connecting creative minds

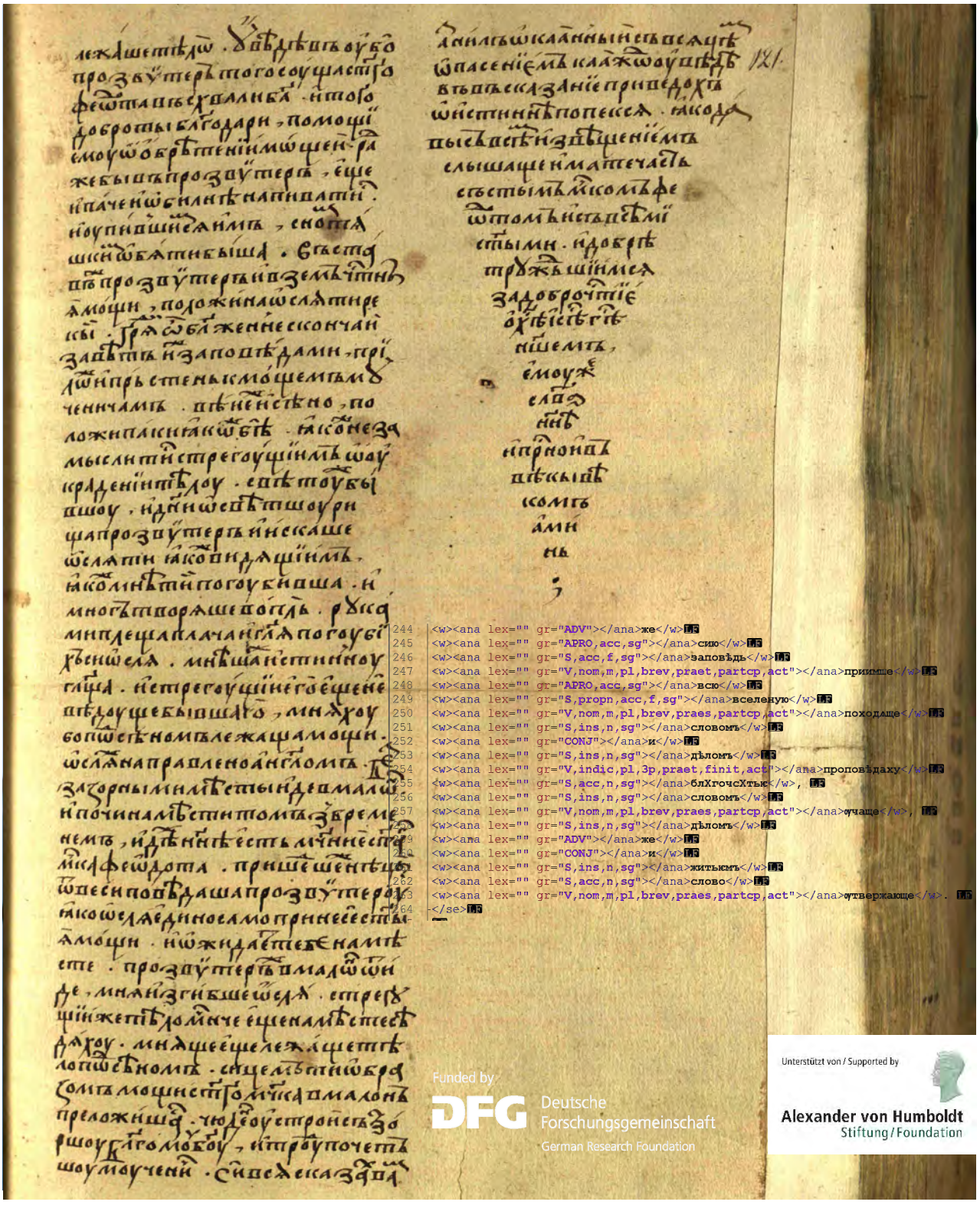
El'Manuscript 2021

8th International Conference on Textual Heritage and Information Technologies

12.04.2021 – 15.04.2021

Albert-Ludwigs-Universität Freiburg

FREIBURG



```
<w><ana lex="" gr="ADV"></ana>же</w>
<w><ana lex="" gr="APRO,acc,sg"></ana>сию</w>
<w><ana lex="" gr="S,acc,f,sg"></ana>заповѣдь</w>
<w><ana lex="" gr="V,nom,m,pl,brev,praet,partcp,act"></ana>приимше</w>
<w><ana lex="" gr="APRO,acc,sg"></ana>всю</w>
<w><ana lex="" gr="S,propn,acc,f,sg"></ana>вселенную</w>
<w><ana lex="" gr="V,nom,m,pl,brev,praes,partcp,act"></ana>походаще</w>
<w><ana lex="" gr="S,ins,n,sg"></ana>словомъ</w>
<w><ana lex="" gr="CONJ"></ana>и</w>
<w><ana lex="" gr="S,ins,n,sg"></ana>дѣломъ</w>
<w><ana lex="" gr="V,indic,pl,3p,praet,finit,act"></ana>проповѣдаху</w>
<w><ana lex="" gr="S,acc,n,sg"></ana>блХгочсХтыѥ</w>,
<w><ana lex="" gr="S,ins,n,sg"></ana>словомъ</w>
<w><ana lex="" gr="V,nom,m,pl,brev,praes,partcp,act"></ana>учаще</w>,
<w><ana lex="" gr="S,ins,n,sg"></ana>дѣломъ</w>
<w><ana lex="" gr="ADV"></ana>же</w>
<w><ana lex="" gr="CONJ"></ana>и</w>
<w><ana lex="" gr="S,ins,n,sg"></ana>житьмъ</w>
<w><ana lex="" gr="S,acc,n,sg"></ana>слово</w>
<w><ana lex="" gr="V,nom,m,pl,brev,praes,partcp,act"></ana>утвержающе</w>.
</se>
```

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation

Unterstützt von / Supported by Alexander von Humboldt Stiftung/Foundation

El'Manuscript 2021

Textual Heritage and Information Technologies

8th International Conference

Freiburg im Breisgau, Germany, 12.04.2021 – 15.04.2021

Booklet of Abstracts

Funded by

DFG Deutsche Forschungsgemeinschaft
German Research Foundation

Unterstützt von / Supported by

Alexander von Humboldt
Stiftung/Foundation

V.i.S.d.P.: Achim Rabus

Organizing Committee: Achim Rabus, Victor Baranov, Heinz Miklas, Aleksandr Moldovan

Editors: Juliane Besters-Dilger, Achim Rabus

Sponsors: Deutsche Forschungsgemeinschaft, Alexander von Humboldt-Stiftung, Albert-Ludwigs-Universität Freiburg

Technical redaction: Kathrin Eisel, Melanie Miller

Contents

Ilia Afanasev

(Saint Petersburg State University)

The Old Church Slavonic Corpus Development: Preprocessing Stage

Keywords**:** Old Church Slavonic corpus, preprocessing, tokenization, stemming, lemmatization, PoS tagging, corpus development

The present paper deals with the problems and techniques of Slavonic historical corpora development. In general, corpus development faces two challenges: linguistic material preprocessing and putting corpus data in the public access. This paper focuses on the first task, presenting a WinForms application, which preprocesses the Old Church Slavonic texts from the collection of TITUS (*Thesaurus indogermanischer Text- und Sprachmaterialien*).

Firstly, the application extracts the required text from TITUS. Then it performs reversed transliteration from the Latin text into Cyrillic (using the special transfer rule system), tokenization, stemming and lemmatization. Finally, PoS tagging is performed, and it is explained why the rule-based approach for it is preferred over the statistical one, and how fine-tuning might provide help in the future.

The result of all preprocessing stages is placed in a database, from where it can be extracted when needed by corpus users. Thus the preprocessing stage of the Old Slavonic corpus creation is executed. The application performance is demonstrated on the Prague Fragments (TITUS) material.

Marina Aksenova

(Minin Nizhny Novgorod State Pedagogical University)

Literary Analysis of Texts in the Context of Changed Orthography and Grammar

Keywords: literary texts, Russian literature, grammar, orthography, analysis

One of the problems connected with older texts is associated with grammatical and orthographical norms that were modified with time. Written Russian language underwent a number of reforms concerning, for example, the correct spelling of the inflections, use of commas. The original text may look rather archaistic for today's readers and so the energy of it may be partially lost. Stylistical association with old-fashioned style may become stronger than the ideas expressed by the author. The paper also tackles the problem of capitalization. Nouns denoting nationalities (Englishman, Frenchman etc.) are capitalised in older texts and the question may arise: is it simple following the grammar rules or the author's desire to make the words more prominent when he speaks about national identities? The research demonstrates the problem by the analysis of Russian printed literary texts of the beginning of the XIX century. The paper discusses how discordancy in language norms should be treated in modern editions of older texts, pros and contras of adaptation to current rules.

Artem Andreev

(Institute for Linguistic Studies, RAS)

How Can Textual Corpora Benefit from Distributed Computing Methods?

Keywords**:** distributed computing, morphological annotation, textual corpora, TEI, digital rights management

There are tasks in corpus linguistics such as unsupervised morphological tagging that are pretty computation-intensive and may therefore benefit from using distributed computing tools. However, existing systems are either strongly oriented towards classical number crunching or are overly generic. Meanwhile, distributed processing in the context of digital humanities poses its unique challenges. First, unlike in natural sciences, the source data need to be passed through several alternative processors and the results should be collated and compared, which also implies automatic conversion between various formats and tag conventions. Second, the data from the corpora may themselves be copyright-protected or contain sensitive personal stories, therefore the distributed system should do it utmost to prevent irresponsive spreading of such data. Even in cases when such issues do not arise, keeping track of provenance is still crucial for digital humanities. In the talk a novel system for distributed corpus processing will be discussed that is

aimed to solve the above issues. The system is implemented in Erlang and relies on detailed TEI markup for its operations. Digital signing and encryption are being employed to ensure proper data distribution.

Alexandre Arkhipov

(Hamburg University, Lomonosov Moscow State University)

Using HTR on Bilingual Evenki-Russian Manuscripts of 1910s

We report on applying Handwritten Text Recognition (HTR) to manuscripts from the archive of Konstantin Rychkov preserved at IOM RAS, St. Petersburg [Рычков n.d.], within the INEL project [Arkhipov and Däbritz 2018]. Folklore texts in Evenki (Tungusic) were collected in Western Siberia in 1910s.

We used services provided by the Transkribus platform [Kahle et al. 2017]. The necessary step of layout analysis proved to be time-consuming due to the organization of the parallel Evenki-Russian text on the page without following a strict separation line. HTR models have been so far trained on 24, 50 and 100 pages (5172 lines, 21204 words).

The best Character Error Rate (CER) attained on validation data is 7.97%. The distribution of errors is non-uniform: most errors are due to just a few problematic issues, especially diacritics. One of them is the acute accent marking stress, which is written quite high above the line, making it fall outside of the line area detected by the preprocessing algorithm. After excluding the acute from training data and recognition, the CER dropped to 4.99%. Ways of bypassing this limitation are being investigated.

Further issues discussed include other diacritics, punctuation, capital vs. small letters, and the usefulness of text2image processing tool.

References:

Arkhipov A. V.and Däbritz C. L. 2018. Hamburg corpora for indigenous Northern Eurasian languages // Tomsk Journal of Linguistics and Anthropology. Issue 3 (21). P. 9–18 (DOI: 10.23951/2307-6119-2018-3-9-18).

Kahle, P., Colutto, S., Hackl, G. and Mühlberger, G. 2017. Transkribus – a Platform for Transcription, Recognition and Retrieval of Document Images // IAPR International Conference on Document Analysis and Recognition (ICDAR), IEEE. P. 19–24 (DOI: 10.1109/ICDAR.2017.307).

Рычков К. М. Образцы материалов по изучению тунгусского языка и фольклора // Архив востоковедов ИВР РАН, Ф. 49, оп. 1, ед. хр. 5. — 322 л. (www.orientalstudies/ru/rus/images/stories/archives/tungus49_1_5.pdf)

This paper has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

Irina Azarova | Elena Alekseeva | Alexey Lavrentev | Elena Rogozina | Konstantin Sipunin

(St Petersburg State University | St Petersburg State University | Centre national de la recherche scientifique, Institut d'Histoire des Représentations et des Idées dans les Modernités, IHRIM | St Petersburg State University | St Petersburg State University)

Content Structuring in St Petersburg Corpus of Hagiographic Texts (SCAT)

Keywords: diachronic text corpus, Old Russian hagiography, TXM technology, content structuring, biblical quotations.

St Petersburg Corpus of Hagiographic Texts (SCAT) has launched two new mark-up formats. The first novation is the comprehensive format developed for the division of hagiographic texts into parts, both explicitly marked as section headings or extrapolated through comparison with texts of the similar genre. The second innovation is an elaborate format of representing the full range of various types of biblical, patristic and liturgical quotations, occurring in the lives of saints. For the time being, three morphologically annotated manuscript texts have been marked up according to these guidelines, and we are planning to add two more texts in the near future. Close cooperation with the IHRIM

research laboratory (Lyon) and wide use of their techniques and technology makes it possible to obtain some illuminating cross-format statistical data and thus offer new insights into the canons and rules of the Old Russian hagiography.

Victor A. Baranov | Roman M. Gnutikov

(Kalashnikov Izhevsk State Technical University | Udmurtia State University)

An Experiment in Eliminating Variation of Data of the Slavonic Historical Corpus to Facilitate Search, Demonstration and Statistical Analysis

Keywords: historical corpus, search and demonstration of data, linguistic statistics

Depending on the principles of publication of a medieval manuscript or the goals for linguistic investigation, a text can be subject to various degrees of unification and/or normalization: from use of textual forms with consideration of all graphic, orthographic and paleographic peculiarities, to not representing graphic variants and reduction of words to the normalized initial form in apparatus.

In the Slavonic historical corpus "Manuscript" (manuscripts.ru) several methods are used for eliminating the variation of the transcription of texts prepared as accurately as possible. These methods enable various techniques: search and demonstration of data in specialized modules, automatic morphological tagging and lemmatization, and preparation of data for quantitative and statistical analysis. When searching in the corpus, diacritics and titlos are ignored, variant letters are replaced with basic ones, and letters in certain positions or combinations of letters are equated; when displaying search results, various types of graphic-orthographic accuracy (from the original transcription to the transliterated form of words) are used; when lemmatizing and encoding morphology, normalized dictionary forms of words are used; etc.

