Дигитални и аналитични подходи към писменото наследство

Материали от 7-мата международна конференция El'Manuscript „Писменото наследство и информационните технологии", 2018

Digital and Analytical Approaches to the Written Heritage

Proceedings of the 7th international conference El'Manuscript „Textual Heritage and Information Technologies", 2018

Digital and Analytical Approaches
to the Written Heritage

Proceedings of the 7th international conference El'Manuscript „Textual Heritage and Information Technologies", 2018

Compilers and Editors:
Anisava Miltenova, Victor Baranov, Heinz Miklas, Kevin Hawkins, Jürgen Fuchsbauer

Дигитални и аналитични подходи към писменото наследство

Материали от 7-мата международна конференция El'Manuscript „Писменото наследство и информационните технологии", 2018

Съставители и редактори:
Анисава Милтенова, Виктор Баранов,
 Хайнц Миклас, Кевин Хаукинс, Юрген Фуксбауер

Книгата се издава със съдействието
на програма „**Помощ за книгата**"

CONTENTS

ПРЕДГОВОР

Сборникът включва статии от участниците в 7-мата международна конференция „Писмено наследство и информационни технологии. El'Manuscript-2018" 14–18 септември 2018 г. във Виена и Кремс (Австрия). Поредицата от тези конференции представляват многогодишен международен проект, в който активно е включена България. Те се организират и провеждат съвместно от европейски и руски университети, от академични институции и обществени организации от 2006 г. насам. Основните теми на конференциите са широк кръг от въпроси за цифровите формати, компютърните езици, технологиите и средствата за дългосрочно съхранение на средновековни ръкописи и старопечатни книги, създаването на компютърни програми и специализирани технически комплекси за техния анализ, подготовка и използване на електронни речници, каталози, бази данни, текстови колекции и корпуси от средновековни текстове и изображения. и др.

Участниците са академични изследователи и университетски преподаватели, работещи в областта на историческата лингвистика, историята на книжовността, историята, археографията, кодикологията, текстологията, палеографията и други области на теоретични и приложни изследвания, специалисти в областта на библиотечната и архивна дейност, както и специалисти, които разработват софтуер, бази данни. формати за съхранение и маркиране на текстове и текстови полета, системи за химически и оптичен анализ на документи и др.

Конференцията за 2018 г. беше подготвена от персонала на Виенския университет, Австрийската академия на науките, Виенския технически университет, Виенската академия по изкуствата, университетите в градовете Кремс и Грац, Ижевския държавен технически университет, Центъра за анализ на изображения и материали на културното наследство във Виена (CIMA) и обществото „Писмено наследство".

Публикациите в този сборник отразяват четири основни области в работата на конференцията през 2018 г. – „Кодикология, палеография, консервация и изготвяне на изображения", „Издания, критика на текста, лингвистика, корпусни изследвания", „Моделиране на данни, кодиране на текстове, анотиране, квантитативен анализ", „Дигитални архиви, бази данни и визуализация на информация" ("Codicology,

Palaeography, Conservation, and Imaging", "Editions, Textual Criticism, Linguistics, Corpus Studies", "Data Modelling, Text Encoding, Annotation, Quantitative Analysis", "Digital Archives, Data bases, and Information Visualization"). Статиите са посветени на реставрацията и съхранението на исторически документи, инструменти за анализ на материали и четене на загубени (изтрити) текстове, теоретични и приложни въпроси на компютърната реконструкция и експониране на средновековни писмени паметници, теоретични, методически и приложни въпроси за създаване на електронни публикации, каталози, библиотеки, корпуси от средновековни документи, резултати от многоизмерни изследвания на писменото наследство и количествен и статистически анализ на големи корпуси от текстови данни, теорията и практиката на стандартизацията, моделиране и маркиране на електронни документи, каталози, речници, резултати от използването на традиционни и интернет ресурси в изследователската работа, средства за визуализация на данни и функционалност на системите за достъп до данни и други иновационни насоки за опазване на културното наследство.

Сборникът включва статии на участници от 12 страни – Австрия, България, Германия, Гърция, Грузия, Италия, Русия, Словения, Франция, Чехия, Швеция и Япония. Статиите са подредени по азбучен ред. Редакционната колегия е в състав: проф. Анисава Милтенова (България), проф. Хайнц Миклас (Австрия), проф. Виктор Баранов (Русия), д-р Кевин Хаукинс (САЩ) и д-р Юрген Фуксбауер (Австрия). Рецензенти са проф. Ахим Рабус (Университет във Фрайбург, Германия) и проф. Андрей Бояджиев (Софийски университет „Св. Климент Охридски").

Анисава Милтенова,
Хайнц Миклас,
Виктор Баранов
Кевин Хоукинс
Юрген Фухсбауер

PREFACE

This volume contains the contributions of participants in the international conference "Textual Heritage and Information Technologies – El'Manuscript 2018"[1], which took place on September 14-18, 2018 at the Department of Slavonic Studies of the University of Vienna and at the Center for the Protection of Cultural Property at the University for Continuing Education, Krems (Austria). This event was the 7th in a series of biennial conferences. It thus represents one part of a long-time scholarly project dedicated to the preservation, examination and utilization of written heritage, which has been launched by Victor A. Baranov of the Technical State University of Izhevsk (Russian Federation) in 2006.

The conference of 2018 was organized by personnel of the University of Vienna, the Vienna University of Technology, the Austrian Academy of Sciences, the Academy of Fine Arts Vienna, the University for Continuing Education Krems, the Centre of Image and Material Analysis in Cultural Heritage (CIMA, Austria), the Technical State University of Izhevsk and the association "Textual Heritage". Academics and university teachers working in the fields of historical linguistics, the history of literature, history, archaeography, codicology, textology, palaeography, conservators, librarians and archivists, as well as specialists creating software, databases, formats for the storage and marking of texts and text fields, devices for the chemical and optical analysis of documents participated in it. Not least because of its interdisciplinary character, it attracted interest all over the world. Papers were read by 133 scholars from 21 countries (Armenia, Austria, Bulgaria, the People's Republic of China, the Czech Republic, Finland, France, Georgia, Germany, Greece, Iceland, Israel, Italy, Japan, the Russian Federation, Slovenia, Spain, Sweden, the UK, Ukraine and the USA).

The papers published in this volume cover the four central fields of "Textual Heritage and Information Technologies – El'Manuscript 2018":

1. Codicology, palaeography, conservation and imaging;
2. Editions, textual criticism, linguistics and corpus studies;
3. Data modelling, text encoding, annotation and quantitative analysis;

4. Digital archives, databases and information visualization.

In particular, the articles are dedicated to the restauration and storage of historical documents, instruments for the analysis of materials and the deciphering of lost (erased) text, theoretical and practical problems of the computer-aided reconstruction and exposure of medieval written monuments, theoretical, methodical and practical questions concerning the creation of electronic publications, catalogues, libraries, corpora of medieval documents, the results of multidimensional research on written heritage, the quantitative and statistical analysis of large corpora and text data, theory and praxis in the standardisation, modelling and mark-up of electronic documents, catalogues, dictionaries, the results of the usage of traditional and Internet resources in the work of researchers, means for the visualization of data, the functioning of systems providing access to data, and other innovative tendencies for the preservation of cultural heritage. The order of the contributions is alphabetical.

Anissava Miltenova,
Heinz Miklas,
Victor Baranov,
Kevin Hawkins,
Jürgen Fuchsbauer

THE STATISTICS AND N-GRAM MODULES
OF THE HISTORICAL CORPUS "MANUSCRIPT"[1]

Victor A. Baranov, Roman M. Gnutikov

Abstract: This article describes the capabilities of two statistical modules of the historical Slavonic corpus "Manuscript" – the n-gram module and the statistics module. It presents their general characteristics and demonstrates that the modules (a) allow the user to create a textual subcorpus based on meta- or analytical characteristics of manuscripts, texts and their fragments and (b) enable the analysis of linguistic forms with graphic-orthographic variation. The article also explains the main parameters for queries in the n-gram module: inputting masks and/or the grammatical meanings of components of the n-gram and choosing statistical measures as well as additional ones, plus indicating the number of components, the distance between them and their order, eliminating stop words, etc. Three modes of the statistics module are described: (a) a mode for analyzing the distribution of queried word forms or lemmas within the manuscript(s), (b) a mode for comparing the absolute and relative number of linguistic units in two or more subcorpora, (c) a mode for statistical analysis of linguistic units of the subcorpus/subcorpora compared with a contrasting subcorpus. The article is illustrated with sample queries.

Keywords: Corpus linguistics, linguistic statistics, medieval Slavonic manuscripts

Today information technologies actively penetrate areas seemingly far from science, such as the investigation of the manuscript heritage. We observe a continuous increase in the number of digital libraries of images of the pages of medieval manuscripts, which previously were accessible to researchers only in libraries and archives but are now provided freely on the internet. The constantly growing active use of digital libraries in the scien-

tific, educational and popularization activities of a wide range of special-
ists, teachers and friends of medieval book-learning naturally increases the
requirements for the functionality of the collections, which in turn results
in the development of new presentation forms for the historical documents.

The area which maximally meets the various requirements and requests
of users is the development and creation of text corpora. The practical and
theoretical significance and effectiveness of their use for the solution of a
wide range of applied problems is proven by many years of experience in the
operation of modern corpora.

Any text corpus has two main components: the corpus manager (the
procedures and programs ensuring the storing, tagging, processing, retrieval
and demonstration of the data) and the tagged data itself; the latter repre-
sents texts or their fragments and linguistic units. As is well known, the aims
of creating such corpora are the following: first, to observe the realization
of linguistic units in speech; second, to compare alternative or similar lin-
guistic forms; and third, to reveal the linguistic features of texts character-
ized by opposing attributes etc. Resources initially created for research now
constitute modern corpora, reaching the volume of hundreds of millions or
more word tokens. They form the basis for the development of information
retrieval systems, automated translation, text clustering and identification,
artificial intelligence and many other applications. It is clear that data pro-
cessing of such volumes is only possible by automated means based on the
quantitative, statistical and distributional characteristics of linguistic units.

In many respects the situation differs in the field of the creation and
use of diachronic (or retrospective, allowing for coordinated retrievals from
modern and historical texts at the same time) and historical corpora (con-
taining ancient and medieval documents). Thus, for instance, the number
of machine-readable digital collections of Slavonic written records is rather
limited, and their size is still relatively small. Great success has already
been achieved in the conversion of early printed books into machine-read-
able format with the aid of automated text recognition.[2] When creating his-
torical corpora, enormous effort is required, however, in the preparation of
electronic machine-readable tagged copies of handwritten medieval codi-

---

[2] For example, the internet resource *Google Ngram Viewer* allows retrieval of Russian
books printed in the first half of the 17th century.

ces; this is the main cause of the significant delay in this field.[3] Another important obstacle for their development and use are the difficulties with developing and creating storage and processing tools for nonstandard data and with access means to the material. At present, there are two possible solutions to this problem: either the use of programs and interfaces designed for a corpus of modern texts or the creation of special tools focusing on the search and analysis of data that differ significantly from the modern graphical and orthographic form and their meta-, analytical and linguistic tagging/markings.

Internet-accessible machine-readable collections and digital publications of Slavonic medieval manuscripts as well as libraries of scanned codices are being used by scientists, teachers, students and post-graduate students with different interests in order to obtain primary material for research, teaching or the preparation of reports. In this process, linguistic data in the format of lists, concordances and other types of representation are analyzed and systematized by the traditional techniques of the appropriate scientific field. At the same time, the text and study of language in medieval Slavonic manuscripts using corpus methods (first of all, mathematical and

---

[3]   Some of the most important projects:

⁻ *Corpus Cyrillo-Methodianum Helsingiense*: An Electronic Corpus of Old Church Slavonic Texts;

⁻ *USC Parsed Corpus of Old South Slavic*: The Historical Syntax of South Slavic;

⁻ *Old Russian Texts*: TITUS;

⁻ *Regensburg Russian Diachronic Corpus*;

⁻ *Povest' vremennyx let*: University of Pittsburgh;

⁻ *Old Church Slavonic texts*: PROIEL;

⁻ *TOROT Treebank*;

⁻ *Old Russian Texts*: Pragmatic Resources in Old Indo-European Languages;

⁻ *Church Slavonic corpus*: National Corpus of the Russian Language;

⁻ *Old Russian Corpus*: National Corpus of the Russian Language;

⁻ *Medieval Russian Corpus*: National Corpus of the Russian Language;

⁻ *Corpus of birchbark manuscripts*: National Corpus of the Russian Language;

⁻ *Corpus of Old Russian birchbark manuscripts*;

⁻ *Sankt-Petersburg Corpus of Hagiographical texts* and others.

For lists of other electronic resources devoted to Slavonic medieval manuscripts see, e.g., "*Общежитие: The World Wide Web portal for the study of Cyrillic and Glagolitic manuscripts and early printed books*" (www.obshtezhitie.net) and "*Inventory of Slavic, East European, and Eurasian Digital Projects*" (www.library.illinois.edu/spx/inventory/projects.html).

statistical) are fundamental. Development along these lines depends directly on an increasing volume of machine-readable resources, on the depth, carefulness and accuracy of their tagging, and on the automation of analysis and the development of special corpus managers and their components capable of preparing queries, in particular, taking into account the textological (analytical) composition of manuscript copies and the variability of linguistic units.

Manuscript Project

One of the corpus projects presently being developed in the areas of Applied Palaeoslavistics and Old Russian studies is called "Manuscript: Slavonic written heritage" (manuscripts.ru). This portal provides access to two kinds of corpora: the historical corpus with tagged transcriptions of Slavonic manuscripts from the 10th to 15th centuries and the corpus of the language of Mikhail Lomonosov. The first corpus contains more than 130 Old Slavonic and Russian manuscripts and their extracts with a volume of more than 3.5 million tokens, the second more than 1100 texts of the 18th century with a volume of more than 1.1 million word forms.[4] While the Lomonosov corpus was created on the basis of his complete works, the full-text data base of the first corpus was created on the basis of photocopies and scans of the manuscripts, which are represented as exactly as possible (folio by folio, line by line, symbol by symbol). Thus, even symbol variants and the text location on the manuscript page are retained.

The database model ensures the storage of structured information on all units of the texts, manuscripts or their parts and symbols and facilitates the processing, search and visualization of results, the procedures of automated transformation, unification and normalization of data, allowing the user to find linguistic material even despite the significant formal variation

---

[4] The transcriptions of the historical corpus and Lomonosov corpus were created and proofreading and tagging carried out in more than 15 years by groups and individual specialists, post-graduate students and students from different universities: Udmurt State University and Izhevsk State Technical University (leader: Prof. Victor A. Baranov), Kazan (Privolzhsky) Federal University (leaders: Professors Oleg F. Zholobov and Maria O. Novak), Vienna University (leader: Prof. Heinz Miklas), Sofia University (leader: Prof. Rumyana Pavlova), Saint-Petersburg State University (Dr Georgy A. Molkov), Moscow Russian Orthodox University (Dr Natalya V. Shalygina) and many others.

of the manuscripts and various forms of data output on the screen. These capabilities make possible the solution of a wide range of problems, including the application of corpus methods.

The main means of access to the corpus data are the one-step and two-step, single-text and multi-text query forms,[5] the parallel corpus module for various copies of the same text,[6] the n-gram module and the statistics module. The query forms for the "Manuscript" corpus (i.e., the search procedures, query forms and the output of retrievals, tag tools and other) both ensure standard operations on data, such as the formation of subcorpora, search of linguistic units on the basis of a sample and morphologic characteristics, output in the form of concordances and lists, and ensure non-standard procedures, such as searching without considering variation of textual precedents, finding linguistic units in a specific list, forming queries on the basis of analytical tagging, demonstrating full texts and their fragments and constructing comparative lists of word forms and lemmas[7].

It is recognized that the possibilities of a text corpus become wider the more varied and exact its data tagging is and the more functional and informative the forms of query and data display are. This is perfectly right. Our corpus has been developed exactly in this way so far.

At the same time the corpus data can be studied with the aid of corpus methods: quantitative and statistical. For this purpose, in the initial stage meta-tagging and division of text into words during the preparation of machine-readable transcripts of the manuscripts are quite sufficient.

While operating a deeply and thoroughly tagged corpus, the corpus manager should have rather standard, well-tested searching procedures for material on the basis of available or added tagging. The implementation of the statistical analysis requires specific tools. For their development the specific character of the historical corpora has to be taken into account.

---

[5] The one-step mode allows for search for linguistic units in all manuscripts of the corpus or one of the collections on the portal without formation of a subcorpus of texts (see, e.g., http://manuscripts.ru/mns/main?P_TEXT=63876685&p_lang=EN). The two-step mode assumes obligatory formation of the subcorpus of copies with the aid of the meta- and analytical characteristics of the manuscripts, texts and their fragments (see, e.g., http://manuscripts.ru/mns/srch.simple?p_lang=EN&p_ed_id=94125050).

[6] See, e.g., the parallel corpus of Slavonic gospels (http://manuscripts.ru/mns/portal.main?p1=30&p_lid=2&p_sid=1).

[7] On the possibilities of the corpus, see, e.g., Baranov 2015.

Having in mind the fast increasing volume of historical machine-readable corpora, on the one hand, and the huge workload needed for their deep tagging, on the other, we believe that the design of specific statistical modules within the corpora which will provide the users with new and efficient tools for the analysis of the textual data must come first.

Let us demonstrate the possibilities of two statistic modules of the corpus "Manuscript" – the module of n-grams and the statistics module– and show the effectiveness of their application to the analysis of medieval Slavonic texts-transcripts.

N-gram module

The first tool for access to the data is the n-gram module.[8]

The n-gram module of the "Manuscript" corpus is designed for searching for collocations and colligations in separate texts or in collections on the basis of quantitative, statistical, linguistic and/or structural parameters.

As is well known, in modern text corpora the search for collocations is done by applying statistical measures to the n-grams (usually bigrams), assessing the proximity (association) of components based on the number of all bigrams in the corpus and on the actual and expected quantities of the analyzed bigram and each of its components.

The distinctive feature of the module for the historical corpus is the availability in the query form of not only the basic parameters – indication of the number of components (Component parameter) and component masks and quantitative or statistical measure (Measure parameter) – but also of additional parameters which allow the user:

a) to take into account or ignore the variability of spelling of the medieval Slavonic texts (the Mask Interpretation and Spelling Consideration parameters) by:

– search based on exact or inaccurate matching,

– the use of regular expressions,

– application of the mask to word forms or lemmas (the Type of Components parameter);

b) to take into account the structural linguistic parameters of n-grams:

– the distance between the components (the Distance parameter: from … to …),

– morphological value of components (the Grammatical Attributes parameter),

– punctuation marks between words (the Boundary of Constructions parameter),

– function words (conjunctions, particles, prepositions) and pronouns (the Exclude Prepositions / Conjunctions / Particles / Pronouns / Non-alphabetic Characters parameter);

c) to limit the retrieval of n-grams with specified characteristics, such as:

– contact and/or distant location of components (the Distance and Contact parameters),

– fixed or free order of components (the Fixedness/Attachment and Sequence parameters),

– limited or unlimited compatibility of components (the Symmetry parameter

and others).

Figure 1: The query form of the n-gram module

Each of these additional parameters can be used both separately and in combination with each other for one single purpose – to search for stable semantic and/or grammatical combinations in large volumes of text.[9]

The basic statistical measures used in the module are the Mutual Information score, T-score, Log-Likelihood score, Dice coefficient, Chi-squared test, C-value, and Inside score[10] (Evert 2004; Evert 2005; Mima et al 1998).

For example, the use of the T-score and Mutual Information measures allowed us to establish that statistical measures successfully reveal colligations in corpus collections (when constructing bigrams/diagrams on the basis of word forms – for example, preposition-case combinations with pronouns) and collocations (in the construction of trigrams based on lemmas – for example, verbal phrases).

up
down
Subcorpus volume - 8550 word forms
Number of n-grams - 57

РУ|EN Help [Extended Query] [My samples]

Show 50 entries

№	Rank	N-gramm	F	F(w₁)...F(wₓ)	T-score	Address
1	1	рабъ твои	25	рабъ(30) твои(288)	4.79789	СВХ (ГИМ, Син. 604): рабъ твоихъ 1.1.1-13.29, рабы твоѧ 1.2.1-22.20, рабъ твоѧ 2.1.1-9.12, ...
2	2	недостоннъ рабъ	8	недостоннъ(10) рабъ(30)	2.81602	СВХ (ГИМ, Син. 604): недостоннъıхъ рабъ 1.1.1-13.17, недостоннъıӕ рабы 2.1.1-8.32, недостоннъıӕ рабы 2.1.1-24.17, ...
3	3	рабъ свои	5	рабъ(30) свои(50)	2.15761	СВХ (ГИМ, Син. 604): рабъ своихъ 12.1.1-20.20, рабы своѧ 16.1.1-4.10, своихъ рабъ 25.2.1-9.12, ...
4	4	въ рабъ	4	въ(156) рабъ(30)	1.72632	СВХ (ГИМ, Син. 604): на рабы 1.2.1-22.18, на рабы 20.2.1-15.10, на рабе 26.2.1-16.6, ...
5	5	нсцѣлити рабъ	2	нсцѣлити(5) рабъ(30)	1.40181	СВХ (ГИМ, Син. 604): нцѣлı раба 26.1.1-22.24, нсцѣан раба 26.2.1-5.28
6	6	доуша рабъ	2	доуша(14) рабъ(30)	1.37948	СВХ (ГИМ, Син. 604): дшю раба 27.1.1-13.14, дшѧ раба 27.1.1-19.11

Figure 2: T-score tests for bigrams with word РАБЪ in the Missal of Varlaam Khutynsky, beginning of the 13th century. (ГИМ Син. 604)

---

[9] On the structural and linguistic parameters of the module's request form, see also Baranov 2016.

[10] https://drive.google.com/file/d/1iE-xlQ_m4MQY8BsPaY5IUsOgLHANiAJw/view

Statistics module

The second module that allows for the analysis of quantitative corpus data is the statistics module.
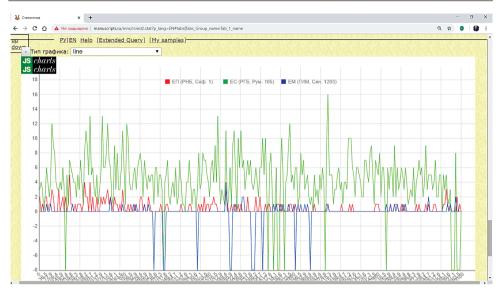
The statistics module[11] was created a few years ago to demonstrate of the quantitative distribution of symbols, word forms, combinations of word forms within the manuscript or several manuscripts. The query form makes it possible to specify a sample of search, the unit of computation step (symbols, pages, folios, fragments) and its length. The result is displayed on the screen in the form of a diagram of the absolute or relative frequency of occurrence of the queried unit in each next step of the manuscript.

Figure 3: The query form of the statistics module

An example of operation of the module is shown in the Figure 4. The letter omega is found at the beginning of words in four gospels of the 12th–13th centuries. The diagram visually demonstrates two orthographic traditions of use of omega in the beginning of words at the minimum.

_____

[11] http://manuscripts.ru/mns/!cred2.stat?p_lang=EN#tabs|Tabs_Group_name:Tab_1_name

Figure 4: The letter omega at the beginning of words in three Gospels of the 12th–13th centuries[12]

The new version of the module has widened possibilities for forming subcorpora for analysis and also provided new possibilities for the quantitative and statistical analysis of linguistic units[13].

In the process of developing the module, special attention was given to the formation of the subcorpora for further analysis of the linguistic units in them.

We aimed to provide the user with the capability of forming a subcorpus not only of texts, but also of manuscripts and fragments of various types.

In addition, it was necessary to allow the user to choose by what parameter of the texts, manuscripts or fragments to group data. In our opinion this should give wider possibilities for the comparative analysis of subcorpora characterized by the distinctive features.

---

[12] "Mstislav Gospel", till 1117 (ГИМ Син. 1203) (blue color); "Panteleymon Gospel", 12th – 13th (РНБ, Соф. 1), (red color); "Simon Gospel", 1270 (РГБ, Рум. 105) (green color).

[13] Statistical measures for searching for keywords and their effectiveness are considered in many papers, see, for example, Salton, G., Yang, C.S. 1973; Sparck, K.J. 1972; Rayson, P., Garside, R. 2000; Robertson, S. 2004; Roelleke, T. 2013 and others.

For example, the public services that display diagrams of the frequency of use of words or word combinations in the texts over time do not allow for grouping texts by genre or the author's gender. Just the same, our previous version of the module that displays the frequency of the analyzed unit within the manuscript did not allow the user to find different frequency of use of the same unit, for example, in fragments of different types, such as in the sticherons and canons of the hymnographic texts.

Of course, absolute quantitative data can be obtained by means of the standard modules of the corpus if the user retrieves results and then performs the necessary computations with the aid of an external program, for example, Excel. However, we wanted to free the user from needing to do this.

Today the statistics module has additional features for preparing sub-corpora for analysis and operations on them:

– widened search on manuscripts, texts and fragments, optionally by using the metatagging of any of these units and logic operations,

– storage and repeated use of the prepared corpora,

– use for analysis of both a single and several subcorpora.
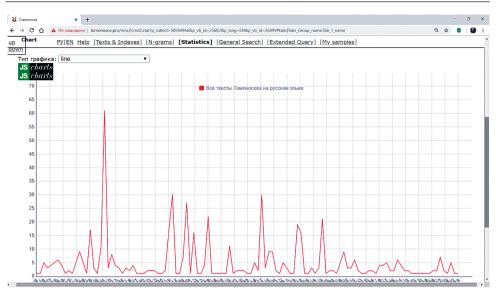
The user can also choose a characteristic of manuscripts, texts or fragments for grouping them and further analysis of the linguistic units in each group.

Let us illustrate the last feature by examples from the corpus of works by M.V. Lomonosov (dating to the 18th century).

The user can choose as a distinctive feature of the texts (for example, letters): the addressee, place of writing, letter subject or another parameter of metatagging.

If the subcorpus is formed from all Russian-language texts by Lomonosov, it is possible to show the distribution of a word both by years and, for example, by text type.

The relative number of instances of the word *коль* in the works by Lomonosov written in different times and in the works of different types is shown in the Figures 5 and 6. The user himself chooses the property by which to group the texts.

Figure 5: The relative number of the word коль in the works by Lomonosov in different years

Figure 6: The relative number of instances of the word коль in the texts by Lomonosov in different genres

A key possibility for analysis is the possibility of indication of the type of operation on the subcorpora: to display data on the linguistic units of

each subcorpus or analyze the linguistic units of one of them against the background of another.

