrūzmane E., Trumpa A., Vanags P. (2024). New Possibilities for Exploring Early Latvian Texts: Switching to the *NoSketchEngine*. *Baltic Journal of Modern Computing*, 12(4), pp. 548–559. DOI: <https://doi.org/10.22364/bjmc.2024.12.4.16>.
- Bollmann, M. (2013). POS Tagging for Historical Texts with Sparse Training Data. *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, pp. 11–18, Sofia, Bulgaria, August 8–9, 2013. URL: <https://aclanthology.org/W13-2302.pdf>.
- Goyman, F. (2023). A grammatically annotated corpus of the old Latvian Postil of Georg Mancelius. URL: <https://zenodo.org/records/7890894>.
- Paikens P., Rituma L., Pretkalniņa L. (2013). Morphological analysis with limited resources: Latvian example. *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA)*. Linköping Electronic Conference Proceedings #85, 2013, pp. 267–277. URL: <http://stp.lingfil.uu.se/nodalida/2013/pdf/NODALIDA24.pdf>.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Association for Computational Linguistics (ACL) System Demonstrations*. 2020. URL: <https://nlp.stanford.edu/pubs/qi2020stanza.pdf>.
- Znotiņš A., Bārzdīņš G. (2020). LVBERT: Transformer-Based Model for Latvian Language Understanding. *Human Language Technologies—The Baltic Perspective*. IOS Press, vol.328, pp.111–115. URL: <https://ebooks.iospress.nl/volumearticle/55531>. DOI: 10.3233/FAIA200610.

LOIČ BOIZOU

Universität zu Köln
lboizou@uni-koeln.de

JOLANTA GELUMBECKAITĖ

Goethe-Universität Frankfurt
gelumbeckaite@em.uni-frankfurt.de

MORTIMER DRACH

Goethe-Universität Frankfurt
Humboldt-Universität zu Berlin
drach@posteo.de

ØYVIND EIDE

Universität zu Köln
oeide@uni-koeln.de

MAXIM IONOV

Universidad Zaragoza

PASCALE BOISVERT

Universität zu Köln

PosTiMe – The Postil Time Machine

Introduction

The *Postil Time Machine* (PosTiMe, 2021–2025) is a novel research project on knowledge transfer in 16th-century Europe; specifically, in Lithuania Minor. Anchored on the Old Lithuanian Lutheran postils – the *Wolfenbüttel Postil* (WP1573/1574) and the *Bretkūnas Postil* (BP1591) – it aims to represent the literary “world” of these two texts. This includes their relations with translation sources and models, sources cited as authorities, the Bible, pre-existing and derived Lithuanian Biblical translations (Bretkūnas’s own 1580 translation of the Psalter and Rehsa’s 1625 print edition, Willent’s 1579 book of pericopes and Sengstock’s revised 1612 edition) and of course with each other as collections of sermons for the the liturgical year. The main objective of PosTiMe is the genetic and historical- linguistic study of these Lithuanian texts and their presentation in a digital *time machine*-type publication (cf. Kaplan 2013).

Edition

This interdisciplinary linguistic-philological and technological project consists of the following tasks: (1) the identification of the construction principles and translation strategies of the Lithuanian texts, (2) the study of the textual relations of the postils, (3) the detection of intra- and intertextual references and the alignment with the so-identified originals using machine translation methods (which necessitates the creation of a language technologies stack for Old Lithuanian), (4) the linguistic-historical interpretation of their contents in the form of a linguistically deeply annotated reference corpus, (5) the modelling as a graph and the development of a Linked Open Data interface in RDFa based on and connected to TEI documents (Eide 2015; Tittel et al. 2018; Chiarcos & Ionov 2018) as well as the representation of these structures as interactive visualizations, and (6) the creation of a platform to make the research results searchable, traversable and generally accessible. The scholarly digital edition of the postils is developed in accordance with state of the art best practice, following the TEI¹ guidelines with a frontend system developed in Vue.js².

Automatic methods

The project enabled the testing of automatic methods for detecting and aligning quotations from other language sources used in Lithuanian postils. The evaluation was performed on WP’s longest sermon (Nativity of Christ) and the Vulgate. Two source-mining approaches were tested: (1) a traditional lexicon-based method (Allen

1 <https://tei-c.org/>

2 <https://vuejs.org/>

et al. 2010, Ganascia et al. 2014, Büchler et al. 2014), and (2) a slightly more recent LLM-based approach (Mikolov et al. 2013, Lample et al. 2017, Feng et al. 2020).

The lexicon-based approach consisted of two steps: (1) transposing the Lithuanian text by replacing each content-word lemma with its Latin translations, and (2) comparing 6-grams from the transposed Lithuanian text with those from the Vulgate. The comparison used a similarity score based on the number of shared words between the two 6-grams, adjusted by the inverse frequency of the words in the corpus to increase the weight of rarer words.

In the LLM-based approach, the only usable results were obtained using Google's LaBSE model via SBERT embeddings. Initial results were poor, but two enhancements led to a major improvement. The most significant was adjusting the segment size to better match the quotation size, achieved by splitting and recombining sentences in WP and verses in the Vulgate. The second enhancement involved lightly fine-tuning the model with the Vulgate and a modern Lithuanian Bible, which led to a concentration of correctly detected segments at the top of the similarity rankings.

