Quotes at the fingertips: The combined approach of the BogoSlov project towards identification of Biblical material in Old Church Slavonic texts

Martin Ruskov[1], Tomáš Mikulka[2], Irina Podtergera[3], Maxim Gavrilkov[3] and Walker Thompson[3]

[1]*Department of Languages, Literatures, Cultures and Mediations, University of Milan, Piazza Sant'Alessandro 1, 20123 Milan, Italy*
[2]*Catholic Theological Faculty, Charles University, Thákurova 3, 16000 Prague, Czech Republic*
[3]*Institute of Slavic Studies, University of Heidelberg, Schulgasse 6, 69117 Heidelberg, Germany*

Abstract

Scriptural quotations shaped and influenced the orthodox Slavic world by laying the groundwork for historical and symbolic exegesis through Old Church Slavonic (OCS) texts. The correct identification of biblical quotations is of the utmost importance for the textological as well as functional analysis of such texts. In this paper, we present a computer-assisted approach towards identifying quotations proposed by the BogoSlov project. This approach aims to combine two distinct methods of quotation identification. These are: 1) explicit rule-based algorithms and 2) quantitative embeddings from implicit language models. To make these accessible to Slavists and theologians, we aim to integrate them into a graphical user interface (GUI) built on best practices from related fields and facilitating the identification and validation of quotations, allowing as short as possible a feedback loop between expert and machine.

Keywords

Palaeoslavistics, computer-supported collaborative work, short text similarity, text reuse identification

1. Introduction

Biblical texts shaped and influenced *Slavia orthodoxa* by laying the groundwork for historical and symbolic exegesis, i.e., the interpretation and understanding of facts and narratives, such that historical events were seen in the light of Scriptural prototypes [1]. The Bible supplied the core quotations through which this exegesis was manifested in Old Church Slavonic (OCS) texts. Therefore, the universal and correct identification of biblical quotations is of the utmost importance for textological as well as functional textual analysis. However, this task is fraught with challenges, demanding a nuanced approach that balances philological rigour with an understanding of the broader intellectual and theological context.

Although biblical quotations have been studied for decades [2, 3], so far attempts to build a software tool for automatic identification of biblical quotations are limited and at best offer support for manual annotation by experts. No standalone solutions are yet available [4, 5], as recently have emerged in other contexts of text reuse [6]. Yet, such technological support for quotation identification could be very impactful to Palaeoslavistics and shed light on previously unknown intertextual relations, text history, and meanings. Among other things, it could facilitate the investigation of the dual phenomena of inter- and hypertextuality, whereby texts drew on common sources and came to reference each other or themselves internally. These phenomena are strongly present in medieval Slavonic Patristic

IRCDL 2025: 21st Conference on Information and Research Science Connecting to Digital and Library Science, February 20-21 2025, Udine, Italy
martin.ruskov@unimi.it (M. Ruskov); mikut5af@ktf.cuni.cz (T. Mikulka); irina.podtergera@slav.uni-heidelberg.de (I. Podtergera)
https://islab.di.unimi.it/team/martin.ruskov@unimi.it (M. Ruskov);
https://www.slav.uni-heidelberg.de/personal/ipodtergera.html (I. Podtergera)
0000-0001-5337-0636 (M. Ruskov); 0000-0003-4362-2531 (T. Mikulka); 0000-0001-9098-2746 (I. Podtergera); 0000-0001-7656-1590 (M. Gavrilkov); 0000-0002-7203-9508 (W. Thompson)

and liturgical texts owing to their extensive citation of the Scriptures; having a tool to identify such quotations would make it possible to study such aspects of these texts more deeply from a semiotic point of view.

In other historical contexts, various techniques of automatic text reuse detection have been employed, from statistical and algorithmic methods [7, 8, 9], to machine learning and transformer models [10]. Only recently, some of these have started to also adopt human-machine interaction techniques [6], but have stopped short of making their approaches adoptable in other domains.