The first of these allows the determination of the absolute or arithmetic average or the median or mode of the data series; the second of these allows the application of the statistical methods to find in a subcorpus being analyzed deviations of the frequency of use of the linguistic units from what is expected, which is obtained on the basis of the average values of the second (contrasting) corpus.
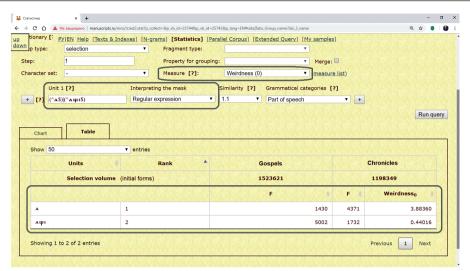
For example, the result of comparison of the frequency of use of the conjunctions *a* and *аще* in the subcorpus of gospels and the subcorpus of chronicles on the screen is shown in the Figures 7.

Figure 7: Conjunctions a and аще in the subcorpus of gospels and the subcorpus of chronicles

The Figure 8 shows the result of the statistical estimation (the Weirdness measure[14]) of the conjunctions *a* and *аще* in the chronicles when the subcorpus of gospels is selected as the contrasting subcorpus.

---

[14] List of measures see is https://drive.google.com/file/d/1hI4hon-Qqe3eI8b-KzXsaBQC3YMiOiHC3/view

Figure 8: The statistical estimation of the conjunctions а and аще in the subcorpus of chronicles

While the first way shows the difference in the number of instances of the conjunction *a* in two subcorpora, the second shows that the large number of instances of the conjunction in the chronicles is not arbitrary and has a considerable deviation from the average value. This means that the large number is the distinctive feature of the subcorpus of chronicles.

Conclusion

As is known, the statistical methods often used to search for keywords in documents that are contained in large corpora are quite effective when applied to the study of small historical corpora when one needs to identify the topic – the content of documents – to compare the contents of documents among themselves.

Furthermore, the relatively small size of the historical corpus also has its advantages: the availability of information about the type and genre of texts allows for analysis at the stage of their selection to assume that they undoubtedly have individual linguistic characteristics, so statistical analysis is necessary not for assigning the text to a particular thematic group but for elucidating quantitative-statistical and thematic-content characteristics that oppose the text (collection / subcorpus) to others[15].

---

[15] See, e.g., Baranov 2017a, 2017b; Baranov 2018a, 2018b, 2018c.

Exactly for this type of analysis, large historical corpora are quite satisfactory, and special multifunctional tools are required. Today one of the main tasks is to make them functional, easy-to-understand and user friendly.

References

Baranov 2015: Baranov, V. A. "Istoricheskij korpus kak cel' i instrument korpusnoj paleoslavistiki." *Scripta & e-Scripta* 14-15 (2015), 39–62.

Baranov 2016: Baranov, V. A. "Modul' n-gramm istoricheskogo korpusa «Manuskript»: strukturnye i lingvisticheskie parametry." In Galiullin K. R., Gorobec E. A., Nikolaev G. A. (eds.): *Nauchnoe nasledie V. A. Bogorodickogo i sovremennyj vektor issledovanij Kazanskoj lingvisticheskoj shkoly*. Kazan': Izd-vo Kazan. un-ta, 2016, 50-61. URL: https://clck.ru/FKqFa.

Baranov 2017a: Baranov, V. A. "Kolichestvennyi i statisticheskii analiz srednevekovykh slavianskikh tekstov: instrumentarii korpusa "Manuskript" i metodika ego ispol'zovaniia." In Kornienko S. I. (ed.): *Tsifrovaia gumanitaristika: resursy, metody, issledovaniia: materialy mezhdunarodnoi nauchnoi konferentsii v dvukh chastiakh*. Perm: Perm State National Research University, 2017, 40-49. URL: https://goo.gl/zyb75Q.

Baranov 2017b: Baranov, V. A. "Statisticheski znachimye slova kak kharakteristika srednevekovogo slavianskogo teksta (na materiale kollektsii Apostolov istoricheskogo korpusa "Manuskript")." In Baranov V. A. (ed.): *Gumanitarnoe obrazovanie i nauka v tekhnicheskom vuze*. Izhevsk: Izhevsk State Technical University, 2017. URL: https://goo.gl/9QegRb.

Baranov 2018a: Baranov, V. A. "Opyt primeneniia kolichestvennykh i statisticheskikh metodov dlia poiska znachimykh slov v istoricheskom korpuse (na materiale srednevekovykh slavianskikh gimnograficheskikh i evangel'skikh kodeksov)." In Rothe, H. (ed.): *Studia Hymnographica, Untersuchungen zum Gottesdienstmenäum nach ostslavischen Handschriften des 11. bis 13. Jh.*, II. Paderborn: Ferdinand Schoningh-Verlag (in press). URL: https://goo.gl/yX4Udd.

Baranov 2018b: Baranov, V. A. "Korrelyacionnyj analiz bigramm russkih evangel'skih spiskov XI–XIV vekov." *Intellektual'nye sistemy v proizvodstve* 16 (4) (2018), 65–74.

Baranov 2018c: Baranov, V. A. "Statistical Analysis of the Slavonic Paraenesis by Ephrem the Syrian (on Three Electronic Copies of the 13–14th Centuries from the Manuscript Corpus)." *Journal of Siberian Federal University. Humanities & Social Sciences* 11/8 (2018), 1211–1228 (DOI: 10.17516/1997-1370-0302).

Evert 2004: Evert, S. "Association Measures." *Computational Approaches to Collocations*. URL: http://collocations.de/AM/index.html.

Evert 2005: Evert, S. *The Statistics of Word Cooccurences Word Pairs and Collocations*. PhD thesis. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 2005.

Mima et al 1998: Mima, H., Frantzi, K. T., Ananiadou, S. "The C-value/ Example-based Approach to the Automatic Recognition of Multi-Word Terms for Cross-Language Terminology." In: *The Pacific Rim International Conferences on Artificial Intelligence (PRICAI'98)*. Singapore, 1998. URL: https:// www.researchgate.net/publication/304254311.

Rayson, Garside 2000: Rayson, P., Garside, R. "Comparing corpora using frequency profiling." In: *Proceedings of the Comparing Corpora Workshop at ACL 2000*. Hong Kong, 2000. URL: http://ucrel.lancs.ac.uk/people/paul/publications/rg_acl2000.pdf.

Robertson 2004: Robertson, S. "Understanding inverse document frequency: on theoretical arguments for idf." *Journal of Documentation* 60 (2004) 503–520.
URL:http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.438.2284 &rep=rep1&type=pdf.

Roelleke 2013: Roelleke, T. "Information Retrieval Models: Foundations and Relationships." In: *Synthesis Lectures on Information Concepts, Retrieval, and Services* 5 (3) 2013, 1–163. URL: https://www.morganclaypool.com/doi/ abs/10.2200/S00494ED1V01Y201304ICR027.

Salton, Yang 1973: Salton, G., Yang, C. S. "On the specification of term values in automatic indexing." *Journal of Documentation* 29 1973, 351–372.

Sparck 1972: Sparck, K. J. "A statistical interpretation of term specificity and its application in retrieval." *Journal of Documentation* 28 (1972), 11–21.

About the Authors

Victor A. Baranov (victor.a.baranov@gmail.com) is a doctor of philology, professor, and head of the linguistics department at Izhevsk Kalashnikov State Technical University. His current research interests are corpus linguistics and statistical analysis of the linguistic data of the medieval Slavonic corpus. He is the project manager for Manuscript.

Roman M. Gnutikov (romaashka@gmail.com) is a programmer for the Manuscript project. He is the main developer and creator of the modules of the historical corpus "Manuscript".

Резюме

Все более активное использование электронных библиотек, содержащих электронные копии средневековых рукописей и старопечатных книг, в научной, образовательной и популяризаторской деятельности широким кругом специалистов, преподавателей вузов и любителей средневековой книжности закономерно повышает требования к функциональным возможностям коллекций. Направлением, максимально отвечающим разнообразным запросам и требованиям пользователей, является разработка и создание текстовых корпусов на основе размеченных машиночитаемых транскрипций. Корпуса должны обеспечивать возможность а) наблюдений за реализацией лингвистических единиц в текстах, б) сопоставления альтернативных или аналогичных лингвистических форм, в) выявления лингвистических особенностей текстов, характеризующихся противопоставленными характеристиками.

Многолетний опыт разработки, создания и эксплуатации исторических корпусов проекта «Манускрипт», содержащих славянские рукописи X–XV вв. и тексты М. В. Ломоносова, позволяет сформулировать пользовательские требования к корпусному менеджеру (к процедурам поиска, формам запросов и демонстрации выборок, к инструментам разметки и др.): обеспечение как стандартных операций работы с данными, таких, как формирование подкорпусов, поиск лингвистических единиц на основе образца и морфологических характеристик, выдачу в виде конкордансов и перечней, так и нестандартных, таких, как нечеткий поиск текстовых прецедентов, формирование запросов на основе аналитической разметки, демонстрация полных текстов и их фрагментов, построение сравнительных перечней словоформ и лемм и др. Основными средствами доступа к данным корпусов являются одношаговая и двухшаговая, однотекстовая и многотекстовая запросные формы, модуль параллельного корпуса списков одного текста, модуль n-грамм и модуль статистики.

Модуль n-грамм корпуса «Манускрипт» (http://manuscripts.ru/mns/cred_ngr.stat) предназначен для поиска коллокаций и коллигаций в отдельных текстах или подкорпусах на основе количественных, статистических, лингвистических и/или структурных параметров. Отличительной особенностью модуля является наличие в запросной форме не только базовых параметров – указание количества компонентов, маски компонентов и количественной или статистической меры, – но и дополнительных характеристик, которые позволяют а) учитывать или не учитывать вариативность словоформ; б) обеспечивать учет структурно-лингвистических параметров n-грамм – расстояния между компонентами, морфологических значений

компонентов, исключение стоп-слов; в) ограничивать выборку n-граммами с определенными характеристиками – с контактным и/или дистантным расположением компонентов, с фиксированным или свободным порядком следования компонентов, с ограниченной или неограниченной сочетаемостью компонентов и др. Все указанные дополнительные параметры могут быть использованы как каждый в отдельности, так и в сочетании друг с другом. В целом возможности модуля обеспечивают поиск устойчивых семантических и/или грамматических сочетаний в подкорпусах.

Модуль статистики (http://manuscripts.ru/mns/cred.stat) имеет гибкие возможности формирования подкорпусов: расширенный поиск рукописей (текстов, фрагментов), хранение и многократное использование подготовленных подкорпусов, анализ как одного, так и одновременно нескольких подкорпусов – и три режима анализа данных. Первый режим предназначен для демонстрации распределения символов, словоформ, сочетаний словоформ в пределах рукописи или нескольких рукописей. Результат выводится на экран в виде графика абсолютной или относительной частоты встречаемости искомой единицы в каждом следующем фрагменте рукописи. Второй – для получения сведений о количестве единиц в частях одного или нескольких подкорпусов. При этом пользователю предоставляется возможность указать единицу группировки данных. Третий – для получения статистических характеристик лингвистических единиц одного подкорпуса на фоне другого. Статистические методы позволяют найти в анализируемом подкорпусе отклонения частоты использования единиц от ожидаемой, полученной на основе средних величин другого – контрастного корпуса.

Приведенные иллюстрации и примеры позволяют говорить о результативности и перспективности использования модулей статистики для выявления тех количественно-статистических и тематико-содержательных характеристик, которые противопоставляют анализируемый текст (подкорпус) другим.

A 'CLOUD' FULL OF DIGITIZED MANUSCRIPTS. THE VENERANDA BIBLIOTECA AMBROSIANA, FROM THE CUSTOS CATALOGI TO THE DATA CURATOR

Fabio Cusimano

Abstract: Digital libraries and digitization projects are not a mere temporary phenomenon, but a real opportunity. Each of us knows how important it is to spread knowledge: today we can do it better than in past decades thanks to new Web-based technologies, to the improved quality of digital objects, to better metadata and retrieval technologies, etc. But the roots of libraries are, of course, older than the so-called *Digital Revolution*, as the case of the Veneranda Biblioteca Ambrosiana in Milan from the first years of the 17th century shows. This prestigious conservation library has managed and curated a special and unique collection of precious manuscripts, and – now as it was then – makes those masterpieces freely available for users from all over the world, both within its reading room and, today, through its new digital initiatives. Many librarians (also known as *Custodes Catalogi*) have succeeded one another over the centuries, always pursuing preservation and conservation targets, and always enriching and curating the catalogs. Today, we are inspired by the work of those masters of the past, but we are also engaged in projects to make these precious sources available for an ever-increasing and global audience thanks to our new, freely-accessible digital library.

Keywords: Veneranda Biblioteca Ambrosiana; digital library; digitization; free access; cloud-based technologies; data curation; IIIF (International Image Interoperability Framework).

Conservation libraries and new technologies: challenge and opportunity at the same time

The main mission of any library is to collect, organize, expand and disseminate knowledge through access to resources stored therein.
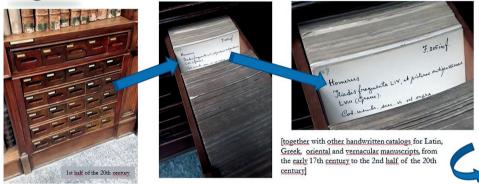
A 'cloud' full of digitized manuscripts. The Veneranda Biblioteca Ambrosiana, from the Custos Catalogi to the Data Curator

IV) The Veneranda Biblioteca Ambrosiana: a real-life approach

The history of libraries began with the first efforts *to organize collections of documents*. Topics of interest include *accessibility of the collection*, acquisition of materials, arrangement and *finding tools* […]. (see: Wikipedia, *History of libraries*)

[together with other handwritten catalogs for Latin, Greek, oriental and vernacular manuscripts, from the early 17th century to the 2nd half of the 20th century]

1st half of the 20th century

See: C. Pasini, *Antichi cataloghi manoscritti dei codici della Biblioteca Ambrosiana*, in «Aevum. Rassegna di scienze storiche, linguistiche e filologiche» (Settembre 1995), 69, 3, pp. 665-695

Image 1: One of the many handwritten catalogs for manuscripts at the Veneranda Biblioteca Ambrosiana (from the author's PowerPoint presentation).

Every library, albeit in the theoretical alternation of its functions between the horizontal activity oriented to the use of books by our contemporaries and the vertical one oriented to their conservation for posterity, is *ipso facto* an organized collection of books; very often, however, this intrinsic function that has always characterized the ideal library is not enough to be able to affirm that every library really acts as a living place of culture.

Although every age has witnessed key moments of evolution and progress across fields of knowledge and technology – often unnoticed by contemporaries – our society has seen so many potentially useful technology tools to spread knowledge and culture that one naturally wonders how it is possible that this did not involve the world of libraries natively!

How come, then, that nowadays we still find ourselves reflecting on what seems to have become a true *cliché*, or the antithetical relationship between library and technology? Each of us could bring his own experience, useful to provide a *quid* related to the perception that users have of the library itself; and precisely this perception would seem to be the key to reading the previous question. In such a context, therefore, the conservation library is often perceived as the refractory place *par excellence* of technology and innovation, destined by definition only to the hoarding of its precious heritage.

The example of the Veneranda Biblioteca Ambrosiana as one of the first 'public libraries' in the 17th century

The Veneranda Biblioteca Ambrosiana of Milan – named after Ambrose, the patron saint of Milan, was solemnly inaugurated on December 8, 1609 by Cardinal Federico Borromeo, Archbishop of Milan from 1595-1631 – is traditionally considered among the first examples of a public library (Rodella 1992, 121-147; Panizza 2012; Rivoltella 2018) in the sense of an institution created with the clear intention of providing access to books to a community of readers as broadly as possible.

Image 2: A bird's eye view of the Veneranda Biblioteca Ambrosiana from Google Maps.

Image 3: Piazza Pio XI and the modern facade (19th century) of the Veneranda Biblioteca Ambrosiana from Google Maps.

Image 4: The ancient entrance of the *Bibliotheca Ambrosiana* **with Cardinale Federico Borromeo's statue (see: https://commons.wikimedia.org/wiki/File:Biblioteca_Ambrosiana_2010.jpg).**

From the *bibliotheca* to the digital library, and from the *bibliothecarius*/*custos catalogi* to the digital librarian/data curator: managing digital collections

Bentivoglio's work *Constitutiones Collegii ac Bibliothecae Ambrosianae* (Bentivoglio 1835; Marcora 1986, 155-164; Annoni 1992, 149-184) devotes an entire chapter to the position of the librarian (*bibliothecarius*): the *Caput X, De Bibliothecario et Bibliotheca*, in which, through twenty-eight paragraphs, we find described the role of the librarian, his duties (as well as those of the personnel reporting to him) and the need to prepare at least two types of catalogs (Bentivoglio 1835, 32-39).

Another figure who, traditionally, joined the *bibliothecarius* at the Veneranda Biblioteca Ambrosiana is that of the *custos catalogi*, the individual with oversight of the catalog (Rodella 2013, 35-36). This job title turns out to be etymologically very interesting and, as we will see, plays an important role in opening up to functions and activities characteristic of the digital age, such as the data curator and the derivative field of data curation.

Content Analysis: "Data Curator"

Sketch of the data lifecycle

Data design and capture	Data storage and management	Data analysis and visualization	Data discovery and consumption
"Data librarian"	"Data manager"	"Data scientist"	"Data services librarian"

"Data curator"

(end of lifecycle is driving everything that proceeds it)

- *Always* concerned about data quality
- *Always* working *directly* with
  - people
  - data/digital objects
  - both
- *Often* working with higher-level
  - policy/planning
  - workflow design

"Curator"
- Latin *curare*: "to take care of"
- Also "to cure"
- "Overseer, manager, guardian"
- Originally caring for minors or disabled persons

Image 5: From the IFLA Library Theory and Research Panel, Data Curation Project, presented at the IFLA Satellite Meeting 2017, Warsaw, August 16 – 17, Data Curator's Roles and Responsibilities: International and Interdisciplinary Perspectives. See: https://ifla.wdib.uw.edu.pl/wp-content/uploads/2017/03/LTR-Panel-presentation-on-the-Data-Curation-Project.pdf.

Very interesting is the etymological link between the data curator and the *custos catalogi*: as in image 5, in fact, the etymology of the English word *curator* is directly derived from the Latin, respectively from the verb *curo, -as, -avi, -atum, -are* and from the noun *curator*, and it is precisely for this reason that it is possible to relate the two figures to each other. The data curator (as well as the digital librarian), in fact, is inspired by the same principles that guided (and still guide) the *custos catalogi*: he works to take care of the catalogs (now predominantly online rather than in print), of the catalog information (today mainly codified using standard formats such as ISO 2709), of the metadata (descriptive, administrative, technical, all united by the XML tags and the metalanguages adopted for their compilation, such as, for example, Dublin Core or specific XML vocabularies), of the digital objects (as well as of their different formats, especially for digital images) and of the various technical procedures to be activated from time to time to start the production of new digital objects through the use of different equipment (digital cameras, scanners, etc.), and to ensure their storage and, together, the persistence of digital information. Another fundamental aspect is that of the global design of digitization interventions and the development of the necessary workflows connected to them.

The digital librarian/data curator at the Veneranda Biblioteca Ambrosiana

What is the data curator responsible for? Taking into account a necessary multidisciplinary approach, we can say – simplifying the issue quite a bit – that the data curator is called upon to act on different and complementary work plans: there cannot be the planning of a digitization intervention without the necessary attention dedicated to the reuse of data inherited from previous digitization projects, to the data quality and to the preparation of a scalable plan for the future reuse of the new data produced today.

Image 6: Digital collection management & data reuse at the Veneranda Biblioteca Ambrosiana (from the author's PowerPoint presentation).

Through the approach adopted at the Veneranda Biblioteca Ambrosiana, we have tried to take all these aspects into due consideration, carefully evaluating what has already been tested in other international institutions,[1] and trying to lay the foundations for the creation of a project that must be scalable, open to productive collaborations, and sharing both technical and scientific aims.

What we have so far cited generically as 'reuse of data' was tested at the Veneranda Biblioteca Ambrosiana in a very concrete way: over ten years of previous digitalization projects have produced, in fact, a very large amount of data (more than 1,800,000 high resolution images in uncompressed 24-bit color TIFF format, yielding 31 terabytes of disk space), which represents a precious nucleus, composed of more than 2,700 fully digitized manuscripts, on which to base the start of a new phase of digitization.

How to manage all this data? How, and where, are the files to be stored? And how to make all this data (and the related metadata) perpetually available?

---

[1] International institutions as, for example, the *Bayerische Staatsbibliothek* (BSB) and the *Münchener DigitalisierungsZentrum* (MDZ), the *University of Notre Dame* (USA), the *Eidgenössische Technische Hochschule Zürich* (ETH Zürich) and its projects *e-codices - Virtual Manuscript Library of Switzerland*, *e-rara* and *e-manuscripta*.

IV) The Veneranda Biblioteca Ambrosiana: a real-life approach
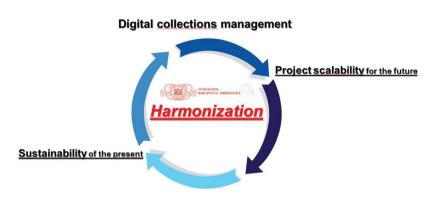
Image 7: A circular, sustainable approach (from the author's PowerPoint presentation).

The current result is summarized in image 8: a general view of the new digital production infrastructure of the Veneranda Biblioteca Ambrosiana, which combines three complementary activities:

1) Data reuse activities;

2) Preparation of two labs for the production of new digital copies:

2a) the Digitization Laboratory equipped with a Zeutschel OS 14000 A1 overhead scanner (with an optical resolution of 600 dpi) and a new lighting system for the rooms by installing UV-free LED lamps with cold light (5000 °K);

2b) the Photographic Laboratory equipped with a medium-format CMOS camera Hasselblad H5D-50c MS 50 megapixel (equal to 300 dpi optical), relative optics in the range of 50 mm, 80 mm and 120 mm and lighting kit with UV-free adjustable LED lamps (also adjustable in the brightness intensity);

3a) Installation of a NAS (*Network Attached Storage*) to be used as a medium-term local storage space, equipped with 8 hard drives of 6 terabytes each, for a total of 48 terabytes, formatted in RAID 6 (*Redundant Array of Inexpensive Disks*) to privilege data security and redundancy (the useful storage space has therefore been reduced to about 32 terabytes);

3b) AWS (Amazon Web Services) S3 cloud storage service: the cloud storage (scalable) solution for all the data and metadata.

All of this is done in order to prepare a consistent base for the new digital library of the Veneranda Biblioteca Ambrosiana (soon to be launched), with the aim of making the digital reproductions of part of the Ambrosian manuscript patrimony freely available online, free of charge and to public audiences.
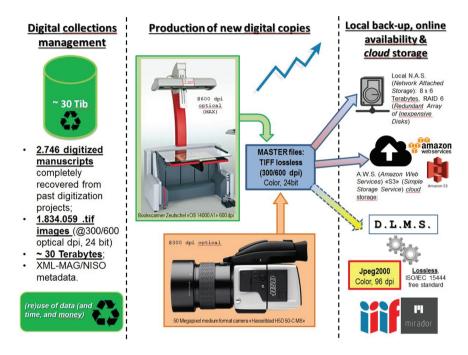
Image 8: An overview of the digital infrastructure at the Veneranda Biblioteca Ambrosiana (from the author's PowerPoint presentation).

The digital library as a knowledge sharing tool

A fundamental aspect for the creation of a new digital library is using the international IIIF-*International Image Interoperability Framework*[2] for viewing high-quality digital content on the Internet (Syndman 2015, 16-21, Brantl 2016, 10-13, Salarelli 2017, 50-66, Magnuson 2018).

---

[2] See "About IIIF": http://iiif.io/about/.

Image 9: The IIIF (*International Interoperability Image Framework***) logo. See: https://commons. wikimedia.org/wiki/File:International_Image_Interoperability_Framework_logo.png.**
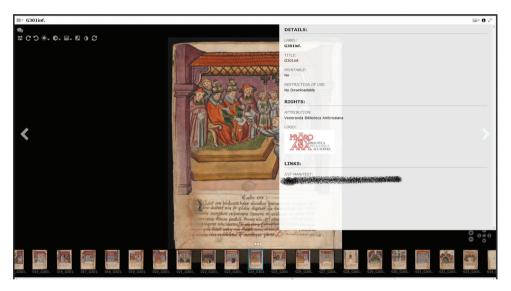
Image 10: Visualization through the Mirador image viewer of a digital copy of the manuscript G 301 inf., f. 9v. © Veneranda Biblioteca Ambrosiana.

Last but not least, the new digital library of the Veneranda Biblioteca Ambrosiana will be interconnected with the current online catalog of the library in order to guarantee a direct link between the bibliographic record and the digital resource: from the bibliographic description present within the online catalog, in fact, a special link will activate (in a way completely transparent for the user) the Mirador image viewer,[3] which will allow the online user an unprecedented viewing experience.

---

[3] See "Mirador": http://projectmirador.org/.

Image 11: free access to the digital copies trough the O.P.A.C. (Online Public Access Catalog).

This new system, therefore, will promote visibility and sharing, as well as providing advantages for libraries and for users – two sides of the same coin, *ad publicum commodum et utilitatem.*

References

Annoni 1992: Annoni, A. "Le Costituzioni e i regolamenti." In *Storia dell'Ambrosiana*. Il Seicento, Milano: Cariplo-Intesa Bci, 1992, 149–184.

Bentivoglio 1835: Bentivoglio, F. *Costituzioni del Collegio e della Biblioteca Ambrosiana volgarizzate dal Dottore Francesco Bentivoglio, bibliotecario della medesima, col testo a fronte*. Milano: G. B. Bianchi e C., 1835.

Brantl 2016: Brantl, M. "Das International Image Interoperability Framework (IIIF): Ein neuer Standard für interoperable Bildrepositorien." *Bibliotheksforum Bayern* 1/10 (2016), 10-13.

Magnuson 2018: Magnuson, L. "Store and display high resolution images with the International Image Interoperability Framework (IIIF)." *ACRL Tech Connect Blog*. URL: https://acrl.ala.org/techconnect/post/store-and-dis-

play-high-resolution-images-with-the-international-image-interoperability-framework-iiif/

Marcora 1986: Marcora, C. "Manoscritti ed edizioni delle 'Constitutiones Collegii ac Bibliothecae Ambrosianae'." In *Accademia di San Carlo. Inaugurazione dell'8° Anno Accademico* (Milano, 16 novembre 1985), Bologna: Cappelli, 1986, 155–164.

Panizza 2012: Panizza, M. *La storia della Biblioteca Ambrosiana*. Novara: De Agostini, 2012.

Rivoltella 2018: Rivoltella, M. (ed.). *Giacomo Filippo Opicelli, Memorie della Biblioteca Ambrosiana*. Milano: Biblioteca Ambrosiana; ITL, 2018.

Rodella 1992: Rodella, M. "Fondazione e organizzazione della Biblioteca." In *Storia dell'Ambrosiana. Il Seicento*, Milano: Cariplo-Intesa Bci, 1992, 121–147.