Use of the modules, ensuring quantitative and statistical evaluation of linguistic units, required creation of new procedures for unifying textual forms, which resulted in reduction of textual forms to one and only one lemma. This paper will show the ways of achieving this goal and present examples of the preparation of data of various degrees of accuracy relative to the original in order to solve various linguistic tasks.

This study is financially supported by the Russian Science Foundation, the project "Distributional-quantitative analysis of semantic changes based in large diachronic text corpora" (project no. 20-18-00206).

Victor A. Baranov | Oksana V. Zuga

(Kalashnikov Izhevsk State Technical University | Udmurtia State University)

Correlation and Cluster Analysis of Fragments of the Earliest Slavonic Gospels

Keywords: Slavonic manuscripts of Gospels, linguistic textual criticism, correlation, cluster analysis

One of the tasks in the field of textual criticism of the Slavonic Gospel is finding efficient methods of classification of the many manuscripts. The majority of scholarship on this topic uses textologically significant variants (variant readings) of contexts, available both in short and complete aprakoses and the four Gospels, as material for studying the topic.

We present the results of a correlation analysis, followed by a clustering of the fragments of several gospel copies corresponding to each other from the historical corpus "Manuscript" (http://manuscripts.ru/mns/portal.main?p1=9&p_lid=1&p_sid=1). Lists of textual forms (bag of words) of the codices parts which match each other textologically (relating to the same cycle of readings written by the same Evangelist) were juxtaposed, and their correlation in pairs was evaluated. The results are used to find the degree of closeness (distance) of the fragments of this type between the manuscripts.

The experiments are based on the following theses: that the text of the compiled manuscripts cannot be studied without consideration of their textual composition, that the degree of closeness of the copies should be revealed, among other things, on the basis of linguistic analysis of textually compared texts (fragments) of different codices, and

that the statistical methods can be applied for comparison of the quantitative characteristics of the lists of words extracted from the fragments of different manuscripts matching each other.

The experimental part of the work was supported by the Russian Foundation for Basic Research (RFBR) in the framework of the project "A linguistic statistical analysis of one- component and multicomponent lexical units of the historical corpus Manuscript" (grant No. 18-012-00463).

Juliane Besters-Dilger | Achim Rabus

(Freiburg University)

Strengths and Weaknesses of Neural Morphological Tagging for Slavic

The neural network tagger CLSTM (described in Scherrer, Mocken and Rabus 2018) has been applied to the Old Russian Zhitie Evfimiya Velikogo (GIM, Chud. 20), translated in the second half of the 14th century. We thank the Vinogradov Russian Language Institute of the Russian Academy of Sciences for providing us the text in Unicode. The strengths of this tagger consist in its ability to automatically annotate an orthographically non-normalized text with dozens of pages within a few minutes, yielding a high accuracy with respect to part of speech and morphological features. Moreover, the tagger is capable of disambiguating case syncretism to a large extent, even in split constructions. Manual correction of the automatic tagging will result in a correctly tagged text considerably faster than when using a rule-based tagger or tagging completely manually.

The weaknesses of the CLSTM-tagger comprise certain examples of wrong POS-tagging, sometimes incomplete attribution of morphological categories to some parts of speech and incorrect analysis of certain morphological categories. Ligatures, superscript letters and punctuation can pose special problems, normalization of punctuation will achieve better tagging results. The rate of correct tags is higher when the token is known from the training data; unknown words (OOV) show a higher error rate.

In the paper, we analyze the strengths and weaknesses of the tagger by providing specific examples. Furthermore, we discuss possible ways of using automatically tagged, un-corrected data for quantitative analysis.

Reference:
Scherrer, Y., Mocken, S., and Rabus, A. 2018. New Developments in Tagging Pre-Modern Orthodox Slavic Texts // SCRIPTA & E-SCRIPTA 18. P. 9–33.

Elias Bounatirou

(University of Bern)

When *telefon* Became *brzoglas*: The Making of Fascist Croatian (1941–1945)

Keywords: Croatian, fascism, language politics, digital text collation, digitization, corpus linguistics

The paper will present results from a research project that analyses features of the Croatian language during the fascist Ustaša regime (1941–1945), and particularly characteristics occurring in fictional prose. The project focuses especially on the novel "Giga Barićeva" by Milan Begović (1876–1948) and the history of linguistic censoring of the text. Unfortunately, Croatian during fascism has so far hardly been investigated due to a research taboo.

In my analysis, centring on the final part of "Giga Barićeva", I will show that fascist Croatian exhibits specific (novel) features, which are even observable in editions of fictional prose, published during the Ustaša reign. In doing so, the presentation will also draw attention to the special interplay of philological (textological) and digital research methods the proposed approach involves. Amongst other things, this includes the process of digitization of the text sources to be analysed and the digital collation of editions of "Giga Barićeva" published before fascism with those that appeared during the Ustaša regime.

Simon Brenner | Fabian Hollaus | Patricia Engel | Heinz Miklas | Manfred Schreiner | Wilfried Vetter

(CIMA, Vienna)

Interdisciplinary Analyses of the Codex Marianus, Vienna Part (Cod. Vind. slav. 146)

Recently the Russian State Library in Moscow finished the restoration of the Old Church Slavonic-Glagolitic Codex Marianus (F. 87, No. 1689), which will now be available for research again. We take the opportunity to acquaint the scientific public with some results of our hitherto research on the Vienna part of this valuable source. Our investigations consisted of four parts: a codicological, multispectral, chemical and philological analysis.

The codicology was examined by our restoration-conservation-specialist P. Engel only recently and was to get as much information as possible about and from the writing material: the source of the parchment, methods of preparation, writing process, deletions and the condition of the fragment. Remarkable are the general good quality of the parchment, probably made from goat, later renewed letters because of water damage, three different reds (one of them as a secondary layer on the brown original ink, the other the bluish red applied throughout the manuscript, partly as the typical "wash" across the relevant letter), deletions concerning entries of all ages (a Glagolitic marginal, Cyrillic as well as modern cursive notes). Further observations showed signs of exposure to fire, impact of liquid, a worm hole, secondary wrinkles, old repairs etc.

The codicological examination was preceded by color and multispectral recordings, which F. Hollaus carried out in 2011. Its aim was to preserve the manuscript at its best and to provide an apt basis for all further investigations, because both the Vienna and the Moscow parts contain passages that are illegible to the naked eye.

The chemical investigation was also executed in 2011 with two portable spectroscopes (XRF and rFTIR, while Raman was not yet available and will be added soon). This complementary analysis served two aims: to get exact information on the parchment and the inks, paints and binders, and to collect data for a comparative study of parchment degradation and its relationship to artificially aged modern parchment. All the investigations were executed comparatively for Glagolitic and other manuscripts from different traditions. According to the received data the closest comparable results showed the Vienna Glagolitic Folia (Cod. Vind. slav. 136), deriving from the second quarter of the 12th c. and usually attributed to either Southern Croatia or Hum, i.e. the neighborhood of Duklja.

In the philological part we first analyzed the newly found Sinaitic manuscripts and then gradually prepared them for edition. Connections within the collection and with other Glagolitic manuscripts forced us to widen the scope of investigation, until finally we decided to classify all OCS-Glagolitic manuscripts. For this we prepared image-templates and questionaries which we filled in with the scribe data: image- and binary rows which allowed to easily confront the results. The most complex one was devoted to dating and localizing the sources by graphemic criteria. According to this method, the Marianus can roughly be attributed to the years 1012-1039, while its redactional area covers present-day Montenegro and its northern and Eastern neighborhood – in those days Duklja. This result is substantiated by the chemical data and a Latin letter on fol. 1 recto (already discovered by F. Repp about 70 years ago, but then forgotten). V. Jagić had already suspected a similar provenance, but could not support his hypothesis due to a lack of sources for comparison. Today, we can refer to the 2011 found Sinai Glagolitic Horologion-fragment from about the same time and area; and to a later writer who immortalized himself in three Glagolitic manuscripts and even revealed his name: the "sinful Dimitrij", who has recently been recognized as the first Serbian writer known by name.

These investigations have brought some more light on the history of the manuscript, but they have to be completed by similar analyses of the main part. This concerns mainly, but not exclusively, phenomena which lack in the Vienna part. Here are to be mentioned the later green additions to ornaments and two of the miniatures, because some Sinaitic Glagolitic and early Latin manuscripts show a similar green – unlike the other Glagolitic sources deriving from the Western Balkans.

Executed within two FWF-projects Nr. P 23133 and P 29892 (https://cvl.tuwien.ac.at/project/the-origin-of-the-glagolitic-old-church-slavonic-manuscripts/) as well as the CIMA-project financed by the Austrian Federal Ministry for Science and Research 2013 (https://cvl.tuwien.ac.at/project/cima/).

George Aaron Broadwell | Brook Danielle Lillehaugen | Xóchitl Flores-Marcial | Eloise Kadlecek | Felipe H. Lopez | Mike Zarafonetis

(Haverford College | University of Florida | California State University, Northridge | Bryn Mawr College | Haverford College | Haverford College Libraries)

Digital Editions of Endangered Language Texts as a Form of Pedagogy, Scholarship, and Resistance

Keywords: digital editions, Zapotec, manuscripts, digital scholarship, language activism

Zapotec communities are undergoing language shift to Spanish due to a violent colonial past and a systemically prejudicial present that positions Zapotec language as of less value than Spanish. In fact, some claim Zapotec isn't even a language or that it can't be written. While having a written form is certainly not a criterion for being a language, the truth is that Zapotec speakers have been writing their language for over 2,500 years. For example, there is a large corpus of alphabetic texts written in Zapotec during the Mexican colonial period.

In this talk, a diverse group of participants involved in an interdisciplinary digital scholarship project to create an online text explorer for these manuscripts show how the creation of digital editions of Zapotec language texts can serve multiple simultaneous roles. When used in classrooms, the project can have pedagogical value, teaching both DH skills and disciplinary content. Scholars involved in the work utilize both methods and results in their research. Finally, Zapotec participants show that re-framing Zapotec language and knowledge through this project is a type of resistance against discriminatory ideologies.

Gijsjan Brouwer

(Royal Netherlands Academy of Arts and Sciences – Humanities Cluster)

Docere: A Generic And Customisable Framework For Publishing and Archiving Digital (Scholarly) Editions

Keywords: digital edition, publishing archive, web infrastructure

In her paper Elena Pierazzo (2019) makes a case for 2 types of digital (scholarly) edition publishing: "Haute Couture" (custom build) and "Prêt-à-Porter" (generic). In short: adjusting the edition to your specific needs requires significant funding. The alternative is using an existing framework to publish which means adopting an encoding standard the tool is able to interpret. Imposing a standard reduces complexity for engineers, but at the same time limits the possibilities of representing a text in digital form. Docere aims to find the sweet spot between "Haute Couture" and "Prêt-à-Porter": a generic solution, however able to interpret any valid XML and highly customisable.

Several concepts underpin Docere's claims. First of all, the XML is the single source of truth, making Docere great for archiving purposes. To render any XML we make use of a preparation function and several data extraction functions. Functions are defined per project or encoding standard. All functions can be run in a web browser, visualising an XML document directly. Instant publishing is a great way to support scholars encoding practice. To render an XML document as an interactive webpage, tag selectors are mapped to web components, which are created per project or reused from other projects. Docere comes with an API and aims to meet the FAIR principles.

Docere is ready to be presented to the wider textual scholarship community and receive feedback for future development.

References:

Pierazzo, Elena. 2019. What future for digital scholarly editions? From Haute Couture to Prêt-à- Porter // International Journal of Digital Humanities 1.2. P. 209–220.

Daniel Bunčić

(Cologne University)

OpenType and the Diversity of Slavic Texts

OpenType is a technology by which fonts can handle complex problems of typography. In this presentation I will discuss a few of the more complex typographical problems with historical and contemporary Slavic texts and how OpenType can be used to solve them: diacritics and superscript letters, ligatures in all three Slavic alphabets, Cyrillic and Glagolitic numbers, and diachronic and diatopic variation of glyphs. I will also discuss limitations of OpenType and problems that should not be solved using OpenType.

Thomas Daiber

(Gießen University)

Orthographic doubts: grammatika and grammatikija

In Vita Constantini-Cyrilli VIII:10 it is related, that Cyril, having come to Kherson, learned the Hebrew language in oral and written form and translated also a grammar of Hebrew. The short time span, Cyril had at his disposal in Kherson for learning a new language and translating a grammatical treatise, had many researchers suggest, that the anecdote should best be explained by assuming, Cyril, when entering Kherson, already possessed some knowledge of Hebrew and only renewed or polished it there. Our paper tries to propose another solution by analysing the ortho graphy of two Old Church Slavonic lexemes "grammatika" and "grammatikija" on the ground of Greek orthography and its respective pronunciation. Do we really deal with a synonymous pair "grammatika/grammatikija" or is there reason to assign another meaning to the latter?

Chiara De Bastiani | Thomas Krause | Martin Klotz | Marco Coniglio | Heike Sahm | Anabel Recker | Jan Christian Schaffert | Svenja Walkenhorst

(Universita degli Studi di Verona | Humboldt-University of Berlin | Humboldt-University of Berlin | Göttingen University | Göttingen University | Göttingen University | Göttingen University | Göttingen University)

Building and Using a Parallel Corpus for Analyzing Translations of High and Low German Incunabula

Keywords: historical texts, literary studies, parallel corpus, corpus linguistics, software infrastructure

The Wiedererzählen im Norden (WiN) project aims to analyze translation strategies, both from a literary and a linguistic perspective, in Early New High German (ENHG) incunabula and their Middle Low German (MLG) translation. In a preliminary examination, MLG texts curiously reveal to be about 10–25% longer than their ENHG counterpart. The reason for this discrepancy may lie in the cultural adaptation of place or person names for the Low German audience, or in systematic syntactical differences, such as the use of full articles instead of their clitic counterparts.

The lexical, morphological and syntactic strategies involved in this translation process have not been analyzed thoroughly in the literature. A corpus of ENHG narratives and their MLG translation could capture both quantitatively and qualitatively the strategies employed in the translation process. Tools like ANNIS [1] can query such parallel corpora, but creating them is not straightforward. Our approach combines the proven tools EXMARaLDA [2] and Pepper [3], but extends the workflow to support token-level text alignment. We present a feasible approach for projects without large infrastructural support and a parallel corpus of eight aligned texts alongside its annotation and investigation.