The evaluation showed that, with these two improvements, the LLM-based approach significantly outperformed the lexicon-based one. While the lexicon-based method identified around one tenth of the Latin quotations, the fine-tuned LaBSE model detected about 60%, with over 40% of the total receiving the highest similarity scores.

Literature

- Allen, Timothy, Cooney, Charles, Douard, Stéphane, Horton, Russell, Morrissey, Robert, Olsen, Mark, Roe, Glenn, Voyer, Robert. 2010. Plundering philosophers: identifying sources of the Encyclopédie. *Journal of the Association for History and Computing* 13(1).
- Büchler, Marco, Mueller, Martin, Burns, Philip R., Franzini, Emily, Franzini, Greta. 2014. Towards a Historical Text Re-use Detection. In: Biemann, Chris, Mehler, Alexander (eds) *Text mining: From Ontology Learning to Automated Text Processing Applications*. Springer, 221-238.
- Chiarcos, Christian, Maxim Ionov. 2018. *Linking the TEI: Approaches, Limitations, Use Cases*. Paper presented at DH2019, Utrecht, Nederlande.
- Eide, Øyvind. 2015. Ontologies, data modelling, and TEI. *Journal of the Text Encoding Initiative*, no. 8 (2015).
- Feng, Fangxiaoyu, Yang, Yinfei, Cer, Daniel, Arivazhagan, Naveen, Wang, Wei. 2020. Language-agnostic BERT Sentence Embedding. *ArXiv*. <https://arxiv.org/abs/2007.01852>
- Ganascia, Jean-Gabriel, Glaudes, Pierre, Del Lungo, Andrea. 2014. Automatic Detection of Reuses and Citations in Literary Texts. *Literary and Linguistic Computing*, 29 (3), pp.412-421. 10.1093/lc/fqu020. Hal-00977310v1
- Gelumbeckaitė, Jolanta. 2017. Predigtkultur in Litauen: Corpus der altilitauischen Postillen. In: Assel, Heinrich, Johann AnselmSteiger, Axel E. Walter (Hg.). *Reformatio Baltica: Kulturwirkungen der Reformation in den Metropolen des Ostseeraums*. Berlin, Boston: De Gruyter, 573–585.
- Kaplan, Frédéric 2013. "How to build an information time machine". Presented at TEDxCaFoscariU, Venice, Italy.
- Lample, Guillaume, Conneau, Alexis, Denoyer, Ludovic, Ranzato, Marc'Aurelio. 2017. Unsupervised Machine Translation With Monolingual Data Only. *ArXiv*. <https://arxiv.org/abs/1711.00043>
- Mikolov, Tomas, Le, Quoc V., Sutskever, Ilya. 2013. Exploiting Similarities among Languages for Machine Translation. *ArXiv*. <https://arxiv.org/abs/1309.4168>
- Tittel, Sabine, Helena Bermúdez-Sabel, Christian Chiarcos. 2018. "Using RDFa to link text and dictionary data for Medieval French." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 7–12. Miyazaki, Japan.

Databases, Models and Corpora

- CorDon, <https://titus.fkidg1.uni-frankfurt.de/cordon/menu/eng/start.html>
- LaBSE, <https://huggingface.co/sentence-transformers/LaBSE>
- Senieji raštai, <http://seniejirastai.lki.lt/home.php>
- SLIEKKAS, <https://titus.uni-frankfurt.de/sliekkas/index.html>
- TITUS, <https://titus.uni-frankfurt.de/indexd.htm>

PIETRO U. DINI

Università di Pisa
pietro.dini@unipi.it

SILVIA PICCINI

Istituto di Linguistica Computazionale "Antonio Zampolli"
silvia.piccini@ilc.cnr.it

ADRIANO CERRI

Università di Pisa
adriano.cerri@unipi.it

Old words for a new world: Lexicon and ontology of Christian terminology in early Baltic catechisms

The Christianization of the Baltic lands between the 13th and 16th centuries triggered profound linguistic and cultural transformations among the indigenous populations, who integrated Christian elements into their traditional pagan worldview (cf., *inter alia*, Vêlius, 1996–2005; Blomkvist 2005; Young 2022). This paper presents a research project funded by the Italian Ministry for Universities and Research (MUR) entitled *Old Words for a New World: Translating Christianity to Baltic Pagans*, which explores this transitional period through the lens of language (cf. Castelli et al.).