Rodella 2013: Rodella, M. "Pietro Mazzucchelli (1762-1829) bibliografo ed erudito ambrosiano." In *Frasso, Giuseppe, and Rodella, Massimo, Pietro Mazzuchelli studioso di Dante. Sondaggi e proposte*. Roma: Edizioni di storia e letteratura, 2013, 35–36.

Salarelli 2017: Salarelli, A. "International Image Interoperability Framework (IIIF): una panoramica." *JLIS* 8, 1 (2017), 50–66. DOI: http://dx.doi.org/10.4403/jlis.it-12090

Syndman 2015: Snydman, S., Sanderson, R., Cramer, T. "The International Image Interoperability Framework (IIIF): A community & technology approach for web-based images." *Archiving Conference* (2015), 16–21.

URL: https://stacks.stanford.edu/file/druid:df650pk4327/2015ARCHIVING_IIIF.pdf

About the author

Fabio Cusimano, Ph.D. (fcusimano@ambrosiana.it) is head of cataloguing and digital asset management (D.A.M.) at the Veneranda Biblioteca Ambrosiana; he is also research affiliate at the *Navari Family Center for Digital Scholarship* (CDS), Hesburgh Libraries, University of Notre Dame (Indiana, USA).

Резюме

Цифровые библиотеки и проекты оцифровки – не просто временное явление, но сегодняшняя реальность. Каждый из нас знает, как важно распространять знания: сегодня мы можем сделать это лучше, чем в прошлые десятилетия, благодаря новым технологиям создания баз данных,

улучшенному качеству цифровых объектов, тщательно подготовленным метаданным, новым возможностям поисковых алгоритмов и т. д. Но корни библиотек уходят, конечно, в далекое прошлое, и примером является Амбросианская библиотека (Veneranda Biblioteca Ambrosiana) в Милане, основанная в первые годы XVII-го века.

Хорошо известно, что основная миссия любой библиотеки – собирать, организовывать, расширять и распространять знания посредством доступа к хранящимся в них ресурсам. Амбросианская библиотека, будучи авторитетным хранилищем и одновременно одной из первых публичных библиотек, владея уникальными коллекциями драгоценных рукописей, и сегодня, как и раньше, делает и эти шедевры доступными пользователям всего мира с помощью новых цифровых возможностей.

Библиотекари (известные также как кураторы каталогов), сменявшие друг друга на протяжении веков, всегда преследовали цели сохранения и консервации каталогов и постоянно обогащали и поддерживали их. Сегодня, вдохновленные работой этих мастеров прошлого, мы осуществляем проекты, цель которых – сделать эти драгоценные источники доступными мировой, постоянно увеличивающейся аудитории с помощью нашей новой, имеющей свободный доступ цифровой библиотеки на основе *IIIF* (*International Image Interoperability Framework*) и средства просмотра изображений *Mirador*. Этот инновационный подход, используемый совместно с действующим публичным онлайн каталогом доступа – *OPAC* (*Online Public Access Catalog*) и обеспечивающий прямую связь между библиографической записью и цифровым ресурсом, предоставит пользователям уникальные возможности просмотра части амброзианской рукописной сокровищницы для создания цифровых репродукций. Онлайн доступ будет свободным и бесплатным.

A NEW ASPECT OF CODICOLOGY: SCIENTIFIC ANALYSIS OF PAPER OF HISTORICALLY IMPORTANT DOCUMENTS AND BOOKS OF THE ANCIENT TO PREMODERN ERAS

Kazuyuki Enami, Yoshihiro Okada, Harumichi Ishizuka, Xu Shaojie

Abstract: Codicology is widely understood as a field under which linguistics, decipherment of important documents and books are subsumed. Yet, less contribution of natural and applied sciences to traditional codicology has been seen. No-one cannot deny that paper is the basic material making the modern and contemporary cultural worlds. Since the very beginning of the invention of paper, in the 2nd century BC, not only texts written on paper but also paper itself has continued to be the living witness of human history. By collaborating with scientific studies of paper, codicology would develop further. We have studied paper analysis during the past 10 years, using scientific techniques: (1) analysis of morphology and fibres of paper by using high-resolution digital microscopy producing 3D images, (2) detection and identification of remains on and in paper, plant, cloth, thread etc. (3) X-ray fluorescent analysis of elements remaining on and in paper. In the present report we will show new findings on the history of papermaking, mainly of Asian paper, witnessed by paper itself using the above novel scientific analysis.

Key words: codicology, scientific analysis of paper, digital microscopy, X-ray fluorescent analysis, Asian paper

Introduction

Traditional and old-fashioned methods of analysis of paper have been practised mainly using ordinary optical microscopy to observe one or more fibres extracted from the documents. In addition to the above methods, one can dye fibre by special distaff or identify the raw materials of paper by colour of fibre after treating. As is well known, paper usually is made from not only a single raw material but rather two or more kinds of raw materials. Each paper has its own morphology depending on the raw material. Therefore, only by observation of the fibres can we get real information on the origin of paper and trace techniques of paper making processes. In recent

decades a novel digital microscope that produces 3D images of surfaces and microstructures of fibres has become available. Here we will show concrete evidence of the usefulness of such experiments and completely new results of analysis of paper used in documents and books from Dunhuang and Central Asia of the Otani Collection of Ryukoku University.

In addition to the above microscopy analysis, X-ray fluorescent analysis was also applied to find metallic elements remaining on or in paper, giving information on the origin of the paper, where it was made, and so on. By these novel scientific methods we were able to obtain new results and new understandings of the history of papermaking techniques and those of historically important documents and books, as shown below.

I. Experimental

In the present study, high-resolution digital microscopes, Keyence VHX 500 and VHX5000, were used for the analysis of paper: their morphology, microstructure, fibres and remains on and in paper (plant debris, thread, cloth, starch and so on). To detect metallic elements on and in paper, the Horiba XGT2700 X-ray fluorescent analyzer (XRF), with Rh target, 30kVx1mA, collimated at 100μm, radiation time 200sec for each point, was used.

II. Results and Discussion

1. Analysis of morphology and fibres of paper by using high-resolution digital microscopy producing 3D images

1.1. Paper used for Buddhists' sutras of Dunhuang manuscripts.

It is now well known that papermaking started in the 2nd century BC in China using rag. Then, 300 years later, in 105 AD Chinese artisans headed by Ts'ai (Cai) Lun invented an entirely new papermaking technique using plant fibres. True old Chinese papers began to appear in documents excavated by European and Japanese expeditions to Dunhuang, Turfan and other Central Asian areas during the end of the 19th to 20th centuries. Hereafter, we will show old Chinese paper used for Dunhuang and Turfan documents of the Otani Collection now preserved in Ryukoku University.

Fig.1 shows the hemp paper used for Buddhists sutra of Tang dynasty Avataṃsaka-sūtra MS00509 (the official registered number from the International Dunhuang Project for the Otani Collection documents). In the Tang

dynasty, higher grade hemp paper was specially prepared for the imperial court Buddhists sutras. But such paper was not used for all Buddhists' sutras during the Tang dynasty: some where written on the typical hemp paper of the era made directly from hemp plant stalks. Hemp fibres have no homogeneous width. So, in order to make paper, paper workers must give enough beating treatments to fibres to obtain good hydrogen bonding between them. As a result, hemp paper fibres on and in paper usually form a bundle (white arrows in Fig.1(a)) and fine microfibrils caused by strong beating are frequently seen (black arrows in Fig.1(b)).

Along with Hemp paper making, another high-grade fine paper using the bast fibre of the Chinese paper mulberry tree (*Broussonetia papyrifera*) began to be made during the 5th century at latest (Enami et al., 2007).
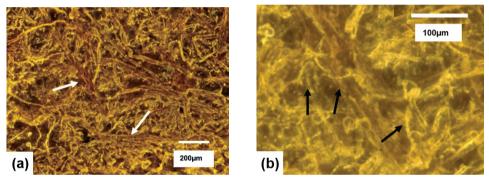
Fig. 1. Microscopy of Buddhists' sutra MS00516 (Avataṃsaka-sūtra of the Tang dynasty written on hemp paper). (a) Bundle of hemp fibres (arrows). (b) Microfibrils of hemp fibres (arrows).
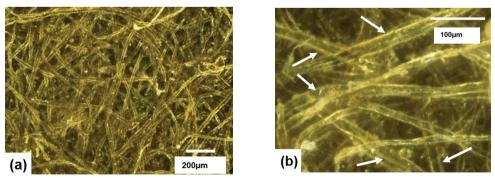
Fig. 2. Microscopy of typical paper made from Paper Mulberry (*Broussonetia papyrifera*) used for Buddhists' sutra MS00503 (the Diamond sutra of the Tang dynasty). (a) Morphology showing Tame-zuki features. (b) Note each fibre is covered by thin sheath (arrows).

Fig. 2 shows the typical paper made from bust fibre of the mulberry tree (*Broussonetia papyrifera*) used for Buddhists sutra MS00503 (the Diamond sutra of the Tang dynasty). As opposed to hemp paper, fibres of mulberry paper are independently distributed (fig.2(a)). In the higher magnification microscopy (b), each mulberry fibre can be seen covered by a thin membrane or sheath (arrows). By this soft membrane, mulberry fibre easily can form a strong hydrogen bond. As a result, fine and dense paper having higher strength compared to hemp paper can be made. In Korea and Japan, there are many Kozo trees (*Broussonetia kazinoki*), and paper continues to be made mainly using fibres extracted from this tree.

1.2. Comparison of Japanese "Nagashi-zuki paper" to Chinese "Tame-zuki paper"

Fig. 3 is a microscopy of paper used for the Selected Works of the poet Bái LèTiān of the Tan Dynasty, re-written in Japan, 1107 AD (Heian period), which was compulsory to study for aristocrats and higher class officers (both men and women) at the Imperial Court during the Heian Period (from the end of the 8th to 12th centuries AD) in Japan. As mentioned above, mulberry (Kozo: *Broussonetia kazinoki*) paper has also been used since the 8th century.
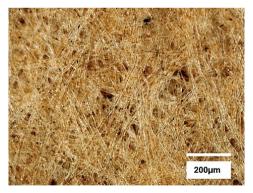
Fig. 3. Microscopy of Japanese Kozo paper used for "The Selected Works of the Poet Bái LèTiān of the Tan Dynasty", "an important cultural asset of Japan (Kyoto National Museum).

Not only the raw material but also the paper-making technique is somewhat different from the Chinese one, especially in the final sheet-making process. Fibre distribution of Chinese mulberry paper shows a randomly oriented feature due to "Tame-zuki"* as seen in Fig. 2 (a), while Japanese Kozo paper made by the "Nagashi-zuki"* technique shows different morphology. Each longer fibre tends to be straightened, and the overall morphology of paper has

a "preferred orientation", as shown in Fig. 3. To distinguish paper used for old texts, whether Chinese or Japanese, is an important problem for codicology. It is confirmed by the present analysis of paper morphology that the text of Fig. 3 was really a hand-copied one for imitating the Tang dynasty's manuscript style in Japan using Japanese Kozo paper.* For the difference in techniques of the "Tame-zuki" and "Nagashi-zuki" methods, see Barrett, 1992.

2. Detection and identification of remains on and in paper: plant, cloth, thread, starch etc.

2.1. The beginning of new a papermaking technique using cereal and grass family plants and sizing treatments using starch started in Central Asia.

Fig. 4 is an image of MS11032, recording the tomb furnishings of a high-rank person's wife of the Qian-Quin (351-394 A.D.) period, during the era of the Sixteen Kingdoms of the Five Barbarians, one of the oldest paper examples of paper of Central Asian origin in the Otani collection.
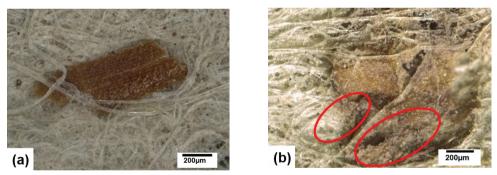
Fig. 4. Microscopy of paper used for the document of tomb furnishings of a high-rank person's wife of the Qian-Quin Kingdom, 385 AD, (a) Fragment of foxtail millet husk. (b) Foxtail millet stalk debris and foxtail millet starch particles (encircled area).

On the paper, a significant amount of plant debris, foxtail millet husk 4 (a) and fragments of stalks and starch particles (encircled area of 4 (b)) with cloth/thread fragments were found to exist. Fibres show the quite different features of those of hemp or mulberry. When compared to the standard sample of starch particles of wheat, foxtail millet and rice, in Fig. 5, it was confirmed that those are identified to be starch of foxtail millet. This is one of earliest examples of sized paper using millet/rice starch. This finding
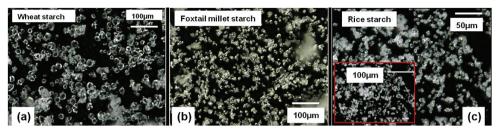
Fig. 5. Microstructure of starches of wheat (3~30µm), foxtail millet (2~10µm) and rice (1~5µm).

suggets that there was a new papermaking technique used in the Central Asian region but not yet in China proper. Since the appearance of paper made from millet stalks during the 4th century, a papermaking technique using grass stalks began to spread widely to the Central Asian region (Enami et al. 2013).

Fig. 6 (a). Mixed paper of foxtail millet fibre and rag used for the Lotus sutra of the 6th century of Northern Wei period (MS08124). (b) Foxtail millet paper for the secular document of Zao family of Turfan region, 741 AD (MS05834).

Fig. 6 (a) is an image of the paper of the Buddhist' sutra MS08124, Lotus sutra, vol. 6, of the Northern Wei of the end of the 6th century. On and in the paper, bristles of foxtail millet with thread fragments were found to exist. A foxtail millet husk remained on the paper of the Zao family (a rich merchant family that lived during 7th to 8th centuries in the Turfan region) in MS05834, a land ownership document (741 AD, Fig. 6 (b)). Paper made from foxtail millet has also been found in several documents collected in the Kucha region of the Tarim basin. These results suggest that during the 4th to 9th centuries, paper made from foxtail millet or other grass stalks had been used for various secular documents by not only ethnic groups of the oasis

states but also by Han emigrants to Turfan and the Tarim Basin regions. The reason why they used millet straws for papermaking is attributed to the fact that the main food stuff of the Central Asian region for more than 4000 years was foxtail millet and common millet (see Enami et al, 2013).
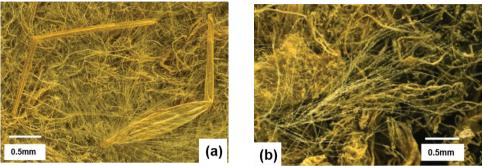
Fig. 7 (a). Reed floret found on the paper of a Uighur document (MS01601), (b) Down of duck found on Uighur reed paper with reed fibres (MS01623).
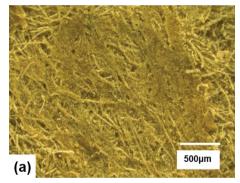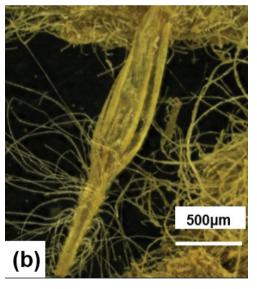
Fig. 8 (a). Reed-origin paper with a beautiful surface treatment suitable for Uighur movable type printing (MS02045). (b) Reed floret found on the same paper, which proves that the paper was really made from reed straw.

Among the Uighur documents in the Otani collection there are a lot of texts using lower-quality paper. Botanical analysis of plant remains found on and in the paper revealed that those are mainly fragments of reed. In addition to the reed-origin debris, bird down fragments were also found as shown in our previous report (Enami et al., 2012). Papermaking involving mixing reed stalk started during the age of Empress Wu Zetian, at the end of the 7th century in Turfan.

Uighurs were nomads, and they had no tradition of cultivating millet. Instead, they were able to get knowledge from texts of the Wu Zetian era using reed-stalk mixed paper and began to make paper using reed stalks because enough reed grew widely in the oases of Turfan. First, they could only make paper ofcomparatively poor quality paper (fig. 7). However, during the 13th to 14th century, when they began to print documents and Buddhists' sutras, they finally could make high-quality fine paper suitable for printing, as seen in fig. 8. It was also found in our former report (Enami et al., 2012) that Uighur-made reed paper became the main paper used by other peoples living in the Central Asian Silk road corridor, including Han emigrants to those areas. In fact, not only the Uighur language but also other languages – Sogdian, Brahmi, and also Chinese – were written on Uighur paper. Uighur paper indeed stimulated the civilization of ethnic groups of the Central Asian region.

2.2. Bamboo paper opened premodern publishing culture during the Song to Ming Dynasties

It should be noticed that the above papermaking techniques using cereal straw and grass straw spread from above mentioned the Central Asian region and then were transferred to China proper during the Wu-dai period until the Song dynasty. Then, Chinese began to make paper using rice straw and bamboo fibres. From the Song dynasty to the Ming dynasty, China became the biggest book publishing country in the world. Among the other various kinds of paper, bamboo paper occupied the first place of paper production and really created the premodern book publishing culture in China. Here some sutras and books were made of bamboo paper. Depending on the unique techniques of the bamboo paper making, plant remains of bamboo or vessel cell debris of bamboo were usually found, as shown in fig. 9. Here again the usefulness of application of the digital microscopy was proved. Such fine bamboo cell structure cannot be found easily using an ordinary microscopy.
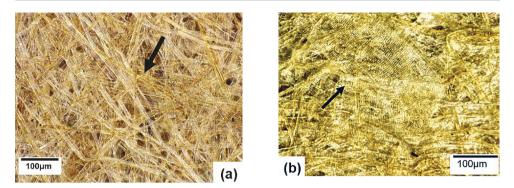
Fig. 9 (a). Buddhist' sutra *Za ahan jing* (Saṃyukta Āgama), Song Huzhou Sī xī Temple Edition, Song Dynasty. (b) Illustrated Encyclopaedia of Cosmos, Earth, Human Life, Ming, 1607, Ryukoku University collection. Vessel cell of bamboo (arrows) is clearly seen on each paper surface.

3. X-ray fluorescent analysis of elements remained on/in paper,

In our former study, it was found using a PIXE (Particle Induced X-ray Emission) analyser that a significant amount of iron can be detected in any Chinese paper, while little iron is contained in Japanese paper (Kohno et.al 2002). On the origin of paper used for the famous Million Charm sutra, which was known to be one of the earliest texts printed in the world, during 764–770 AD in Japan, there has been long debate among Japanese historians about whether the paper was Chinese or Japanese. By digital microscope observation, it was found that the morphology of paper used for the texts showed "Tame-zuki" features (Enami et al 2010). During the very early time of papermaking in Japan, the beginning of the 8th century, paper was made using the Chinese style "Tame-zuki" method and therefore, using only microscopy, it was impossible to identify the origin of the paper. Then, we tried to give an answer to this debate with the help of XRF analysis by searching metallic elements in the paper of the Million Charm texts. We found that little iron could be detected in the paper used for Million Charm texts (A: pure Kozo paper, B:Kozo + Hemp mixed paper) (fig. 10(a) and 10(b)), while in the paper specimens used for two Dunhuang sutras MA00508 of the Otani Collection and Stein Collection (Or. 8210 and S.2077) large amounts of iron was found (fig. 10(c) and fig. 10(d)). Finally, we identified that paper used for the Million Charm sutras as Japanese paper. These results have showed
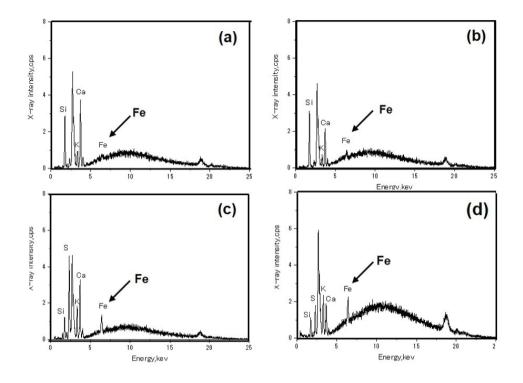
Fig. 10. XRF spectrum of metallic elements of (a) Million Charm A(Kozo), (b) Million Charm (Kozo+Hemp), 764~770 AD, (c) MS00508 Diamond sutra, Tang, Mulberry and (d) Stein 2077, Mulberry, 930 AD. Note the peaks of iron in (a) and (b) were hidden in the back-ground level.

that XRF analysis is surely a useful technique in codicology. The difference in iron content between Chinese and Japanese paper might be attributed to a difference in iron content in the water in China and Japan.

References

Barrett 1992: Barrett T., *Japanese Paper Making*, Wether Hill, NY. and Tokyo, 1983, First paperback ed. 1992.

Enami et al 2007: Enami K., Kato M., Yano T., Kohno M, Barnard M., Matsuoka K. "Analysis of morphology of and elements on the paper specimens of the Stein collection of the British Library." In Lín S., Morrison A. (eds). *Tra-*

dition and Innovation: proceedings of the 6th IDP conservation conference".
The Beijing Library Press, 45–51.

Enami et al 2010: Enami K., Sakamoto S., Okada Y., Masuda K., Kohno
M., "Origin of the differences in papermaking technologies between those
transferred to the East and the West from the motherland China." *Journal of
the Association of paper Historians* 14/2 (2010), 12–22.

Enami et al 2012: Enami K., Sakamoto. S., Okada Y., Kohono M., "Ap-
proach to the History of Social and Cultural life in Medieval China and Central
Asian through the Scientific Analysis of Paper." In Popova I., Liu Y. (eds).
*Dunhuang Studies : Prospects and Problems for the coming Second Century
of Research*. Institute of Oriental Manuscripts, Russian Acad. of Sciences. St.
Petersburg, 39–48.

Enami et al 2013: Enami K., Sakamoto S., Okada Y., Mori M., Masuda K.,
"Paper made from Millet and Grass Fibre found in the Secular Documents of
pre-Tang and Tang dynasty." *IPH 2012 Congress Book* 18 (2013), 149-159.

Kohno et al 2002: Kohno M., Yoshida K., Enami K., Kasajima H., Ueyama
D., "Physical analysis of ancient manuscripts of the Otani Collection." In Whit-
field S., (ed). *Dunhuang Manuscript Forgeries*. London: The British Library,
2002, 268–290.

Acknowledgement:

About the author:

Enami Kazuyuki Born in 1940, Tokyo. Prof. Emeritus, Research Fel-
low of the Digital Archives Research Centre, Ryukoku University. Dr. of
Engineering. Member of the International Association of Paper Historians
since 2010. E-mail muscariae_1114@yahoo.co.jp

Summary
The analysis of the morphology, the fibres, of paper as well as of remains
on and in paper, plant, cloth, thread, etc. in paper by using high resolution dig-
ital microscopy producing 3D images and XRF analysis of elements on and in
paper gave completely new understandings of paper used for old Chinese and

Japanese texts, Sacred sutras and secular documents as shown in the present analysis. The present results suggest that a combination of traditional ways of codicology and novel scientific approaches is indispensable for the further development of the field.

CONSERVATION OF WRITTEN HERITAGE 0.1 – PROSPECTS, CHANCES AND WORK TO BE DONE

Patricia Engel

Abstract: Advances in digital technologies and the progressive digitization of our written cultural heritage have radically changed the nature of, and the outlook on the professional effort for, its conservation in terms of

– the perception of written cultural heritage by human beings (including professionals)

– the content of the conservation work itself.

Digitization – that is, the transmission of information into binary code – has only slowly entered the world of cultural heritage conservation and, in some areas, quite unostentatiously. Therefore it seems meaningful to start with a brief overview of the area where we have already got used to employing digitization in one way or another as a means of conservation of our written cultural heritage. Digital technology also raises certain hopes for new opportunities not only for preserving written heritage but also for understanding it better. The question, however, is in which way we can use digitization from now on in a realistic, effective and most meaningful way to the best benefit of our common human patrimony.

This contribution is based on a series of statistically relevant interviews performed by the author. It also formulates some answers, and the author hopes that it could become a starting point for a targeted interdisciplinary discussion.

Keywords: conservation, material, written heritage, digitization

1. *Introduction*

Augmented reality, animated video projection,[1] 3D public storage facilities – our machine-assisted museums, archives, libraries and collections tend to become a playground and test area for technical possibilities and make us believe that by using these devices and assistance we are able to better our understanding of our cultural heritage items and our history. Yet, this is clearly not something which can be achieved miraculously by the

---

[1] https://www.berlin.de/ausstellungen/archiv/4903983-3238788-von-monet-bis-kandinsky.html.

sheer force of technology. Human cognition is a very complex process, and understanding the ways in which our imagination and perception are influenced by these technologies and whether or not our grasp of things, including our common past, always automatically improves in the process would, of course, require substantial research in psychology and other humanistic sciences; specifically in conservation, which is our concern here, the benefits of these methods must still be balanced against their possible disadvantages.

More "traditional" digital tools provide clear and undisputed advantages in conservation. Digital climate monitoring could be one example, digital description protocols, another. In any case, the actual question and objective must be kept in mind, and the digital solution is not the best by default in each and every case.

2. *The perception of written cultural heritage by human beings (including professionals)*

Digitization makes objects more easily accessible. Information, both as text and as picture, can be obtained via smartphone and the Internet, no matter where we are or when we look it up. We have already become used to having it always ready at hand. As for other items, this is also true for manuscripts. Digitization with the associated democratization of knowledge has enabled a large number of private researchers to dig into their families' histories with the help of digitalized archival material from churches and towns. It is not even clear whether digitization has just met the needs of the hobby researchers or has created new research opportunities.

Cameras of constantly improving quality fitted in our telephones and their availability have made modern people into photographic chroniclers, and they add to the results of science by depicting sights and places to accompany records in registries of births, marriages and deaths. Everyone is taking digital images of everything today, and we find it normal to be allowed access to images of virtually everything.

With regard to images of gospels and other manuscripts which are not classical archival material, we may have become so used to them that in a sense we regard the digital image as the original codex. This process will continue, with photography reaching a level of quality and resolution far beyond our eyesight. It can be assumed that attitudes toward these new tools

differ depending on the age of the person concerned, as the older ones grew up without this technology. At the same time the author observes a growing fascination with "the original", the "real", "the authentic", "an item that was held in a particular person's hand". This does not necessarily always mean a respect for the heritage *per se*, but an emotional satisfaction and a "kick" from the original over a copy that is characteristic of an age in which we gradually move away from the original matter in most things.

Concerning manuscripts, the question is: Where can I find the originals? The easiest way to see manuscripts is to go to an exhibition. For most visitors of an exhibition it is impossible to distinguish a good reproduction from an original manuscript. It needs to be discussed amongst manuscript experts, museum staff, exhibition designers and conservators what it actually means to expose an original manuscript in an exhibition where very likely not even 1% of the visitors can judge whether a photograph is a reproduction or the original. What we are talking about here is not the exposition to light and any degree of potential "attrition" of the material, but access – meaningful access. Of the estimated 1% of visitors who can see the difference between a photograph and a manuscript, only very few content themselves with the single two-page opening exposed through the glass of the showcase. We must ask whether it would not make much more sense for those wishing to see the original to go to the library and access it as a reader, without the glass of the showcase and in silence and concentration as well as with the possibility to see also other folios beyond the two particular leaves shown in the exhibition.