References:

[1] Krause, Thomas. 2019. ANNIS: A graph-based query system for deeply annotated text corpora. Ph.D. Thesis, Humboldt-Universität zu Berlin, Mathematisch- naturwissenschaftliche Fakultät.

[2] Schmidt, Thomas, and Wörner, Kai. 2014. "EXMARaLDA." In The Oxford Handbook of Corpus Phonology.

[3] Zipser, Florian, and Romary, Laurent. 2010. A Model Oriented Approach to the Mapping of Annotation Formats Using Standards. In Workshop on Language Resource and Language Technology Standards, LREC http://hal.archives-ouvertes.fr/inria-00527799/

Tsvetana Dimitrova

(Institute for Bulgarian Language, Bulgarian Academy of Sciences)

On Sentence Segmentation in (Historical) Corpora

Keywords: historical corpora, sentence segmentation, linguistic annotation, syntax, discourse

The paper addresses the issue of sentence segmentation of medieval Slavic texts (in the meaning of boundaries of the sentence unit which may or may not differ from the boundaries of a clause unit and a statement unit) that lack explicit markers for the beginning and the end of a sentence. This segmentation is also one of the main steps in text segmentation for the purpose of linguistic annotation (and syntactic annotation, in particular).

Different approaches to be discussed take into account the syntactic structure (the presence of a subject – explicit, implicit, null subject, etc.; the presence of a finite verb form; elements that appear in specific (more or less fixed) positions within a sentence, a clause or a phrase – such as the clitics) and the semantic content (and discourse structure).

Dmitriy Dobrowolski

(Higher School of Economics Moscow)

Style of Old Russian Sermons in the 11–12th Centuries: A Quantitative Approach

Despite their "traditional" content Old Russian sermons represent the mindset of the society they circulated in and thus form a promising kind of historical sources. However, it is difficult to interpret Old Russian sermons as cultural phenomena, because we know little about authors and circumstances under which these texts were created. And as the possibilities of traditional attribution techniques are in many cases exhausted, we have to resort to the formal analysis of the style to determine the date and authorship of the studied texts.

Most of the sermons in question are too short to apply classical methods of stylometry. However, some of them prove to have distinctive linguistic features (eg. high frequency of particles *bo* and *li*), which allow to determine a style, characteristic for the texts written in the Caves monastery in the 11th century. In my talk I plan to concentrate on the attribution of Sermon on Christian and Latin Faith and The Word on God's Punishments (both allegedly by St. Theodosius of the Caves), and discuss the computer tools applicable to stylistic analysis of Old Russian sermons and the historical meaning of the results obtained with those tools.

Quinn Dombrowski

(Stanford University)

From Annotation to Modeling: Computational Horizons for Medieval Slavic Studies

For over 30 years, digital tools and methods have held the promise of transforming medieval Slavic studies. Online publishing, databases, and dictionaries have significantly lowered the barrier for scholars in the field to work more efficiently and collaboratively. Manuscript materials with rich annotation and commentary are now freely available online, for anyone to access, rather than being bound up in expensive, limited-run print volumes. And yet, while digital resources have impacted how Slavic medievalists access and produce scholarship, the methods and outputs are still fundamentally similar: painstaking, detail-oriented close reading and analysis of small numbers of texts. Until recently, the cost to digitize and transcribe a manuscript text was significant, and a close analysis is a way to make use of that investment of time and resources. But as machine transcription tools like Transkribus enable an influx of newly-

digitized texts, close reading is no longer feasible for every text we have access to. This talk will survey recent developments at the intersection of natural language processing (NLP) and medieval studies, considering what would be involved for an international community of medieval Slavists to build infrastructure to support similar work.

Nataliia Drozhashchikh | Elena Efimova

(University of Tyumen)

Lemmatization for Old English: Building Text Analysis Application

The present paper focuses on a software performing task within the framework of computational linguistics and NLP. It aims at programming Old English lemmatizer for fundamental and applied research in diachronic studies. We provide evidence for the automatic morphological analysis, review the existing tools for text analysis, and build our own application, encompassing three functions – lemmatization, stemming, and POS tagging.

We have not succeeded in finding any appropriate applications for Old English morphological analysis. To build a single page application (SPA) we propose to implement an NLP model of a bidirectional LSTM as the effective method for lemmatization of Old English texts. This method may prove useful since it combines language-specific instruments and novel computational tools. The model will be realized in Old English lemmatizer (OEL, see an attached SPA preview). Lemmatization model will be trained on The York-Helsinki Parsed Corpus of Old English Poetry and The York-Toronto-Helsinki Parsed Corpus of Old English Prose. Besides lemmatization, the program will embed stemming and POS tagging. The data for the application are the lexemes collected from the open resources (Wiktionary) and the digital edition of the Bosworth-Toller Anglo-Saxon Dictionary developed by Ondřej Tichý (http://bosworth.ff.cuni.cz). The source code will be available on GitHub.

Lemmatization is an algorithmic process that enables researchers to extract the basic forms of a word, group similar orthographic variants, and assort inflected forms. Lemmatizer is a computational tool relevant for corpus analysis. As such, lemmatization belongs to automatic morphological analysis tasks that represent the biggest challenge facing NLP specialists. In modern linguistics, corpus analysis resources for lemmatization are considered efficient and well-developed. Most of them work as the rule-based systems that are not always applicable for languages at earlier stages of development and low-resource languages. Typically, such languages are morphologically rich and phonologically inconsistent. Lemmatizers for such languages are not accurate. That poses a serious challenge for NLP and computational analysis. For example, Old English lemmatizer presented at http://cltk.org/, though fits many various lexical and morphological tasks, unpredictably disregards certain chunks of vocabulary. It regularly misses lexemes typical for Old English vocabulary.

In the pilot study we have faced many problems determined by the lack of free dictionary resources. In future, we plan to measure the accuracy of the lemmatizer and consider the problem of word sense disambiguation.

Hanne Eckhoff

(Oxford University)

Aligning the Psalterium Sinaiticum with the Greek Psalter

In this paper I describe the work on automatically aligning the Psalterium Sinaiticum with the Septuagint Psalms in the Tromsø Old Russian and OCS Treebank (TOROT). I will briefly account for the transcription, text processing and manual annotation of the Psalterium Sinaiticum itself, and then describe the choice of Greek text (Swete) and the automatic lemmatization and morphological analysis. Next the algorithm for automatic token-level alignment of texts is briefly described, and the success rate calculated and analysed. The results seem quite good from a quantitative perspective (over 90% accuracy in most cases), and it may seem tempting to try to use the data directly. However, a pilot study of aspect in the Greek and OCS text shows that the automatically processed Greek parallel text leads to considerable data loss, and that much manual sifting of apparent mismatch examples is necessary to arrive at a preliminary analysis. In a low-resourced historical language such as Old Church Slavonic we cannot afford working with this amount of noise and data loss. We can use automatic tagging and alignment to ease our workload, but we have to manually post-correct the output.

Kazuyuki Enami | Yoshihiro Okada | Taketoshi Hibiya | Satoru Sato | Takashi Yokota | Shigeru Sawayama

(Ryukoku University, Kyoto | Ryukoku University, Kyoto | Prof. Jissen Women's University, Tokyo | Prof. Jissen Women's University, Tokyo | Prof. Jissen Women's University, Tokyo | Prof. Jissen Women's University, Tokyo)

Scientific Study of Paper Used for Ukiyoe Pictures Published in the Edo-Era by High-Resolution Digital Microscope

Keywords: Ukiyoe, price of Ukiyoe, kozo and mitsumata mixed paper, mitsumata fibre, rice straw fibre, rice powder

Paper used for over 50 works of Ukiyoe published during 1780s to 1860s of artists, Hokusai, Toyokuni, Hiroshige, Kunisada, Kuniyoshi and others was analysed using digital microscope Keyence VHX-5000. It was found that paper used for all works was mixed paper Kozo (*Broussonetia kazinoki*) with Mitsumata (*Edogeworthia papyrifera*) or rice straw fibres, filled with rich amount of rice powder, different from high price thick pure Kozo paper "Hosho" generally believed to be used for Ukiyoe printing. Ukiyoe is world collectors' item today, but the price of an Ukiyoe picture at that time was very cheap only 480 ~720 JPYen equivalent in today's price level. For them to use higher price (400JPYen or more/sheet; as above) "Hosho" paper was absurd. Above mentioned mixed paper used for Ukiyoe is never lower grade paper, but suitable paper for printing 20 or more colours. Fine (8~15μm width), curved and twisted Mitsumata or rice straw fibres effectively fill opening between straight Kozo fibres (16~30μm width) and rich rice powder makes paper much whiter. Ukiyoe publishers chose cheaper and suitable paper giving ordinary citizens opportunities to access easily full colour Pop arts.

Jürgen Fuchsbauer | Fabio Maion | Lara Sels

(University of Innsbruck | KU Leuven)

Towards a Database of Older Slavonic Literature: First Steps Towards the Digitisation and Publication of Francis J. Thomson's Card Index

Keywords**:** Database, Handwritten Text Recognition, Incipitarium of Older Slavonic Texts, Normalization of Church Slavonic

During 6 decades of research work Professor Francis J. Thomson collected an incredible amount of detailed information about older Slavonic literature, which he laid down in his famous card index. In order to make this huge collection accessible to the present and future generations of researchers, the cards shall be digitized, tagged, and made available on-line. In a second step, we plan to apply HTR software (presumably Transkribus) to the scanned images, thus laying the foundation for a digital database of older Slavonic literature.

In our paper we shall give an approximate overview of the extent of Thomson's card index as well as of the thematic fields it consists of. One of them is an incipitarium of older Slavonic texts. This shall be the first part to be edited and published. The entries in the card index, like the incipits themselves, are not normalized with regard to grammar and orthography. In spite of the possibility of fuzzy search, normalization is a prerequisite for a searchable database. We therefore consider it necessary to discuss in advance the options for normalizing Church Slavonic phrases of diverse origins with a larger audience, which we would like to do at El'Manuscript 2021.

Oksana Gorban | Marina Kosova | Elena Sheptukhina | Andrey Svetlov

(Volgograd State University)

Don Cossack Army Official Documents of the 18-19th Centuries: Compiling a Linguistic Corpus

Keywords: historical corpus, official documents, Don Cossack Army, metadata annotation, morphological markup, structural markup

The report summarizes the results of the study (project 19-012-00246, supported by the Russian Foundation for Basic Research), which solves a number of linguistic issues associated with building a diachronic linguistic corpus of official documents of the Mikhailovsky Stanichny Ataman Fund (GAVO, Volgograd, Russia).

The necessity of definitive presentation of documents is expanded.

A list of relevant metadata parameters of the documentary fund, enclosing the title, addresser, addressee, location and date of compiling, place and date of delivery, authenticity, is defined.

Verbal markers that are necessary and sufficient for automatic meta-marking of the document type, addresser, addressee, and others are detected and used for exemplifying some results of the test mode of the Python application developed by the project participants, in particular the document type. The application uses a regular expression library to analyze texts and detect certain patterns.

The report describes the textual features that make it difficult to divide the text, and therefore, perform the morphological and structural markup, including archaic and dialect words, grammatical forms that lack the modern Russian standard language, spelling variants, syntactic specificity and the absence of punctuation marks. The directions of the software product adaptation for the morphological and structural markup of the texts in the documentary fund are provided.

Lena Henningsen | Duncan Paterson

(Freiburg University)

Manuscript Culture during the Cultural Revolution

Keywords: CJK, TEI, PRC, XML, Cultural Revolution, Fiction, Exist-db

Description: Digital Scholarly Editions of entertainment novels in manuscript form created in China during the 1960s

The conceptual tools of manuscript studies are rooted in the study of medieval European objects. The circulation of hastily and clandestinely produced manuscripts of entertainment fiction (手抄本 shouchaoben) shaped the transmission of literature in the PRC. Moreover, the materiality of these sources, including the conditions of their production, circulation and consumption significantly impacted on the process in which the textst were (re-)created, often reflection the scribes experiences during the Cultural Revolution. Presenting the story through diverging witnesses, illuminates the creative rewriting that occurred as the text circulated. Our editions make these works available to researchers interested in the textual heritage of PRC literature. Using the relatively short Three Times to Nanjing as a case study, we present our editorial pipeline for processing and publishing tei-xml editions of CJK manuscripts. This allows us to critically explore points of tension between East-Asian codicological conventions and existing standards. We include a discussion of our code contributions to exist-db, and the TEI Guidelines targeting editors of non-Latin source languages.

Kameliya Hristova-Yordanova

(Bulgarian Academy of Sciences)

Electronic Descriptions of Slavonic Manuscripts at the Bulgarian National Library "St. St. Cyril and Methodius" – Problems and Perspectives

Keywords: Slavonic manuscripts, National Library "St. St. Cyril and Methodius", MARC, COBISS, electronic description.

The paper deals with attempts to provide the public with electronic descriptions of Slavonic, Greek, Latin, etc. manuscripts held at the Manuscripts and Rare Books Department of the Bulgarian National Library "St. St. Cyril and Methodius". Currently archivists are working on the manual input of data regarding the manuscripts in question into the Co-operative Online Bibliographic System & Services (COBISS). Several problems are to be presented and discussed at the conference, for instance what characteristics of manuscripts are worth being provided in the aforementioned system and which ones are unnecessary since there are accessible published descriptions of most manuscripts. The public will be acquainted with the most recent work regarding the matter.

Angelina Kalashnikova

(Saint Petersburg Institute of History of the RAS)

Infrared Digital Scanning of the 15th – the First Half of the 16th Centuries Russian Judicial Documents' Watermarks

Keywords: land trials, watermarks, infrared photography.

Russian judicial charters surviving from 15th – 16th c. were detailed court records; protocols of judicial procedure that contain information about judges, litigants and the subject of the deed. The aim of the project is to collect the data-base of watermarks of Russian judicial charters. Infrared digital scanning of judicial documents allows one to obtain precise high- definition images of watermarks as well as paper's layout. Comparison of digital copies of judicial documents' and manuscripts' watermarks of the same monastery will show if the same paper was used in judicial process and monastery paperwork. In case watermarks will be the same, it is possible to presume that monasteries were engaged in the process of judicial paperwork and were interested that the process was written down. Moreover, comparison of watermarks of judicial documents produced by the same scribe and/or in the same year allow one to shed a light on the court bureaucracy and circulation of paper.