Focusing on a selected corpus of five 16th-century catechisms—representing three languages (Old Prussian, Old Lithuanian, and Old Latvian) and two Christian confessions (Catholicism and Protestantism)—this pilot project investigates how key Christian concepts were rendered and understood within Baltic linguistic frameworks. A summary of the corpus is provided in Table 1. The project comprises three interconnected research lines, drawing on historical linguistics, comparative religion, computational lexicography, and knowledge representation:

- **Translation Strategies and Lexical Analysis:** Extraction and analysis of lemmas, collocations, and expressions referring to core Christian concepts. The study examines both literal (word-for-word) translation practices and instances of creative adaptation, comparing strategies across the corpus and identifying source models (cf. Dini 2014).
- **Theological and Cultural Interpretation:** By applying typologies developed in the historical study of religions (cf. e.g. Salzmann et al. 2016; Dal Covolo, Sfamini Gasparro 2021), the project aims to reconstruct a Baltic-specific conceptual landscape of Christian theology as it emerged in the 16th–17th centuries.
- **Computational Lexicography and Ontological Modelling:** Development of a diachronic, multilingual lexicon of Baltic religious terminology structured according to Mel'čuk's Explanatory and Combinatorial Lexicology (ECD), cf. Mel'čuk et al. (1995); Mel'čuk (2006). The project also involves modelling a domain ontology (Guarino et al. 2009) of Christian key concepts to enable semantically rich access to the corpus via advanced querying mechanisms.

The presentation highlights the methodological innovation of applying the ECD model to ancient languages for the first time. Selected examples from the corpus will illustrate how religious terms are formalized both lexically and ontologically enabling new forms of textual and conceptual analysis compared to previous traditional studies on the Baltic religious lexicon (e.g., Borowska 1957; Polomé 1995). The digital resource—comprising the multilingual lexicon and ontology—is currently under development and is designed following Semantic Web technologies (Berners-Lee et al. 2001) and the FAIR principles (Wilkinson et al. 2016), ensuring data that are Findable, Accessible,

Interoperable, and Reusable. In accordance with the principles of open science, these technologies enable the resource to be shared, integrated, and reused across diverse research contexts and multidisciplinary perspectives by the scholarly community.

Language	Translator(s), Title	Year	Confession	Nr. of Tokens
Old Prussian	A. Will – P. Megott, <i>Enchiridion</i>	1561	Lutheran	6.712
Old Lithuanian	B. Vilentas, <i>Enchiridion</i>	1579	Lutheran	8.987
	M. Daukša, <i>Kathechismas arba mokslas kiekvienam krikščioni priwalvs</i>	1595	Catholic	13.052
Old Latvian	Anon., <i>Catechismvs Catholicorum</i>	1585	Catholic	4.869
	Rivius et al., <i>Enchiridion</i>	1586	Lutheran	7.513

Table 1. The Selected Corpus of Early Baltic Catechisms

References

- Berners-Lee, T., Hendler, J., and Lassila O., 2001. The Semantic Web. *Scientific American*, Vol. 284, No. 5, 34–43.
- Blomkvist, N., 2005. *The Discovery of the Baltic. The Reception of a Catholic World-System in the European North (AD 1075-1225)*. Leiden / Boston: Brill.
- Borowska, N., 1957. Wpływy słowiańskie na litewską terminologię kościelną na podstawie Dictionarium Szyrwida. *Studia z Filologii Polskiej i Słowiańskiej*, Kraków / Warszawa, 2, 320–365.
- Castelli, E., Cerri, A., Dini, P. U., Monaca, M., Mormino, M., Piccini, S., forthcoming. Early Baltic Catechisms in the Transition from Paganism to Christianity: Showcasing an Interdisciplinary Research Project. In D. Bronner, E. Parina (eds.), *Contents of Faith in Transfer*, J.B. Metzler.
- dal Covolo, E., Sfameni Gasparro, G. (eds.) 2021. *Pagani e cristiani: conflitto, confronto, dialogo. Le trasformazioni di un modello storiografico*. Città del Vaticano: Libreria Editrice Vaticana.
- Dini, P.U. 2014. *Ins undevdfche gebracht. Sprachgebrauch und Übersetzungs-verfahren im alt-preußischen Kleinen Katechismus*. Akademie der Wissenschaft zu Göttingen, Berlin: de Gruyter.
- Guarino, N., Oberle, D., Staab, S., 2009. What is an ontology? In S. Staab, R. Studer (eds.), *Handbook on Ontologies*, Springer, 1–17.
- Mel'čuk, I. A., Clas, A., Polguère, A., 1995. *Introduction à la lexicologie explicative et combinatoire*. Bruxelles: Duculot.
- Mel'čuk, I. A., 2006. Explanatory Combinatorial Dictionary. In G. Sica (ed.), *Open Problems in Linguistics and Lexicography*, Monza: Polimetrica Publisher, 225–355.
- Polomé, E., 1995. Notes on the Baltic religious vocabulary. *Linguistica Baltica*, 4, 34–40.
- Salzman, M., Ságghy, M., Lizzi Testa, R. (eds.), 2016. *Pagans and Christians in Late Antique Rome. Conflict, Competition, and Coexistence in the Fourth Century*. Cambridge: Cambridge University Press.
- Vėlius, N., 1996–2005. *Baltų religijos ir mitologijos šaltiniai*, 4 vols. Vilnius: Mokslo ir enciklopedijų leidybos institutas.
- Wilkinson, M. D., Dumontier, M., IJsbrand J. A., Appleton, G., Axton, M., Baak, A., Blomberg N. et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 1–9.
- Young, F. (ed.), 2022. *Pagans in the Early Modern Baltic: Sixteenth-Century Ethnographic Accounts of Baltic Paganism*. Leeds: Arc Humanities Press.