The author once saw an exhibition in which 50% of the manuscripts were copies. Used to seeing, identifying and understanding manuscript material, when she suspected that the exposed objects were not the original books, she tried to analyse their surfaces in raking light from various points to verify her suspicion. It was finally confirmed by the head conservator of this important library, although the description in the showcase did not state this fact. Thus, experience had clearly shown to the author that (1) because of the distance created by the showcase, the dim lighting, and other factors it was really difficult to guess that the manuscripts did not represent originals, and (2) obviously no one complained about it; otherwise, the library could not have displayed this sort of exhibition over decades.

That digitization can pose a serious problem for researchers was demonstrated by the author in her publication on the Cod. Guelf. 84.5 Aug. 2, where the facsimile edition was retouched in such a way that the stitches of the byssi became completely invisible. The author had published articles on this manuscript (see, e.g., Engel & Gallistl 2009: 129–178) and knew its material and technique of production as well as the changes of the material over time. Therefore, when she saw the facsimile print, she immediately recognized that the stitches were not reproduced. The presence or absence of byssi is an important feature and so is the sort of material and style of the sewing with which they are fixed. Consequently, using ordinary photographs or facsimiles for studies, which is often done, does not meet the requirements of a serious researcher if he/she has to do with a codex whose reproduction was retouched in such a manner. Only after having studied the underdrawings in manuscripts of the Reichenau school, partly on the basis of facsimiles and after having written about it, the author became aware of this problem herself.

Finally, there is another, positive aspect related to the digital imaging of manuscripts: scholars become more clearly aware of the value of the material as a source of information about the history of the manuscripts. Of course, digital photos for this kind of use have to be multi-spectral, i.e., made in various wave lengths, to be presented in various false colours or processed in some other way. The usefulness of transmitted light images is still underestimated in this respect. The author had once obtained a small black and white image of a papyrus scroll in transmitted light which she showed to Myriam Krutzsch, who dedicated her life to the conservation of papyri. Although her colleague did not have any other information, she immediately defined the object as a death book (which was in fact the case) and was able to judge its proper date and place of origin, having "read" the material in transmitted light. The use of transmitted light images by Jiří Vnouček (Vnouček 2017) for the identification of the number of skins used for a certain codex and the manner of production based on the scraping patterns of the skins whose traces are preserved on the parchment represents another way of using photos made under transmitted light. It almost goes without saying that raking light is also a very useful method to survey material.

3. *The work of conservation*

Conservation, conceived primarily as the understanding of a work of art, for example a manuscript in its artistic and historic dimension as well as in its material aspects as to bring it into future, starts with the description of the material and the meaning of the object in question by the conservator as a result of his/her own survey and study of information from other fields of knowledge. Building on this ground, damages can be identified, and a conservation concept can be developed.

In the old days photo-documentations were done in black and white, because their images were considered more durable than colour prints. Usually, there was a considerable time span between taking the photographs and receiving them in hand as prints in order to see if all were properly focused, so the practical conservation could begin. This procedure was essential to ensure that the status of the manuscript at the moment it came into the hands of the conservator was documented sufficiently. Today, with the use of digital methods, quality control of the images is done within seconds, and their costs are so low that it is possible to make ten times more photographs without cost.

Multi-spectral imaging, especially in non-visible light, can assist the conservator in his/her material survey, and image processing software allows for extracting more information from the images and comparing their features.

This method speeds up the conservation procedure significantly and permits the conservator to stay with one manuscript at a time, while in earlier times he/she had first to take all necessary photographs of all items in his/her studio and then return to the first object to start work; thus, a different time management is established.

The fact that today cameras are omnipresent in our daily life lets us also suggest ideas for their more meaningful use in our professional tasks. Naturally, this was also done before, but today it can be established in a much broader fashion. There are several setups in conservation where we use images as a tool of observation, namely

– if we do not want to interfere too quickly, but create some time for observation, and

– if we do not know *ad hoc* which would be the best conservation measure for a specific problem at hand.

In both cases digital images can be a most meaningful tool.

For example, in the course of the development of a conservation concept for the Hildesheim manuscripts kept in the Hildesheim treasury, areas of manuscripts decorated with silver leaf showed significant darkening, both as a halo around the areas on the recto sides and as a darkening due to iron migration on the verso. Older pictures could be used to compare the recent situation with the previous one. In such a case digital monitoring is a useful tool to document the decay occurring over time.

The same method can be used for ink corrosion decay, but it was also used by the author for a survey of insect infestation. As for the latter, hundreds of frass holes are difficult to monitor otherwise.

The degree of gelatinization of parchment is also very difficult to describe without the help of images combined with a colour card. Its yellowish glassy state cannot be described in detail, unless the shrinkage temperature is measured; which again requires a serious effort for the extraction of fibres. For a simple documentation of the speed of decay a digital image seems as a good choice.

In general, conservation work involves documentation of the state of an item before conservation begins, of all individual steps taken and the state of an item after restoration. Digital imaging allows for a much more detailed photographic depiction of all these states than was possible before. While one picture before and one after the conservation-restoration process was never sufficient, taking so few images was common practice.

We can expect that all of our collected digital data will one day be integrated into big data. They show and preserve many details of the material and the techniques of its production (parchment, paper, inks, pigments and dyes), as well as the making of the codices. Attempts have already been made to construct a complete digital and uniform conservator's documentation set (Ravenberg 2016: 251-268), and there are numerous databases and portals striving to achieve a uniform material documentation.

All these data should not only be collected but also connected to library catalogue entries. Presently, library catalogue entries have not yet reached the depth they could achieve if the descriptions of conservators were taken into consideration. Such connections would enhance the understanding both

of the librarians and the scholarly community, ensure their deeper under-
standing of the material and the suggestions of the conservators. Then, we
all would be better equipped to estimate the quality of a conservation plan,
which eventually contributes to the well-being of the object at hand.

4. *Conclusion*

Based on her observations, interviews and discussions, the author could
derive five premises concerning public and digital images of manuscripts:

1. We should use the interest in the "original" as an instrument to raise
awareness of the delicacy and fragility of original manuscripts, the value
they actually represent for society and the public resources needed for re-
search into conservation and the practical application of conservation mea-
sures. Digital media and the Internet can also be used for this purpose.

2. We should start a discussion about current exhibition practices. For
years museum experts have been leading such discussions about new didac-
tic concepts in museums, but we also need discussions specifically dedicat-
ed to manuscripts.

3. Digital imaging and post-processing of the images by computer vision
technology are excellent research tools. Commercially produced facsimiles
can no longer achieve a level which is needed for serious research. Also, the
value of transmitted-light images and raking-light images is still underesti-
mated; because it is very difficult to place the viewing axis of the camera at a
shallow angle to the folio surface in the same manner as the eye of the observer.
Raking-light perspective often reveals information only if the observation is
done from various angles and the results can be compared in the observer's
brain. Here lies a challenge for computer scientists to simulate this human
ability with digital machines (cf. the Light Dome developed for CIMA by
Simon Brenner[2]).

4. In conservation slow decay processes which cannot be monitored
can easily be documented by digital images; this holds true for metals which
destroy the parchment or paper support, insect frass and gelatinization (to
name only a few examples).

5. Linking the information about material which was collected in count-
less conservators' descriptions of manuscripts and conservation measures

---

[2] See the article of H. Miklas in this volume.

will represent a knowledge base that would serve the conservators and the scholars of other fields and surely lead to new insights of various kinds.

References

Engel 2011: Engel, P. "Digital Images: A Valuable Scholar's Tool or Misleading Material?" in International Conference on Integrated Information (IC-ININFO) 2011, Kos, Greece, 2011, 272 –276.

Ravenberg 2016: Ravenberg, H. "'ARM': A Data Model for Book Conservation Treatment and Knowledge Sharing." in Historical Book Binding Techniques in Conservation, ed. G. Boudalis, M. Ciechanska, P. Engel, R. Ion, I. Kecskemeti, E. Moussakova, F. Pinzari, J. Schiro, J. Vodopivec. Horn, Wien: Berger Verlag, 2016 (ISBN 978-3-85028-785-2, 251–268).

Engel & Gallistl 2009: Engel, P., B. Gallistl. "Die Reichenauer Handschriften der Dombibliothek Hildesheim und der Herzog August Bibliothek Wolfenbüttel im Vergleich." Wolfenbütteler Beiträge 15 (2009), 129–178.

Vnouček 2017: Vnouček, J. "The 12th Century Glossed Gospel of St. Luke from St. Augustine´s Abbey in Canterbury – How to understand the production of the manuscripts with the help of visual and biomolecular analyses of parchment preservation." Paper, given at the WAF general meeting, 16th January 2017 (https://waf.cvl.tuwien.ac.at/news.php – 1st Feb. 2019).

About the author

Patricia Engel holds magister, doctorate and habilitation degrees in conservation-restoration of cultural heritage of the universities of fine art in Vienna and Warsaw. She worked as assistant professor in the Hildesheim HAWK, Germany, and from 2010 has headed the European Research Centre for Book and Paper Conservation-Restoration affiliated to the Center for Cultural Property Protection at the Department for Building and Environment at the University for Continuing Education, Krems, Austria. Her main focus is on key questions in written heritage conservation. She initiated and lead international projects (Getty, EU, FFG) on mould, paper-deacidification and ink corrosion. E-mail: patricia.engel@donau-uni.ac.at.

Summary

This contribution takes as its starting point the position of both a researcher in the humanities and a conservator of written heritage. It is based on a series of statistically relevant interviews performed by the author and illustrates using new technologies for attracting an audience by meaningful choice of digital

tools. Digital climate monitoring could be one example; digital description protocols, another.

Next the perception of written cultural heritage by human beings (including professionals) is surveyed. The omnipresence of information, both as picture and text, obtainable via the internet and portable devices, is described as having changed our expectations and demands which also apply to old texts, such as archival data and manuscripts. On the other hand, the wider population has evolved into a contributor of data, with their technical tools such as cameras in smartphones.

With regard to images of the Gospels and other manuscripts which are not classical archival material, we may have become so used to them that, in a sense, we regard the digital image as the original codex. This process will continue, with photography reaching a level of quality and resolution far beyond our eyesight. It can be assumed that the attitude to these new tools differs depending on the age of the person concerned, as older persons grew up without this technology. A trend towards a growing fascination with "the original", the "real", "the authentic", or "the item which was once held in a particular person's hand" – emerging as if to balance the technical nature of our environment – is then described. Finally, the question of when to use the original manuscript and when to use a photo of it is put forward.

The next chapter is concerned with digitization of the material items in practical conservation. The benefit of digital over analogue photography in this respect is described in detail and with various examples. Furthermore, photography as a tool of observation is explained, including its many additional options, such as multi-spectral imaging and processing digital pictures in various ways.

Finally, the author advocates using digital means for raising awareness about the value of cultural heritage and the need to preserve it, and suggests a new discussion about current exhibition practices, the introduction of simple yet informative survey methods in conservation for the digital world, and linking together various pieces of information about material which was collected in countless conservators' descriptions of manuscripts and conservation measures, representing a knowledge base that would serve the conservators and scholars of other fields and lead to new insights of various kinds.

CODICOLOGICAL AND PALEOGRAPHICAL ASPECTS IN CONSERVATION
(Based on data from the Tsaishi Gospel)

Darejan Gogashvili

Abstract: The conservation of the *Tsaishi* Gospel (N14, preserved at the Zugdidi Museum, Georgia) is based not only on knowledge obtained in the field of restoration-conservation but also on that acquired through codicology and paleography. The only published description of the manuscript is a brief one from the early 20th century, which dates the Gospel to the 18th century. More thorough examination was impossible because serious damage rendered the manuscript unfit to be studied: the leaves of the book were stuck together and represented a single mass. An appropriate conservation methodology was eventually developed, and, as a result, the Gospel has now been partially separated, and fragments of three miniatures inserted into the manuscript were discovered and cleaned. Comparative iconographic analysis determined the place of each fragment and resulted in the reconstruction, identification, and dating of the previously unknown miniatures. One of them was executed on paper; the remaining two on parchment.

Codicological and paleographic analysis reveals that the *Tsaishi* Gospel is a compilation of two different manuscripts, dated to the 13th and 17th centuries. The miniatures come from two different genres of manuscripts, gospel and liturgy; they are not original to the *Tsaishi* Gospel, and they were placed between the pages for protection or preservation.

The example of the *Tsaishi* Gospel demonstrates the use of codicological-philological research and systematic analysis in conservation to create a solid basis for interdisciplinary collaborative work, and it broadens the area of activities in conservation.

Keywords: Manuscript, miniature, codicology, palaeography, conservation, Identification

Introduction

The *Tsaishi* Gospel (N14) is preserved in Georgia at the Dadiani Palaces History and Architectural Museum. The manuscript received its name after its last location of safekeeping. Before its relocation to the Museum, the manuscript was kept at *Tsaishi Monastery* (at *Tsaishi*) *Dormition Cathe-*

dral), which for centuries was a center of education, served as a cathedral church, and represented one of the oldest episcopal seats of Georgia.[1]

The only published description of the manuscript with colophons is a brief one from the early 20th century, which dates the Gospel to the 18th century. According to the description, the manuscript was commissioned by Archbishop of *Tsaishi Maxime*.[2] More thorough examination was impossible due to serious damage which rendered the manuscript unfit to be studied.

The Manuscript Description before Conservation Treatment

The text of the Gospel is written on paper in *Georgian Nuskhuri script*. The manuscript was badly damaged: the pages of the book were stuck to one another and represented a single mass. The Gospel was accompanied by small pieces of paper fragments and two severely damaged miniatures: one of them is executed on parchment and other one on paper, and a significant part of both miniatures' surface was covered by small paper fragments glued to each other.

The Sewing Station

The book block consists of quires. Due to damage, it was not possible to identify the exact number of folios or quires, the sewing type of the endband, the method of attachment of the textblock to the board and so on, but I was able to investigate the spine of the textblock and sewing structure (Fig. 1).

Observation has shown:

• The manuscript spine is not smooth and has a supported sewing structure;

• The textblock is sewn using a flexible sewing method on single raised cords, passing the thread entirely round all the cord;

• The distance between the sewing stations is almost the same, changing from 50 to 52 mm;

The outermost segment is of a different size – 30–32 mm smaller than the distance between the main stations;

---

[1] The Monastery is located in western Georgia, in the Samegrelo-Zemo Svaneti region. It was built in the 13th-14th centuries and restored in the 17th century. The oldest part of the church dates back to the 11th century. Zakaraia: 1981, pp. 89-99; 126- 134

[2] Taq'aishvili: 1913-1914, pp. 187-189

• The sewing thread is of a diameter of approx. 1 mm, twisted in an S direction and composed of three z-spun yarns (3z-S);

• The diameter of the cord of the sewing support is approx. 4 mm, has re-plied threads with a final S twist, the structure of which is 4s-2z-S.

Conservation Treatment

The pre-conservation study of the manuscript has been carried out. The causes and degree of the damage have been identified. The study has shown that the damage is caused by heterotrophic microorganisms such as bacteria and microscopic fungi. As a result of their impact, the paper pages of the manuscript are adhered together, and represent a single mass. A significant part of the paper leaves are disintegrated, the middle of the manuscript has turned into dust and it is no longer possible to save it. Some pigment spots of various colours caused by the microorganisms are noticeable on the surface[3].

On the basis of the conducted tests, the appropriate conservation-restoration materials and method have been worked out and selected, the stages and specificity of work have been determined. As a result, the manuscript has been partially separated, the pages which were adhered to each other have been brought apart, treated, and consolidated. The surfaces of miniatures and small-sized fragments have been cleaned, flattened and temporarily fixed. The place of each fragment in the general schematic drawings was defined by the comparative analysis of miniature paintings inserted in illuminated Georgian gospels and liturgies.

Results and Discussions
I. Codicology and Paleography

On the basis of codicological and paleographic analysis, it has been revealed that:

1. The text of the Gospel is a compilation of two different manuscripts, dated to the 13th and 17th centuries; it is of the redaction of *Giorgi the Hagiorite*, accompanied by the Index of Yearly Lessons from the Gospels[4];

---

[3] Gogashvili: 2010, pp. 88, 89; fig. 1 – 4
[4] Gogashvili: 2010, pp. 89, 95

The text of the Gospel is executed on paper in calligraphic *Nuskhuri script*. Due to the damage, it is not possible to identify the number of folios or quires. The text is written in two columns with black ink (writing ink is stable and insoluble in water); the headings and some of the initials are in *Asomtavruli script* and in red ink (vermilion);

The size of the folio is 220 × 175 mm; text size is 155 × 105 mm (reconstructed); width of the columns varies from 45 to 47 mm; the space between columns is 15/14 mm; every column is made up of 24 lines (reconstructed); the space between the lines varies from 6 to 8 mm; margins: top: varies from 23 to 28 mm; bottom: 30/31 mm; right: varies from 33 to 36 mm; left: varies from 26 to 36 mm.

II. Iconographic and artistic-stylistic analysis

The manuscript was accompanied by two damaged and deformed miniatures. One of them is executed on paper, and the other on parchment. Another third miniature on parchment and small-sized fragments belonging to all three miniatures were found during the special conservation work and partial separation.

One of the miniatures painted on parchment was seriously destroyed, damaged by microorganisms; the other was entirely covered with paper fragments stuck to each other. We were able to determine the two portraits of the sitting Evangelists by cleaning the miniatures and constructing schematic drawings with the small fragments. The Evangelists are represented in the process of work, in accordance with the manuscript decoration iconographic models that have been accepted since the 11th century.[5] With iconographic and artistic-stylistic analysis it was determined that one of the miniatures represents St. Mark (194/192 × 141/143 mm) and the other St. Luke (186/188 × 143/145 mm) (figs. 2 and 3); their date of execution is at the turn of the 11th – 12th centuries. Further analysis of the painting technique, style and artistic level has determined that the miniatures likely belonged to a richly decorated manuscript executed on parchment.[6]

It is established that among the illuminated manuscripts, the gospels are distinguished by their principles and specific character of decoration,

---

[5] Шмерлинг: 1967, p. 188-189
[6] Gogashvili: 2010, pp. 90-95; 104-105; fig. 5, 6, 8; Gogashvili & Kavtaria: 2018, pp. 159 – 162; 166; fig. 2-9

adopted a well-developed artistic pattern by the 11th century. According to this pattern, Eusebius's Letter and Canon Tables are exhibited at the beginning of the text of the Gospels. The text itself is divided into four chapters, the Evangelists' figures are arranged in accordance to the corresponding chapters; each Evangelist's portrait is placed on the verso page opposite the title page of the corresponding Gospel.[7] As for the *Tsaishi* Gospel, the place of the portrait of the Evangelist does not correspond to the text: St. Luke's illustration was placed before St. Mark's Gospel, and furthermore the writing material and sizes of folio are different. This all brought us to the conclusion that the portraits are not the part of *Tsaishi* manuscript decoration; rather, they were placed in-between the pages for protection and preservation purposes.

The third miniature executed on paper was severely damaged: it accompanied the manuscript separately, a significant part of its surface was covered with fragments of paper glued to each other (fig. 4).

Small fragments belonging to the miniature were found after the special conservation treatment of the Gospel; this gives us reason to think that the miniature was inserted in the manuscript (fig. 6). The surface of the miniature and its fragments were cleaned (fig. 5). The miniature was found to have structural damage as well as biological damage caused by microorganisms. As a result of their operation, pigments used in miniature painting in some cases almost completely disintegrated or became discolored, and only fragments are visible on the surface of the paper.

Regardless of the severity of the damage, the figure on the illustration is still visible; the garment (phelonion, omophorion decorated with crosses), pose, golden nimbus, gold leaf background, portrait marks (elongated facial contours, short beard, dark hair), gives us reason to assume that the miniature illustrates one of the three Holy Hierarchs (St. John Chrysostom, St. Basil the Great, or St. Gregory the Theologian) – likely the image of St. John Chrysostom[8].

The place of each fragment in the overall scheme was determined: the details of the painting layers preserved on the miniature and the shape of the

---

[7] Шмерлинг: 1967, p. 189; Mach'avariani: 2012, p. 225.

[8] For St. John Chrysostom' s iconography see Osepashvili: 2002 Osepashvili, L. pp. 29-31

fragments helped us to position six fragments out of seven[9] (fig. 8). As to the seventh fragment, despite the significant damage to the painting layer, it was possible through iconographic analysis to identify and determine its place in the general schematic drawing. There is a detail of a small table and sacred vessels used during the Divine Liturgy depicted on it (fig. 7).

On the basis of iconographic analysis it has been possible to reconstruct the schematic drawing and determine that the figure of St. John Chrysostom represented on the miniature is on the verso page (fig. 9). The miniature is not part of the *Tsaishi* Gospel's illumination; instead, it is a portrait of a completely different origin, belonging to a manuscript of the liturgical genre.

The image's iconography demonstrate characteristic resemblance with the manuscripts of the 18th century (A-123) and 17th-18th centuries (A-1555) in particular,[10] with the image of John Chrysostom included in the decoration of the Divine Liturgy. It should also be noted that the size of the illustrated page of St. John Chrysostom (from the *Tsaishi* Gospel) is almost the same as the Holy Hierarchs' portraits from A-1555. This gives us grounds to assume that the manuscript to which the miniature from the *Tsaishi* Gospel belonged and the manuscript A-1555 should be identical in size.

Dimension	Miniature from Tsaishi Gospel	A 123, Liturgy, 17th-18th centuries.	A 1555, Liturgy, 18th century.
I	*St. John Chrysostom*	*St. John Chrysostom*	*St. John Chrysostom*
Folia size		194/195 × 145 mm.	212 × 164 mm.
Miniature size	162/160 × 120/118 mm.	133 × 94/132 × 93 mm.	165 × 124 mm.

---

[9] Sizes of fragments: fr.1. 9 × 5 mm; fr.2. 12 × 5 mm; fr.3. 24 × 21 mm; fr.4. 19 × 8 mm; fr.5. 3 × 3 mm; fr. 6. 9 × 4 mm; fr.7. 32 × 20 mm.

[10] A-123 and A-1555 preserved at Korneli Kekelidze Georgian National Center of Manuscripts.

A-123. Liturgy. 17th-18th centuries. 11+122 folios. 19.5 × 15. Paper. Contents: Liturgy of St. John Chrysostom; Liturgy of St. Basil; Liturgy of St. Gregory. See Kartul khelnats'erta aghts'eriloba 1976: pp. 101-103. A-1555. Liturgy. 18th century. 142 folios. 21.5 × 16.4. Paper. Contents: Liturgy of St. John Chrysostom; Liturgy of St. Basil; Liturgy of St. Gregory. See Kartul khelnats'erta aghts'eriloba 1955: pp. 73-74.

Dimension	Miniature from Tsaishi Gospel	A 123, Liturgy, 17th-18th centuries.	A 1555, Liturgy, 18th century.
Margins: top, bottom, right, left	2/1 mm; 2/0 mm; 0 mm; 6/4 mm.[11]	27 mm; 34/33 mm; 22 mm; 30 mm.	14 mm; 34/33 mm; 14/13 mm; 26 mm.
II		*St. Basil the Great*	*St. Basil the Great*
Folia size		194/193 ×146 mm.	213 × 164 mm.
Miniature size		139 × 97 mm.	163 × 120 mm.
Margins: top, bottom, right, left		22/20 mm; 33 mm; 17 mm; 34 cm.	15 mm; 35/ 34 mm; 15/17 mm; 30 mm.
III		*St. Gregory the Theologian*	*St. Gregory the Theologian*
Folia size		193 ×145 mm.	213/212 × 165 mm.
Miniature size		137 × 96/95 mm.	160 ×122 mm.
Margins: top, bottom, right, left		21/20 mm; 35/34 mm; 20 mm; 31/30 mm.	20 mm; 34/33 mm; 15 mm; 28 mm.
IV		*A 123, Liturgy*	*A 1555, Liturgy*
Folia size		195/194 × 145 mm.	213 × 165 mm.
Text size		140/138 × 105 mm.	164 × 125 mm.

Conclusions

Codicological and paleographic analysis revealed that the *Tsaishi Gospel* is a compilation of two different manuscripts dating from the 13th and 17th centuries.

The miniatures come from two different genres of manuscripts, gospel and liturgy; according to artistic-stylistic analysis, the portraits of Evan-

---

[11] The margins (top, bottom, right, left) of the miniature from the *Tsaishi* Gospel are cut off (fig. 5).

gelists (St. Mark and St. Luke) should belong to a richly decorated manuscript executed on parchment; St. John Chrysostom's portrait belonged to a genre completely different from the gospels: liturgy. The miniatures are not original to the *Tsaishi Gospel*, and they were placed between the pages for protection or preservation.

The example of the *Tsaishi Gospels* demonstrates the use of codicological-philological research and systematic analysis in conservation to create a solid basis for interdisciplinary collaborative work, and broadens the area of activities in conservation.

References

1. Gogashvili & Kavtaria 2018: Godashvili, D. Kavtaria, N. ''Peculiarities of Tao-Klarjeti Artistic School of the Miniatures from Tsaishi Gospels.'' *Pro Georgia, Journal of Kartvelological Studies* 28 (2018), 158-174.

2. Gogashvili 2010: Gogashvili, d. „tsaishis otkhtavze dartuli miniat'urebi (k'onservatsia, ident'ipik'atsia)", *Saertashoriso k'onperentsia t'ao k'larjeti, masalebi, Tbilisi 2010, 87-105* [გოგაშვილი:2010. გოგაშვილი, დ. „ცაიშის ოთხთავზე დართული მინიატურები (კონსერვაცია, იდენტიფიკაცია)", *საერთაშორისო კონფერენცია ტაო-კლარჯეთი, მასალები,* თბ. 2010, 87-105].

3. Kartul khelnats'erta aghts'eriloba 1955: *(A) k'olektsia, t'. V, Tbilisi 1955* [ქართულ ხელნაწერთა აღწერილობა 1955: *(A) კოლექცია ტ. V,* თბ. 1955].

4. Kartul khelnats'erta aghts'eriloba 1976: *(A) k'olektsia, t' 1₂ Tbilisi 1976* [ქართულ ხელნაწერთა აღწერილობა 1976: *(A) კოლექცია ტ. 1₂* თბ. 1976].

5. Mach'avariani 2012: Mach'avariani, E. "kartuli khelnats'eri ts'ignis dek'oratsiuli sist'emis ganvitareba (IX-XIV ss.)." *Kartuli khelnats'erebi, damts'erlobisa da shemk'ulobis sak'itkhebi*, Tbilisi. 2012, 221-234 [მაჭავარიანი 2012: მაჭავარიანი, ე. "ქართული ხელნაწერი წიგნის დეკორაციული სისტემის განვითარება (IX-XIV სს.)". *ქართული ხელნაწერები, დამწერლობისა და შემკულობის საკითხები,* თბ. 2012, 221-234].