Sebastian Kempgen

(Bamberg University)

Unicode and Slavic Philology - Status quo and Open Questions

Unicode (now at version 13 [as of 2020] with a regular release on a yearly basis) has been a very stable basis for Slavic Philology during the past few years. From time to time, minor additions have made their way into the standard. However, not all areas can be considered to have been fully addressed. The presentation will feature observations from manuscripts and early printed books regarding the Cyrillic and the Glagolitic scripts and thus contribute and stimulate future proposals for additions to the standard.

Viacheslav Kozak | Anastasia Makarova | Diana Balashevich | Anastasiia Kharlamova | Andrey Sobolev

(Institute for Linguistic Studies, RAS | University Zürich / Institute for Linguistic Studies, RAS | Saint Petersburg State University | Institute for Linguistic Studies, RAS | Institute for Linguistic Studies, RAS)

Linguistic Description of the Glagolitic Fragment from the 14th c. (NLR, Berčić II 36)

Keywords: 14th century, Croatian Glagolitic script, missal, Church Slavonic, Old Croatian, diglossia

The paper discusses textual and linguistic features of the fragment of a Glagolitic missal from the collection of the National library of Russia (NLR, Berčić II 36, 14th c.) against the background of the general linguistic situation in Medieval Croatia. The manuscript contains a part of the *Proprium de tempore* from the Ash Wednesday (*Feria IV. cinerum*) to the 6th Saturday of Lent (*Sabbato post Domenicam de Passione*). The carried out analysis shows, that the liturgical texts of the missal (Scripture readings, psalms and prayers) are written in good Church Slavonic language of the Croatian recension with few vernacular (mainly orthographic) elements: cf. forms *častnêi* (*čьst-), *početakъ* (*počętъkъ), *priêti* (*prijęti), *otrenetь* (*otъgьnati*, 3sg), *bespečalnu* (f.ins.sg). The rubrics are written in Old Croatian, what can be shown on the example of the morphological forms of the present: cf. *dê* (*dêêti*, 3sg), *dosvrši* (*dosvьršiti*, 3sg), *dêlamo* (*dêlati*, 1pl), *dêmo* (*dêêti*, 1pl). At the same time the orthographic features of the rubrics are mainly Church Slavonic: cf. *kъda*, *vъz'ma*, *nedêle*, *prêkъlonête*. In general, the distribution of the allogenetic elements in the manuscript shows the interaction of the Church Slavonic and vernacular elements within the situation of diglossia.

The study was carried out on the electronic text of the manuscript, that has been created in the Internet application "bercic.stin", developed in the Old Church Slavonic Institute of Zagreb. The application at the current stage of development permits to perform the manual graphic and morphological markup of Glagolitic manuscripts, the search and the example selection within the marked up features. An electronic edition of the Glagolitic fragment is being prepared.

Kristijan Kuhar

(Old Church Slavonic Institute, Zagreb)

In the Quest for Theological and Liturgical Texts in Glagolitic Miscellanies (15th c.)

Keywords: Miscellanies, Glagolitic literature, Theology, Liturgy, Liturgical Textology

In the medieval Croatian Glagolitic Literature genre and typology of a manuscript entitled "Miscellany" is well known. Such manuscripts develop through the 14th and 15th century and contain various translations of Christian literature, mostly translated from Latin or Italian language. Literary contents of Glagolitic miscellanies are recently scientifically researched in the field of linguistics and the history of literacy. However, detailed analysis of liturgical and theological texts contained in such manuscripts is only bibliographically denoted, but not analysed.

Within this research, the author intends to make a quest for liturgical and theological texts in four Glagolitic miscellanies from the 14th and 15th century: Cod. Slave 11, Cod. Slave 73, Ms. Canon. Lit. 412 and Ms. Canon. Lit. 414. The goal of the research is the detailed denotation of theological and liturgical texts and their function in the miscellanies. The presumption is that liturgical texts in miscellanies are an appendix to the Glagolitic breviaries and missals, and theological texts are used as a literature for the education of Glagolitic Clergy.

This research is part of the project "Research of the Old Croatian Glagolitic Miscellany Heritage" (IP-2019-04-5942, Croatian Science Foundation).

Natalia Kuzemko

(Student of Saint-Petersburg State University)

Binding Stamping Research Methodology (Trasological Analysis)

Keywords: binding techniques, trasology, trace, digital modeling, trasological analysis

Binding embossing is an important factor in the attribution of historical books, according to which the scientist can determine a place, time of the creation and a workshop. The current problem of correlation between binding and a written document can be solved by creating a methodology for attributing binding to embossing. An analysis of the features of the tool provides accurate identification of the binding by embossing.

The main goal of my research is to combine existing developments into a common methodology based on the trasological analysis. At the heart of this approach is the perception of binding as a trace system and identification bookbinder skills. The research was based on the identification and analysis of groups of signs, such as signs of the tool and signs of its use, signs of a trace-recognizing object and signs of a technological model. Were applied digital modeling, trasological and identification study of stamp impressions on Old Russian manuscripts of the XV–XVI centuries.

The development of such an approach to the research of documents can be important in the treasure in the development of source study and restoration methodology.

Gerhard Lauer

(Basel University)

Do Manuscripts Have Style? Digital Stylometry for Manuscript Studies

Since manuscripts are available at large numbers at our fingertips, new, computer-based ways to edit and to explore manuscripts gain momentum. But medieval manuscripts could not be handled like today's language on Twitter, instead they need more thorough investigation informed by specific domain knowledge, but also by an increasing knowledge about quantitative approaches in the humanities and on stylometry in particular. The talk will discuss the opportunities and limits of digital stylometry for manuscripts studies

Alexei Lavrentiev | Liubov Kurysheva

(IHRIM Research Laboratory, CNRS, France | Institute for Philology, Siberian Branch, RAS)

A Bilingual Digital Edition of La Belle et la Bête and its Russian Translation by Kh. Demidova

Keywords: translation studies, digital editing, TXM, Russian literature of the 18th century, fairy tales, Russian-French literary contacts

This paper presents a digital edition of the manuscript of the first Russian translation of Leprince de Beaumont's The Beauty and the Beast fairy tale (1756), aligned to its French original. The translation was made in 1758 by a twelve-year-old girl, Khionia Demidova (1746-1792) and dedicated to her elder brother, and its original manuscript is conserved at the scientific library of the Saratov State University (no. 456). This document is interesting from several points of view: the "naive" translation made by a young girl allows us to understand how the French literature was perceived in the 18th century Russia, what aspects of the French language and socio-cultural phenomena of the Western Europe were difficult to understand, and how the socio-cultural phenomena of the Western Europe were perceived. The peculiarities of Khionia's spelling and punctuation provide data on her knowledge of Russian grammar and orthography.

A general overview of the digital edition is provided in (Kurysheva and Lavrentiev 2019). In this paper we focus on the technical side of the project: its workflow, dealing with irregular spelling, NLP tools and alignment to the French text. The edition is currently available on TXM demoportal (http://portal.textometrie.org/demo/?command=documentation&path=/LABELLE).

Reference:

Kurysheva L. A. and Lavrentiev A. M. 2019. The Story of Labelle and the Beast: a digital edition of a manuscript of the first Russian translation of the Beauty and the Beast fairy-tale powered by the TXM demo portal // Siberian Journal of Philology, 1. P. 54–61. DOI:10.17223/18137083/66/4 (in Russ.)

Alexei Lavrentiev | Elena Markova

(IHRIM Research Laboratory, CNRS, France | Belgorod State University)

Medea Project: Digital Editing and Analysis of Medieval and Renaissance Medical Texts in France

Medieval and Renaissance medical literature is a vast domain which provides valuable data for research in history of sciences, philosophy and linguistics (Nicoud 2007). However, high- quality digital editions of primary sources in this field are still rare. The Medea project (Medieval and Renaissance Medical texts: Editing and Analysis) aims at creating and investigating a corpus of digital editions for multi-disciplinary research on medical literature. The corpus will include original digital editions of primary sources (manuscripts and early printed books) and digitized critical editions. All texts will be encoded in TEI XML and available in an open archive. TXM free and open-source software platform (Heiden 2010 ; http://textometrie.org) will be used for both online publication and qualitative and quantitative analysis of the corpus. Additional resources, such as specialized vocabulary and proper names indexes will be linked to the editions.

The project is currently at an early stage, however the prototypes of digital editions of Guido Parato's *Régime de santé* (ca. 1460, *cf.* Lavrentiev *et al.* 2014) and Laurent Joubert's *Erreurs populaires* (1579) are already available on BFM and TXM web portals.

Marija Lazar | Michael Hoffert

(Saxon Academy of Sciences of Leipzig)

Constructing a Time Machine to Surf the Past: Compiling and Individualizing a Historical Legal Glossary

Keywords: e-lexicography, ontology, needs-adapted data, legal language, Slavic, German, Latin

Reassessing the role of terminology as an integrative part of digital ontologies (Jannidis et col. 2017, 164–166) has led to rethinking the form and the processing of terminology glossaries and databases as a substantial module in digital editing.

This paper presents the preliminary findings of compiling a legal multilingual glossary based on the text corpus of the German Law. The German (alias Saxon-Magdeburg) Law was a successful model of municipal legal codex, spreading from Saxony across nearly the entire East Central Europe from the 13th century onwards until the 19th century. This spectacular example of the transfer of legal culture was accompanied by translations into the East and West Slavic languages and by the creation and modernization of the legal terminology used there in. Against this highly dynamic background, it is not surprising to encounter diachronic and diatopic discrepancies in the orthographic representation of terms and in their semantic realization. This raises reasonable doubts about the appropriateness of the static representational forms mostly practiced therein – transliteration or transcription – as a sustainable lexicographic strategy (Sahle 2013/I, 243–253; cf. our criticism of [*Wörterverzeichnis der Rechtstermini*]; Bergenholtz 2013, 30-31). The lemmatizing of the terms, their TEI-encoding, and hyperlinking with the full-text editions are the steps required to adapt the glossary to the needs of different user categories, as well as to meet the claim of 'pluri-monofunctionality' (Spohr 2013, 103–104).

Reference:

Bergenholtz, Henning. 2013. Access to and Presentation of Needs-Adapted Data in Monofunctional Internet Dictionaries // Pedro Antonio Fuertes-Olivera, Henning Bergenholtz (eds.): *e-Lexicography. The Internet, Digital Initiatives and Lexicography*. London: Bloomsbury. P. 30-53.

Timm Lehmberg | Anne Ferger | Daniel Jettka

(Hamburg University)

Cross-Resource Exploitation of Manuscript and Spoken Language Data

Keywords**:** Language Documentation, Corpora, Cross-Resource Exploitation, Geovisualization, Graph Visualization

The contribution will present solutions for the collaboration and integration of manuscript data into deeply annotated spoken language corpora of lesser resourced languages, developed by the long-term project INEL ("Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages", see also Arkhipov/Däbritz 2018).

In the field of language-documentation traditionally exist considerable amounts of manuscript data (cp. Sanjek 1990, Sanjek et al. 2016), ranging from collections of transcriptions and field notes including sketches, (see Fig 1) to large scale collections of lists and index cards, containing vast amounts of lexical or ethnological knowledge (see Fig 2).

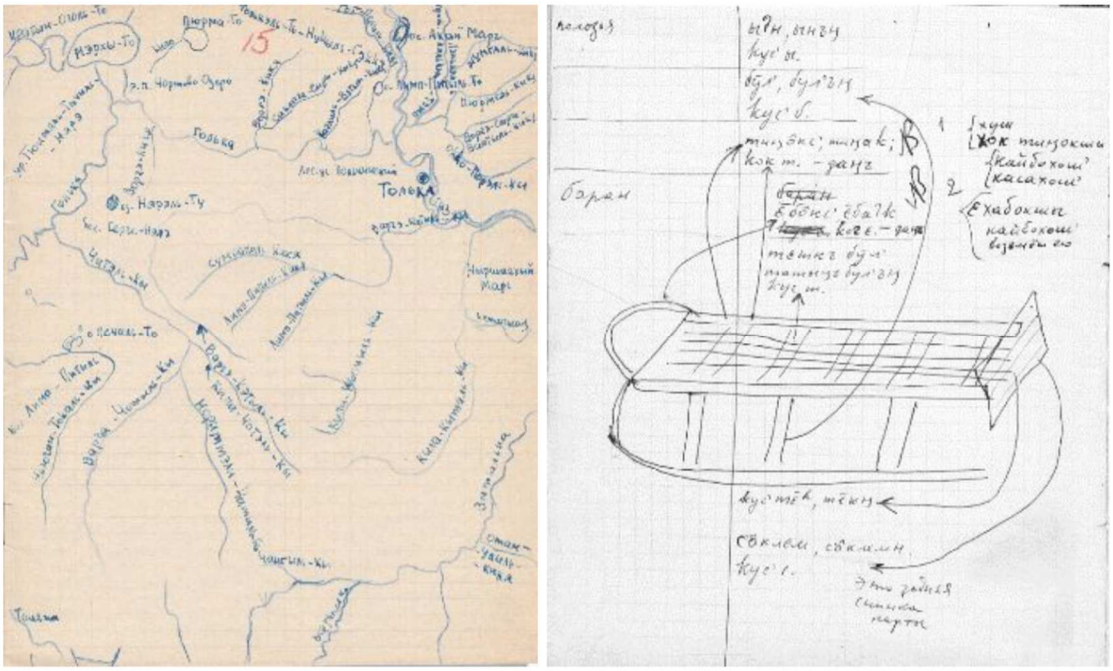


Fig 1: Manuscript fieldnotes from 'Kuzmina Archive' (volume 1, textbook 3, page 2) and Werner Archive (Ket-Yugh materials volume 2, page 3a) containing figurative information on toponymy and semantics

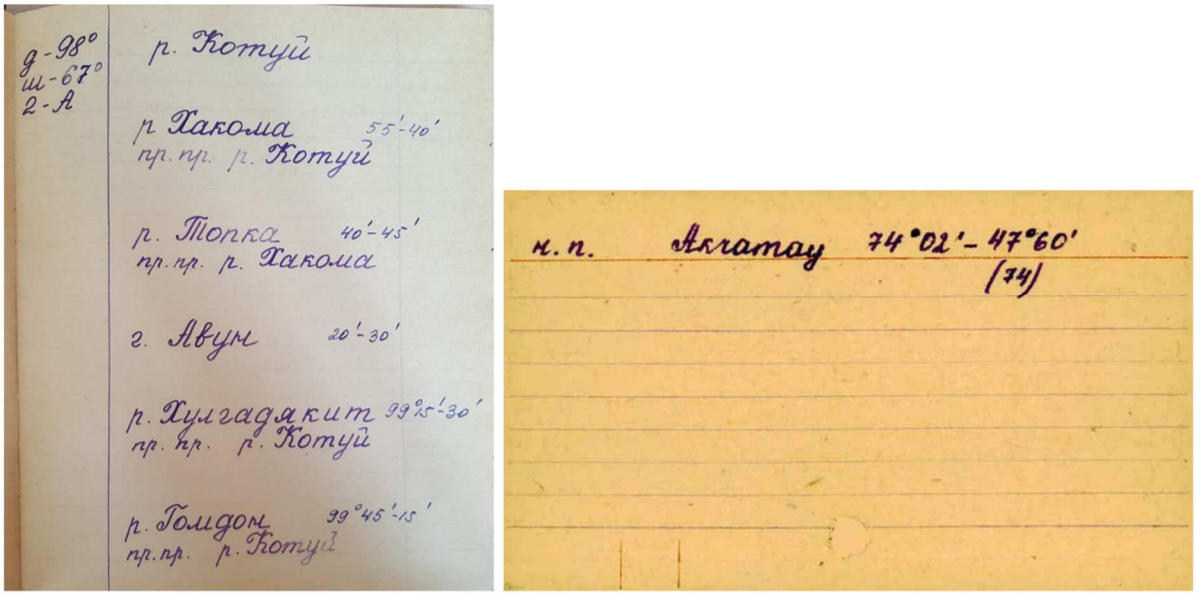


Fig 2: Index Cards and Manuscript notes from 'Tomsk Toponyme Archive' at 'Tomsk State Pedagogical University' (TSPU) holding geolocation name and spatial data.