ANNA HELENE FEULNER

Humboldt-Universität zu Berlin
anna.helene.feulner@hu-berlin.de

HENRIK HORNECKER

Humboldt-Universität zu Berlin
henrik.hornecker.1@hu-berlin.de

Das Altlitauische Quellenverzeichnis (ALQ): aktueller Stand und digitale Zukunft

Das ALQ als geplantes Gesamtverzeichnis altlitauischer Quellen ist seit seinen Anfängen vor über 20 Jahren stetig angewachsen und hat aus eigener und fremder Forschung neue Beiträge zur bibliographischen Erschließung des Altlitauischen aufgenommen. Mit den Publikationsvorbereitungen haben wir bisher gewartet, da wir den letztgültigen Informationsstand des DFG-geförderten Editionsprojekt sämtlicher altlitauischer Kleintexte (ALKT) an der Humboldt-Universität zu Berlin noch einspeisen wollen, an dem wir in unterschiedlicher Funktion beteiligt sind bzw. waren.

Mittlerweile steht das ALKT kurz vor dem Abschluss. Im Rahmen dieses Projekts gelang es dem Team nicht nur, eine erfreuliche Zahl an verloren geglaubten Sprachdenkmälern aufzuspüren; wir verzeichnen auch etliche Neufunde von Quellen, Textexemplaren und Begleitmaterialien, die unser Wissen vom Altlitauischen des 16. und 17. Jahrhunderts erweitert und vertieft haben.

Als im Rahmen der Vorbereitungen für das *Altlitauische etymologische Wörterbuch* (ALEW) ab 2004 in Listenform die ersten Versionen des ALQ entstanden, war nur eine herkömmliche Buchpublikation geplant. Die 'Digital Humanities' – als Begriff wie als Forschungsansatz – waren noch nicht in aller Munde. Die rasanten Fortschritte des Informationszeitalters ermöglichen es uns nun jedoch einerseits, der Tatsache Rechnung zu tragen, dass jedes "Gesamtverzeichnis" altlitauischer Textzeugen immer nur einen momentanen Stand wiedergeben kann: Neue Funde werden gemacht und weitere Quellen erschlossen werden. Andererseits können wir mit dem ALQ von den neuen Möglichkeiten, Informationen zu organisieren und zu publizieren, profitieren.

Die nach wie vor avisierte gedruckte Buchform wird den Wissensstand des ALQ bei Projektende sichern, was wegen der Vergänglichkeit im digitalen Raum zwingend erforderlich ist. Um nach Abschluss der Buchpublikation den erwartbaren weiteren Wissenszuwachs einarbeiten zu können und um einen besseren Zugriff auf unsere bibliographischen Daten zu ermöglichen, planen wir nun zusätzlich zum gedruckten ALQ eine erweiterbare digitale Komponente. Wir werden die Ergebnisse des ALQ-Projekts in einer öffentlich zugänglichen Zotero-Gruppe organisieren, deren Inhalte durch uns als Bearbeiter aktualisiert werden können. Zotero ist ein freies, weit verbreitetes und bedienerfreundliches Instrument für die digitale Bibliographieverwaltung, das wir unseren Zwecken sehr gut anpassen können.

In unserem Vortrag werden wir erläutern, wie wir diese Zotero-Gruppe organisieren, um ein digitales Äquivalent zur Buchpublikation zu schaffen, das – soweit zumindest das Ziel – dieselben Informationen wie das Buch enthalten soll. Über das gedruckte Buch hinaus bietet Zotero neben der Aktualisierbarkeit praktische Such- und Filteroptionen. Wir werden den aktuellen Stand der Zotero-Gruppe und die dadurch eröffneten Möglichkeiten vorstellen. Außerdem werden wir den Erstellungsprozess erläutern. Wir erhoffen uns von der Präsentation und dem Austausch mit Fachkollegen auf dem Workshop nützliches Feedback für die weitere Entwicklung und Verbesserung des digitalen ALQ.

Literatur

ALEW = Hock, Wolfgang; Fecht, Rainer; Feulner, Anna Helene; Hill, Eugen; Wodtko, Dagmar S. (2015), *Altlitauisches etymologisches Wörterbuch (ALEW)*, 3 Bde., Hamburg: Baar. – Version 2.0 online, URL: <https://alew.hu-berlin.de>.

ALKT = Hock, Wolfgang / Feulner, Anna Helene, DFG-Projekt “Kritische Edition altlitauischer Kleintexte vom Überlieferungsbeginn bis 1700”.