6. Osepashvili: 2002 Osepashvili, L. „Ts'minda ioane okrop'iris ik'onograpia kartul khelnats'erta miniat'q'rebze." *Zhurnali, religia* 4-5-6 (2002), 28-36 [ოსეფაშვილი: 2002, ოსეფაშვილი, ლ. „წმინდა იოანე ოქროპირის იკონოგრაფია ქართულ ხელნაწერთა მინიატურებზე" *ჟურნალი, რელიგია,* № 4-5-6, 28-36].

7. Шмерлинг 1967: Шмерлинг, R. *Художественное оформление грузинской рукописной книги IX-XI столетии.* кн. I, Тбилиси, 1967 [Шмерлинг 1967: Шмерлинг, R. *Художественное оформление грузинской рукописной книги IX-XI столетии, кн.I,* Тб., 1967].

8. Taqʻaishvili 1913-1914: Taqʻaishvili, E. *dzveli sakartvelo, tʻomi III*, Tbilisi 1913-1914 [თაყაიშვილი 1913-1914: თაყაიშვილი, ე. *ძველი საქართველო, ტომი 3,* თბ. 1913-1914].

9. Zakaraia 1981: Zakaraia, P. *Kartuli tsentʻralur-gumbatovani arkitʻektʻura, XI-XVIII ss. №3 XIII-XVIII ss.* Tbilisi, 1981 [ზაქარაია: 1981: ზაქარაია პ. *ქართული ცენტრალურ გუმბათოვანი არქიტექტურა, XI-XVIII სს. №3 XIII-XVIII სს.* თბ. 1981].

About the author

Darejan Gogashvili, deputy head of the Museum of Fine Arts, Georgian National Museum. Graduated in chemistry and art history in Georgia, in Tbilisi State University, in 2005 completed her PhD. Has been an intern in the British Library (2000) and the Conservation Department of the Berlin State Library (2001, 2010). Has been working as a manuscript conservator and senior scientist (2008 – 2016) in the Georgian National Museum. Participated in various forums and projects and has been the author of more than 20 publications. Email: gogashvilid@gmail.com

რეზიუმე

ცაიშის ოთხთავის (№14. დაცულია ზუგდიდის ისტორიულ მუზეუმში) კვლევისა და კონსერვაციის მეთოდი ეყრდნობა, როგორც რესტავრაციის სფეროში არსებულ ცოდნას, ასევე ფილოლოგიური მეცნიერების ერთ-ერთი მიმართულების, კოდიკოლოგიის დარგში არსებულ სამეცნიერო ცოდნას.

დაზიანების გამო, ხელნაწერი არ ექვემდებარებოდა კვლევასა და შესწავლას. მისი ერთადერთი მოკლე აღწერილობა გამოქვეყნებული აქვს ექ. თაყაიშვილს, იგი ოთხთავს XVIII ს. ათარიღებს. ხელნაწერის კვლევა შესაძლებელი გახდა, მხოლოდ მისი ნაწილობრივი დაშლისა და წინასწარი საკონსერვაციო სამუშაოების ჩატარების შემდეგ, რომლის დროსაც გაიწმინდა და დამუშავდა ერთმანეთს შეწებებული ფურცლის ცალკეული გვერდები, მინიატურები და მცირე ზომის ფრაგმენტები. გამოვლენილი მასალის კოდიკოლოგიური, იკონოგრაფიული და მხატვრულ-სტილური ანალიზის საფუძველზე

დადგინდა რომ: სახარების ტექსტი წინასწარი პალეოგრაფიული მონაცემების მიხედვით, ორი XIII ს. და XVII ს. ხელნაწერების კომპილაციას წარმოადგენს; ტექსტში ჩართული სამი მინიატურიდან, ორი ეტრატზე შესრულებული მახარებელთა მჯდომარე პორტრეტებია (წმ. მარკოზი, წმ. ლუკა), მათი შესრულების თარიღი XI-XII სს. მიჯნით განისაზღვრა; მესამე, ქაღალდზე შესრულებულ მინიატურაზე წმ. იოანე ოქროპირია გამოსახული, რომელიც ოთხთავისაგან სრულიად განსხვავებული შინაარსის – ლიტურგიკული ჟანრის ხელნაწერს, ჭამისწირვას ეკუთვნოდა; მინიატურები არ წარმოადგენდა ციაიშის ოთხთავის გაფორმებას, ისინი მხოლოდ დაცვის ან შენახვის მიზნით დაურთეს მას.

Fig. 1. A) The Spine of the text block with four sewing stations; B) Remain of natural coloure sewing thread
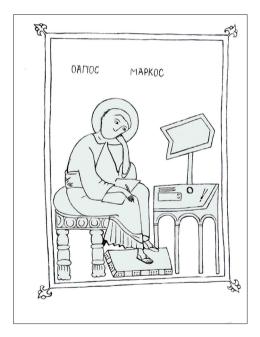
Fig. 2. St. Mark the Evangelist. Drawing, reconstruction
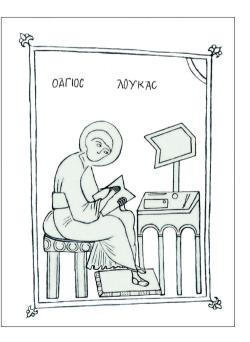
Fig. 2. St. Mark the Evangelist. Drawing, reconstruction

Fig. 4. Miniature on paper before conservation treatment

Fig. 5. Miniature on paper after conservation treatment

Fig. 6. The small fragment (size: 24 × 21 mm) [B] has been discovered due to the partial disassembling of the Gospel [A], linking to the fact that the miniature was inserted in the manuscript
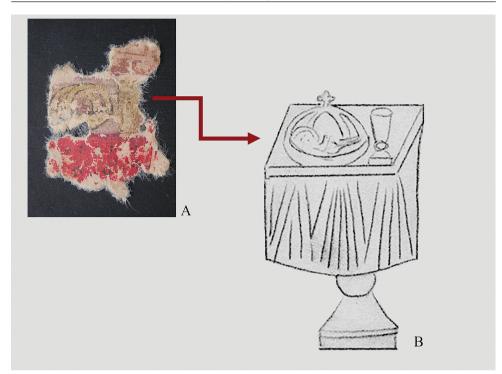
Fig. 7. A) Fragment 7 (size: 32 × 20 mm), detail of a small table and sacred vessels used during the Divine Liturgy; B) Small table and sacred vessels. Drawing, reconstruction
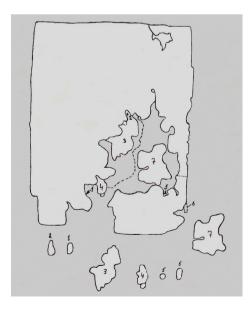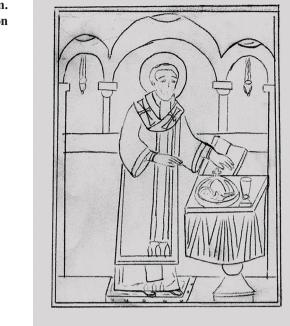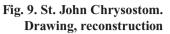
Fig. 8. Schematic drawing of the miniature and its fragments

Fig. 9. St. John Chrysostom.
Drawing, reconstruction

Eig. 10. A-123. Liturgy. 17th-18th cc. St. John Chrysostom.
Title page of the of St. John Chrysostom's Liturgy

Eig. 11. A-1555. Liturgy. 18ᵗʰ c. St. John Chrysostom.

THE IDENTIFICATION, DEFINITION
AND TYPOLOGY OF MANUSCRIPT LACUNAE

Elisabet Göransson

Abstract: In this article, the question of how to identify and define lacunae is discussed, both in general and exemplified in manuscripts containing collections of monastic sayings of the Desert Fathers. These sayings, organized in different ways, were probably written down in Greek during the early fifth century and were soon translated into many languages. In a relational MySQL database, the sayings as they appear in the manuscripts are given a unique identifier. The structure and the contents of each manuscript can be compared through this reference system, which builds on the unique structure of each manuscript. However, when a manuscript contains lacunae, the results when comparing the structures are not trustworthy. This article discusses how to deal with such problems. A general definition of the notion of *lacuna* and a typology for lacunae that could be used both by cataloguers of manuscripts and by scholars are suggested. Moreover, a model for how to calculate the parts of a text that are missing in smaller lacunae is presented.

Keywords: lacuna, MySQL relational database, Apophthegmata Patrum, Sayings of the Desert Fathers

Introduction and background

The sayings of the Desert Fathers, also known as "Apophthegmata patrum" or "Verba seniorum," are the Christian versions of the classical rhetorical *chreiai*, the school exercises, that were attributed to the early Desert Fathers and Mothers. A *chreia* is a very broad category of either sayings or actions. This type of narrative was defined by Theon of Alexandria in his *Progymnasmata* with the following: "A concise statement or action which is attributed with aptness to some specified character or to something analogous to a character" (Hock & O'Neil 1986: 82–83).

The sayings were preserved and transmitted in different collections, primarily alphabetically or thematically organized, and they were translated very early on into the languages used in early and medieval Christian monastic contexts. The text traditions including these collections are thus very

wide and multifaceted indeed, and, moreover, multilingual, since translations were made in different stages as well in the transmission process, during which the collections themselves also developed in different ways.

In previous scholarship it was taken for granted that what has later been acknowledged as the sayings of the Desert Fathers would first have circulated only as orally transmitted texts, and later on were collected and written down (Harmless 2004: 248–251; Burton-Christie 1993: 76–81). This has been questioned more recently since the collections of sayings also include material from letters, lives and sermons that were circulated in written form at an early stage. Scholars, however, concur in the view that the monastic sayings were first written down and organized in collections in Greek, probably during the early fifth century (Faraggiana 1997; Rubenson 1995: 152–157.) It has been suggested that the first stages in this process took place not in Egypt, but in Palestine (Regnault 1987:73–83). Scholarship on the collections of sayings in recent years has been focused on the relationship between early monasticism and classical *paideia*; the didactic function and use in relation to the *Progymnasmata* tradition have been studied and discussed by several scholars.[1]

Even if scholars in general agree that the collection or collections were first written down in Greek, it is still not clear how the collections were organized when they were first translated into different languages.

At least five different types of organization of the collections of sayings can be defined:[2]

a) alphabetic, in which the sayings are attributed to certain monks,

b) anonymous, in which the sayings are not attributed,

c) systematic, in which the sayings have been organized under certain themes and divided into chapters,

d) mixed, including, e.g., a combination of an alphabetical and an anonymous part

e) "wild" collections, in which no form of organization can be identified.

---

[1] See Mc Vey 1998; Larsen 2006; Rönnegård 2010; Rapp, 2010.

[2] Apart from the five types listed here, a "biographic" type has also been identified by Dr. Britt Dahlman. In the biographic type of collection, sayings attributed to certain fathers have been grouped together but not organized alphabetically. An example can be found in the Collectio Caracallensis (BHG 1445a), extant in the manuscript Athos, Karakallou 251.

All the different types of collections are represented in later Greek manuscripts, whereas in other languages mainly the systematically or alphabetically organized types are extant.

The earliest text witnesses, however, are not in Greek but in Syriac. There are around ten manuscripts preserved from the early 6th century, and at least five more from before 1000.[3] The earliest Syriac manuscripts are seemingly unorganized, which might strengthen the assumption that one of the earliest collections was neither strictly alphabetically nor systematically organized.

The Greek collections that are systematically organized have previously been subject to studies on their genealogy, in which mainly the actual sets of sayings, their redactions and their order of appearance have been studied. Scholars defined different stages in this process. What has been considered to be the earliest systematically organized Greek collection may have been derived from an early alphabetical collection in Greek. Translations of the Greek systematical collection were then early on made into Syriac, Latin, Coptic and Palestinian Aramaic. The Latin manuscripts contain only the systematic and a few more or less unorganized collections. The earliest ones in Latin have been dated to ca. 650, and there are around 15 preserved manuscripts copied before the year 1000. The Latin early manuscripts are also older than the earliest Greek ones that are preserved today: there are two very early Greek fragments from the 5th to the 7th centuries,[4] but otherwise the earliest preserved Greek manuscripts are from the 9th century. There are also preserved papyrus fragments dated to the 7th century in Coptic. The only manuscript in Sahidic Coptic including a larger part of the systematic collections is dated to the 9th century and thus constitutes an early witness to the systematic type that can be paralleled to the Latin collections (Elanskaya 1994: 11).[5]

---

[3] See Holmberg 2013 with references to the inventories made by René Draguet and William Wright.

[4] Cairo fragment of a long saying (ca 5th century), preserved both in the alphabetical collection under Makarios no. 3 and as no. 9 in the systematical collection: see Galazzi 1990; and a fragment of the systematical collection, MS Oxon. Bodl. Gr. th. g. 8 (ca 8th century): see Bagnall & Gonis 2003.

[5] The folia in the manuscript have been scattered; the fragments were edited by Chaîne 1960, Elanskaya 1994 and Lucchesi 2004.

Rather early, probably during the late 8th century, collections of sayings were also translated into Armenian, Arabic (and via Arabic into Ethiopic), There were surely no Old Slavonic translations in the 8th century. There is, Georgian and Sogdian.[6] There is one manuscript in Georgian dated to 907, and one Arabic manuscript dated to 901.[7] The oldest preserved manuscripts in Armenian and Ethiopic are from the 12th century and onwards. From the beginning of the 13th century, the text collections were translated into vernacular languages all over Europe. The text collections can be found in text witnesses in many languages during the medieval age, as well as in early modern times: in addition to the languages mentioned above, there are manuscripts with translations into other languages such as early English and French, Norse and German.

The textual traditions of the monastic sayings have typically been associated with the collections as they were defined in the early standard editions (Rubenson 2011, 2013; Dahlman 2013). These editions have been based upon different selections: some are monotypic editions – that is, editions based on a single manuscript – while others aim to give a critical text of manuscripts contained in one specific type of organization in a certain language (even though it is very hard to base such editions upon the common errors method). Most of the previous editions, however, present an "all-inclusive" collection in which all sayings that have been found in different manuscripts containing a certain type of collection of sayings in one language have been included. The established text in the editions have in general been either a majority-based text or a text mainly based upon one selected text witness or upon a certain preferred group in a stemma. However, since the variation in the appearance of the sayings in the collection and their order of appearance is high, so that no manuscript actually contains the exact same set of sayings, none of these types of editions gives a fair representation of the text traditions.

---

[6] According to Outtier 1980, the translation to Georgian would be from the end of the 8th century.

[7] For more information about the oldest identified Arabic manuscript, dated to 901 (Strassb_4225), see Oestrup 1897.

Collecting and analyzing the collections of monastic sayings in a relational database

Even if it seems impossible to get a clear, definite picture of the early development of sayings of the Desert Fathers due to the lack of preserved early manuscript witnesses in Greek, the reception of the collections in different languages can certainly be studied much more. A prerequisite for studying and comparing relations between the textual traditions is for larger amounts of relevant data to be assembled and compared. In a current project at Lund University, the structure and the text of editions and also of modern translations in different languages (Greek, Latin, Arabic, Coptic, Georgian, Slavonic and Syriac) have been added to a relational MySQL database.[8] The narrative of the individual saying – that is, the smallest separate text units found in the sources – have been identified with a unique id, making it possible to relate the different sets of sayings and the appearance of them, as well as the order of their appearance in different collections when the structure and text of manuscripts in different languages are also added, which was the next step taken in the development of the database. The established text in the editions is used as a basis for the references.[9] The structure and XML/TEI-encoded transcriptions of manuscripts that contain the textual traditions in different languages are thus being continuously added. The process of adding the structure and contents of the sometimes quite varied manuscripts is aided by the possibility of performing free text searches in the database.[10] Newly added manuscripts can be compared with the entire textual tradition in different languages. By choosing different

---

[8] The current project is entitled "Formative Wisdom – The Reception of Monastic Sayings in European Culture. Scholarly Collaboration on a Digital Platform." The online platform and interface, *Monastica* (https://monastica.ht.lu.se), including e.g. tools for adding new TEI encoded transcriptions of manuscripts, is a further development (by Samuel Rubenson and Johan Åhlfeldt) of a complex relational off-line MySQL database and program *Apophthegmata Patrum Database* (APDB), constructed by Kenneth Berg and Samuel Rubenson (project leader).

[9] The variant readings in the critical apparatuses have however not been added. – Not only complete editions and/or translations of collections are included in the database but also smaller parts that have been published as part of, for example, an article. This explains the large number of editions and translations listed: no less than 134 editions and 51 modern translations.

[10] The contributions by the individual scholars are specified in detail in the metadata for each manuscript.

manuscripts for comparison, it is possible to study relationships in the same language, over time and across language borders, which was not possible before.

It has sometimes been said that the sayings of the Desert Fathers is an example of a fluid text genre. This is both true and false. The fluid element in this genre consists mainly in what sayings are included in a collection and what sayings are not, plus their order of appearance in a collection. The main organization of the chapters, normally either alphabetically or systematically, has generally been kept in the copying process, but there are always some changes. Each manuscript containing collection(s) of sayings is therefore a unique composition in that there are always some differences in the choice of sayings. On the other hand, the texts in the sayings themselves are not as fluid or variable as their appearance or order of appearance. The narrative in one saying is almost without exception directly identifiable, even across the language barriers. The main difference between the collections therefore lies in the different sets of sayings that were included and how they were organized. Therefore, it is very useful for scholars working with these text traditions for as much data as possible from earlier scholarship to be assembled in the database. Since the collections of sayings apparently were organized differently very early on, if not already from the start, it is highly relevant to study further the relations across the language barriers, both concerning the actual sets of sayings – that is, the content itself and its structure – and the variation present, including with regard to the textual variation. The structural variation is therefore one factor to consider, while the internal textual variation is another.

In order to compare this in the best way, missing parts of the texts also need to be considered. There are some questions to ask here. If structures are compared where there are lacunae in the manuscripts, this might be misleading. In this article, therefore, some questions concerning lacunae in a manuscript will be discussed, both in general and in the context of collections of monastic sayings. What significance can the lacunae in manuscripts have when studying a text tradition? Can a typology of lacunae be established that also could take into account the significance of different genres and how different types of narratives might be more prone to the occurence of lacunae? Moreover, how could the existence of lacunae best be described and considered when working with a database (or a detailed catalogue for

that matter) in which the entire contents and the structure of a manuscript is to be included and analyzed?

A suggested definition of a manuscript lacuna

First of all: how does the scholar discover a lacuna in a manuscript? Whenever the manuscript contains a work that is rather stable, the lacuna can easily be seen when a quire starts in the middle of a narrative that is not a continuation of the narrative in the previous quire, or when one verso page ends in the middle of the narrative and the next folio does not belong to the same part of the text. Such lacunae most often occur in the very beginning or end of a manuscript: a narrative starts "in medias res", or an interrupted narrative ends the manuscript. However, lacunae of course also appear in the middle of manuscripts, especially in between the quires. They might have been caused by misunderstandings in the binding of the quires or by mistakes in the later rebindings of manuscripts. Moreover, one or several folio pages with written text can have been cut out for different reasons.

A lacuna is normally described only in general terms, as a "gap" or "missing part". I suggest that a more specific definition is to be used when it comes to a lacuna in the manuscript tradition. The general definition of the lacuna could be specified as follows: *an interrupted narrative in an otherwise continuous text.*

In most cases, whenever a sentence evidently is a continuation of a sentence other than the one found on the previous page or abruptly ends a manuscript or a quire, the lacuna is evident. However, if a lacuna was already there in the exemplar – that is, the manuscript that was copied by a scribe – this means that the interrupted narrative was copied automatically by the scribe. The scholar today might therefore find the interrupted narrative not in a discontinuous narrative between one manuscript page and the following one, but in the middle of the text on a page, since no manuscript is copied strictly line by line so that each line ends with the same word. The exemplar in which the lacuna occurred in the first place might not be extant at all today, but later copies of that exemplar in the manuscript generations to come might transmit the interrupted narrative and the lacuna in a text that would seem uninterrupted in the later manuscripts. Thus, there is a difference between the lacunae that evidently occurred in one specific manuscript, indicated by the fact that the narrative is interrupted at the very end

of a manuscript page and/or begins with a resumed narrative inside a saying on the first line of another manuscript page, and the cases in which the narrative is interrupted in the middle of one manuscript page, since that would be more indicative of a *transmitted* lacuna. Consequently, if manuscripts are found in which the last line of a page ends at the place in the narrative where it is found as an interrupted narrative in the middle of a page in another manuscript, this could prove to be an important link between the two manuscripts. Possible transmitted lacunae should therefore be noted as such.

There could be cases of what could have been smaller lacunae, that is, clearly interrupted narratives that could later have become part of the text tradition, e.g., if the corrector of a manuscript added the missing part in the margin, which was then copied into the next manuscript generation. Then, the interrupted narrative signaling a lacuna would disappear in the next manuscript generation, containing only the continuous text again. Of course, the scholar of today then remains ignorant of what happened since the evidence of a previous lacuna is not there anymore: such situations must remain hypothetical.

Possible lacunae with non-interrupted narratives

Occasionally, an editor might have reason to suspect that a large omission of a part of a text in one manuscript in a rather stable text genre – like in the systematical collections of sayings – did not occur in this manuscript first, and was not deliberate, but that the omission was there in the exemplar, that is, in the manuscript that the present manuscript was copied from. A Greek manuscript preserved in the BnF, Par. gr. 2474, presents an example of this on f. 80v. The text is continuous and there is no interrupted narrative, but almost seven chapters in this stable text tradition have been omitted so that one saying that normally is the first one in chapter 8 is immediately followed (on the same page) by sayings from the middle of chapter 15 onwards with no interruptions in the normal "set" of sayings; otherwise, the standard set of sayings both before and after this "omission" are there. What could be the reason for this? Four possible reasons present themselves, the two first of which I would consider as most probable: 1. The manuscript was copied quire by quire and a subset of several quires were omitted by mistake in the copying process; 2. The large (accidental) omission was already there in the exemplar, and was thus transmitted into the next manuscript generation; 3.

The scribe deliberately omitted the large missing part, 4. A deliberate omission was integrated in the manuscript tradition when a copy was made from an exemplar including a deliberate omission.

The last two possible reasons seem the most illogical, given the stability in the text tradition otherwise, whereas the first option is perhaps the most probable one. Anyway, this large omission may in itself be considered as a significant "error" in the transmission process in the sense that if exactly the same large set of sayings would be missing in other manuscripts (in the same language or in other languages), it would be clear that they are closely related.[11] This is not the obvious interrupted narrative but might be another type of lacuna, of a non-interrupted narrative. This type of plausible unintentional omission that could have been integrated into the manuscript transmission thus might be a deliberate variation, rendering a new redaction of a collection.

In the first two possible situations listed above, the omission would occur by accident in the process of the actual copying of the manuscript. As mentioned before, this is a type of omission that would also in principle be considered equal to a significant error, at least in some genres. In the collections of sayings an unintentional omission of an entire section could very easily take place. An omission of a few sayings on the same page with an identical or almost identical introduction – e.g., "Abbas dixit", which is quite common – is then a scribal mistake that could become part of the textual tradition since the new "set" of sayings thus appearing in the manuscript could be transmitted into the next manuscript generation, and so on. The lacuna that is in reality a case of *saut du même au même* thus becomes part of the text tradition.

So what should then be considered to be a lacuna, and at what point is it no longer a lacuna but a part of a changed text that has become part of a later text tradition? I suggest that the term lacuna should be confined to either very clearly interrupted narratives as stated in the definition above and otherwise only as *possible* lacunae with non-interrupted narratives. The latter would only be relevant in rather stable text traditions.

---

[11] I wish to thank Dr. Britt Dahlman for drawing my attention to this example, for valuable observations and a fruitful discussion.

A suggested typology of lacunae

Having proposed this definition of what is to be considered to be a lacuna, I would like to suggest that the way to describe a manuscript could be somewhat more nuanced when it comes to the type of the lacuna. Whenever a scholar is working with a more or less stable or closed text tradition and the actual missing text in a lacuna is possible to ascertain, this could be specifically noted. It could help in the further analysis of the manuscripts and their relations. Thus, I believe that there are reasons to distinguish between two main different types of lacunae. I will give examples from the present database project.

Type 1: "Missing text": the missing text parts can be identified

In a stable text tradition, when one verso page ends in the middle of the narrative, and the next folio page starts in the middle of another part of the narrative – whether it is a matter of one or two leaves that have been cut out later on, or of the first folium in the next quire missed in the copying process but meant to be included – the specific text that is missing is actually possible to deduce. The missing part can be specified by the number of words – in the present case by the number of missing sayings and the number of words in them – which is actually possible to deduce since the volume of the text omitted is not bigger than what scholars can be certain should have been there. The fact that the number of lines in medieval manuscripts is almost always consistent makes these assumptions rather secure. This could then be defined as Type 1: a lacuna that can be identified with specifically the text that is missing but would have been there from the beginning. I will give an example of the usefulness of making a calculation of the exact contents of what is missing. When I studied an early Latin manuscript only via the image gallery displayed online, the last few folios in the manuscript in question did not fit into the normal pattern of the particular collection of systematically organized sayings that this manuscript includes. What seemed at first to be lacunae I suspected to just be an omission in the upload or during the process of scanning the manuscript. I made some calculations given that the number of lines and average number of words on each manuscript page is almost always quite consistent (which goes for most medieval manuscripts). Since the text in the sayings is quite stable, an average calculation of the number of words in each saying that is normally given in a specific order

in the more stable collections showed rather exactly what sayings would been there in the missing folio(s). As suspected, it proved to be a mistake in the upload of the scans, so the missing folios were not in fact missing, and the exact sayings that according to the calculations would be extant on the different folio pages were indeed there in the original. Thus, this way of deducting a lacuna of Type 1 – that is, what apparently is not there but "should" be there – is quite accurate and proves that concerning stable or closed text traditions the scholar can deduce what would have been in place in situations where the manuscript only lacks one or possibly also two folio pages in a row. However, if the lacunae are larger than maximum two folio pages, I consider it too risky to make those assumptions.

Type 2: "Proper lacuna"

The second type is the "proper" lacuna – that is, when scholars cannot know how much text, and specifically what text, is missing, identified by their interrupted narratives, whether in the beginning, inside or at the end of a manuscript.

How to account for lacunae of Type 1 and Type 2
Type 1

When adding information about the structure of each manuscript to the relational database, lacunae of type 1 ("missing text") can be noted not as lacunae without any possibility of identifying the contents further, since these are in fact cases where the scholar can know what is actually missing. Therefore, this type of missing information can be specifically noted as such in a relational database by adding "M" for missing and putting this information together with the specific IDs of the missing sayings in a separate column, thus treating it as an addition to a source, or by adding this information directly in the structure of the manuscript with the suffix "M". By specifying exactly the sayings that are missing, it is possible to take the full structure into account when comparing the sets of sayings internally within one specific language but also across the languages. It is also possible to just search out the parts that are thus marked in order to be able to compare manuscripts and see whether there are manuscripts that might have been the exemplar or exemplars from which the present manuscript was copied, since the folio and line numbers are also entered in the relational database input data. In that way, it is easier to see

where one full quire or one full folio page is missing since that folium should start with the line that is missing in the manuscript with the lacuna.