In the BogoSlov project (acronym for *Biblical OriGins in slavOnic texts – Systems for Language-modelled Observation and Verification*), we set out to develop experimental software infrastructure to assist the identification and listing of Biblical quotations in OCS texts. This infrastructure is intended to include tools for annotation, algorithms and language models, combined in a toolset software bundle to support scholars in their efforts of identification and verification Biblical quotations.

However, due to the varying degree of literality, identification of Biblical quotations is not a straightforward task. Regardless of their origin, i.e. direct or indirect quotation (imported by the patristic source, which can be difficult to identify), quotations in general can vary in their fidelity to the original. This variation could range from exact reproductions, explicitly marked by the author or slightly modified versions that reflect either a conscious adaptation to fit the surrounding text or a transmission error. The before mentioned variation could include subtle and barely noticeable thematic allusions, which invoke Biblical themes or ideas without directly quoting the text. As a consequence for this less direct range of the spectrum, two specific challenges emerge. First, such allusions are hardly recognisable due to their divergence from the precise language of the original Biblical text. Secondly, there is a great deal of lexical and grammatical variation in translations of biblical texts and quotations. These issues delineate the limits of any attempt of identifying quotations merely by comparing texts.

2. Project Overview

Since it would require considerable effort to build a universal tool that could identify all types of biblical quotations in any type of OCS text, the BogoSlov project plans to work around two pilot studies, in which the technology would be developed and tested on well-researched types of text, rich in biblical quotations, namely well-known homilies and hagiography. Thus, one pilot case study will focus on *Vita Constantini* and *Vita Methodii*, and in a second case study, a small number of less researched texts [11] would be analysed to identify citation-related clues that might help better contextualise the texts historically. To accomplish this, within the project two datasets will be created and integrated into an interlinked database. The first of these datasets is a corpus of known biblical quotations, an idea first elaborated by Naumow [2]. The second corpus is a machine-readable, tagged database of OCS biblical texts.

Within the project, two major biblical sources of quotations (primary texts) will be examined: the Psalter and the Gospels, both of which are frequently quoted in medieval texts. The Psalter text is quite stable and was often known by heart at the time of writing these texts, suggesting a well-established and consistent tradition. Due to this stability, quotations from the Psalter are commonly easier to identify, making them suitable for an approach utilising explicit algorithmic modelling based on string similarity and longest common subsequence [12, 13].

In contrast, Gospel quotations present a more complex challenge. This complexity arises first from the nature of the synoptic Gospels, which share many common passages, allowing the wording to be influenced by prevailing traditions. As a result, within the Christian oral tradition, there exists a form of "evangelical harmony", where it becomes difficult to strictly separate nuances originating from specific Gospels. A second source of lexical variation that is valid not only for Gospel quotation is the fact that an important part of OCS literature was translated from Greek (and less often Latin) texts. In the process of translation, lexical variation was introduced, even within individual texts. To identify quotations with such variation or allusions, the word embeddings of language models allow for the study of semantic, rather than syntactic similarity.

3. Proposed Approach

The BogoSlov project builds around the idea that quotation identification should be a computer-assisted process, meaning that this work should still be driven by theologians, philologists and medievalists, but whenever possible, tedious tasks should be (semi-)automated. With this premise, we set out to research three parallel directions: 1) helping experts efficiently perform manual identification and validation of quotations on one hand, and two automated approaches to detection of potential quotations on the other: 2) explicit algorithms for rule-based suggestion for possible quotations, and 3) language models for suggestions for potential allusions via implicit semantic modelling. These three will be brought together through a common data representation that would allow exchange of results between them. This combined approach aims to allow for a feedback loop between expert and machine that is as short as possible.

3.1. Data Model

Combining the three aforementioned approaches, that are detailed in further sections, is only feasible if it is supported by a reproducible and interpretable quotation representation format. A database needs to allow for the systematic storage, querying and user annotation of the identified quotation candidates.