Whereas in the case of object language data the aims of the INEL-project determine proven workflows and methods (cp. Wagner-Nagy et al. 2018, Hedeland/Ferger forthcoming) that include metadata enrichment, transcription, glossing and annotation, the curation of the peripheral manuscript resources has to be performed in a tension field between feasibility and best practice of manuscript research (cp. Jettka/Lehmberg forthcoming).

Therefore the solutions shown in the presentation aim at

long-term availability which includes digitization following best practise guidelines, metadata enrichment and long-term storage in sustainable repositories under open- access conditions,

persistent identification of manuscript resources including part identification with best possible granularity,

correlation and integration of relevant information which, among other things, includes the identification of implicitly connective information parameters and graphical visualization.

All methods will be illustrated with concrete examples.

References:

Arkhipov A. and Däbritz C. L. (2018). Hamburg corpora for indigenous Northern Eurasian languages // Tomsk Journal of Linguistics and Anthropology, (3). P. 9–18. (https://doi.org/10.23951/2307-6119-2018-3-9-18)

Jettka, D. and Lehmberg. T. (Forthcoming). Towards Flexible Cross-Resource Exploitation of Heterogeneous Language Documentation Data // Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020). Marseille.

Hedeland, H. and Ferger, A. (Forthcoming). Towards Continuous Quality Control for Spoken Language Corpora // International Journal of Digital Curation.

Sanjek, R. (ed.). 1990. Fieldnotes. The Makings of Anthropology. Ithaca, London: Cornell University Press.

Sanjek, R. Tratner, and Susan W. (eds.). 2016. Fieldnotes. The Makings of Anthropology in the digital world. Philadelphia: University of Pennsylvania Press.

Wagner-Nagy, B., Szeverényi S. and Gusev, V. 2018. User's Guide to Nganasan Spoken Language Corpus // Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology.

Yuan Li | Guanwei Liu | Shoju Ikeda

(Hokkaido University)

Issues around Glyph Creation of Undefined Chinese Characters in Kanchi'inbon Ruijumyogisho

This study focuses on glyph creation of undefined Chinese characters in *Kanchi'inbon Ruijumyogisho* 観智院本類聚名義抄 and aims to build a glyph list comprising the full headwords in the dictionary using GlyphWiki.

Chinese character dictionaries, such as *Kanchi'inbon Ruijumyogisho*, are important documents in Japanese linguistics and literature. As they have been preserved as old manuscripts, problems like variant characters and errata prevent headwords forming the framework of dictionaries from being expressed elaborately through Unicode CJK characters. The ownership of the documents prohibits their free publication and distribution.

A database named Integrated Database of Hanzi Dictionary in Early Japan (abbreviated HDIC, https://hdic.jp) is being constructed by the research group that the authors are affiliated to. Kanchi'inbon Ruijumyogisho database is a part of HDIC. The full-text has been deciphered and the data will be made public soon.

GlyphWiki is community-driven and stores structural information on Chinese characters. It creates and displays undefined Chinese characters in character-coded sets used in digitising Chinese text resources. It can generate font files and images from the glyph stored on the sever. A test version of our glyph list is available on the GlyphWiki group page (https://glyphwiki.org/wiki/Group:toyjack_krm-ninbu).

Olga Lyashevskaya | Dmitri Sitchinava | Maria Skachedubova

(Higher School of Economics Moscow| Russian language institute, RAS / Higher School of Economics | Russian language institute RAS / Higher School of Economics)

Lemmatization of the Middle Russian Corpus within the RNC: Choice of Solutions

Keywords: Middle Russian, lemmatization, proper names, hyperlemma

The morphological markup of the Middle Russian corpus is now augmented by the Universal Dependencies syntactic annotation, following the general lines and POS selection of this format. The lemmatization basically follows the lexical, phonetical and orthographical conventions of the Dictionary of the 11th–17th centuries. However, there is a challenge to resolve many issues that were not addressed at all or described inconclusively in the bulk of this edition.

We use the notion of hyperlemma that unites lemmas with different paradigms, whereas paradigms where just the representing word forms differ are analyzed as instantiating the same lemma (Nom. sg. красивой/красивый).

A telling example are proper names that are not described in the Dictionary but pose a list of problems of their own. We use and extend in particular the solutions proposed in the index of the Old Novgorod Dialect by A. A. Zalizniak. They include the normalization of dialectal morphology (as Уласке = Власко) and of the variativity of the kind Григорий/Григорей/Григорьй. An additional problem not addressed by Zalizniak are the syntactically complex proper names typical for Middle Russian (as Василей Петрова сына Ленина).

Olga Lyashevskaya

(Higher School of Economics Moscow, Vinogradov Russian Language Institute, RAS)

Universal Dependencies for Pre-Modern Russian: Morphology

We present the RNC Middle Russian corpus, a historical corpus of the Russian National Corpus, annotated according to the Universal Dependencies (UD v.2) scheme (Zeman et al. 2019).

Firstly, we outline the UD tagset for parts of speech and morphological features and discuss key differences between the UD and RNC tagsets. As an illustration, we address several annotation templates such as named entities and periphrastic verb forms.

Secondly, for the corpus of ca. 10 million tokens, we propose a model for incremental semi-manual annotation. Starting from a small number of 'seed' texts annotated manually from scratch and the remaining part annotated automatically by a 'noisy' model, we apply a correction strategy in order to:

(i) improve the quality of the training data up to the 'silver' level via correction templates; (ii) improve the consistency of manual annotations;

(iii) improve the accuracy of the tagging model; and

(iv) increase the amount of 'gold' data.

As a result, we gradually tip the balance between the poorly annotated training data and more and more reliably annotated test data in favor of the latter. Moreover, as the behavior of the accuracy metric clearly depends on the quality, amount, and representativeness of the test data with respect to the training set, our approach makes this metric more conservative yet more reliable in the task of tagging raw texts.

Ludmila Mazur | Oleg Gorbachev

(Ural Federal University)

Soviet All-Union Census of 1959 as a Source of Data for Family Reconstruction Studies: Challenges of Developing and Normalization of a Network Data Model

Keywords: All-Union Census 1959, network data model, primary data, census schedule, city family, Ekaterinburg-Sverdlovsk

In the State Archive of Sverdlovsk Region there are files containing schedules of the All-Union Census of 1959. Creation of the information resource on the basis of these primary data will require building a dataset that will allow researchers to analyze both explicit and implicit information that the data contain. A census schedule comprises an alphabetical list of residents of a housing unit (house or apartment), which can be used for family reconstruction studies. To systematize the primary data, we developed a data model combining the principles of individual- and household-level recording. The data model consists of two relational tables: 'Population' and 'Family'. The first table contains nominative information in accordance with the census program. The second table includes the information about the family composition, type, number of children, dependents and workers, and living conditions. The relationships between the tables enable a researcher to move from individual to household records and back, to select samples of families of certain types and, accordingly, individual records. The project is currently realized at the Ural Federal University and is aimed at developing a database to facilitate the use of historical data on the urban family in Sverdlovsk and Sverdlovsk region.

Elena Mikhailova

(The National Library of Russia)

Manuscript Monuments of the Secular Musical Culture of Peter the Great Times (Late 17th – First Quarter of the 18th Century): A Complex Study and a Digital Presentation Project

Keywords: Musical manuscript, Peter I, integrated research, digital presentation

The era of the reign of Peter I in Russia is a difficult transitional period in various fields. Secular genres are actively developing in musical culture, mainly the cant (three-voice song). There have been changes in the musical writing: five-line notation and special musical paper came into use.

Separate scholars — literary critics, musicologists, historians — turned to musical manuscripts from the time of Peter the Great, but with specific, local goals. A complex study, including musical and literary analysis, methods of codicology, paleography, filigree studies and other disciplines, has not yet been conducted.

The Manuscripts Department of The National Library of Russia holds many monuments of the musical culture of Peter the Great. A complex study of these manuscripts will provide an opportunity to get the most complete picture of these monuments. For the digital representation of musical monuments, it is planned to develop a special electronic resource.

Alexandra Milanova

(Institute of Balkan Studies & Centre of Thracology, Bulgarian Academy of Sciences)

Digital History in the Making: Case Study from Bulgaria

Nowadays, much of documentary and historical heritage of Bulgaria and the Balkans rests in vulnerable and often difficult to access archives. Some of these archives preserve information that may cast new light on historical figures, places, and events and may lead to their reinterpretation. Despite their great significance to scholars, teachers and general public, such collections are often at risk of being lost or damaged before the history they keep is retrieved, analyzed and used for further research.

Realizing the danger we face and our responsibility towards future generations, a group of researchers (including me) from the Institute of Balkan Studies at the Bulgarian Academy of Sciences joined forces with the librarians in order to document and publish online previously inaccessible and neglected archives from the history of Bulgaria and South-East Europe.

The paper will showcase the process of identifying, cataloguing, digitising, and contextualizing various archives. It will also present the results of several Digital Humanities projects that the Institute of Balkan Studies is currently implementing. Last but not least, it will demonstrate the range of materials documented. Many of them will be published for the first time, illustrating the potential these collections have to further our understanding of history.

Anissava Miltenova | Ivan I. Iliev

(Institute for Literature, Bulgarian Academy of Sciences | Institute for Literature, Bulgarian Academy of Sciences / Freiburg University)

Computer Corpora and Public Challenges

The presentation is an overview of the completed and widely used international and national projects for digital corpora with a main working team in the Department for Old Bulgarian Literature, Institute of Literature at BAS. In the last two decades they have been the following: South and East Slavs: Diversity and Interaction of Written Cultures 11th-20th c. ERA NET Plus (2017-2020) – Bulgaria, Russia, Belgium; Scripta Bulgarica: Digital Library of Medieval Bulgarian Literature (2014-2017), financed by Bulgarian Scientific Fund.

The authors outline the aims and scope of the projects and their results in the context of the needs of society and the challenges facing the humanities in our time – in education, research, museum and library work, the media, as well as knowledge of the history of literature over the centuries.

Ekaterina Mishina

(Vinogradov Russian Language Institute, RAS)

The Annotation of Verbal Aspect in Diachrony: Parameters, Algorithms and Problems

Keywords: verbal aspect, digital grammatical annotation, Old Russian, Old Church Slavonic

Digital annotation of verbal aspect in Old Russian and Church Slavonic texts is a challenging and complicated task that requires a complex approach. Studying modern Russian, we always know the aspect of a verb, whereas this is not the case in diachrony. The determination of an aspectual status (perfective, imperfective, biaspectual) of a particular verb for earlier stages is possible only after considering together different parameters such as: actionality, lexical semantics, morphology, functional distribution, syntactic restrictions, collocations, statistics etc. For an effective use of corpora all essential parameters should be annotated sufficiently to enable researchers to quickly collect the information necessary to build an aspectual profile of a verb. It is also important to understand the hierarchy of the parameters, as they might have different degrees of importance, and for this purpose a special algorithm should be developed. The preliminary results, related to the parameters of annotation, the algorithms and some annotated data (using 'Morphy', the System for Digital Morphological annotation of Old Russian and Church Slavonic Texts, developed at Vinogradov Russian Language Institute RAS) are going to be presented.

Alexandr Moldovan

(Vinogradov Russian Language Institute, RAS)

Textual Information in the Linguistic Corpus

Keywords**:** Old Slavic Manuscripts, Textual Criticism

Modern web-based databases reproduce all the features of hardcopy-based critical editions. A special section of the *TEI* provides recommendations on the coding and design of the critical apparatus module. This technology allows the user to add additional information about the text: spelling options, textual discrepancies, etc. All these features allow to comprehensively study a particular spelling. Nonetheless, the digital databases do not provide additional software that would employ the capacities of the user's computer to perform any type of the textual search.

However, the studies of the linguistic history of the text and the understanding of the linguistic patterns generally require not only information about individual spellings, but also about the spellings' positions in a number of other similar instances. To overcome this limitation, we need a universal search engine that would allow us to extract different types of hypertext data, making textual information available for automatic search.

A promising direction in the development of representations of such information is the technology of parallel corpora, which allows for a pairwise comparison of all copies of the text. The uniqueness of each copy is taken into account, and each copy is considered not only as a simple copy of the source text, but also as an independent text, reflecting the individual characteristics of the scribe and the general historical and linguistic context.