An example of how this "missing" function can be used is an early Latin manuscript, Brussels 9850-52 from the early 8th century. In chapter five in the collection translated by "Pelagius and John" according to the tradition, the sayings that are normally numbers 19 to 22 in the collection are missing. When counting an average of the number of words for 5 pages (ca 170) and comparing with the number of words that are normally extant in these specific sayings in other manuscripts (665 words), it is quite clear that the text of two folio pages are missing here, which adds up to that number of words, keeping in mind that the textual variation in the sayings themselves is quite small. In Figure 1, a yellow line for the Type 1 sequence of missing sayings in the manuscript Brussels 9850-52 fills out the missing part (the information was included in a parallel table). As can be seen, this is the only difference in the sequence of sayings in this particular chapter when comparing the manuscripts Brussels 8216-18, Cologne 165 and Brussels 9850-52.
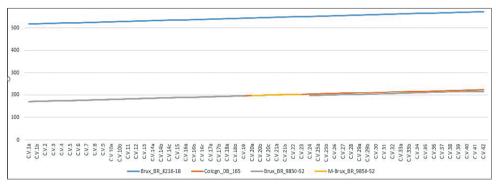
Figure 1. Comparison of section of sequence of sayings in Brussels 8216-18, Cologne 165, Brussels 9850-52 and a separate column for the missing sequence M in Brussels 9850-52.

Type 2

These proper lacunae are not uncommon in the manuscripts. When comparing the similarities and differences between the different sets of sayings, lacunae could alter the big picture so that we would see another pattern than the one that has been there from the start. Since we in these cases cannot and should not guess the contents, this problem cannot so easily be solved. Visualizations including manuscripts with lacunae are then more

accurately used if the data disturbance caused by the lacunae is eliminated somehow, or, at least, if the different aspects revealed by the existence of the lacunae are analyzed case by case.

When a certain collection of sayings is quite stable and the chapters always appear in the same order, it is also quite safe to assume that a manuscript that has a proper lacuna in such cases did include the missing chapters as well. However, since the text volume in these cases is much larger and a certain fluidity in the set of sayings can be seen, the scholar could only mark the chapters themselves as missing, but not the actual sayings within each chapter. Therefore, when making comparisons of structure and also of textual variation within certain sayings, it is important to make comparisons where the parts that are affected by lacunae are not included, so that only the full parts or chapters including the actual text of preserved sayings are compared. An example of how the selection of only the comparable parts is vital can be given here: first (figure 2), information containing all the preserved parts of four selected manuscripts without the extraction of only the parts that have not been affected by the presence of the lacunae, and then (figure 3), a comparison including only the chapters that are common for the two Latin and two Greek manuscripts and not affected by lacunae.
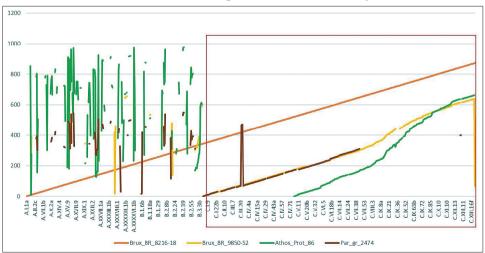

Figure 2. The occurrence and sequence of sayings in the entire Latin manuscript Brussels 8216-18 compared with the Latin manuscript Brussels 9850-52 and the Greek manuscripts Athos Prot 86 and Par Gr 2474. The comparable collection (the PJ collection) is marked with the red square. See further in Figure 3.

Figure 3. Enlargement of the PJ collection shown in Figure 2. The parts not affected by lacunae are marked with the red square.

Conclusion

In this article, it has been pointed out that lacunae themselves can give evidence of direct links between manuscripts, both what is in this article described as proper lacunae, defined by their interrupted narrative, and possible lacunae with a non-interrupted narrative. Moreover, it is possible to identify exactly what text is missing in smaller lacunae in a manuscript with a stable or closed text tradition. Information about the type of lacunae should therefore be added in manuscript catalogues and critical editions of texts, and also in the metadata section on the manuscripts included in digital corpora, editions or relational databases of manuscripts. In that way each scholar can make use of the information and refine it further, while considering what is not there when making comparisons with other manuscripts containing the same text, and in some cases also concluding what might have been there.

List of manuscripts
Athos: Bibl. Tou Prôtatou
Athos Prot. 86
Brussels: Bibliothèque royale de Belgique
Br. 9850-52,
Br. 8216-18
Cologne: Erzbischöfliche Diözesan- und Dombibliothek
Col. 165
Paris: Bibliothèque Nationale de France
Par. Gr. 2474

References

Bagnall & Gonis 2003: Bagnall, R. S. & Gonis, N. "An Early Fragment of the Greek *Apophthegmata Patrum*". *Archiv für Religionsgeschichte* 5. Bd., Heft 1 (2003), 275–278.

Burton-Christie 1993: Burton-Christie, D. *The Word in the Desert. Scripture and the Quest for Holiness in Early Christian Monasticism*. New York & Oxford: OUP, 1993, 76–81.

Chaîne 1960: Chaîne, M. *Le manuscrit de la version copte en dialecte sahidique des "Apophthegmata Patrum"* (Bibliothèque d'études coptes 6). Cairo: Imprimerie de l'IFAO, 1960.

Dahlman 2013: Dahlman, B. "The *Collectio Scorialensis Parva*: An Alphabetical Collection of Old Apophthegmatic and Hagiographic Material." In Rubenson, Samuel (ed.): *Early Monasticism and Classical Paideia* (*Studia Patristica* LV:3). Leuven/Paris: Peeters, 2013, 23–34.

Elanskaya 1994: Elanskaya, A. *The Literary Coptic Manuscripts in the A. S. Pushkin State Fine Arts Museum in Moscow* (Supplements to Vigiliae Christianae 18). Leiden: Brill, 1994.

Faraggiana 1997: Faraggiana, Chiara. "*Apophthegmata Patrum:* Some Crucial Points of Their Textual Transmission and the Problem of a Critical Edition." In Livingstone, E.A. (ed.): *Papers presented at the Twelfth International Conference on Patristic Studies held in Oxford 1995. 1: Historica, theologica et philosophica, critica et philologica* (Studia Patristica, 29). Leuven: Peeters, 1997, 455–467.

Gallazzi 1990: Gallazzi, C. "P. Cair. SR 3726: Frammento degli Apophthegmata Patrum." *Zeitschrift für Papyrologie und Epigraphik* 84 (1990), 53–56.

Harmless 2004: Harmless, W. *Desert Christians. An Introduction to the Literature of Early Monasticism.* Oxford: OUP, 2004, 248–251.

Hock & O'Neil 1986: Hock, R. F. & O'Neil, E. N. *The Chreia in Ancient Rhetoric, vol. I: The Progymnasmata.* (Texts and Translations 27. Graeco-Roman Religion Series 9). Atlanta: Brill, 1986.

Holmberg 2013: Holmberg, B. "The Syriac collection of *Apophthegmata Patrum* in MS. Sin. Syr. 46." In Rubenson, Samuel (ed.): *Early Monasticism and Classical Paideia* (*Studia Patristica* LV:3). Leuven/Paris: Peeters, 2013, 35–58.

Larsen 2006: Larsen, L. I. *Pedagogical Parallels: Re-reading the Apophthegmata Patrum.* (Ph. D. Dissertation). Columbia University, 2006.

Lucchesi 2004: Lucchesi, E. "Un petit complément au manuscrit de la version copte en dialecte sahidique des 'Apophthegmata Patrum.'" In Zanetti, U.& Lucchesi E. (ed.), *Aegyptus Christiana. Mélanges d'hagiographie égyptienne et orientale dédiés à la mémoire du P. Paul Devos Bollandiste* (Cahiers d'orientalisme, 25). Geneva: P. Cramer 2004, 163–164.

Mc Vey 1998: Mc Vey, K. "Chreia in the Desert: Rhetoric and the Bible in the Apophthegmata Patrum." In Malherbe, J, Norris, F. W. & Thompson, W. (ed.): *The Early Church in its Context: Essays in Honour of Everett Ferguson* (Novum Testamentum, Supplements, 90). Leiden: Brill, 1998, 245–255.

Monastica: https://monastica.ht.lu.se

Oestrup 1897: Oestrup, J. "Über zwei arabische Codices sinaitici der Strassburger Universitäts- und Landesbibliothek." *Zeitschrift der Deutschen Morgenländischen Gesellschaft* 51 (1897), 453–471.

Outtier 1980: Outtier, B. "La plus ancienne traduction géorgienne des apophtegmes: son étendue et son origine." *Mravaltavi* 7, 7–17.

Rapp 2010: Rapp, C. "The origins of hagiography and the literature of early monasticism: purpose and genre between tradition and innovation." In Kelly, Christopher, Flower, Richard & Williams, Michael (ed.): *Unclassical Traditions, Vol. I: Alternatives to the Classical Past in Late Antiquity* (Cambridge Classical Journal. Proceedings of the Cambridge Philological Society. Supplementary Vol. 34 (2010), 119–130.

Regnault 1987: Regnault, L. *Les Pères du desert à travers leurs Apophthegms*. Sablé-sur-Sarthe: Abbaye Saint-Pierre Solesmes, 1987, 73–83.

Rubenson 1995: Rubenson, S. *The Letters of St. Antony. Monasticism and Making of a Saint* (Studies in Antiquity and Christianity). Minneapolis: Fortress Press, 1995, 152–157.

Rubenson 2011: Rubenson, S. "The Apophthegmata Patrum in Syriac, Arabic and Ethiopic. Status questions." *Parole de l'Orient* 36 (2011), 305–313.

Rubenson 2013: Rubenson, S. "The Formation and Re-formations of the Sayings of the Desert Fathers." In Rubenson, Samuel (ed.): *Early Monasticism and Classical Paideia* (*Studia Patristica* LV:3). Leuven/Paris: Peeters, 2013, 5–22.

Rönnegård 2010: Rönnegård, P. *Threads and Images: The Use of Scripture in Apophthegmata Patrum* (Coniectanea Biblica. New Testament Series 44). Winona Lake, Indiana: Eisenbrauns, 2010.

About the author

Elisabet Göransson is an Associate Professor of Latin at Lund University, Sweden. Her research interests concern methods and tools for investigating, analyzing, editing and visualizing text traditions. E-mail: elisabet.goransson@ctr.lu.se

Summary

In this article the question of how to identify and define lacunae is discussed, both in general and exemplified in manuscripts containing collections of monastic sayings of the Desert Fathers. These sayings, organized in different ways, were probably written down in Greek during the early fifth century. They were soon translated into Latin, Syriac, Coptic, Palestinian Aramaic, Arabic and Ethiopic, Old Slavonic Georgian and Sogdian, and had a large reception in the vernaculars later on. In a relational MySQL database, the sayings as they appear in the manuscripts are given a unique identifier. The structure and the contents of each manuscript can be compared through this reference system, which builds on the unique structure of each manuscript. However, when a manuscript contains lacunae, the results when comparing the structures are not trustworthy. This problem is discussed and possible solutions are presented. I suggest a definition of the notion of a lacuna in general: an interrupted narrative in an otherwise continuous text.

There are also manuscripts in which there may be reason to suspect possible lacunae that are in fact not signaled by interrupted narratives. The collections of sayings include many brief narratives, and when the "typical" sequence of these narratives is broken in a manuscript, this may be an effect of a lacuna in the exemplar – that is, the manuscript that was the model for a copy – or of an actual lacuna that occurred when there was a mistake in the copying of quires. These possible lacunae with non-interrupted narratives are harder to detect, but nevertheless, I argue that when a scholar or a cataloguer suspects such an "non-interrupted narrative" type of lacuna, this should be noted.

Furthermore, for the obvious lacunae, a typology is suggested that could be used both by cataloguers of manuscripts and by scholars. The first lacuna type, "missing text specified", is relevant for texts in a closed or stable text tradition: the exact missing text section in a smaller lacuna can then be identified by calculating an average number of words on each manuscript page and the number of words missing, taking into account the fixed number of lines on each manuscript page. In that way it is also possible to see if the missing section was due to a loss of one or two folio pages. The second lacuna type is the "proper lacuna", concerning which the scholar cannot know what or how much text is missing. The possibility of adding information about what text is missing in a lacuna (type 1) is higher in closed/stable text genres.

PARADIGMATIC N-GRAM APPROACH IN STUDYING LINGUISTIC STANDARDIZATION

Michael Hoffert and Marija Lazar[1]

Abstract: Linguistic regularity and repetitiveness are widely studied in data-driven linguistic research by applying the n-gram method. Together with language variation, they comprise key features of a natural language. The present paper discusses the applicability of the n-gram method on historical data in the legal domain, for the consolidation of linguistic constituents and their pleonastic usage are indicators of the professional register development.

Due to a restricted stock of n-grams, the paradigmatic data arrangement has been selected as opposed to the existing practice of the syntagmatic one. Not only has this procedure impacted the frequency and the accuracy of those counts compared with previous research, but also paved the way for combining qualitative and quantitative investigation of n-grams. Finally, we explain the advantages of applying the proposed approach to n-gram analysis in the given settings, which allows for studying domain-specific linguistic regularity beyond the lexis.

Keywords: n-grams; parallel corpora; low-resource language; legal language; standardization; Czech; Slovak

1. Linguistic standardization from the n-gram perspective

Linguistic regularity and repetitiveness are widely studied in data-driven linguistic research by applying the n-gram method on big data (Biber, Conrad & Cortes 2004: 280-281; Kopaczyk 2013: 60-62; 145). Other than an antipodal feature of a natural language, the variation, regularity and repetitiveness result in reduced variability of language constituents and are distinctive for professional varieties, which represent in essence the principles of linguistic standardization and thus serve as particularly rewarding material for the study of standardization. In this paper, *n-grams* are understood as automatically extracted linear sequences that regularly occur in particular genres or registers, mostly in larger corpora (Wood 2015: 45).

---

[1] The paper has been written in the context of the long-term academy project "The Saxon and Magdeburg Law as a Cultural Link between the Legal Orders in Eastern and Central Europe" at the Saxon Academy of Sciences, Leipzig, Germany.

This paper addresses the issue of the applicability of the n-gram method on historical data in the legal domain, which is acknowledged as one of the most formalized and regular professional languages; it apparently already exhibited those features in its early developmental stages (Kopaczyk 2013: 22-24; 51-71). This makes legal texts particularly suitable for exploring linguistic standardization, for they not only show the onset of the tendency of constituent selection that gives rise to the reduction of variation preceding the standardization process, but also offer an opportunity to quantify this process. In what follows, we set out to prove that appropriate customization of the n-gram method can advance research on low-resource and sparsely digitized languages, which is the case with the majority of Slavic languages at a historic stage.

The present study was based on two legal codices from the latter half of the 15th century, written in Northern Bohemia and Upper Hungary in two different varieties of Czech – *Práwa saszká* (the *Saxon Laws*, 1470-1471) and a part of *Žilinská kniha* (the *Žilina Law Book,* 1473). The correspondent corpora encompass 11,200 and 47,746 tokens, respectively, and are aligned with parallel German source texts.[2] Both codices emerged as a result of the urbanization of East Central Europe; during this fundamental change, the norms of German municipal law were the cornerstone of new cities' foundations and of the reshaping of existing ones. This urbanization process was also labeled as a *legal transfer* of German law to East Central Europe (Lück 2013: 298-300). Both the *Saxon Laws* and the *Žilina Law Book* were thus pioneering translations from Early New High German sources. In both cases one can speak of the legal register *in statu nascendi*, whose progress as that of stages and of intellectual movements in legal transfer from that era will be explored. Because of their functionality, both translations were not only used as an integrative means of German law on Czech- and Slovak-speaking territories, but they were also an example of how Early New High German impacted Slavic professional varieties used in the region.

---

[2] For the *Saxon Laws*, the Early New High German pendant is *Sächsisches Weichbild* (*Saxon Weichbild*, 15th c.), however, it certainly seems that this was not the exact source text for the translation. The *Žilina Law Book* contains the Early New High German source text that has been apparently translated in the Slavic part of the same manuscript. Both parallel corpora will be made available for public use in 2020.

To make the above-mentioned texts conform with n-gram-research, they had to be converted to a uniform digitized shape, with each text deserving an individual treatment. The *Saxon Laws* required less effort, for a transcribed and digitized text was already at our disposal and its orthography was nearly consistent; only the editorial markings were filtered out. As for the *Žilina Law Book*, the whole digitalization process was accomplished by converting the PDF file we received from the publishing house into an editable format and deleting all editorial markings in the file. The file was subsequently subdivided into two parts, into passages written by scribe A and by scribe B, for paleographical analysis indicated different patterns in their highly variable orthography.[3] Finally, TUSTEP routines for semi-automatic orthography standardization were applied to both parts. Upon completion of these preparatory operations, the n-grams were extracted using our own tool written in TUSCRIPT and bash (shell) script, which was tailor-made for paradigmatic n-gram extraction. As a result, two n-gram-lists were generated – the first for scribe A and the second for scribe B – that were matched in the final step to get a shortlist of the most frequent n-grams. This shortlist was the subject of subsequent analysis.[4]

2. Extracting and clustering the n-grams paradigmatically

The existing practice of syntagmatic data arrangement was expanded and enriched with a paradigmatic one on account of data sparseness. Instead of analyzing separate groups of n-grams according to their length (2-, 3-, 4-grams etc.; short/long n-grams), the n-grams were clustered according to their relations to each other across all available n-gram lengths:

```
column    1  | 2 | 3    | 4 | 5a     | 5b       | 6
         4-17  3   TM-X                            přijdu oba před pravo
                2   EK           [7–2]             ranita a přijdu oba před pravo a
                3   TMEX  1      [6–0]             a přijdu oba před pravo a
                2   TM           [6–5]             ranita a přijdu oba před pravo
                3   TM-X         [5–3]             přijdu oba před pravo a
                3   TM-X         [5–6]             a přijdu oba před pravo
                3   TM-X                 [3–140]   přijdu oba před
                3   TM-X                 [3–146]   oba před pravo
               46   TMEX 12               [2–8]    před pravo
                4   TMEX  1              [2–453]   oba před
                3   TM-X                 [2–713]   přijdu oba
```

Figure 1: Paradigmatic clustering of the extracted n-grams.

---

[3] For more detailed information, see Sopko 1984.

[4] The source code and outcome of n-gram extraction can be found at http://smr.saw-leipzig.de/materialien.html.

Figure 1 plots the data output format after n-gram clustering, with the columns containing the following information: (1) a unique n-gram identifier, while "4" stands for an n-gram length and "17" for an n-gram position within the list of 4-grams; (2) frequency of n-grams; (3) n-gram type; (4) the number of final n-grams within the TMEX type; (5a-b) the unique identifiers of related n-grams, while (5a) shows n-grams with higher cardinality and (5b) n-grams with lower cardinality than the stem n-gram; (6) n-grams in the n-gram cluster.

This procedure allowed, on one hand, for combining a quantitative and qualitative analysis of data, which is indispensable when dealing with language in a historical context (see List 2016: 9). On the other, it deepened our understanding of the relationship network between n-grams, with a range of basic and complex relations.

To depict basic relations between n-grams, the concepts *"final node" (EK = Endknoten)* and *"partial node" (TM = Teilmenge)* were introduced, which implies variability of the string size – i.e., an n-gram can be expanded or shrunk toward a unit with sharp limits. It is assumed that every n-gram has its minimal and maximal possible representation in a given corpus, and that those limits depend on how the parameters "frequency" and "n-gram length" are set. Both depend on the corpus size and its structure, as well as on the aims of the particular study, which renders this framework highly flexible. Given that a rather small data sample was examined and our objective was to gain the maximum from the output, the minimal frequency threshold was delimited to 2 ($f_{min}=2$) in the current extraction, and the length of the extracted n-grams spanned from 2 to 12 ($2 \leq n \leq 12$), the maximal n-gram length found in our corpus. In linguistic terms, we were interested in those constituents which were repeated at least twice in order to be able to also include the less frequent non-functional units, e.g., phraseology. For the same reason, the usual delimitation of the n-gram-length from 2- up to 4-grams was exceeded, and longer recurrent units were also looked at.[5]

Going back to the clustering principles of n-grams, to arrange the data paradigmatically, a distinction was drawn between those units that were the longest possible sequence of given tokens in the corpus (final nodes) and those units representing a chunk of a final node consisting of at least 2

---

[5] Cf. Biber 2009: 282-286; 301-302 for more arguments in favour of studying longer n-grams.

tokens and repeated at least twice (partial nodes and their modifications (TM-X and TMEX)):

Figure 2: Relationship between a final node (EK, a tri-gram) and partial nodes (TM, bi-grams), the *Žilina Law Book*, scribe A.

At this stage, we will not exhaustively comment on n-grams' different structural groups, but some of them will be referred to during the demonstration of our case study.[6]

The customized n-gram arrangement also raises questions about the suitable methodology to be employed in their analysis. Given that the current study deals with standardization of legal language, the pigeonholing of n-grams into rhetorical categories appeared to be less instructive; instead of that, the genuine textual categories ascribed to legal language attracted our attention and have been verified empirically.[7] The following sections will focus on topical n-grams that belong to the core n-grams of legal language.

3. Topical n-grams as a semantic, structural, and textual group

Topical n-grams are initial strings of legal paragraphs defining their subject matter. In the present study, they are usually extracted as bi-grams following the O + PREP pattern ["about" / "regarding" / "considering" + prepositional case] and corresponding to the Early New High German VON + DAT pattern ["about" / "regarding" / "considering" + dative case] in German source texts. An example would be the n-gram *o zboži* "regarding the goods / the property" [2-454]$_A$;[2-48]$_B$[8] with ten hits in different relational categories in the *Žilina Law Book*, cf.:

---

[6] A more detailed account of n-gram extraction procedures and of n-gram types can be found in Bily, Carls et al. (to appear).

[7] See Biber, Conrad & Cortes 2004: 381; 389-396.

[8] In square brackets, we refer to unique n-gram identifiers from our extract derived from the sub-corpora of scribes A and B in the *Žilina Law Book,* respectively.

(1a) **O zbozij**, czo se dwa czlowieky nato zgimatu.
 About the property that two people lay claim to.

(ŽLB, § 230, f. 114r; scribe B)

(1b) **Von gute** daz zwen man an ſprechen
 About the property that two people lay claim to.

(ŽLB, § 288, f. 140v)

However, the expression of topical n-grams in the *Saxon Laws* and the *Žilina Law Book* differs in some cases; this will be discussed in the following.

The tri-gram *o zboži čo* "regarding the goods / the property that" ([3-46], only scribe B) neatly illustrates how topical bi-grams can be modified through determiners in postposition and thus become extended from 3- to 7-grams even within our restricted n-gram pool:

Figure 3: Bi-gram *o zboži* "considering the goods / the property" and its cluster.

As Figure 3 suggests, due to the aforementioned functional properties, topical n-grams have a fixed initial position within the final node.[9] According to this regularity, in the n-gram extraction the shortest possible final strings are always those with the initial o + PREP pattern, whereas the other short strings are partial and thus merely functional ones. Furthermore, the o + PREP pattern, completed with a determiner in postposition, usually forms a final node (see Figure 3, heterogeneous tri-gram [3-46]). Hence, the status of topical n-grams as final or heterogeneous nodes suggests their completeness and thus a fixedness, which is not found in partial nodes.

The further formal variation of topical n-grams that leads to the production of longer final n-grams was possible through the production of binomials, an important constituent of legal register (Ivanov & Toporov 1978: 229-237; Kopaczyk 2013: 188-207; Kopaczyk & Sauer 2017: 20; Schmidt-Wiegand 1998: 281; Zhivov 2008: 322-327). Among our topical

---

[9] A few examples with an initial noun in the final node were possible due to an algorithm that skips punctuation. E.g., [7-11]_B *smrt o tom ktož bude obžalovan o* "death [.] [Rr] egarding the one to be accused of."

n-grams in the matching pool, we extracted one single binomial *o sjednani a o smluvie oč by* "considering the contract what [it] would [be] about" ([7-5]$_A$;[7-29]$_B$), which is the longest n-gram (a 7-gram) among the short-listed bundles. In this case, a binomial unites two coexisting Czech legal terms denoting contract: *"sjednání/sgednani"* and *"smlouva/smluwa."* The binomials could unite either divergent regional and chronological variants, or a Slavic and a German term that were subsequently used as a text-interpretative (glossing) strategy.

Comparing with the *Žilina Law Book,*[10] *the n-gram output in the Saxon Laws* with 10 bi-grams and 2 tri-grams shows a lower n-gram cardinality, length, and frequency. The reasons for this discrepancy are: the *Saxon Laws'* corpus is four times smaller than the *Žilina Law Book* corpus, the source texts differ in their structure regarding the inclusion-exclusion of the index, and the translational and textual strategies of both codices deviate from each other, which will be explained in the following section.

4. Topical patterns as products of translation

As discussed above, the German sources of both Slavic translations played a decisive role in how they were shaped since the translators mostly followed a literal translation strategy, which resulted in the virtual relocation of German texts into Slavic.[11] With this in mind, we focused on textualization strategies in Early New High German laws and compared them with the topical patterns in Slavic translations with regard to their symmetry and asymmetry. While Example 1a-b displays symmetry between the source and target text, the following Example 2 is an instance of sense-for-sense translation and, subsequently, of asymmetry between the codex written in Early New High German and its translation into Slovakized Czech:

---

[10] The shortlisted n-grams from the *Žilina Law Book* encompass the following topical units: 49 bi-grams, 31 tri-grams, 15 4-grams, four 5-grams, three 6-grams and one 7-gram.

[11] In the discussion about the preciseness of translation, we will operate with traditional categories such as word-for-word, sense-for-sense, and free translation (see Robinson 1998), for we are interested in the verbal form of the source and target text and their symmetry/asymmetry. However, a conceptual interpretation framework of cultural translation can be found in Burke (2009) und will be referred to in section 5.