Approaches in related research [3, 14] tend to gravitate towards the use of XML-formats, inspired by TEI (Text Encoding Initiative). While at this point this exact format does not present particular advantages to the current effort, compatibility with it would guarantee possible future exchange of results. Of particular interest to us is following a related established standard for text localisation, such as CTS URN [14]. These are strings (Uniform Resource Names) that contain a document identifier in their first part and text localisation (e.g. through the specification of a range) in the second. The CTS URN features the following components: urn:cts:NAMESPACE:WORK:PASSAGE, where PASSAGE could contain indications of start and end, separated by dash ("-") or optionally specific strings indicating more precise subsection in the passage, separated by the at-sign ("@")[1]. A few examples from our context follow:

1. `urn:cts:proiel:bible.marianus.matt:5.48` for the Gospel of Matthew 5:48 in the Codex Marianus manuscript as processed for the PROIEL treebank project[2].
2. `urn:cts:titus:bible.zograph.jo:4.5@-4.5@` for a specific text in the Gospel of John 4:5 in the Codex Zographensis as presented in the TITUS corpus[3].
3. `urn:cts:scripta-bulgarica:uchitelno-evangelie:5.20.cd@-5.20.cd@` for a specific text on page 20, columns C and D in the 5th sermon of Constantine of Preslav's Didactic Gospel (Uchitelno Evangelie) as presented by the Scripta Bulgarica corpus, and externally proposed as corresponding to the previous example[4].

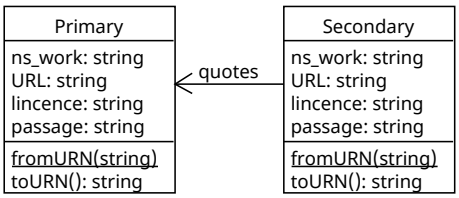

Figure 1: The conceptual data model, representing quotes in our database.

Notice that the online sources for the above examples do not support CTS URN, so it will be our responsibility to define the corresponding namespaces and link them to the original sources. With this

[1] i.e. urn:cts:NAMESPACE:WORK:PASSAGE@SUBSECTION
[2] https://syntacticus.org/sentence/proiel:20180408:marianus:38407
[3] https://titus.uni-frankfurt.de/texte/etcs/slav/aksl/zograph/zogra.htm?zogra071.htm#NT_Jo._4
[4] http://scripta-bulgarica.eu/bg/sources/uchitelno-evangelie-na-konstantin-preslavski-tlkuvanie-vrhu-gl-4-ot-evangelie-na-yoan

premise, as indicated in Figure 2, quotations are required to be pairs of two similar objects representing text segments: one biblical, and one medieval, e.g. the pair between examples 2 and 3 above. Each of these objects needs to be serialisable and deserialisable to URN, which combines a representation of the document and snippet attributes.

3.2. Annotation

Due to the error-prone nature of the task of quotation identification, of key importance to the process is an user experience (UX) that allows experts to 1) browse for hypothetical partial (i.e. local only) alignments around the use of particular words or phrases, and 2) visualise and refine in context already identified quotations and allusions. Here, viewing a partial alignment is synonymous with contextual visualisation of a quotation.

Previous research has focused on algorithms for automatic identification, but has stopped short of providing humanities researchers an accessible interface to validate and refine automatically identified quotations or allusions [7]. In contrast, we believe that an efficient semi-automated text reuse detection process is only possible with a graphical user interface (GUI) which allows side-by-side viewing of biblical (primary) texts on one side, and medieval texts (secondary) on the other, much along the lines of how this is done in text alignment software [15]. Furthermore, the parallel search of a specific vocabulary should be possible.