Klaus Müller | Aleksej Tikhonov | Roland Meyer

(Musterfabrik Berlin/Fraunhofer IPK | Humboldt-University of Berlin, Institute for Slavic Studies and Hungarology)

Scribe vs. Authorship Clustering in Historic Czech Manuscripts with LiViTo: A Case Study with Visual and Linguistic Features

Keywords**:** AI, Digital Humanities, machine learning, linguistics, author, scribe, Czech

Manuscripts in small archives may contain reports by personal witnesses, new information on everyday culture and language. Above all, the linguistic characteristics can bring new aspects of the history of the documents. A mixed method approach for clustering potential scribes and authors of handwritten documents will be presented by introducing LiViTo which combines linguistic insights and computer vision techniques in order to assist researchers in the analysis of handwritten historical documents. This talk should show that it is feasible to use data mining and statistics for analysing written cultural heritage and Artificial Intelligence to access the material documents and use the data as input for further linguistic analysis. LiViTo was trained with historical Czech texts by 18th century immigrants to Berlin, a total of 564 pages from a corpus of about 5,000 handwritten pages without indication of author or scribe. An overview of the development of LiViTo and an introduction into its methodology and its functions will be provided. Findings concerning the corpus of Berlin Czech manuscripts and possible further usage scenarios will be discussed.

Kirill Nazarenko

(St. Petersburg State University)

Digital Visualization Based on the Analysis of Handwritten Texts: Reconstruction of the Russian Naval Costume of the Early 18th Century

Keywords: Visualization, visual reconstruction, analysis of handwritten texts, early 18th century, Russian naval costume

The study of the costume requires a complete visual reconstruction even if the garments are preserved. The situation is even more difficult when genuine objects are not preserved and the researcher has only texts and an extremely limited number of fine arts. This fact complicates visual reconstruction noticeably. The key issue is the development of methods for analyzing available information. It is necessary to find all fragmentary evidences, divide them into information blocks, converse quantitative data into the metric system, develop methods for estimating the sizes of various objects based on indirect data (such as the size and quality of fabrics and other used materials, average height of men of 18th c. etc.) To solve these problems, a database for each subject of the costume was created, reflecting its size and patterns. The analysis of such data allows to make a visual reconstruction of the image of Russian sailors and to find out in what combination certain items of clothing were worn and how often they were in use.

Vladimir Neumann

(State Library of Berlin)

"Deep Mining" of the Collection of Old Prints "Kirchenslavica Digital"

Keywords: Old prints, Church Slavonic, full texts, Transkribus, Text recognition models, data retrieval

The lecture deals with efforts of the "Staatsbibliothek zu Berlin - Preußischer Kulturbesitz" (SBB-PK) that aim to make its collection of about 250 Church Slavonic prints from the 17th to the 19th century accessible in terms of content using the methods of modern information technology from the Digital Humanities sector. The focus is on the full-text indexing of the heterogeneous Church Slavonic old prints mainly from East Slavic origin using (HTR+) language models from the "Transkribus" programme. Depending on whether they are Moscow, Kiev, Vilnius or Old Believer prints, these models require different approaches and corresponding adaptations that take into account the printing area and printing period. Prints such as "Kirillova kniga" (1644), "Kniga o věre" (1875) or "Gistorija Ioanna Damaskina" (1637) and many others are processed on a big scale, whereby the developed OPTR models are constantly refined by training new verified data. The full texts generated in this way are permanently stored in various XML

formats (ALTO, TEI) on the one hand in a central repository for subsequent use, and on the other hand they are merged with original digital copies in the IIIF-compatible digital library of the SBB-PK. Additionally, the Church Slavonic full texts will be indexed using special SOLR analyzers for efficient searches (N- Grams, Stemming, Translit, Fuzzy) and made searchable in subject portals (including the "Slavistik-Portal") using modern text-image web design.

Ekaterina Nosevich | Ivan Poliakov | Denis Tsypkin

(A.P. Karpinsky Russian Geological Research Institute Saint Petersburg | National Library of Russia Saint Petersburg | Saint Petersburg State University)

Modern Methods of Manuscripts Marks Studying: Wax Drops

Keywords: wax, pollen, Russia, medieval manuscripts

Wax candles drops on the manuscript pages are used by the authors for indication of the type of the wax lighter and the position of the book, the candle and the writer during the process of reading and writing. Wax drops also contain more information, which can be extracted with specific methods. We present preliminary results of our investigation on pollen analysis of wax candles drops on medieval Russian manuscripts. Pollen grains in the natural wax belongs to regional vegetation plants and, comprising pollen spectra can be used to identify the location, where the wax candle was produced. As we suggest that the production of wax was commonly entertained and candles were not the object of export in medieval Russia, thus we can track down where the manuscript was created and then read. The development of this method can provide new data about the history of the manuscript as the physical object of art.

Ekaterina Nosova | Dmitriy Weber

(St.-Petersburg Institute of History RAS | St. Petersburg State University)

Study of Applied Black Seals from the Collection of Nikolay Likhachev: Preliminary Observations

The study is aimed at examining peculiarities of applied seals of the 16th – 19th centuries made from black sealing wax, which were relatively rarely studied. 56 seals dated by the 16th – early 19th c. were selected for research.

When examined under ultraviolet light, black seals acquire a dark red colour with irregular black segments. Infrared investigations allowed assuming that this heterogeneity is due to the basic material of the seal and not to distribution of a dye. Red and golden crystals were found under the microscope. SEM-EDX analysis showed that they contain mercury and sulphur (red crystals) and arsenic and sulphur (golden crystals), i.e. cinnabar and orpiment or realgar. Since the crystals are invisible to the naked eye, it is impossible to assume that this admixture has an aesthetic function. They may appear from another seal on the same document that contains mercury and sulphur. However, microscopic investigation proves that the crystals are deeply immersed in the seal. Therefore, they can be explained as impurities.

The data obtained during the study indicate the presence of three phases in the objects under discussion: the main seal mass of two types and the foreign pigment particles. This heterogeneity has to be taken into account when interpreting the data obtained by other natural science methods, because otherwise the researcher risks to spread the information obtained from one area to the whole object.

This work was supported by Russian Foundation of Basic Research in frame of the Project No. 18-00-00292 included in 18-00-00429 (K).

Mariia Novak

(Vinogradov Russian Language Institute, RAS)

Cyril of Jerusalem Catechetical Lectures in the 13th-Century Old-Russian Tolstovskii Sbornik: A Textological Study

Keywords*:* Cyril of Jerusalem, Old Slavonic versions, Tolstovskii Sbornik, lexis, grammar, textology

This study is a part of the project devoted to the Old-Russian 13th-century Tolstovskii Sbornik (TS) online machine-readable edition and its comprehensive philological study. It focuses on the textual features of Cyril of Jerusalem catechetical lectures in the mentioned source, compared with two other manuscripts: Sin (SHM, #478, the 11–12th cent.), Rum (RSL, #194, the 16–17th cent.). The analysis of a random sample (containing 50 lexical and grammatical units from lectures 4–12) shows five types of correlation between the text copies: units in all copies differ (4 grammatical items out of total 5); units in all copies coincide (9 lexical items out of total 12); TS = Sin ≠ Rum (14 lexical items out of total 21); TS = Rum ≠ Sin (5 grammatical items out of total 8); Sin = Rum ≠ TS (3 grammatical items out of total 4). The prevalence of matches in TS : Sin correlation confirms their origin from one archetype, what was the matter of dispute before. Among coincidences of all three codices, lexical units predominate, and among their differences, otherwise, more grammatical features occur. Further peculiarities will be discussed during the conference presentation.

The study was accomplished with the financial support of the Russian Foundation of Basic Research (no. 18-012-00428a)

Matija Ogrin

(Research Centre of Slovenian Academy of Sciences and Arts)

The Repertorium of Slovenian Early Modern Manuscripts. A Place of Cultural Memory and Textual Research

Keywords: Manuscripts, Slovenian Literature, Baroque, TEI

In Slovenian literature and culture, many texts could come into existence and circulate only by means of manuscript culture – even late in the era of print. To study these materials, the *Repertorium of Early Modern Slovenian Manuscripts* was launched in 2011 as a small-scale research infrastructure, providing detailed, TEI-encoded manuscript descriptions and related digital facsimiles (http://ezb.ijs.si/nrss/).

The *Repertorium* facilitates research of manuscript texts, which differ significantly from the Early Modern Slovenian printed texts in language, genres and contents. In the periods of Baroque and Enlightenment, a variety of textual genres and forms evolved, where Slovenian manuscripts emerged in several, most distinct socio-cultural contexts, reaching from writings of civil and ecclesiastic persons to texts of self-educated peasants.

In the paper, we present the structure of the *Repertorium*, together with formalised, TEI- encoded typology of textual genres and another typology of socio-cultural contexts of the manuscript writers. We also present the recent technological upgrade of the entire collection and offer some critical remarks after ten years of endeavours.

Nilo Pedrazzini

(University of Oxford)

Tackling Lack of Linguistic Data with HTR: A Specialized Model for the Transcription of Serbian Church Slavonic Manuscripts

Keywords**:** Church Slavonic, Handwritten Text Recognition, Transkribus, neural networks, corpus linguistics, digitization

The paper presents an HTR+ model for the automatic transcription of Church Slavonic (CS) manuscripts of the Serbian recension. I provide a detailed description of the training process (through the software platform Transkribus)

and its challenges, with the objective of making the method maximally reproducible. The undertaken project stemmed from the need to dramatically expand digital annotated corpora of early Slavic (TOROT Treebank).

Recently, HTR+ models were trained for a variety of pre-modern Cyrillic hands, with impressive CER reached for manuscripts written in ducti and linguistic varieties similar to those contained in the training sets (Rabus 2019).

The purpose of the new specialized model is twofold:

1. To automatically transcribe texts written in a Serbian recension of CS with a high success rate and minimal post-correction.
2. To boost the generic CS model already available in Transkribus, thus approaching the ultimate goal of attaining a single powerful tool able to cope with most types of pre-modern Cyrillic hands.

The transcriptions thereby produced meet different needs, including more time allowance for detailed mark-up in TEI editions or deep linguistic annotation in digital corpora, and, more generally, increased accessibility to Slavic textual heritage through faster digitizations.

Yana Penkova | Achim Rabus

(Vinogradov Russian Language Institute, RAS | Freiburg University)

Indefinite Pronouns in Old East Slavic: A Corpus Approach

The paper focuses on the development and functional distribution of indefinite pronouns in Old East Slavic, taking into account different genres and registers. Unlike the indefinite pronouns in present-day Russian, indefinites in Old East Slavic have hardly ever been subject to corpus-based research. However, the system of indefinites in a language is volatile in the distribution of different items, which is difficult to trace using traditional, qualitative methods.

All the examples in the collected data set were driven from the Old Russian corpus (a subcorpus of RNC). They were tagged for *the type of indefinite marker*, the source including its *originality* and *date*, *type of reference* of the indefinite marker and its semantics (see M. Haspelmath), *syntactic information*, *type of discourse passage* (discourse or narration) and the *degree of formality* (formal or informal) represented in the context. Then, we applied advanced statistical methods such as Conditional Inference Tree and Random Forest analysis, as well as LOESS and multinomial logistic regression. Our analysis allowed for identifying the main predictors in the choice of a particular indefinite marker and tracing the diachronic evolution in the functional distribution of indefinites according to formality and originality factors.

Tatiana Pentkovskaya

(Lomonosov Moscow State University)

О текстологии печатного перевода Корана 1716 г.

Keywords: Quran, Petrine Era, Russian literary language, books of the civil press, manuscripts, textual criticism

Первый печатный перевод Корана («Алкоран о Магомете») был издан в Санкт-Петербурге по указанию Петра в 1716 г. Оригиналом послужил французский перевод Андре Дю Рие, впервые опубликованный в 1647 г. («L'Alcoran de Mahomet»).

Сохранился кавычный экземпляр для печатного издания – рукопись РГАДА, ф. 381 (Син. Тип.) No 1034. Она содержит систематическую правку языка, сделанную справщиками санкт-петербургской типографии Михаилом Волковым и Иваном Кременецким (Круминг, 1994: 233), и является важнейшим источником сведений о характере литературного языка петровской эпохи. Эта правка перенесена в корректурный экземпляр Корана – РГАДА, БМСТ/ГП, No 3 (36). До настоящего времени она не являлась предметом подробного лингвистического описания. На титульном листе корректурного экземпляра прил. «славенский» заменяется на «российский» при указании на язык перевода (Круминг, 1994: 232), однако характеристики этого языка не свидетельствуют об однозначной русификации текста. В докладе будут проанализированы основные направления исправлений, внесенных при подготовке к изданию текста в кавычный экземпляр, в сопоставлении с французским оригиналом. Исправления охватывают не только орфографию, но и лексику и грамматику (здесь особо выделяется правка окончаний прилагательных и замена глагольных форм), включая перестройку фразы.

Круминг А. А. Первые русские переводы Корана, выполненные при Петре Великом // Архив русской истории. Вып. 5. М., 1994. С. 227–239.

Štefan Pilát

(Charles University Prague)

GORAZD: The Old Church Slavonic Digital Hub – New Developments

The Gorazd project was successfully finished in 2020. The main goal of the project was digitization and on-line publication of Old Church Slavonic dictionaries which were created by the Slavonic Institute of the Czech Academy of Sciences, primarily four-part Old Church Slavonic dictionary, and Old Church Slavonic card index. The paper will introduce new results of the project reached over the past two years and it will show possibilities of their scientific application. The options of future development of the project will be outlined too. These options include completion and expanding the material base of the dictionaries, new implementation of the Etymological Dictionary of the Old Church Slavonic Language and possibility of separating the Czech Church Slavonic vocabulary to its own lexical database.

Anna Ptentsova

(Moscow State University)

Доумати (dumati), гадати (gadati) and their Aspectual Pairs in the Historical Subcorpus of the Russian National Corpus

Keywords: Old Russian, Russian National Corpus, lexical semantics, aspectology

The talk is about semantic features of two Old and Middle Russian verbs which were synonyms with very similar meanings in many contexts (and even could be used next to each other within the same phrase, e.g. а ты брате <...> старѣи еси насъ а доумаи гадаи о русскои земли (Kiev Chronicle, 236г: 28) – ‘You brother <...> are older than us, deliberate then (with us) about Russian land’).

I shall also describe the process of formation of aspectual pairs for these verbs (съдоумати / оудоумати and съгадати / оугадати respectively).

The study is based primarily on the Old and Middle Russian subcorpora of the Russian National Corpus (www.ruscorpora.ru); in addition, I consider tagged texts from the collection of Old Russian manuscripts (http://www.lrc-lib.ru/) and historical dictionaries.