ILJA LEMEŠKIN

Lietuvių kalbos institutas
Pražský lingvistický kroužek
ilja.lemeskin@seznam.cz

M. Mažvydo kaip knygų savininko ir skaitytojo įrašai. Dėl duomenų bazės „Spausdintų knygų glosynas“

Per pastaruosius kelis dešimtmečius Lietuvoje ir užsienyje buvo įgyvendinta nemažai projektų, skirtų seniesiems baltų rašto paminklams dokumentuoti bei tyrinėti: „Senosios lietuvių kalbos tekstynas“, „Senieji raštai“, „Latviešu valodas seno tekstu korpuss“, „Prūsų kalbos paveldo duomenų bazė“, „Samuelio Boguslavo Chylinskio Naujasis Testamentas“, „Digital Old Lithuanian: Corpus of Kristijonas Donelaitis (1714–1780)“, „Postil Time Machine“ ir kt. Taip buvo sukaupiti ir apdoroti tekstai (daugiausia spausdinti bei verstiniai), kurių skaičius yra baigtinis ir nuo XIX a. vidurio mažai kinta. Naujų spausdintų šaltinių atradimas yra mažai tikėtinas ir lauktinas, todėl ateities lituanistikos (baltistikos) žvilgsnį prasminga kreipti į kitą plotmę ir į kitus žanrus. Visų pirma tai yra nuosavybės įrašai ir glosos, kurių yra didelė gausa ir įvairovė XV–XVII a. leidiniuose ir kurie, deja, nebuvo sistemingai kaupiami, apdorojami bei tyrinėjami. Tik laiko klausimas, kada lietuviškų glosų gretas (Narbutas, Zinkevičius 1989; Subačius, Leńczuk, Wydra 2010) praplės įrašai prūsų bei latvių kalbomis, tačiau negalima pamiršti glosų ir su jų tyrimu glaudžiai susijusių nuosavybės įrašų (Feigelmanas 1981; Lemeškin 2025), daugiausia lotynų kalba padarytų lietuvių (baltų) raštijos pradininkų bei jų tęsėjų.

Pranešimo tikslas – M. Mažvydo nuosavybės įrašo bei glosų analizės pagrindu aktualizuoti šios žanrinės prigimties tekstų svarbą baltistikai, pristatyti projektą „Duomenų bazė *Spausdintų knygų glosynas*: M. Mažvydo kaip knygų savininko ir skaitytojo įrašai“, kurį Lietuvių kalbos institutas 2025-04-15 pateikė Lietuvos mokslo tarybai. Projektą, t. y. duomenų bazės „Spausdintų knygų glosynas“ kūrimą, ketinama pradėti nuo trijų paleotipų konvoliuto (VUB RSS: II-4996–4997), kuris yra kilęs iš asmeninės M. Mažvydo bibliotekos. Šaltinis tam tikslui tinka, nes turi skirtingų asmenų įrašytų skirtingos paskirties glosų (be nuosavybės įrašų, konvoliute yra dar 36 glosos). Jo pagrindu bus nustatyti bendri rėmai ir taisyklės, reglamentuojantys, kaip dirbti su šiuo žanru. Toliau projektą bus galima tęsti ir plėtoti: į duomenų bazę įtraukti glosas įrašytas baltų kalbomis; apdoroti glosas tų knygų, kurių XVI–XVII a. savininkai yra žinomi pagal nuosavybės įrašus; įtraukti įrašus knygose, kurių savininkai mums lieka nežinomi.

Literatūra

- Feigelmanas, Nojus. M. Mažvydo autografas universiteto bibliotekoje. *Literatūra ir menas*, 1981, nr. 24 (1802), p. 16.
- Lemeškin, Ilja. „Pabažnysta visam naudinga est“. M. Mažvydo devizas. *Knygotyra*, 2025, t. 84.
- Narbutas, Sigitas; Zinkevičius, Zigmas. Lietuviškos glosos 1501 m. mišiose. *Baltistica*, 1989, t. 3, p. 325–341.
- Subačius, Giedrius; Leńczuk, Mariusz; Wydra, Wiesław. The Earliest Known Lithuanian Glosses (~1520–1530). *Archivum Lithuanicum*, 2010, t. 12, p. 31–70.

SIGNE RIRDANCE

Stockholms universitet
signe.rirdance@balt.su.se

EVERITA ANDRONOVA

Stockholms universitet
everita.andronova@balt.su.se

Towards a digital philology of Early Latvian: Observations and possibilities

The growing interest in interdisciplinary research and digital philology, along with the gradually increasing digital availability of sources of early Latvian, create a promising environment, but a broader application of NLP methods and other advanced tools to early Latvian texts has so far remained limited.

This presentation explores emerging directions in the digital study of early Latvian written language. Building on the foundations of neohumanist philology, we outline how new technologies, such as machine learning, large language models and AI, can support both longstanding and evolving philological tasks: textual criticism, reconstruction, authorship attribution, distant reading and others. Drawing from hands-on experience, current research and best practices, we reflect on the possibilities and limitations in selected areas, such as handwritten text recognition (HTR), digital editions of early texts and application of crowdsourcing.

HTR remains a challenge, especially for under-resourced languages like early Latvian. While tools like Transkribus offer promising functionality, high-quality results still require significant training data, and current models struggle with variation and diacritics. These issues will be illustrated through examples from ongoing work in digitising the handwritten Latvian translation of the Proverbs of Solomon in the manuscript dated 1628 by Andreas Getzel (Gezelius).