This GUI-enabled tool should be able to flag potential quotations or allusions in a large input text, as well as allow for the search for short phrases or biblical references and return a list of known quotations, as discussed by Hue-Gay et al [3]. This is intended to significantly speed up the initial stages of analysis as it is currently performed by experts. Yet, the tool is still intended to be used by expert users providing only suggestions to experts that still would need to further verify possible quotation candidates. Thus, the tool is only semi-automatic and expert validation and detailed philological work remain indispensable. Human expertise is required to confirm the presence of a quotation, assess its degree of literality, and understand its function within the text. In BogoSlov we will employ usability research to ensure that the interface of this tool is both efficient to use [16] and cognitively undemanding [17].

The collaborative workflow of existing tools, like MoreEver [18] and UE extractor [19], served as a starting point for the discussion of a possible interface that should allow users to manually locate

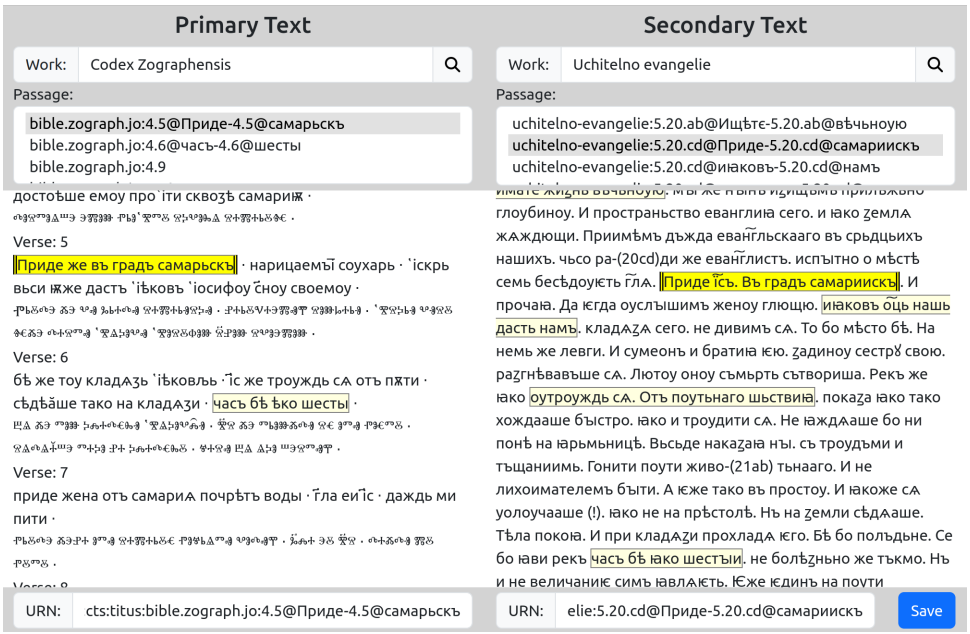

Figure 2: The proposed review GUI which allows experts to select a quotation and refine the corresponding alignment.

and annotate potential quotations to support the training and verification of the language models. An early mock version of a quotation validation screen is shown in Figure 1. The two panes have similar functionalities, with the left one referring to biblical corpora (primary texts) and the right one to medieval (secondary texts). In the top-most search bar, users could search the text of interest from the corpora. When at least three characters are typed, an autocomplete drop-down list of available matching texts would appear to ease selection. The selection of text specifies an URN up to the WORK section. Once a text is selected, the list below gets populated with addresses of citation candidates from the text. Clicking on one of them, updates both the selection in the text below, along with its alignment on the other side, and the URN in the text box that represents the exact text selection. The text boxes for the two texts include WORK and PASSAGE components of URNs, i.e. complete addresses of text ranges. Users will be able to make selections on the two texts or manually edit URNs. These actions should be interactively synchronised, so that when changing the selection, the URN gets updated, and when changing the URN, the selection gets updated. In other words, the text preview and the URN box offer two different affordances for quotation review and refinement.

3.3. Algorithms

There is a long-standing research tradition around text reuse identification through short text similarity algorithms [12, 13]. In it typically longer texts are broken down into sentences or text snippets of comparable length and then these short texts between the two corpora are compared one at a time.