(2a) **O erbu** zenskem.
　　　Umrze li muzy <ſwoy> zenia, kteraz dieti nema […]
　　　About woman's heritage.
　　　If a married woman dies childless […]

　　　　　　　　　　　　　　　　　　　　　　　(ŽLB, § 61, 114r)[12]

(2b) **Dicz erbet** daz weyp wol wen ſy ſtirbet
　　　STirbt eim man ſeyn weyp dy nicht kinder hot […]
　　　This is how it has to be inherited lawfully from a woman who dies.
　　　If a married woman dies childless […]

　　　　　　　　　　　　　　　　　　　　　　　(ŽLB, §119, 34r)

In the German source text (*2b*) a sentence expressing a statement is used while in its Slavic translation it was replaced by the O + PREP pattern as discussed in section 3:

(2c) **O**　　　**erbu**　　　　　zenskem
　　about　　heritage-PREP.SG　woman's-M.SG.PREP
　　Dicz　**erbet**　　　　　daz weyp　wol wen　ſy　　ſtirbet
　　DEM　　hand down-3.SG.PRS the woman　lawfully if　she　die-3.SG.PRS
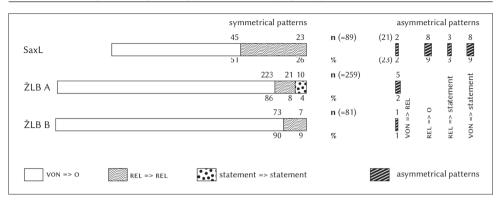
In order to place these findings in relation to the whole texts, all instances of topical patterns have been examined in terms of contrasts within the Early New High German-Slavic parallel corpora[13] and the outcomes of this comparison were mutually compared. All instances have been subdivided according to symmetrical and asymmetrical relations between the source and the target text, and particular attention was paid to topical constituents with the pattern VON + DAT translated either symmetrically or asymmetrically:

Figure 4 illustrates an important difference in the use of different types of topical patterns between the *Saxon Laws* and both scribes in the *Žilina Law Book*: while the scribes who compiled the *Žilina Law Book* adhered predominantly to symmetrical translation and to the O + PREP pattern (scribe A 86%; scribe B 90%), the scribe(s) who worked on the *Saxon Laws* did so only in 51% of instances. This fact explains, inter alia, the higher cardinality of topical n-grams in the *Žilina Law Book* than in the *Saxon Laws*.

---

[12] *o erbu* "about / concerning the heritage" is a bi-gram with four hits in the *Žilina Law Book* ([2-1584, EK]_A; [2-593, TM]_B).

[13] For more information on the corpora refer to section 1 and Footnote 2.

Figure content:

	symmetrical patterns				asymmetrical patterns			
				n	(VON => REL)	(REL => O)	(REL => statement)	(VON => statement)
SaxL	45	23		(=89)	(21) 2	8	3	8
	51	26	%		(23) 2	9	3	9
ŽLB A	223	21	10	(=259)	5			
	86	8	4	%	2			
ŽLB B	73	7		(=81)	1			
	90	9		%	1			

Legend: □ VON => O   ▨ REL => REL   ⬚ statement => statement   ▨ asymmetrical patterns

Figure 4: Topical pattern realization in the parallel legal corpora.[14]

Among the other symmetrical topicality patterns, the relativization pattern played a far less important role in the *Žilina Law Book*, with respectively 8% and 9% of all instances, while, in contrast, in the *Saxon Laws* it was exhibited in 26% of all instances and thus was the second most important means of topicality expression:

(3a)   **kterak** geden czlowiek ſyna ſwe*ho* zaſtati moze przed prawem
        How one can appear for one's son in court.

(SaxL, Ch. 89, f. 174r)

(3b)   **Wie** eyn man ſinen ſon vortreten ſal etc
        How one has to appear for one's son in court etc.

(SaxWB, Ch. LXXV, f. 173v)

As initially stated, the choice of an appropriate pattern relied on the source text; hence, the given discrepancy goes back to the shape of the source text and empirically reveals the word-for-word translation technique that has been primarily used. The same conclusion can be also drawn about a small group of topical statements,[15] found in 4% of instances in the *Žilina Law Book*, scribe A.

The asymmetrical translations with topicality patterns form a rather small group of instances, particularly in the *Žilina Law Book,* whose cases are irrelevant with 2% of all cases in scribe A's sub-corpus and with 1% in scribe B's sub-corpus. In contrast, the *Saxon Laws* exhibits 23% of such

---

[14] In this overview, we excluded some exceptional cases that occurred only once: nine instances in the *Saxon Laws* and eleven instances in the *Žilina Law Book*, scribe A.

[15] See Example 2b.

cases, and this group comprises the topical patterns described among the symmetrical patterns: however, in this case, the most prominent position was taken by statements with 12%; 9% by the o + PREP pattern; and, finally, 2% of the instances by the topical relativization pattern. Interestingly, the last of these was shown in all asymmetrical cases found in the *Žilina Law Book*.

This obvious discrepancy in asymmetrical translation of topicality patterns in the given corpora might have a twofold explanation. On the one hand, using asymmetrical translation can deliver an argument in favour of the onset of the sense-for-sense translation strategy that can be supported with strong evidence on different levels of legal language in the *Saxon Laws* (Lazar 2018). On the other, however, in the parallel corpus of the *Saxon Weichbild* and the *Saxon Laws* more *comparable* than *parallel* texts have been matched, which might explain the discrepancies that are coined above as asymmetrical translations.[16] Conversely, in the case of the *Žilina Law Book*, genuinely parallel texts have been aligned, which has obviously resulted in the finding of predominantly symmetrical translations.

5. Standardization of professional varieties: between n-grams and consolidated legal patterns

Based upon our analysis, the following conclusions have to be drawn:

(1) The topical n-grams could be automatically extracted from our corpora due to their formal and functional symmetry, and they are subsequently to be recognized as well-established units of textual structure of city laws in the 15th century. They have been relocated from German to Slavic legal texts, applying the word-for-word translation principle. However, the quality of the automatic recognition had to be improved manually in order to disambiguate the n-gram homonymy.

(2) The topical n-grams are extracted as final nodes that could be extended by at least one determiner in postposition or modified as binomials. These longer n-grams appear to be distinctive in legal language, whereas binomials have already been particularly acknowledged by previous scholarship. Conversely, the n-grams allow for understanding binomial formation

---

[16] The source and target text analysis regarding their symmetry-asymmetry can be proposed as a method to detect parallel texts in case of unknown pendants.

as syntactic and textual units, particularly if the contrastive material from the Early New High German source text is applied. The fact that the topical n-grams were repeated in the quality of final nodes stresses the importance of the distinction between the merely functional units (partial (TM) and shared (TM-X) nodes) and the units relevant for information processing (final (EK) and heterogeneous (TMEX) nodes). This distinction has to be considered in further n-gram research.

(3) n-gram extraction showed discrepancies in topicality marking between the *Saxon Laws* and the *Žilina Law Book* that have been explained using parallel corpora of these two legal translations and their Early New High German pendants. While topicality marking with O + PREP pattern dominates in the *Žilina Law Book*, the *Saxon Laws* exhibits a higher diversity of topical patterns, including a comparably large group of asymmetrical ones (23%). The latter partially goes back to the German source text's peculiarities, which have been faithfully transferred to Slavic. The existence of this group seems to be influenced either by the changing translational practice toward sense-for-sense translation or by the fuzziness of the alignment within the comparable corpus. This type of unit comparison in parallel and comparable texts can be used as a method of source text detection on a larger scale.

If we agree – at least partially – with an explanation of diversity in topical marking that resulted from applying the sense-for-sense translation technique, this new approach to the source and target language should be acknowledged as an extraordinarily progressive one, even in the European context of the humanist movement that intellectually underpinned it.[17] The asymmetrical translation patterns (**Figure 4**), primarily in the *Saxon Laws*, seem to be particularly instructive here. The choice was made in favour of topical patterns that already existed in symmetrical translation – their use was expanded toward instances that apparently varied with the patterns in Early New High German. The preferred substitutive patterns were O + PREP and statements, and they mostly stood for relativization patterns as well as for VON + DAT patterns. Interestingly enough, we do not encounter the state-

---

[17] Robinson 1998: 126 mentions Leonardo Bruni (1370-1444) among the first humanists who ridiculed the tradition of literary translation and who followed the practice of sense-for-sense translation. Although the amount of comparable studies is not considerable, for a more exhaustive bibliography on this topic, with emphasis on early humanist translational practices in Bohemian Lands, see Lazar 2018.

ment pattern in the *Saxon Laws* among the symmetrical translation cases, but we have observed it among the symmetrical patterns in the *Žilina Law Book*, scribe A. A small group that showed intersections between the *Saxon Laws* and the *Žilina Law Book* was the one where the VON + DAT pattern was substituted by a relativization pattern. The reason for translational asymmetry and variation within the patterns remains unclear, but we will stick to the explanation that a stock of topical patterns existed that could be used and varied.[18] It shows that legal texts tended to be standardized, but at the same time allowed for variation within this reduced framework.

6. Concluding Remarks

These observations show the interplay between Sinclairian idiomacity and open-choice principle that were discussed at the beginning of the paper; the idiomacity becomes visible in the consistent n-gram output and the open-choice principle by using a restricted pool of varying, interchangeable topical patterns. In the current case study, the O + PREP pattern, as well as its extensions with postpositional determiners and binomials, demonstrated the uniformity and, thus, qualified for n-grams. The analysis of parallel and comparable corpora revealed the full scope of the varying patterns in the given texts. The existence of both indicated groups demonstrates the effect of the principle of linguistic economy that was a prerequisite for the formation of professional varieties and for their ongoing standardization.

Hence, Slavic legal translations of German law were a multi-tasking accomplishment that encompassed substantial relocation of legal norms along with their cultural accommodation within the realm of Slavic-speaking urban communities, the creation of their own legal varieties based on Early New High German, and their standardization to make them professional. Between the *Saxon Laws* and the *Žilina Law Book*, we could trace some quantitative differences in groups of symmetrical topical patterns in parallel corpora, but at the same time we observed a similar degree of their consolidation and their cross-regional consistence, which attests to cross-cultural ties between Northern Bohemia and Upper Hungary, as well as to an in-

---

[18] For comparison's sake, see Zhivov 2008: 319 on variation in different types of conditional sentences in legal Old/Middle Russian.

ter-regional meaning of legal Czech – both, probably, based on the shared legal culture.

However, the instances of asymmetrical (non-literal, sense-for-sense) translation show the upcoming independent development of the legal register in Slavic, which anticipated the subsequent stage of the legal transfer when compilations[19] and completely new and original texts were created within the German legal tradition. Regarding asymmetrical translation patterns, the examined sources showed different progress in linguistic relocation and, subsequently, in legal transfer, while the *Saxon Laws* demonstrated a substantially larger segment of asymmetrical (i.e. independent from the German text) topical patterns than its counterpart, the *Žilina Law Book*.[20] This implies a more profound, meta-linguistic occupation with the German source text in the *Saxon Laws* compared to the *Žilina Law Book* that showed an initial attempt at transferring legal contents to the local Slavic variety and deviated only sporadically from the German text form. Hence, the linguistic emancipation of the Northern Bohemian Czech allowed for emancipation from the German source texts and for the development of an independent legal tradition that included learned glossing of customary legal texts, their compilations, their reshaping according to the tradition of the Roman law, creation of original texts and, finally, mediation to other legal cultures, e.g., to the Polish-Lithuanian Commonwealth. Conversely, in Upper Hungary the demonstrated trend in legal language emancipation was substantially weaker, which explains a strictly local scenario in its development in subsequent decades separate from the European mainstream in legal culture and a more essential role of local customary law and its language. This shows that not only translation was decisive for legal transfer but also that there was a legal practice that ensured reusability of translations and led to linguistic standardization of vernaculars. The common root of East Central European legal vernaculars, the Saxon-Magdeburg law, thus fostered a common legal identity.

---

[19] Including Romanization – a learned systematization of and commentary on the sources of customary law based on the prototype of Roman law.

[20] Given that the asymmetry did not result from the discrepancy in the selected German pendant and the real source text.

Glossing conventions

1, 2, 3	person
DAT	dative
DEM	demonstrative
M	masculine
PREP	prepositional case
PRS	present
REL	relativizer
SG	singular

n-gram types

EK	final node
TM	partial node
TM-X	shared node
TMEX	heterogeneous node

List of manuscripts

SaxL   *Práwa saszká* (the *Saxon Laws*, 1469-1470), Parlamentary Library Prague, without signature

SaxWB *Sächsisches Weichbild* (*Saxon Weichbild*, 15th c.), Staatsbibliothek zu Berlin Preußischer Kulturbesitz, Ms. germ. fol. 389 (Opp.-Nr. 118, HSC 12050)

ŽLB   *Žilinská kniha* (the *Žilina Law Book*, 1369-1561), State Archive in Žilina, without signature

References

Biber, Conrad & Cortes 2004: Biber, D., Conrad, S., Cortes, V. "*If you look at…* Lexical Bundles in University Teaching and Textbooks." *Applied Linguistics* 25.3 (2004), 371-405.

Biber 2009: Biber, D. "A Corpus-driven Approach to Formulaic Language in English. Multi-word Patterns in Speech and Writing." *International Journal of Corpus Linguistics* 14.3 (2009), 275-311.

Bily, Carls et al. (to appear): Bily, I., Carls, W., Gönczi, K., Lazar, M. *Sächsisch-magdeburgisches Recht in Tschechien und in der Slowakei. Untersuchungen zur Geschichte des Rechts und seiner Sprache* (IVS SAXONICO-MAIDEBVRGENSE IN ORIENTE, 5). Berlin, Boston: de Gruyter (to appear).

Burke 2009: Burke, P. "Cultures of Translation in Early Modern Europe." In Burke, P., Po-chia Hsia, R. (eds.): *Cultural Translation in Early Modern Europe.* Repr., digitally printed version. Cambridge, New York, Melbourne etc.: Cambridge University Press, 2009, 7–38.

Ivanov & Toporov 1978: Ivanov, V. V., Toporov, V. N. "O yazyke drevnego slavyanskogo prava. K analizu neskol'kikh klyuchevykh terminov." In Borkovskiĭ, V. I. (ed.): *Slavyanskoe yazykoznanie. VIII mezhdunarodnyĭ s"ezd slavistov. Zagreb, Ljubljana, sentyabr' 1978 g. Doklady sovetskoj delegacii.* Moskva: Nauka, 1978, 221–240.

Kopaczyk 2013: Kopaczyk, J. *The Legal Language of Scottish Burghs. Standardization and Lexical Bundles (1380-1560).* Oxford: Oxford University Press, 2013.

Kopaczyk & Sauer 2017: Kopaczyk, J., Sauer, H. "Defining and Exploring Binomials." In Kopaczyk, J., Sauer, H. (eds.): *Binomials in the History of English. Fixed and Flexible. International Conference on English Historical Linguistics 18 (Louvain, Belgium) 2014.* Cambridge: Cambridge University Press, 2017, 1–23.

Lazar 2018: Lazar, M. "Übersetzen, übertragen, deuten. Übersetzungspraktiken als Einflussfaktor für den Transfer des sächsisch-magdeburgischen Rechts in Ostmitteleuropa." In Köster, G., Link, Ch., Lück, H. (eds.): *Kulturelle Vernetzung in Europa. Das Magdeburger Recht und seine Städte. Wissenschaftlicher Begleitband zur Ausstellung „Faszination Stadt".* Dresden: Sandstein Verlag, 2018, 195–214.

List 2016: List, J.-M. *Computer-Assisted Language Comparison: Reconciling Computational and Classical Approaches in Historical Linguistics.* Jena: Max Planck Institute for the Science of Human History, 2016.

Lück 2013: Lück, H. "Rechtstransfer und Rechtsverwandtschaft. Zum Einfluss des Magdeburger Stadtrechts im Königreich Böhmen." In Malý, K., Šouša, J. (eds.): *Městské právo ve střední Evropě. Sborník příspěvků z mezinárodní právnické konference.* Praha: Karolinum, 2013, 298–317.

Schmidt-Wiegand 1998: Schmidt-Wiegand, R. "Anwendungsmöglichkeiten und bisherige Anwendung von philologisch-historischen Methoden bei der Erforschung der älteren Rechtssprache." In: Hoffmann, L., Kalverkämper, H., Wiegand, H. E. (eds.): *Fachsprachen / Languages for Special Purposes. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft / An International Handbook of Special-Language and Terminology Research.* (Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science (HSK), 14/1). Berlin, New York: de Gruyter, 1998, 277–283.

Sopko 1984: Sopko, J. "Pisári a vznik Žilinskej knihy". In Marsina, R. (ed.): *Národnostný vývoj miest na Slovensku do roku 1918. K 600. výročiu vydania výsad pre žilinských Slovákov.* Martin: Osveta, 1984, 41–71.

Robinson 1998: Robinson, D. "Literal Translation." In: Baker, M. (ed.): *Routledge Encyclopedia of Translation Studies.* London, New York: Routledge, 1998, 125–127.

Wood 2015: Wood, D. *Fundamentals of Formulaic Language. An Introduction.* London, New York: Bloomsbury Academic, 2015.

Zhivov 2008: Zhivov, V. M. "Yuridicheskie kodeksy i rezhim interpretacii." In Bondarko, A. V., Kustova, G. I., Rozina, R. I. (eds.): *Dinamicheskie modeli. Slovo. Predlozhenie. Tekst. Sbornik stateĭ v chest' E. V. Paduchevoĭ* (Studia Philologica). Moskva: Yazyki slavyanskikh kul'tur, 2008, 309–328.

About the Authors

Michael Hoffert (hoffert@saw-leipzig.de) holds a Master's degree in Logic and Science Theory and is employed as a technical research staff at the Saxon Academy of Sciences for the long-term project "The Saxon and Magdeburg Law as a Cultural Link between the Legal Orders in Eastern and Central Europe."

Marija Lazar (lazar@saw-leipzig.de) holds her PhD in Slavic Studies from the University of Hamburg. Her monograph is devoted to text typology of the Russian business letter from the 12th till the 18th century, and she specializes on critical text analysis, language history, legilinguistics, corpus linguistics, and digital humanities. She is currently employed at the Saxon Academy of Sciences for the long-term project "The Saxon and Magdeburg Law as a Cultural Link between the Legal Orders in Eastern and Central Europe."

Резюме

Языковая регулярность и повторяемость, будучи наряду с вариативностью основными характеристиками естественного языка, широко исследуются в эмпирической лингвистике при помощи метода н-грамов. В рамках статьи обсуждается применительность данного метода по отношению к данным, извлечённым из историко-правовых текстов, поскольку консолидация лингвистических единиц и их плеонастическое применение являются показателями развития профессиональных регистров.

Оперируя ограниченным количеством н-грамов, авторы статьи предлагают парадигматический метод их группировки в отличие от синтагма-

тического метода, применяемого в предыдущих работах. Такая процедура извлечения н-грамов не только влияет на подсчёты их частотности и улучшает точность этих показателей по сравнению с предыдущими исследованиями, но и позволяет сочетать квалитативные и квантитативные методы их изучения. В заключении рассматриваются преимущества применения нашего метода на диахронных данных, что позволяет расширить изучение регулярности профессиональных лектов за пределы лексики.

GORAZD:
AN OLD CHURCH SLAVONIC DIGITAL HUB

Vladislav Knoll

Abstract: The *GORAZD: An Old Church Slavonic Digital Hub* project, realized by the Department of Old Slavonic and Byzantine Studies of the Institute of Slavonic Studies of the Czech Academy of Sciences, Prague, is entering its final phase. The aim of the project is the digitization of work done by generations of Prague scholars dedicated to Old Church Slavonic lexicography (among them the unique *Old Church Slavonic Dictionary*/*Slovník jazyka staroslověnského* published in four volumes in 1966–1997) as well as the creation of software tools enabling user-friendly processing of digital historical dictionaries. As the software part of the project was completed by the end of 2018, the further goal is to make the digitized works accessible until the end of 2020. This paper sums up the Prague Old Church Slavonic lexicographic tradition, provides an updated report on the status of the project and lists the benefits of the created digital Old Church Slavonic databases in comparison with the traditional tools.

Keywords: Old Church Slavonic, digital dictionaries, INVENIO, historical dictionaries, Old Church Slavonic dictionaries, lexicography

Introduction

The Institute of Slavonic Studies of the Czech Academy of Sciences is a place with a long tradition of research on Old Church Slavonic lexicography. The research is concentrated in its Department of Old Slavonic and Byzantine Studies, most of whose members are currently involved in the project *GORAZD: An Old Church Slavonic Digital Hub*. For the period of 2016–2020, this project has received financial support from the Ministry of Culture of the Czech Republic's NAKI II programme (project number DG16P02H024). The goals of the project are quite ambitious, including both scholarship and software results. Among the software results, the basic ones are the creation of software tools for digitization of historical dictionaries as well as a presenta-

tion interface enabling advanced search functions integrating digital dictionaries and their card index. At the moment, when this paper is being written, the software part of the project is completed, and the project is approaching its final scholarship phase. This includes the digitization and online accessibility of the heritage of more than a half century of work by lexicographers of our institute consisting of several dictionaries (the most prominent of them being the large *Old Church Slavonic Dictionary*/*Slovník jazyka staroslověnského*) and a unique card index. Besides this main target, separate scholarly works are also being published in the framework of the project, which will also comprise the publication of four monographs by the end of 2020.

In this paper, the lexicographic tradition of the Institute of Slavonic Studies is briefly described, along with its achievements, which are compared with its digital adaptation made in the *GORAZD Project*.

Lexicographic Tradition of the Institute for Slavonic Studies

The pre-history of the *Old Church Slavonic Dictionary* can be traced to the first half of the 20th century, when the interest of specialists of Czech Old Slavonic studies was focused on the definition of the Old Church Slavonic canonical text corpus. The first attempt to excerpt such material belongs to Václav Vondrák (1859–1925), professor of the newly established Masaryk University of Brno, whose card index is stored in the Institute of Slavonic Studies. His work was revised and restarted by Miloš Weingart (1890–1939), professor of Comenius University in Bratislava and later of Charles University in Prague. The results of his excerption activity were lost. Both scholars also involved students in this activity. However, the results were significantly impacted by the fact that the excerption methodology was neither strict enough nor well followed by the collaborators (Bláhová 2016: 13–15).

The history of the current *Old Church Slavonic Dictionary* (*OCSD*) starts during World War II, when (in 1943) a Committee for an Old Church Slavonic dictionary was created in Prague by initiative of Bohuslav Havránek (1893-1978). The famous Slavist and Bohemist, then professor of Masaryk University of Brno, managed to assemble leading specialists for the task. This time, the text corpus and excerption rules were elaborated very thoroughly, and work on the creation of the card index was started (*SJS* I: III, Bláhová 2016: 19–23). From the very beginning, two types of card index were produced simultaneously (Čermák 2016: 109–111): a more detailed

Old Church Slavonic Card Index (*OCSCI*) containing 841,362 card slips[1] and a series of card indices ordered according to the equivalents in the language of the original text. Of these card indices, the largest is the *Greek-Old Church Slavonic Card Index,* containing roughly 550,000 card slips, followed by the *Latin-Old Church Slavonic Card index* comprising 96,281 card slips (82 % of which belong to *Forty Homilies on the Gospels* by Gregory the Great).[2] The smallest one is the *Old High German – Old Church Slavonic Card index* consisting of only 21 card slips.[3]

From 1953, the team of the *Old Church Slavonic Dictionary* was transferred to the Institute of Slavonic Studies of the Czechoslovak Academy of Sciences. The position of the main editor was given to Josef Kurz (1901–1972), one of the main authors of the dictioary's methodology. After the end of the main excerption phase (Čermák 2016: 111), a sample issue of the dictionary (*Specimen*) was published in 1956. The first issue appeared in 1958, and its methodology and quality were acknowledged by the scholarly public. In that period, the *OCSD* team initiated the activity of the Committee for an Church Slavonic Dictionary at the International Committee of Slavists (Bláhová 2016: 42). After the death of Josef Kurz, the editorial board was led by Zoe Hauptová (1929–2012). Despite unfavourable circumstances caused by the political climate, the dissolution of the Institute of Slavonic Studies (1963–1992) and moving the team across various workplaces,[4] the full dictionary (*SJS*) was published until 1997 in 52 issues, 4 volumes and 3,204 pages of lexicographic material. The result of the effort is a unique lexicographic work representing the largest dictionary of the oldest written Slavonic language. The main specifics of its entry structure comprise:

• A list of texts containing the lexeme at the beginning of each entry

• Multilingualism consisting in providing equivalents in up to six languages including three living languages and in up to three further languages of the model texts

• Marking the entries containing the complete list of all occurrences of the lexeme

---

[1] 13 % of this amount covers the additionally excerpted card slips, which were done after the main excerption activity in the 40ies and 50ies.

[2] The exact numbers are based on the digitized card indices.

[3] This amount represents just the corresponding parts of Freising Fragments.

[4] This period was thoroughly described by Bláhová (2016).

● The encyclopaedic character of the work, consisting in offering short characterizations of proper names (both names of persons and toponyms) and definition of the Church terminology present in the texts

● Cross-references to synonyms

Picture 1: Entry in SJS

It is worth mentioning that the text corpus of the dictionary (in the original *SJS* 83 texts, now 89 texts) can be divided into the following three groups (Pilát et al. 2018: 7):

1. Canonical Old Church Slavonic manuscripts

2. Texts whose origin is dated to the Cyrilo-Methodian period but were preserved in more recent manuscripts (e.g. *Vitae* of Constantine the Philosopher and Methodius)

3. Texts likely composed on Czech territory before the end of 11th century but mostly documented in later copies (the largest one being the Slavonic translation of *Forty Homilies on the Gospels* by Gregory the Great)

Besides *OCSD*, there is a project focused just on the vocabulary of canonical manuscripts. This activity dates back to 1972, with the cooperation with the Academy of Sciences of the USSR (Chromá 2016: 130). The outcome of this cooperation was then published in Moscow in 1994 and reprinted in 1999 (*SS X-XI*). This dictionary was originally conceived as a

practical handbook for students, containing both Russian and Czech equivalents. Its scholarly added value consists in mentioning the frequency of each lexeme in the text corpus.

Already in the early phase of the composition of *SJS* (Čajka 2006: 88-89), it became clear that revision and completion of the work would be needed. During the forty years of publishing *SJS,* the methodology not only evolved, but also new canonical manuscripts were identified and new text editions appeared. The revision of the text corpus and its excerption started in the new millennium. Its first result are the *Indices to the Old Church Slavonic Dictionary*, which appeared in 2003 (Ribarova 2003). In parallel with the preparation of the 2nd enlarged edition of *SS X-XI*, hereafter referred to as the *Dictionary of the Oldest Old Church Slavonic Manuscripts* (*DOOCSM*), the revision of *SJS* started. Between 2008 and 2016 the complete revision of the Volume I of *SJS* (*Addenda*) was published in eight issues (with 504 pages in total), its last main editor being Štefan Pilát.

The last lexicographic project managed by the Institute of Slavonic Studies is the *Greek-Old Church Slavonic Index*. Its preparation phase started back in 1989, while the work was initiated in 2004 under the leadership of Emílie Bláhová (1931–2016), an experienced co-author of all the previously mentioned dictionaries. The volume I of the *Index* (published in eight issues comprising 522 pages), an invaluable tool for study of translation techniques, was completed in 2014; the work on the next volume is being continued by a team led by Václav Čermák.