Yet another challenge to address is named entity recognition (NER), which has not been applied to early Latvian prints extensively. As reported, NER based on the Stanford CRF classifier has primarily been applied to newspaper texts on periodika.lv (Paikens, Auzina et al. 2012). We will present initial observations on using an online NER tool available from the NLP pipeline for Latvian (<https://nlp.ailab.lv/>) on sources from the Corpus of early Latvian texts (https://nosketch.korpuss.lv/#dashboard?corpname=senie_unicode) in both the original and the normalised historical writing.

We argue for collaborative, interdisciplinary approaches to address these constraints and point to possible directions for advancing the study of Latvian historical texts using digital methods.

References

- Paikens P., Auzina I., Garkaje G., Paegle M. (2012). Towards named entity annotation of Latvian National Library corpus. *Human Language Technologies — The Baltic Perspective*. IOS Press, vol. 247, pp. 169–175. DOI: 10.3233/978-1-61499-133-5-169.

MINDAUGAS ŠINKŪNAS

Lietuvių kalbos institutas
mindaugas.sinkunas@lki.lt

ONA ALEKNAVIČIENĖ

Lietuvių kalbos institutas
ona.aleknaviciene@lki.lt

LKI Senųjų raštų duomenų bazei – 30: darbai, galimybės, perspektyvos

1995 m. susibūrė lituanistikos šaltinių skaitmeninimo iniciatyvinė grupė, kurią sudarė kalbininkai habil. dr. Saulius Ambrazas (1957–2010), dr. Ona Aleknavičienė ir programuotojas Vytautas Zinkevičius. Anuomet užsibrėžti tikslai tebėra aktualūs ir šiandien: saugoti kalbinę, kultūrinę ir istorinę informaciją skaitmenine forma, publikuoti retus spaudinius ir rankraščius mokslo visuomenei lengvai prieinama forma, skatinti modernius rašomosios lietuvių kalbos tyrimus ir spartinti lituanistikos plėtrą.

Skaitmeniniai tekstai rengiami pagal dokumentinio perrašo principus, atspindima puslapio makrostruktūra, koduojami metaduomenys (nuo 2013 m. ir autorystė, regionas, konfesija, struktūrinės šaltinio dalys, intertekstiniai ryšiai, žymimi tekstologiniai komentarai). Darbai paspartėjo įsibėgėjus pasaulio atminties institucijų fondų skaitmeninimui, kito ir bazės kūrimo metodika: spaudinių nuotraukos dabar dažniausiai apdorojamos grafinių ženklų atpažinimo priemonėmis, suliejamos ir lyginamos kelių išlikusių egzempliorių nuotraukos, skaitoma korektūra, o rankraščiai rengiami pritaikant dar ir kritinio leidimo principus.

Per 30 metų pavyko parengti didžiąją daugumą XVI–XVII a. lietuvių kalbos šaltinių, tačiau vėlesnių amžių raštų perrašų vis dar trūksta. Skaitmeninių perrašų rinkinį šiuo metu sudaro 6,38 mln. žodžių: XVI a. – 1,08, XVII a. – 1,78, XVIII a. – 1,84, XIX a. – 1,68 mln. Parengtos ir tos lituanistikos šaltinių dalys, kurios parašytos kitomis kalbomis, taip pat kai kurie svarbesni vertimų šaltiniai. Sukaupę kitakalbių tekstų apimtys gana didelės: XVI a. – 0,91, XVII ir XVIII a. – po 0,81, XIX a. – 0,14 mln. žodžių (iš viso – 2,67 mln. žodžių). Atskirą vienetą sudaro senųjų ir dabartinių Biblijų įvairiomis kalbomis rinkinys (7,44 mln. žodžių).

Kai kurių šaltinių žodžiai yra anotuoti: nustatytas antraštinis žodis (lema ir hiperlema – apytikslei paieškai skirta lema be priešdėlio, fonetinis variantas ar tikrinio žodžio atitikmuo), žodžio forma perrašyta dabartine rašyba bei pateikta gramatinė jos informacija (kalbos dalis, morfologinės kategorijos). Ši informacija nuo 2021 m. naudojama mašiniam mokymuisi: sudaromi neuroniniais tinklais paremti modeliai, kurie generuoja reikiamas anotacijas naujiems šaltiniams. Indeksų rengimą spartina bei anotacijos sluoksnius kuria ir kiti LKI sukurti ar pritaikyti kalbos technologijų įrankiai (dabartinė žodžio forma sintezuojama senųjų raštų Lemuokliu, taisykliniais algoritmais niveliuojama rašyba ir apskaičiuojamas teksto fragmentų panašumas).

2008 m. pradėjo veikti interneto svetainė „Senieji raštai“, skirta šaltinių ir žodžių formų kontekstinių konkordancijų peržiūrai, failų atsisiuntimui (<https://seniejirastai.lki.lt>). 2015 m. svetainė buvo papildyta paieškos sistema (ieškoma tarp 1,20 mln. žodžių) ir gretinamuoju Biblijos leidimu. Nuo 2023 m. kuriama nauja LKI Senųjų raštų svetainė, kurioje išnaudojamos metaduomenų ir dirbtinio intelekto automatinio anotavimo teikiamos galimybės. Pranešime bus pristatytas naujosios svetainės maketas, aptartos šaltinių peržiūros, žodžių paieškos, chronologinių ir kitų filtrų funkcijos ir atsiveriančios naujų mokslinių tyrimų perspektyvos.