Due to OCS being a late (compared to classic languages) established written tradition spanning very broad language variation, and using both the Glagolitic and Cyrillic alphabets, there are a number of accepted redactions. In other words, even common sounds and words have varying established spellings, which makes even simple lexical correspondence a challenging task. This calls for methods that allow greater flexibility than typical short text similarity techniques used in the context of contemporary languages, allowing for tolerance to variation both in orthography and morphology.

3.4. Language Models

Even short text similarity algorithms that exhibit tolerance for greater variation cannot cope for situations of rephrased texts or summarisation, typical for allusions [14]. In this context, recent developments in the field of language models could be helpful. In particular, semantic embeddings have emerged as a useful text quantification technique, used in language models. Commonly these are word embeddings, but more recently also other types, such as sentence embeddings have been proven to yield interesting results in text reuse detection [20, 10, 21].

On the downside, language models require corpora sizes that are unavailable for under-resourced languages, such as OCS [22]. A viable alternative is training a multilingual model including only the limited OCS resources available. Particularly interesting are results with multilingual models which exhibit improved performance over single-language ones [23, 24, 25]. Theoretically, this opens the possibility to address situations very typical in our context where a Greek homily – that contains Biblical quotations – was later translated to OCS. Such translations lead to the need to partially align them to the OCS translation of a biblical text, which would be an even more ambitious text reuse detection task.

However, the choices surrounding training a multilingual model are not trivial. On one hand, including contemporary texts introduces unwanted socio-historical biases [26]. This particular problem could be addressed by training a model dedicated to classical languages, making sure to include only corpora from relevant historical periods. On the other hand, models trained on languages using multiple alphabets (such is the case of a mixed-classical model) are known to achieve lower performance on the minority alphabets [27]. We are looking for ways to train RoBERTa [28] and Sentence BERT [29] from scratch on corpora from the late antiquity and early middle ages. One particular challenge is that OCS was written in Glagolitic and Cyrillic alphabets with further variability, related to the specific historical circumstances of the language. As a consequence, we are interested in possibilities to further explore how to reduce variation and/or bootstrap data by tackling orthographic variation, related to the

different alphabets used in the languages.

4. Further Work

As previously mentioned, OCS texts of the Gospels underwent various revisions and refinements, sometimes aligning closely with the Greek standard and sometimes not. This led to a complicated array of textual variants circulating within the Slavic world. A comprehensive analysis of Biblical quotations in OCS texts often necessitates a comparative approach, involving not only the OCS translations themselves but also their Greek and Latin sources. By comparing the OCS text with these sources, researchers can identify the specific Biblical texts that were used, discern the influences and editorial decisions that shaped the text, and uncover any redactional layers that indicate later modifications. Although not considered part of our project, this comparative work is essential for understanding the intellectual and theological formation of the author or translator. The degree to which an author or translator accurately quotes or adapts a Biblical text can provide insights into their level of education, familiarity with the sources, and the theological or rhetorical objectives that guided their work. The program developed within this project would help identify the affinities of these quotations with specific Gospels. However, detailed analysis and interpretation will still be necessary to fully understand the nuances and origins of these quotations.

In conclusion, the identification and analysis of Biblical quotations in OCS texts are complex tasks that require a combination of technological assistance and deep philological and theological expertise. While semi-automatic tools can greatly aid in the process, they cannot replace the nuanced analysis that only a trained scholar can provide. Through careful comparison with Greek and Latin sources, researchers can uncover the intricate layers of influence and modification that shape these medieval texts, offering valuable insights into the intellectual world of the time.

Whereas, our case studies focus on Old Church Slavonic (OCS) texts dated around the 9th and 10th centuries, we expect that the developed methodology and algorithms are adaptable to applications for Church Slavonic (12th century and beyond) and possibly other medieval languages, not excluding Latin and Greek.