Alla Polianina

(Lobachevsky State University of Nizhny Novgorod)

The Problem of Age Classification when Publishing Texts of Historical Sources and Documentary Records

Keywords: age classification, cultural value of information products, age marking of texts

Turbulent changes in the formats of the public information space cause problems in regulating the publication of tests in order to protect children. The international experience of such regulation is represented by several models, including special approaches to the age classification of text information. There are various principles that allow you to exclude certain texts from the age-marking system, for example, if the text belongs to a cultural heritage. An analysis of international practice regarding criteria for cultural values and cultural heritage demonstrates different approaches and rules

for classifying information as having historical, artistic or cultural value for society. In this regard, there are many issues related to the understanding and discussion of general approaches.

Among them:

- interpretation of historical texts for information that is restricted (prohibited) for children;
- selection of a regulatory model from existing international practice;
- development of a specific age marking system for historical sources, manuscripts (their digitized copies), and documentary records;
- development of a specific approach to the use of age classification of historical texts and cultural heritage;
- development of criteria for texts belonging to the national cultural heritage and the cultural heritage of mankind to exclude texts from the age-marking system;
- technical issues of age classification and limits (difficulties) of morphological filtering of historical texts for the purpose of age restriction;
- solving the problem of providing access to cultural heritage.

Vladimir Polomac

(University of Kragujevac)

Compiling a Diachronic Corpus of Serbian: Theoretical and Methodological Challenges

The paper analyzes the most significant theoretical and methodological difficulties faced in compiling an electronic corpus of Serbian from the 12th to the 16th century: a) the difficulty of defining the corpus of Serbian in the aforementioned period having the concept of homogeneous diglossia in mind (the complementary use of Old Serbian and Serbian Church Slavonic in the manuscripts belonging to various genres), b) the difficulties pertaining to preparing and collating different genre manuscripts for the corpus (especially when it comes to compiling the text registers, manuscript metadata, principles of transferring texts into an electronic form and the possibilities of automatic text recognition), and c) the difficulties related to text annotation (especially regarding annotation level determination and text lemmatization principles). The above stated fundamental theoretical and methodological issues, as well as possible solutions are discussed and exemplified by a sample electronic corpus of Serbian charters and letters from the 12th and 13th century created for this occasion by means of a software platform known as Sketch Engine.

Olga Semenova | Egor Salnikov | Lidiia Ostyakova

(Higher School of Economics Moscow)

Morphological Tagging for 17th Century Russian

Keywords: corpus linguistics, POS tagging, morphological tagging, NLP evaluation, corpus heterogeneity, Middle Russian

The variability of orthography, grammar, and genre in pre-modern Russian texts makes it difficult to build models for part-of-speech and morphology tagging (Berdičevskis et al. 2016, Gavrilova et al. 2017). The processing of the 17th century Russian is of particular interest, since it is an intermediate state between Old Russian and modern Russian. In our study, we conduct a series of computational experiments in which we use two models trained on the corpus data from the 15th and 16th cc. to tag the 17th c. texts. We use both qualitative and quantitative methods of analysis in order to evaluate the performance of each model and examine their comparative accuracy. These experiments provide insight into whether and to what extent orthography, inflection and word patterns change over time. Using confusion matrices for parts of speech and grammatical features we observe patterns related to the historical development of the language, such as a gradual decrease of conjunctive verbs frequency and even their partial disappearance, change in occurrence of short adjectives, etc.

More importantly, this approach allows us to explore the potential use of models trained on a certain time period for tagging texts of another period. It could be especially useful if there is a lack of training data among texts of a certain historical era.

Anna Senina

(Higher School of Economics Perm)

Perm Zemstvo and the Right to Public Opinion in the Newspaper "Permskaya Zemskaya Nedelya"

Keywords: Text sources, newspapers, visual turn, content analysis, public opinion, political discussions, censorship

Representing "buildings without a foundation and a roof", the Zemstvos at the beginning of the 20th century could not openly engage in politics. Under the conditions of censorship, official Zemstvo newspapers could not cross the boundaries of the political space, however, the editorial board of the Zemstvo newspaper "Permskaya zemskaya nedelya" began to look for strategies to circumvent this ban. New results were obtained through the use of information technology. The key tool for pronouncing political opinions, assessments and hopes on the pages of the newspaper were letters to the editorial office from the Zemstvo community, which were submitted to the audience as private opinions, but initiated a public discussion about the place and role of the Zemstvo in the social life of the Russian Empire. The content analysis of the texts revealed the moods and expectations of the Perm provincial zemstvo in the process of transformation of the society of the Russian Empire in the period 1907-1917. In the newspaper we can also find another way to talk about the "political": visual strategies. Considering the source in the context of a visual, or iconic turn in historical science, we can single out the key image that the Zemstvo tried to create – the image of the people's representative. With the beginning of the First World War, the images of the Zemstvo were supplanted from the newspaper by general imperial themes, as a result of which, even in the midst of revolutionary events, the "Permskaya zemskaya nedelya" could no longer take its former place in social life. The struggle for the right of the Zemstvo to opinion as a public historical project is presented on the timeline.

The study was supported by a grant from the Russian Foundation for Basic Research, № 20-011-33059 "Zemstvo self-government and parliamentary representation as the key phenomena of sociocultural and political transformation of Russia in the second half of the XIX – early XX centuries"; № 20-09-00443 "Ideological and political propaganda discourses of "whites" and "reds" in the information struggle on the Eastern Front of the Civil War (according to the newspaper periodicals 1918–1922)"

Dmitri Sitchinava | Anton Dyshkant

(Russian Language Institute, RAS / Higher School of Economics Moscow | Independent researcher)

Integration of the Old East Slavic Epigraphical Databases, Corpora and Indices

Keywords: birchbark letters, epigraphy, databases of archeological findings, corpora, word indices

The paper presents work on two databases of vernacular Old East Slavic writing, viz. the databases of birchbark letters and epigraphy. The task is to link the information of archaeological/historical and linguistic character, namely, the possibility of simultaneous expansion and updating of the database and of a linguistic corpus that enables grammatical and lexical search. This includes, in particular, creating an online workstation for the morphological markup of texts linked to the database entries, the possibility of exporting XML-databases with morphological markup to include them in the RNC, the possibility of automatic generation of forward and backward word indices to the database on the basis of the marked corpus. Creation of a new database on the Old East Slavic epigraphy is designed to overcome the fragmented state of the publications and research in the field, combining the material accumulated by the scholars into a single electronic resource. From the point of view of programming, the architecture of the epigraphy database is analogous to the one of the birchbark database, which makes it possible to create an annotated corpus of the Old East Slavic epigraphy.

Daniil Skorinkin

(Higher School of Economics Moscow)

Russian Digital Humanities – a View from Inside

The history of 'Digital' Humanities in Russia dates back to the Formalist movement in the early XX century. The list of Russian 'DH forerunners' includes a number of XX-century literary scholars, historians, and linguists. Today there

are several DH centres in Russia that take up the baton and use modern digital tools and methods to study literature, history, and culture. I will talk about DH research, digital preservation, and the related public engagement projects in Russia. I will focus mainly on my own 6 years' experience as a DH researcher/educator at the HSE University DH centre. I will also talk briefly about our ongoing effort to increase public awareness of DH in Russia through building a media outlet reaching 100,000 readers.

Maria Smirnova | Ivan Poliakov

(National Library of Russia)

Corpus of Autobiographical Notes in Russian Manuscripts in XVII– XVIII Centuries: Methods of Search and Study

Keywords: autobiographical notes, manuscript culture, Early modern period, manuscript studies, Russian history

The memoir genre originated in Russia in the late XVII – early XVIII centuries. However, some autobiographical motives appeared in manuscripts during the whole XVII century. Sketchy records concerning authors'/readers' biographies may be found on the pages of various kinds of handwritten books: liturgical, commercial, genealogical, medical and miscellanies etc.

Despite the growing interest in history of self-consciousness of the Russian individual of the Early modern period, and in origin of new secular genres in manuscript culture in particular, these notes have never become a subject of study.

The study focuses on complete searching of manuscripts of XVII–XVIII centuries containing various kinds of autobiographical notes. The main result of the study will be the creation of the database including the whole corpus of the autobiographical notes in Russian manuscripts of the Early modern period.

The reported study was funded by RFBR according to the research project No. 20-39-70005.

Andrey Svetlov | Anatoly Komendantov | Alexander Matveev

(Volgograd State University)

On Software Development for a Corpus of Archival Fund Documents

Keywords: corpus software, historical corpus, stemming, morphological tagging, Don Cossack Army.

Over the past few years, a group of philologists from Volgograd State University has been actively working with documents of the archival fund "Mikhailovsky Stanichny Ataman" (1734–1836). Now, the preliminary manual processing and digitalization of documents has almost been completed, and the following work demands the automatization of some tasks. One of the first problems proposed to our IT group was the morphological analysis of texts. We created an application with a graphical interface for morphological analysis and text tagging. The application is based on the stemming tool MyStem with no graphical interface and quite limited functionality. In addition, we started our work on the second task related to creating a corpus of archival texts. We considered an available free software, but came to conclusion that there is no application or service that would fully satisfy our needs. So, we chose to extend the functionality of previous application to a combined web service. The concept of the service is to integrate the morphological analysis application and corpus software into a single web interface, which combines all the necessary functionality. A user can upload new text for morphological analysis and tagging, the results are editable, and then you can add them to the corpus database. To develop this service, we use the Spring Java framework, Swagger, MongoDB (it can be replaced with another NoSQL DBMS) and the Thymeleaf template engine.

Walker R. Thompson

(Heidelberg University)

Using Text Recognition Tools to Transcribe Multilingual Lexica

Keywords: Epifanii Slavinetskii's Greek-Slavic-Latin dictionary, Fedor Polikarpov's Dictionarium trilingue, Transkribus, OCR, multilingual pre-modern texts

The proposed talk will present the preliminary results of efforts to produce digital editions of a 17th century trilingual manuscript (Epifanii Slavinetskii's autograph of his Greek-Slavonic-Latin dictionary), as well as of an early 18th century printed dictionary (Fedor Polikarpov's Dictionarium trilingue), using handwritten text recognition technologies (Transkribus). The talk will provide preliminary assessments of the error rate in the automatic transcriptions as well as highlighting particular problems that have arisen during the process of transcription, with special reference to superscript characters and diacritics. It will be argued that, despite these challenges, there are considerable advantages to employing such tools in terms of time savings and the simplification of work on edition projects, especially when one is dealing with very large manuscripts that are also visually and structurally complicated.

Cynthia M. Vakareliyska

(University of Oregon)

Tweaking the Digital Menology Collation

This paper is an update on the long-term project of developing an expansive searchable digital collation of Slavic, Greek, Latin, and other calendars of saints, both menologies and other genres. The digital blueprint of the collation, and its maintenance, are by David J. Birnbaum (University of Pittsburgh).

Over the years since the project first began, in the 1990's, we have made numerous changes to the blueprint in order to accommodate methodological issues in the tagging of calendar entries, as they have arisen. Most recently, while beginning the revision of the file containing the Latin Martyrologium Hieronymianum (MH), I gained access to Delehaye's 1931 commentary to the MH, which rightfully questions and often outright debunks the identification of numerous saints and sets of saints on almost every day of the year, particularly with respect to the many martyrs in Africa, who were incorporated ineptly into the MH from the Syrian Martyrology and the Carthaginian calendar, together with many mistranslations, repetitions, garblings, and misplacements of saints' names and toponyms. Delehaye's 718-page work has enormous implications not only for the MH, which was discredited by the Catholic Church many centuries ago, but also for Cardinal Baronius's later Martyrologium Romanum (MR), which relies to a large extent on the MH, and contributes to the confusion by interpreting lists of individual saints in the MH as companions of each other, martyred at the same time and place. Some of these errors continue on in medieval Slavic and Greek calendars of saints that contain Western saints' entries.

Delehaye's comments on the original misnaming of saints and their locations in the MH raise philosophical issues for the digital collation as a whole. Up to now, where a Slavic calendar has the wrong saint's name or description for a clearly identifiable figure, the tag name for the saint replaces the error with the name of the figure who would have been originally intended in an earlier source for the calendar, even though there is usually no way to tell whether the error was introduced by the copyist or by a predecessor in an earlier antigraph. Otherwise, I have followed the philosophy of not intervening with judgments about the historical existence or sainthood of figures listed in any given calendar. But when identification errors in a Slavic calendar predate that calendar by seven or eight hundred years, as a result of a mistranslation into the Latin MH, should they be corrected in the tag name to match the original saint from the Syrian Martyrology or the Carthaginian calendar, as I have been doing now in the MH tag names? How far back chronologically are identification corrections appropriate, if ever? And if a medieval Greek or Slavic calendar lists a series of unrelated martyrs from the Syrian Martyrology as a set of companions because the MR did so, in misinterpretation of the MH, should that series of saints be tagged as a set in the collation too? Since the medieval Eastern Orthodox Church had no centralized beatification or canonization procedure, and since these martyrs come into Eastern Orthodox calendars through the Western tradition, there is no established Eastern Orthodox view regarding their validity.

The paper explores these issues and reports on how they have been resolved for purposes of the digital calendar collation.

Allison Vanouse

(Boston University)

Transport Protocols and the "Attacker" in Digital Preservation: the Case of the Très Riches Heures du Duc de Berry

Keywords: information theory, post-media, open access, electronic literature, text encoding, media transfer protocols

This paper suggests adopting the metaphor of "attacks" on transport service protocols for contexts germane to archivists, preservation specialists, and other cultural workers with an interest in maintaining access to works of art in digital presentations. Proceeding from the example of the *Très Riches Heures du Duc de Berr,y* an argument is presented for expanding or refining the definition of "digital editions" to better suit a results-oriented, post-media framework: all digital editions must aim to enable functional access to versions that users can describe as "legible"; furthermore, institutional presentations that fail to meet this objective—such as the prominent, search-optimized link to the *Très Riches Heures* hosted by the University of Chicago Division of the Humanities—should be read by users as a form of failed digital edition. A reading of internal memos of the Internet Assigned Numbers Authority (IANA) related to media transport evaluates this institution's use of terms such as "attack" and "vulnerability" with respect to media transmission, and suggests an iterative, interoperable protocol for digital humanists and researchers who face these perils without the terminology of conflict.