ANTA TRUMPA

LU HZF Latviešu valodas institūts
anta.trumpa@lu.lv

Dar kartā apie Georgo Elgerio žodyną „Dictionarium Polono-Latino-Lottaucum“ (1683)

Georgo Elgerio žodynas, išleistas keliolika metų po autoriaus mirties, jau daug kartų buvo apžvelgiamas įvairiais aspektais. Keletas tyrėjų kaip galimus šaltinius analizavo kitakalbius žodynus (pvz., Judžentytė, Zubaitienė 2014, 2015, 2016), taip pat aptarta kitų latvių kalbos žodynų, pavyzdžiui, Mancelio „Lettus“, įtaka (Beitiņa 2013b).

Elgerio žodynas nuo kitų XVII a. latvių leksikografijos šaltinių skiriasi tiek rašybos, tiek turinio požiūriu. To šimtmečio latviškuose žodynuose buvo įprasta vartoti vokišką rašybą, o Elgerio žodyne, kaip ir kituose jo darbuose, ji lenkiška. Taip pat esama tyrimų, skirtų keliems kartams leisti G. Elgerio giesmynų rašybai ir netgi akcentografijai (pvz., Andronov 2014, Vanags 2018), bei publikacijų apie Elgerio evangelijų ir epistolų vertimų rašybą (pvz., Christophe 2002).

Nors Elgeris (po jo mirties – knygų leidėjai) žodyne ir naudojami kitų žodynų duomenimis (apie tiesioginius perrašus iš Mancelio žodyno liudija atskiruose fragmentuose pasitaikanti šio autoriaus rašyba), tačiau šių žodynų turinys skirtingas – tai ne vokiečių-latvių ar latvių-vokiečių, o lenkų-lotynų-latvių žodynas, jame gerokai daugiau žodžių nei kituose XVII a. amžiaus latviškuose žodynuose. Pažymėtina, kad ir pats Elgeris gimė Česyse, Vidžemėje – kitame regione nei, pavyzdžiui, Žiemgaloje gimęs Georgas Mancelis, Sėloje gimęs ir Žiemgaloje dirbęs Christophas Füreckeris ar Johannas Langijus iš Kuršo. Žodynų tyrinėtoja Daina Zemzarė pabrėžia, kad Elgerio žodyno „latvių kalba atstovauja vidurio tarpei“ (Zemzare 1961, 64), tačiau vėliau ji tekste pripažįsta, kad „Elgerio leksika labiau atspindi pirmą plačiausią Rytų Latvijoje paplitusių žodžių sąrašą“ (Zemzare 1961, 68).

Elgerio žodyno latvių kalbos ypatumams taip pat skirta keletas tyrimų: aptarta sintaksė (pvz., Beitina 2012, 2013a), morfologija (pvz., Pavela 1998), leksika (pvz., Trumpa 2014).

LU HMF Latvių kalbos institutas ir LU MII Dirbtinio intelekto laboratorija, modernizuodami senųjų raštų tekstyną (projektas DHELI, No. VPP-IZM-DH-2022/1-0002), visus tekстыne esančius tekstus verčia į dabartinę rašybą. Elgerio žodyno konvertavimo taisyklių kūrimas yra sudėtingas dėl didelės žodyno apimties, nenuoseklios rašybos ir daugybės klaidų. Tuo pat metu, kuriant konvertavimo taisykles, registruojami ir žodyno rašybos ypatumai, skirtumai, pavyzdžiui, nuo 1621 m. giesmyno rašybos, taip pat galima rasti trūkumų ankstesniuose Elgerio žodyno rašybą apibūdinančiuose aprašuose. Pavyzdžiui, D. Zemzarė teigė, kad raide *ß* žymimas garsas [š] (Zemzare 1961, 64), tačiau šis nedidelis tyrimas rodo, kad ja atspindimi ir garsai [ž], [s] bei [z].

Nors daugybė latviškame žodyno tekste vartojamų raidžių ir jų derinių yra daugiaprasmiai, vis dėlto esama tam tikro nuoseklumo, leidžiančio formose, kurios skiriasi nuo šiuolaikinės latvių kalbos, išžvelgti ne klaidingus užrašymus, o tarmių ypatumus, pvz., įrašas *czurka* (*Szczurek Glis, Lêla pælle / czurka*, Elger 1683, 542A) konvertuojant virsta forma *čurka*, kuri, labai tikėtina, yra ne klaida, o tarminis žodžio variantas, plg. *čurka* ‘žurka’ (EH I 296).

Pranešime bus aptariami Elgerio žodyno rašybos pavyzdžiai, apžvelgiami ir analizuojami atskiri mūsų laikų požiūriu tarminiais laikytini žodžiai.