Acknowledgments

This contribution has received funding from the SEED4EU+ collaborative research programme under the project “Biblical OriGins in slavOnic texts – Systems for Language-modelled Observation and Verification” (acronym: BogoSlov), scheduled to run throughout the year 2025.

References

[1] R. Picchio, The Function of Biblical Thematic Clues in the Literary Code of “Slavia Orthodoxa”, in: Slavica Hierosolymitana, volume 1, Magnes Press, Hebrew University, Jerusalem, 1977, pp. 1–31.

[2] A. Naumov, kartotece cerkiewnosłowiańskich użyć biblijnych. Cytaty biblijne w staroruskiej części Kodeksu Uspienskiego, Rocznik Slawistyczny 44 (1983) 21–29.

[3] E. Hue-Gay, L. Mellerin, E. Morlock, TEI-encoding of text reuses in the BIBLINDEX Project, Journal of Data Mining & Digital Humanities Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages (2017). doi:10.46298/jdmdh.3989.

[4] M. Moritz, A. Wiederhold, B. Pavlek, Y. Bizzoni, M. Büchler, Non-Literal Text Reuse in Historical Texts: An Approach to Identify Reuse Transformations and its Application to Bible Reuse, in: J. Su, K. Duh, X. Carreras (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 1849–1859. doi:10.18653/v1/D16-1190.

[5] H. Eckhoff, Automatic Alignment of the Psalterium Sinaiticum and the Septuagint Psalms, - (2021) 71–90. URL: https://www.ceeol.com/search/article-detail?id=1005901.