Liudmila Varlamova

(Russian State University for the Humanities, Moscow)

Standardization of the Long Term Preservation of Digital Documents and their Formats

Keywords: цифровой документ, стандартизация, хранение, долгосрочное сохранение, формат, стандарты ИСО

Обеспечение долгосрочного сохранения цифровых документов является одним из важнейших вопросов современности, которым занимаются многие страны мира. Международная организация по стандартизации (ИСО), изучая передовой опыт этих стран, разрабатывает международные стандарты, регулирующие эти вопросы. Процесс долгосрочного сохранения цифровых документов включает в себя множество аспектов, наиболее существенными из которых являются носители и форматы этих документов. В данном исследовании будут представлены основные стандарты ИСО, регулирующие процессы долгосрочного хранения документов с акцентом на цифровые, а также на форматы их хранения. Учитывая универсальность этих технологий и международное признание стандартов ИСО, исследование будет представлять интерес для широкого круга специалистов, работающих в области сохранения культурного наследия на национальном и международном уровнях.

Regina A. Vernyaeva

(Kalashnikov Izhevsk State Technical University)

Collocations with the Component -*ьп(о)* in Russian Chronicles: Quantitative and Statistical Analysis (on the Basis of the Corpus of Russian Chronicles of the Historical Corpus "Manuscript")

Keywords: Quantitative and statistical research, Chronicles, collocations.

Currently, quantitative and statistical methods are relevant in various areas of textual research. This technique is especially frequently used when working with modern texts. However, works aimed at analyzing linguistic data in ancient texts by means of quantitative and statistical techniques are scarce.

The research material for this work was three lists of the most ancient Russian Chronicles published in the corpus of Russian Chronicles of the historical corpus "Manuscript" (manuscripts.ru) – Lavrentievskaya, Ipatievskaya and Radzivillovskaya.

This study is aimed at identifying collocations that contain the *-вп(о)* component using the n-gram module of the historical corpus "Manuscript" in order to identify rare and frequent stable collocations in the Chronicles texts, analyse and interpret the statistical data on the use of *-вп(о)* forms using statistical methods. It is important to define the syntactic environment of the forms with *-вп(о)*, determine the part of speech and analyse of the topics of the selected contexts. The results of the study are: 1) a list of the rarest and most frequent collocations with the component *-вп(о)*; 2) a statistical analysis of the forms with *-вп(о)*, comparing the results of this analysis with previously obtained data on the use of these forms in the Chronicles; 3) the definition of the syntactic functions of forms with *-вп(о)* (predicative / attributive); 4) determination of the subject of text fragments of the Chronicles in which collocations with the component *-вп(о)* are recorded.

Cristina Vertan | Walther v. Hahn

(University of Hamburg)

Annotation of Vague and Uncertain Places and Events in Historical Texts

Keywords: fuzzy ontology, annotation, vagueness, uncertainty, data modeling, reasoning

Annotation and interpretation of vagueness is a central issue in digital processing of historical texts. However, this issue was completely neglected until now, and has as consequence often distorted interpretation of digitized historical texts. In this article we presented the current state of the art on vagueness annotation and introduce an approach for mark-up of vague and uncertain words and expressions at several levels, among them linguistic and domain specific. We present in detail the way how vague and uncertain time events and geographical places are represented into a fuzzy ontology. Consequently, instances of this ontology (individuals) define our mark-up in text. The ontology is realised in OWL 2, use the <annotation> feature in order to mark-up fuzzy properties and concepts. Later a convertor transforms this ontology in a fuzzy representation which can be connected with a reasoner.

References:

1. Bobillo, Fernando and Delgado, Miguel and Gomez-Romero, Juan, "Reasoning in Fuzzy OWL 2 with DeLorean, in Uncertainity Reasoning for the Semantic Web II", Bobillo, F., Costa, P.C.G., d'Amato, C., Fanizzi, N., Laskey, K.B., Laskey, K.J., Lukasiewicz, Th., Nickles, M., Pool, M. (Eds.), Lecture Notes in Artificial Intelligence, Springer Verlag, (2013)
2. Bobillo, Fernando and Straccia, Umberto, "fuzzyDL: An Expressive Fuzzy Description Logic" http://www.umbertostraccia.it/cs/software/fuzzyDL/download/old/documents/fuzzyDL.pdf (last retrieved 29.03.2020)
3. Güney, A. and Vertan, C. and von Hahn, W. "Combining hermeneutic and computer based methods for investigating reliability of historical texts", Proceedings of the "Twin Talks" workshop collocated with DHN 2019, Steven Krauwer and Darja Fišer (Eds.), University of Copenhagen https://cst.dk/DHN2019Pro/TwinTalksWorkshopProceedings.pdf, pp. 25-38, (2019).

Xiaojie Xu | Kazuyuki Enami

(Toyo Bunko, The Oriental Library, Tokyo)

Analysis of Incunabula Paper Quality: Beginning from "The Travels of Marco Polo"

Keywords: printing paper, *The Travels of Marco Polo,* high-resolution digital microscope, early printing

Toyo Bunko is home to 77 different *The Travels of Marco Polo*, including the Latin version published in 1485. This report will analyze the printing papers of *The Travels of Marco Polo* published up to the 17th century in Europe.

As a method, we used the high-resolution digital microscope, a scientific and nondestructive technique. Through analysis of paper manufacture material, we try to shed light on the procurement method of paper and the characteristics of printing papers in early printing in various regions of Europe.

Valentina Yakunina | Elizaveta Popova

(Yaroslav-the-Wise Novgorod State University)

Economic, Political, and Sociocultural Communications between Russia and Baltic Region in the 15–17th Centuries based on the Archival Collections of Tallinn, Lübeck, Berlin, Stockholm and St. Petersburg

Keywords: Key words: Baltic region, Middle Low German manuscripts, Russian-Hanseatic relations, database

The project "Economic, political, and sociocultural communications between Russia and Baltic region in the 15–17th centuries based on the archival collections of Tallinn, Lübeck, Berlin, Stockholm and St. Petersburg" has started in 2019 directed by Prof. M. Bessudnova. The main directions of the project participants' study are research and analysis of handwritten documents of the 15–17th centuries stored in the archives of Tallinn, Lübeck, Berlin and St. Petersburg. It is carried out the scientific processing of archival materials in particular transcribing of handwritten texts, dating, annotating, identifying of the available information and the presence of publications, complete or partial translation, arrangement on the web-portal, introducing into scientific use by way of publications, reports and issuing of sources in Russian. It is expected to present an integrated model based on the acquired information of the relationship between the Baltic Sea countries and the Hanseatic League at the turn of the medieval and early modern period.

The scientific novelty of this study consists in the refusal of historiographic clichés, first of all, in the refusal of the ideas of confrontation between the Baltic Sea countries (Livonia) and Russia before and in the beginning of the modern period, and in an attempt to design a much more integrated model of their relationship. The modeling method proposed by the project developers is based on a detailed research study of various thematic and problem units, after that the study results are brought together.

The report will also demonstrate the results of the already done study on the example of several cases that allow understanding the features of working methods.

This study is done with the support from the Russian Science Foundation (RSF) project no. 19-18-00183

Svetlana Zemicheva | Maxim Gromov

(Tomsk State University)

Tomsk Dialect Corpus as a Universal Information Search System

Keywords: electronic dictionary, dialectology, dialect corpus, Russian Siberian dialects

Presentation of texts of folk speech culture in digital format is an important linguistics task, as illustrated by the creation of Russian and foreign electronic dictionaries, databases and dialect speech corpora. At the same time, there are still few corpus resources created on the basis of Russian dialect speech and they have a small size and / or limited number of annotation types.

The Tomsk dialect corpus is the first Russian complex resource combining representativeness and different types of annotation. The corpus counts more than 1.6 million tokens and includes transcriptions of speech recordings which were collected over the last 70 years from nearly 400 settlements. The novelty of the project is also determined by the regional specifics (recordings of the speech of Russian Siberian dialect of the Middle Ob region). The goal of the resource development project is to create a dynamic model of dialect communication reflecting its specifics, taking into account linguistic and extralinguistic parameters. The corpus as a universal information search system includes 5 elements: 1) digitalized texts, 2) annotation and search by extralinguistic parameters, 3) annotation and search by morphological parameters, 4) annotation and search by texts parameters (thematic and genre), 5) definitions of dialect lexemes.

The study was funded by the Russian Science Foundation, project No. 19-78-10015

Ekaterina Zhdanova

(Kalashnikov Izhevsk State Technical University)

Texts of the Corpus of Russian Dialects of Udmurtia as a Source of Linguistic and Culturological Information

Keywords: linguistic corps, Russian dialects of Udmurtia, LGIS "Dialect", history, cultural science, ethnography

The Corpus of Russian dialects of Udmurtia created on the basis of the linguistic-geographical information system (LGIS) "Dialect" (URL: http://dialect.manuscripts.ru/), contains spoken recordings from 166 settlements of Udmurtia of 1970-1990's. The texts are mainly presented in the form of scanned copies of notebooks' pages. There are 9,300 scanned copies of the pages in corpus. All records are certified. The markup provides the token creation and visualization of contexts by user (http://dialect.manuscripts.ru/Lexical/FindQuestPage). It allows to analyze the features of the vocabulary, as well as some phonetic and grammatical features of Russian dialects of Udmurtia.

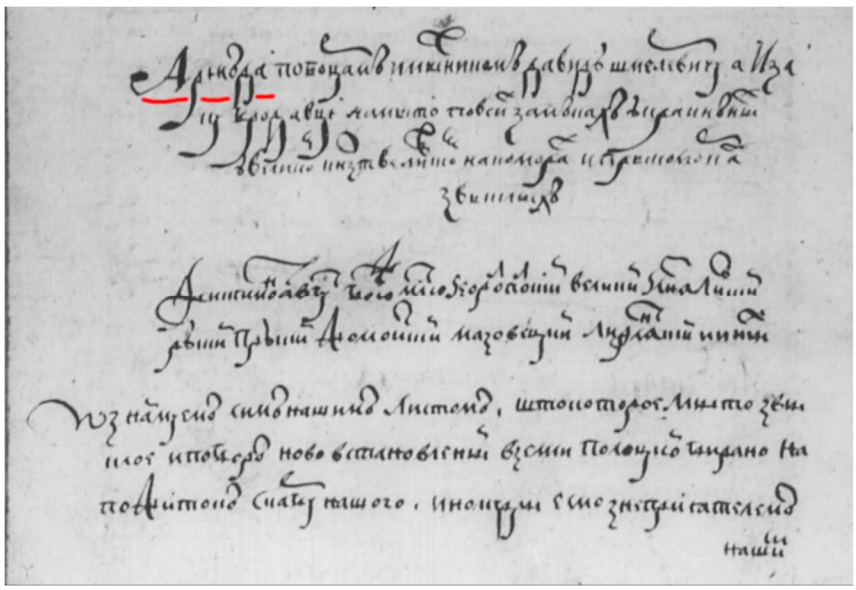
Now the texts of the corpus are in the public domain (http://manuscripts.ru/dialect-test/notebooks). Dialect speech recordings can serve as a source of non-linguistic information about historical events and personalities, material and spiritual culture, customs and traditions of the local population, national composition and interethnic relations in Udmurtia of the 20th century.

Larysa Zherebtsova

(Oles Honchar Dnipro National University)

The Markup of the Act Documents of the 16th Century from the Lithuanian Metrica Concerning the Customs Systems' History with CEI

The article deals with the marking of leases from the end of the 15th century until the 1560s from the Lithuanian Metric (LM) using CEI tags. In our case, CEI tags better reflect the specifics of the text structure of acts. LM is a complex of archival manuscript books created in the office of the Grand Duchy of Lithuania. The documents of leases are one of the most numeral and informative among the LM acts, which reflected the main aspects of the customs system development in the GDL. The markup with CEI tags allows us to select the typical Urkundenbuch structure from the documents of the act and highlight the main terms of the lease agreements that governed the work of customs officers in Lithuania. The analysis showed that among the conditions some were mandatory and characteristic of all lease agreements. We highlighted the internal structure of the lease and observe which parts of its form contain a certain condition.



Oleg Zholobov

(Kazan Federal University)

Online Publication, Fragmentation and Linguistic-Statistical Study of a Megatext from the 13th Century

Keywords: online publication, fragmentation, linguistic-statistical study, Tolstovskii Sbornik

Presently, our research team is finishing the online edition of the tenth manuscript known as the 13th century Tolstoy collection (Tolstovskii Sbornik). This collection is a megatext with a complex composition. The online edition of the Tolstovskii Sbornik and its comprehensive study are supported by the Russian Foundation of Basic Research (project no. 18-012-00428).

A fragmentation module has been created for a comparative study of the texts included in the collection. As the analysis shows, various grammar subsystems exist in separate texts in the anthology. The heterogeneity of the compilation allows us to identify diagnostic grammatical forms with differentiating capacity. The statistical distribution of variable grammatical forms is also of differentiating importance. The megatext grammar, thus, consists of separate subsystems. The identified subsystems diagnose the proximity or remoteness of the Sbornik's texts to other sources.

A comparison study of the most frequent words in Cyril's of Turov homilies with 12 Old Russian codices from the 11th century, using the rank correlation method and the construction of dendrograms, based on the obtained results, made it possible to establish a hierarchical, three-level nature of relations between them: structural, structural-thematic, and lexical-thematic.

Svetlana D. Zlivko | Liudmila S. Shatunova

(Kalashnikov Izhevsk State Technical University)

The English-Language Versions of Multicomponent Terms in the Electronic Linguistic Dictionary of M. V. Lomonosov

Keywords: M. V. Lomonosov, electronic dictionary, foreign-language equivalent, multicomponent term

The electronic linguistic dictionary of M. V. Lomonosov is an integral part of the Lomonosov corpus (lomonosov.pro). The dictionary macrostructure may represent static and dynamic characteristics of a linguistic term (in the dictionary and scientific texts). The key parameters of a dictionary definition are: 1) a scientific area, 2) the extent of being a term, 3) a meaning and interpretation of a term, 4) etymology, 5) foreign-language equivalents of a term (in Latin, Greek, English, German).

This research explores the problem of selecting the foreign-language equivalents of multicomponent terminological naming units which are basis of scientific discourse including important evolutionary stages of linguistic knowledge. This issue is especially relevant for terminography and computer lexicography.

The research interests also include an inventory of multicomponent terms-neologisms (authorial terminological naming units) which demonstrate the range of linguistic development in different fields.