Literatūra ir šaltiniai

Andronov, Aleksey (2014). On the problem of Georg Elgers accentography (Geistliche Catholische Gesange, 1621). *The 3rd “Slavistična srečanja” Meeting: 10th International Workshop on Balto-Slavic Accentology (IWoBA X)*. Ljubljana, 16th–18th

October 2014. University of Ljubljana, Faculty of Arts, Department of Slavistics. Ljubljana: University of Ljubljana, pp. 15–16.

Beitiņa, Maigone (2012). Nominālās prepozicionālās vārdkopas

- Georga Elgera vārdnīcā „*Dictionarium Polono-Latino-Lottaicum*”(1683). *Vārds un tā pētišanas aspekti. Rakstu krājums, 16 (1)*. Liepāja: LiePA, 28.–41. lpp.
- Beitiņa, Maigone (2013a). Sintaktiskā sinonīmija Georga Elgera vārdnīcā „*Dictionarium Polono-Latino-Lottaicum*”(1683). *Baltu filoloģija, XXI (1)*, 2013. Baltu valodniecības žurnāls. Rīga: LU Akadēmiskais apgāds, 5.–30. lpp.
- Beitiņa, Maigone (2013b). Vai Georgs Elgers ir lietā līcis Georga Manceļa vārdnīcu „*Lettus*”? *Baltu valodas: vēsture un aktuālie procesi. Akadēmika Jāņa Endzelīna 140. dzimšanas dienas atceres starptautiskās zinātniskās konferences materiāli*. Rīga: LU
- Latviešu valodas institūts, 2013, 14. lpp.
- Christophe, Bernardo (2002). Zur Orthographie der *Evangelien und Episteln* von Georg Elger. *Res Balticae, 8. Miscellanea italiana di studi baltistici*. Pisa: ECIG, S. 107–130.
- EH – Endzelīns, Jānis, Edīte Hauzenberga. *Papildinājumi un labojumi K. Mūlenbacha „Latviešu valodas vārdnīcai”*. 1.–2. sēj. Rīga: Kultūras fonds, 1934–1938; Grāmatu apgāds, 1946.
- Elger, Georg 1683, *Dictionarium Polono-Latino-Lottaicum. Opus posthumum*. Vilnae: Typis Academicis Soc. Iesu.
- Judžentytė, Gintarė; Zubaitienė, Vilma (2014). Dēl Konstantino Sirvydo „*Dictionarium trium linguarum*” (1642) ir Georgo Elgerio „*Dictionarium Polono-Latino-Lottaicum*” (1683) lietuviškosios resp. latviškosios dalies. *Starptautiska konference „Aktuāli baltu valodu vēstures un dialektoloģijas jautājumi”*. Veltījums valodnieces Martas Rudzītes (1924–1996) piemiņai. Referātu kopsavilkumi. Rīga, 2014. gada 7. –8. novembris. Latvijas Universitāte, Humanitāro zinātņu fakultāte, Latvistikas un baltistikas nodaļa, Baltu valodniecības katedra. Rīga: LU Baltu valodniecības katedra, p. 26–27.
- Judžentytė, Gintarė; Zubaitienė, Vilma (2015). Konstantino Sirvydo „*Dictionarium trium linguarum*” (1642) ir Georgo Elgerio „*Dictionarium Polono-Latino-Lottaicum*” (1683): jų santykis ir leksikografinių metodų skirtybės. *Baltu filoloģija, XXIV (1)*, 2015. Baltu valodniecības žurnāls. Latvijas Universitāte, Humanitāro zinātņu fakultāte, Baltu valodniecības katedra. Rīga: LU Akadēmiskais apgāds, p. 89–141.
- Judžentytė, Gintarė; Zubaitienė, Vilma (2016). Dar dēl Georgo Elgerio „*Dictionarium Polono-Latino-Lottaicum*” (1683) registro šaltinio. *Baltu filoloģija, XXV (1)*, 2016. Baltu valodniecības žurnāls. Rīga: LU Akadēmiskais apgāds, p. 37–58.
- Pavela, Justina (1998). Par dažām verbu formu īpatnībām G. Elgera *Dictionarium Polono-Latino-Lottaicum*. *Linguistica Lettica, 2*. Latviešu valodas institūta žurnāls. Rīga: LU Latviešu valodas institūts, 180.–187. lpp.
- Trumpa, Anta (2014). Augu nosaukumi Georga Manceļa vārdnīcā „*Lettus*” (1638) un Georga Elgera vārdnīcā „*Dictionarium polono-latino-lottaicum*” (1683). *Vārds un tā pētišanas aspekti. Rakstu krājums, 18 (1)*. Liepāja: LiePA, 286.–297. lpp.
- Vanags, Pēteris (2018). Dziesmu grāmatas rakstība. Georgius Elger. *Geistliche Catholische Gesänge... (Braunsberg 1621)*. Ed. facs. Māra Grudule, Justyna Prusinowska, Mateusz Solarz. Warszawa: Wydawnictwo Naukowe Sub Lupa, LX–LXI lpp.
- Zemzare, Daina (1961). *Latviešu vārdnīcas (līdz 1900. gadam)*. Rīga: Latvijas PSR Zinātņu akadēmijas izdevniecība.