[6] M. Düring, M. Romanello, M. Ehrmann, K. Beelen, D. Guido, B. Deseure, E. Bunout, J. Keck, P. Apostolopoulos, impresso Text Reuse at Scale, 2023. URL: https://hal.science/hal-04151808.
[7] G. Franzini, M. Passarotti, M. Moritz, M. Büchler, Using and Evaluating TRACER for an Index Fontium Computatus of the Summa contra Gentiles of Thomas Aquinas, in: E. Cabrio, A. Mazzei, F. Tamburini (Eds.), Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), volume 2253, CEUR-WS, Torino, Italy, 2018. URL: https://ceur-ws.org/Vol-2253/#paper22.
[8] D. A. Smith, R. Cordell, A. Mullen, Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers, American Literary History 27 (2015) E1–E15. doi:10.1093/alh/ajv029.
[9] A. Vesanto, A. Nivala, H. Rantala, T. Salakoski, H. Salmi, F. Ginter, Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771-1910, in: G. Bouma, Y. Adesam (Eds.), Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language, Linköping University Electronic Press, Gothenburg, 2017, pp. 54–58. URL: https://aclanthology.org/W17-0510/.
[10] F. Periti, P. Cassotti, S. Montanelli, N. Tahmasebi, D. Schlechtweg, TRoTR: A Framework for Evaluating the Re-contextualization of Text Reuse, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 13972–13990. URL: https://aclanthology.org/2024.emnlp-main.774.
[11] T. Mikulka, Textological, linguistic and theological features of the newly identified corpus of Old Church Slavonic homilies, Slavia 92 (2023) 610–624. doi:10.58377/slav.2023.5.05.
[12] A. Islam, D. Inkpen, Semantic text similarity using corpus-based word similarity and string similarity, ACM Trans. Knowl. Discov. Data 2 (2008) 10:1–10:25. doi:10.1145/1376815.1376819.
[13] D. W. Prakoso, A. Abdi, C. Amrit, Short text similarity measurement methods: a review, Soft Computing 25 (2021) 4699–4723. doi:10.1007/s00500-020-05479-2.
[14] M. Berti, C. Blackwell, M. Daniels, S. Strickland, K. Vincent-Dobbins, Documenting Homeric Text-Reuse in the Deipnosophistae of Athenaeus of Naucratis, Bulletin of the Institute of Classical Studies 59 (2016) 121–139. doi:10.1111/j.2041-5370.2016.12042.x.
[15] T. Yousef, S. Janicke, A Survey of Text Alignment Visualization, IEEE Transactions on Visualization and Computer Graphics 27 (2021) 1149–1159. doi:10.1109/TVCG.2020.3028975.
[16] K. Hornbæk, M. Hertzum, Technology Acceptance and User Experience: A Review of the Experiential Component in HCI, ACM Transactions on Computer-Human Interaction 24 (2017) 1–30. doi:10.1145/3127358.
[17] F. Paas, J. J. G. Van Merriënboer, Cognitive-Load Theory: Methods to Manage Working Memory Load in the Learning of Complex Tasks, Current Directions in Psychological Science 29 (2020) 394–398. doi:10.1177/0963721420922183.
[18] A. Morollon Diaz-Faes, C. S. R. Murteira, M. Ruskov, Values That Are Explicitly Present in Fairy Tales: Comparing Samples from German, Italian and Portuguese Traditions, Journal of Data Mining & Digital Humanities NLP4DH (2024). doi:10.46298/jdmdh.13120.
[19] M. Ruskov, L. Taseva, Computer-Aided Modelling of the Bilingual Word Indices to the Ninth-Century Uchitel'noe evangelie, in: Proceedings of the Workshops and Doctoral Consortium of the 26th International Conference on Theory and Practice of Digital Libraries, 2022, pp. 19–30. URL: http://ceur-ws.org/Vol-3246/03_paper-6921.pdf.
[20] P. Cassotti, L. Siciliani, M. DeGemmis, G. Semeraro, P. Basile, XL-LEXEME: WiC Pretrained Model for Cross-Lingual LEXical sEMantic changE, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1577–1585. doi:10.18653/v1/2023.acl-short.135.
[21] I. Muneer, R. M. A. Nawab, Cross-Lingual Text Reuse Detection at sentence level for English–Urdu language pair, Computer Speech & Language 75 (2022) 101381. doi:10.1016/j.csl.2022.101381.
[22] Q. Dombrowski, From Annotation to Modeling: Computational Horizons for Medieval Slavic Studies, Scripta & e-Scripta (2021) 11–21. URL: https://www.ceeol.com/search/article-detail?id=

994247.

[23] Z. Wang, K. K, S. Mayhew, D. Roth, Extending Multilingual BERT to Low-Resource Languages, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2649–2656. doi:10.18653/v1/2020.findings-emnlp.240.

[24] P. Rust, J. Pfeiffer, I. Vulić, S. Ruder, I. Gurevych, How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 3118–3135. doi:10.18653/v1/2021.acl-long.243.

[25] P. Singh, A. Maladry, E. Lefever, Too Many Cooks Spoil the Model: Are Bilingual Models for Slovene Better than a Large Multilingual Model?, in: J. Piskorski, M. Marcińczuk, P. Nakov, M. Ogrodniczuk, S. Pollak, P. Přibáň, P. Rybak, J. Steinberger, R. Yangarber (Eds.), Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023), Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 32–39. doi:10.18653/v1/2023.bsnlp-1.5.

[26] M. Cuscito, A. Ferrara, M. Ruskov, How BERT Speaks Shakespearean English? Evaluating Historical Bias in Contextual Language Models, 2024. URL: http://arxiv.org/abs/2402.05034, arXiv:2402.05034 [cs].

[27] J. Pfeiffer, I. Vulić, I. Gurevych, S. Ruder, UNKs Everywhere: Adapting Multilingual Language Models to New Scripts, 2021. doi:10.48550/arXiv.2012.15562, arXiv:2012.15562.

[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019. doi:10.48550/arXiv.1907.11692, arXiv:1907.11692.

[29] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, 2019. doi:10.48550/arXiv.1908.10084, arXiv:1908.10084.