Besides these projects done in the Institute of Slavonic Studies, the prestige of *SJS* influenced other projects abroad, which have been done with the direct participation of Prague scholars. Thus, the methodology of *SJS* is being used in the the Croatian (*RCJHR*, published since 1991) and Macedonian (*RSJMK*, published since 2000) Church Slavonic dictionaries as well as in the *Comparative Index of Dictionaries Produced under the Commission for Old Church Slavonic Lexicology and Lexicography* (*CompIndex*) led by Zdenka Ribarova (1945), daughter of Josef Kurz. In the Czech Republic, the material of *SJS* was used by the recently completed (2018) *Old Church Slavonic Etymological Dictionary* (*ESJS*) composed by the Etymological Department of the Institute for Czech Language of the Czech Academy of Sciences in Brno. The *SJS* itself was reprinted in Russia in 2006.

The end of the *analogue phase* of the Prague Old Church Slavonic Lexicography is marked by the publication of a large volume dedicated to its history and personalities (Čajka & Černý 2016), which is an invaluable source for the history of the Old Slavonic studies in Prague in the 2nd half of the 20th century in general.

The Gorazd Project

The idea of digitizing *SJS* is not new. It appeared for the first time in 1992 as an initiative of a private publishing house. However, that time it was not realized (Čajka 2016: 90). The start of the digital era of the Prague Old Church Slavonic Lexicography is linked to the *GORAZD Project*, led from 2016 by a young but experienced lexicographer, Š. Pilát. As mentioned in the introduction, the ambition of the *GORAZD Project* is higher than just digitization. From the possible methods, the implementation team chose the option to integrate all the lexicographic works of the Institute, including the card index, into one cross-referenced system accessible via a uniform presentation interface that enables advanced search solution.
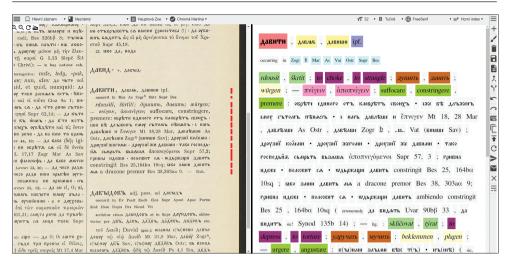
Thus, the lexicographic part of the project comprises the:
1. *Old Church Slavonic Dictionary*
2. *Dictionary of the Oldest Old Church Slavonic Manuscripts*
3. *Greek-Old Church Slavonic Index*
4. *Old Church Slavonic Card Index*

The digitization of the *Old Church Slavonic Dictionary* is being done in five phases. The first phase was the transfer of the large paper dictionary, using four different alphabets, six different languages and various font types, into the digital environment, for which we chose *ABBYY FineReader 12* and *ABBYY Recognition Server*. To increase the efficiency of this task, we designed the *OCR Vocabulary of Old Church Slavonic*. The result of this first phase was a raw digital copy of *SJS* in ALTO/XML format.

The 2nd phase consisted of processing these data by newly developed tools within the open-source digital-library solution INVENIO. We designed software called *Gorazd Generator*, which cut the raw material into entries and created their basic structure according to pre-defined criteria. The output of this tool is displayed in a web application called *Gorazd Editor,* enabling very user-friendly visualisation of both the microstructure and original scanned pages of the paper dictionary.

Picture 2: Gorazd Editor

The 3rd phase involves checking and correcting all entries created by the *Gorazd Generator*, which includes accurate correction of the text in all dictionary languages and emendation of the microstructure mark-up. It is to be added that the use of the *Gorazd Editor* does not suppose knowledge of any mark-up language since it works with well distinguishable colours. Thanks to thorough work done on the development of the software tools, the work is going quicker than it may seem – e.g., all mark-up can be emended with keyboard shortcuts. In order to get a satisfactory result, to each record (corresponding to an entry) a different status may be attributed recording the progress of work. The statuses, combined with various levels of access rights into the *Gorazd Editor,* enable the supervisor control of each emended entry.

The 4th phase of the digitization consists in updating the "digitized copy of *SJS*". This task comprises the integration of the *Addenda* directly to the entries corresponding to the Volume I of *SJS*, a change of the metalanguage and adding equivalents in English. In order to make the *Old Church Slavonic Dictionary* available to the broadest public, we are translating all Latin explanations (thus the grammar and the encyclopaedic definitions) into English.

The 5th phase implicates the integration of the new digital dictionary into a presentation interface, which is, in our case, powered by Gulliver, a

software solution developed by the company AiP Beroun Ltd. In this interface, we are also integrating the remaining two lexicographic works: The *Dictionary of the Oldest Old Church Slavonic Manuscripts* and the *Greek-Old Church Slavonic Index*. The tagging enables the interlinking of the contexts of all dictionaries and the Card Index, plus providing advanced search options. These include the possibility of a combined search by grammatical categories, texts or group of texts, equivalents in modern (English, Russian, German, and Czech) and historical (Greek, Latin, and Old High German) languages[5] and, of course, full text. Besides this, the presentation interface is able to provide quick statistical data on entries according to the chosen criteria (e.g. number/list of o-stem substantives in a certain group of texts). In order to better address user needs, the interface has three language versions (English, Russian and Czech), supports creation of personalized user collections and includes a user feedback tool for each entry.

Picture 3: Old Church Slavonic Dictionary in Gulliver

---

[5] The dictionary is thus usable also as e.g. an English-Old Church Slavonic dictionary, which opens completely new options to use the lexicographic material. A virtual keyboard for both old Cyrillic and Greek alphabets is available (Pilát & Knoll 2017: 340).

Digitization of the *DOOCSM* and the *Index* was easier as these projects had started in this millennium and thus the input data was already in a digital format. A larger update has been undertaken for the *DOOCSM*: the source of this digital dictionary is the newly revised and enlarged version of the original *SS X–XI*, which has not yet been published and which includes about 200 new entries. A specific feature of the digital *DOOCSM* is the capability to search according to lexeme frequency. After several discussions, we have decided to focus on the Czech reading public (scholars and students) that led to the change of metalanguage (from Russian to "internationalized" Czech). Within the current NAKI II programme, we are planning to publish a sample part of this dictionary online. The rest will be available after the appearance of the printed (2nd) edition of *DOOCSM*.

Picture 4: Dictionary of Oldest Church Slavonic Manuscripts in Gulliver

An important part of the *GORAZD Project* is the digitization of *OCSCI* (the largest card index). This unique database, accessible for the first time to scholars beyond the walls of the Institute of Slavonic Studies and containing all the preserved Old Church Slavonic forms, will be published online in two modes. Besides the possibility to find concrete card slips via lemmas of *OCSD*, the user will have the possibility to browse the complete scanned card index in a separate user-friendly interface. We do not plan to publish the Greek-Old Church Slavonic, Latin-Old Church Slavonic or Old High German-Old Church Slavonic Card Indices before finishing lexicographic works based on them.

Picture 5: Old Church Slavonic Card Index Interface

Recently, the project team has prepared an electronic database of *OCDS* sources, which is based on a significantly updated and enlarged version of the section *Fontes,* including both *SJS* (*SJS* I: LXII–LXX) and *Addenda* (*Addenda* 25–27). The database includes information on every text of the excerpted text corpus in English, Russian and Czech, enriched with links to publicly accessible digital manuscript facsimiles and editions. The database will be published on the project page (http://gorazd.org/). The project team is also involved in public engagement powered by the project's Facebook profile (https://www.facebook.com/projekt.gorazd/), through which materials to motivate students of Old Church Slavonic are shared.

Finally, let me sum up the software products of the project (Pilát et al. 2018: 3–4). We have already mentioned the *OCR Vocabulary of Old Church*

ation of projects dedicated to historical or encyclopedic dictionaries by other project teams, providing the possibility to reformat a paper dictionary into an editable electronic database and back to a new printed edition, and eventually to create a brand new marked-up dictionary both for online and printed use.

By the end of 2020, the scholarly part of the project will be offered to the broad public, and the mentioned lexicographic databases will become accessible via the project page (http://gorazd.org/).

After the end of the NAKI II programme, we plan to continue developing the database. The enlarging of the *Old Church Slavonic Dictionary* and the follow-up of the *Index* will be done directly in INVENIO. The digital environment enables a variety of new strategies for further development of *OCSD* (e.g., integration of lexicographic material text by text, Pilát 2016: 120), which will definitely replace the traditional *Addenda*. It will also be possible to start new lexicographic projects based on the collected material (Chromá 2017: 180) or to enrich the digital hub with multimedia records and other teaching material (Pilát et al. 2018: 16). Our *Old Church Slavonic Digital Hub* can also be opened for integration of further related lexicographic databases, which may enhance the *GORAZD project* on national or international levels.

References

Addenda: Hauptová, Z., Konzal, V., Pilát, Š. (ed.) *Slovník jazyka staroslověnského – Lexicon linguae palaeoslovenicae V. Addenda et corrigenda* Praha: Slovanský ústav AV ČR, Euroslavica, 2016.

Bláhová 2016: Bláhová, E. "Kronika Slovníku jazyka staroslověnského." In Čajka, F., Černý, M. (ed.): *Slovník jazyka staroslověnského. Historie, osobnosti a perspektivy*. Praha: Slovanský ústav AV ČR, 2016, 13–86.

Čajka & Černý 2016: Čajka, F., Černý, M. (ed.): *Slovník jazyka staroslověnského. Historie, osobnosti a perspektivy*. Praha: Slovanský ústav AV ČR, 2016.

Čajka 2016: Čajka, F. "Dodatky ke Slovníku jazyka staroslověnského (Lexicon linguae palaeslovenicae V. – Addenda et corrigenda." In Čajka, F., Černý, M. (ed.): *Slovník jazyka staroslověnského. Historie, osobnosti a perspektivy*. Praha: Slovanský ústav AV ČR, 2016, 87–108.

Čermák 2016: Čermák, V. "Řecko-staroslověnský index (Index verborum graeco-palaeoslovenicus)." In Čajka, F., Černý, M. (ed.): *Slovník jazyka staroslověnského. Historie, osobnosti a perspektivy.* Praha: Slovanský ústav AV ČR, 2016, 109-116.

Chromá 2016: Chromá, M. "Slovník nejstarších staroslověnských památek – Staroslavjanskij slovar' (po rukopisjam X-XI vekov)." In Čajka, F., Černý, M. (ed.): *Slovník jazyka staroslověnského. Historie, osobnosti a perspektivy.* Praha: Slovanský ústav AV ČR, 2016, 127-138.

Chromá 2017: Chroma, M. "Projekt Gorazd: virtual'nyj portal staroslavjanskogo jazyka i buduščeje češskoj staroslavjanskoj leksikografii." In Osipova, Je. P. (ed.): *I. I. Sreznevskij i russkoje istoričeskoje jazykoznanije: opyt i perspektivy. K 205-letiju so dnja roždenija I. I. Sreznevskogo.* Rjazan': Rjaz. gos. un-t imeni S. A. Jesenina, 2017.

CompIndex: Ribarova, Z. (ed.): *Comparative Index towards the Dictionaries Produced under the Commission for Old Church Slavonic Lexicology and Lexicography at ICS.* Skopje: Macedonian Academy of Sciences and Arts, Vol. I. 2015, Vol. II. 2018.

ESJS: Havlová, E., Erhart, A., Janyšková, I. (ed.): *Etymologický slovník jazyka staroslověnského.* Seš. 1-19. Praha: Academia, Brno: Tribun EU 1989-2018.

GORAZD: The Old Church Slavonic Digital Hub, http://gorazd.org/=en. Accessed 21 February 2019.

Index: Bláhová, E. (ed.) *Řecko-staroslověnský index – Index verborum graeco-palaeoslovenicus.* Praha: Slovanský ústav AV ČR, Euroslavica, fasc. 1-8, 2008-2014, vol. I. 2014.

Pilát & Knoll 2017: Pilát, Š., Knoll, V. "Bericht vom Projekt GORAZD: Digitales Portal des Altkirchenslawischen" *Byzantinoslavica* 75, 1-2, 338-340.

Pilát 2016: Pilát, Š. "Projekt GORAZD: Digitální portál staroslověnštiny." In Čajka, F., Černý, M. (ed.): *Slovník jazyka staroslověnského. Historie, osobnosti a perspektivy.* Praha: Slovanský ústav AV ČR, 2016, 117-126.

Pilát et al. 2018: Čajka, F., Chromá, M., Knoll, V., Pilát, Š., Spurná, K. *The Gorazd Project: An Old Church Slavonic Digital Hub.* Praha: Academia, 2018, http://www.academia.cz/edice/kniha/the-gorazd-project-an-old-church-slavonic-digital-hub. Accessed 21 February 2019.

RCJHR: Grabar, B., Hauptová, T., Mareš, F. V., Ribarova, Z. (ed.): *Rječnik crkvenoslavenskoga jezika hrvatske redakcije*. Zagreb: Staroslavenski institut, 1–20 1991–2014, vol. I. 2000, vol. II. 2015.

Ribarova 2003: Ribarova, Z. *Indexy k staroslověnskému slovníku.* Praha: Slovanský ústav AV ČR, Euroslavica, 2003.

RSJMK: Ribarova, Z. (ed.): *Rečnik na crkovnoslovenskijot jazik od makedonska redakcija Skopje: Institut za makedonski jazik „Krste Misirkov", 1-13, 2000-2018, vol. I. 2006.*

SJS: Kurz, J., Hauptová, Z. (ed.) *Slovník jazyka staroslověnského – Lexicon linguae palaeoslovenicae* Praha: Academia, Euroslavica, fasc. 1-52, 1958-1997, vol. I. 1966, vol. II. 1973, vol. III 1982, vol. IV. 1997. Reprint: *Staroslavjanskij slovar' 1-4.* Sankt Peterburg: Izdatel'stvo Sankt-Peterburskogo universiteta, 2006.

Specimen: Kurz, J. (ed.): *Slovník jazyka staroslověnského – Lexicon linguae palaeoslovenicae. Ukázkový sešit – Specimen.* Praha: Nakladatelství Československé akademie věd 1956.

SS X-XI: Cejtlin, R. M., Večerka, R., Bláhová, E. *Staroslavjanskij slovar' (po rukopisjam X-XI vekov).* Moskva: Russkij jazyk, 1994. Reprint: 1999.

About the author
Vladislav Knoll, knoll@slu.cas.cz
PhD, Department of Old Slavonic and Byzantine Studies of the Institute of Slavonic Studies of the Czech Academy of Sciences, Prague.

Резюме

Проект *ГОРАЗДЪ: Цифровой портал старославянского языка* на сайте http://gorazd.org/ является достойным продолжением долгой традиции чешской палеославистической лексикографии в цифровой эпохе. Первые попытки создания словаря древнейших славянских текстов на основе строго определенного корпуса начались уже в начале XX века. Разработка картотеки настоящего *Словаря старославянского языка* (*ССЯ*) началась в 1943 г. при содействии крупнейших чешских славистов. В печатном виде словарь выходил с 1958 по 1997 гг. На основе его материала возникли последующие лексикографические проекты, такие как *Словарь древнейших старославянских памятников* (с 1994 г.), *Греческо-старославянский указатель* (с 2008 г.) и *Дополнения к первому тому Словаря старославянского языка* (с 2008 г.).

Создание оцифрованной версии перечисленных трудов является одной из основных задач проекта ГОРАЗДЪ, который реализуется сотрудниками Славянского института Чешской академии наук. Помимо использования преимуществ виртуального пространства, включающих продвинутые возможности поиска, персонализацию лексикографических данных, создание статистик и, благодаря множеству взаимных ссылок, быстрое передви-

жение по словарям и картотеке, электронные версии словарей были актуализованы и усовершенствованы. Что касается основного труда – *ССЯ*, мы решили изменить метаязык, которым являлась в печатном издании латынь. Чтобы упростить доступ для экспертов на глобальном уровне, все грамматические и энциклопедические комментарии были переведены на английский язык, который таким образом стал четвертым живым языком, на котором приведены эквиваленты старославянских лемм. *Дополнения к первому тому Словаря старославянского языка* были включены прямо в структуру словарных статей *ССЯ*. Важные изменения ожидают также *Словарь древнейших старославянских памятников*. Он основан на второй, расширенной версии первоначального словаря, изданного 1994 г. в Москве под названием *Старославянский словарь (по рукописям X–XI веков)*. Помимо тщательного пересмотра словаря и дополнения материала из новых источников, в новую версию были добавлены английские эквиваленты; метаязыком стал чешский.

Важной задачей проекта является также разработка софтверных инструментов, которые способны создать из многоязычного печатного словаря толкового или энциклопедического характера электронную маркированную базу данных, поддающуюся удобному редактированию, которое не предполагает знаний разметки XML. Более того, один из разработанных инструментов позволяет быстрое создание файла, предназначенного для подготовки печатного издания словаря, основанного на электронной базе данных.

Софтверный этап был завершен в конце 2018 г., а его результаты в виде исходных кодов софтверных инструментов будут в ближайшее время представлены широкой аудитории. Опубликование электронных словарей планируется на конец 2020 г. Создание *Цифрового портала старославянского языка ГОРАЗДЪ* открывает большие перспективы и для продолжения лексикографических проектов Славянского института прямо на электронной платформе с возможностью простого генерирования новых печатных томов или изданий отдельных словарей и, не в последнюю очередь, интеграцию похожих лексикографических проектов.

DIGITAL EDITIONS AND CORPORA OF FRANCOPHONE DIARIES BY ALEXANDRE CHICHERIN AND OLGA ORLOVA-DAVYDOVA[1]

Alexei Lavrentiev, Michèle Debrenne,
Nina Panina, Dmitry Dolgushin, Andrey Borodikhin

Abstract: This paper presents a project of digital editions and research corpora of francophone diaries written by two Russian aristocrats, Alexandre Chicherine (1793–1813) and Olga Orlova-Davydova (1814–1876). The original French texts of these diaries have never been published in spite of their considerable value for research in linguistics, literature, social history and history of the arts. Alexandre Chicherin's diary contains a great number of the author's drawings that are closely related to the text of the diary entries. Olga Orlova-Davydova's diary contains interesting details on the everyday life of the Russian aristocracy and on some historical events and well-known figures. Both diaries provide rich data on the interaction of the Russian and French languages and cultures. The workflow of the project consists of a primary transcription with Microsoft Word using styles and some special characters (micro-syntax). This transcription is automatically converted to TEI XML and then imported into the TXM corpus analysis and publication platform (http://textometrie. org). Prototype corpora of both diaries are available on the TXM demo portal (http://portal.textometrie.org/demo/?command=page&path=/JournauxFrancophones).

Keywords: diaristic literature, French, TXM, TEI, cultural interaction

1. Introduction

Personal diaries written in French by Russian aristocrats in the 19th century are an interesting object for research in the fields of linguistics, literature, social history and history of the arts. Whereas studies of diaries created in monolingual contexts (both Russian and French) have been carried out for a few decades,[2] only in recent years have the first works fo-

[2] See Egorov 2018 for bibliographic references.

cused on the bilingual aspect of this literature been published (Gretchnaia & Viollet 2008; Gretchnaia et al. 2012). Many valuable documents are still unpublished, and technologies developed in digital humanities offer a unique opportunity to make these documents available to the academic community in the form of complex digital resources including facsimile, multi-layer transcription, critical apparatus, linguistic annotation and tools for search and visualization. In this article we present the work done on a corpus of two diaries written in the beginning and in the middle of the 19th centuries that were analysed and published using the TXM corpus analysis software.

2. Source documents

The first diary of the project corpus was written by Alexandre Chicherin (1793–1813, referred to as "AC" hereafter), a young Russian officer who took part in the war against Napoleon. He kept his diary from 6 September 1812 to 13 August 1813, when he was mortally wounded in a battle. The manuscript is held in the Russian Public Historical Library. It consists of 270 pages of text and illustrations (83 in total). These illustrations play a very important role in the diary because they interact with the text in many ways, and these relations are particularly interesting to study (Panina 2018). A Russian translation of the diary was published in 1966 (Engel & Perper 1966), but only a few of the illustrations were reproduced. In some cases the translation is not quite faithful to the original, and some passages were deliberately omitted (Mischenko 2001). Therefore, a digital edition of the original French text accompanied by all the illustrations and equipped with tools for analysis of text-to-image relations would be of great use to the research community.

The second diary belongs to Olga Orlova-Davydova (born Bariatinskaya, 1814–1876, referred to as "OD" hereafter), a member of a well-known Russian aristocratic family. Most of the entries correspond to the period from 1830 to 1847, with just one note added in 1849. (OD kept a separate diary, written in Russian, from 1869 to 1870, but this document is beyond the scope of our project.) The original manuscript is stored in the Russian State Library (Orlov-Davydov fund, F. 219, box 92), and there exists a manuscript copy which was probably ordered and revised by her daughter Maria Orlova-Davydova in the late 1890s or early 1900s. This

copy is stored in Novosibirsk State Public and Technical Library of the Siberian Branch of the Russian Academy of Sciences as a part of the Tikhomirov collection.[3] This copy is an interesting document for the study of Franco-Russian bilingualism, as it contains traces of language interference of both the author and the copyist. It also has the advantage of being more legible and easier to access for the project team based in Novosibirsk. For this reason the project started with the transcription of the copy, whereas the transcription of the original was added at a later stage.

3. Workflow

The workflow for both diaries was generally the same, although each source required some special processing. The first, and the most important and time-consuming, stage consisted of the transcription of primary sources with Microsoft Word. At this stage customized styles and special characters were applied to prepare automatic conversion to TEI XML and import into TXM. At the second stage the document files were converted to TEI XML using the OxGarage service,[4] and finally a series of XSLT transformations was applied during the TXM import process. The choice to do all the work of editing and pre-annotating the transcriptions with Microsoft Word (and not directly in TEI XML) was due to the fact that it was impossible to organize a training in this technology for the Russian team within the deadlines of the project. The high cost of user-friendly XML editing software like Oxygen XML Editor also influenced the decision to use Microsoft Word. The project was an occasion to test how far it is possible to go in text annotation using Microsoft Word.

4. Document structure

Microsoft Word's "Heading X" styles were used to encode the text structure. The basic level in both diaries is a daily entry, but there are some differences at higher levels of the structure.

AC's diary is contained in a single volume, so encoding the structure was rather straightforward: a single file for the diary and a division for

---

[3] See Borodikhin & Dolgushin Acc. for a detailed presentation of the source documents.

[4] http://oxgarage.tei-c.org.

each daily entry. The situation is much more complex in the case of OD. Both the original and the copy were written in several notebooks, and the chronological order of the entries does not always correspond to the archival order of the documents. In 1834 OD wrote a long autobiographical chapter starting with her birth and ending in 1833, so there is an overlap with some earlier entries. This chapter marks, in our opinion, the beginning of a "new diary" for several reasons: the story of OD's life starts from the beginning, she tries to organize records in chapters (although she abandons this effort some years later), and she calls the chapter "Chapter 1" and writes it (and the following chapters) in a new notebook. In addition to regular entries, OD used a separate notebook to write down some spiritual thoughts in 1843, 1847 and 1849. Therefore, we believe that this diary consists of three separate "works": (1) the early diary (1830-1834), (2) the new diary (1834-1845) and (3) spiritual thoughts (1843, 1847, 1849).

The primary transcription of the diary was made in the archival order of the source documents. Each notebook was transcribed in a separate document file. The call number is given in "Heading 1" style in the beginning of each document. "Heading 2" is used to indicate to which of the three "works" of OD's the following entries belong.

A normalized date in {YYYY-MM-DD} format (according to the Julian calendar used in Russia until 1918) was added in "Heading 3" style for each daily entry in both diaries. This was necessary to allow automated analysis and comparison of records according to their date.

In OD's diary, some dates were given in the Gregorian calendar (especially during her trips to Western Europe), and some dates are erroneous. A correct normalized date according to the Julian calendar was provided in all cases. Some records describe events of several days; in this case a "+" sign was added to the normalized date (e.g. {1830-11-23+}). The dates as they appear in the source document were transcribed and marked up with "date" character style.

In AC's diary most of the records were accompanied by illustration. For each image, we added a table with some metadata including the title of the drawing, the date of the drawing and the corresponding entry, the drawing genre and style, the people and places depicted, etc.). Whenever possible, the inner structure of the entries was marked up using custom-

ized paragraph styles such as "1_preamble", "2_action", "3_retrospection", "4_comparison" and "5_conclusion".

5. Segment-level markup

Names of people and places were marked up using character styles named "persName" and "placeName" respectively. An "unclear" style was applied to hardly legible segments. All these styles are automatically converted to proper TEI XML tags by OxGarage. In other cases we had to use project-specific style names and some special characters to ensure more precise TEI XML markup. These are processed and converted to TEI XML by XSLT scripts during import into TXM.

For instance, the styles named "sic-ortho", "sic-lex" and "sic-gramm" were used to mark spelling, vocabulary and grammatical errors, respectively. If necessary, a particular letter or group of letters was marked as erroneous using curly braces: e.g., *G{é}nève* (the whole word is marked using "sic-ortho" style).

If correction marks were present in the source document, the styles "add", "del" and "subst" were used for additions, deletions and replacements, respectively, and square brackets indicate the added or deleted characters. For instance, *pren[aie/a]nt* means that the word form *prenaient* was changed to *prenant*.

Russian words were marked with "lang-ru" if written in Cyrillic and "translit" if they were transliterated (e.g., *téléga*, Russian for 'cart').

The situation is more complicated where several different phenomena co-occur. For example, a personal name could be spelled in Cyrillic and a place name could contain an error. In this case combined style names such as "placeName-sic" were used, but it is clear that the use of styles in Microsoft Word meets its limits here, as unlike TEI XML elements, the character style of word-processing software cannot nest.

6. Pagination

In order to connect the transcription to the facsimiles of the source documents and to facilitate browsing through the document, special markup was used for page breaks. Some pages are numbered in the source documents, some are not, and in some cases the numbers appearing in the source document are erroneous. Therefore, in the digital edition each page had two

numbers: an "original" number, included whenever present in the source document, and an "archival" number systematically added. If both numbers are identical, simple notation in angle brackets was used: e.g., <10>. If the original number is different from the archival one, the latter was given after a vertical line: e.g., <14|15>. If the original page is unnumbered, the following code was used : e.g., <|12>.

If a page break occurred inside a word in the source document, the mark is placed exactly where it was found in the source document: e.g., *der-<11>nière*.

7. Import into TXM

TXM is a free open-source software platform designed for corpus compilation, analysis and publication (Heiden 2010). Thanks to its "XML XTZ + CSV" import module, it is possible to create corpora from any XML-encoded source documents through a series of XSLT transformations and under certain conditions generate synoptic pagination displaying side-by-side the text and a facsimile of the source document.

XSLT processing is available at different import stages:

1. Splitting or merging XML source files in order to optimize the corpus structure;

2. Pre-processing XML files to prepare for tokenization;

3. Post-processing XML files after tokenization to fix possible errors and create token-level annotations;

4. Generating one or several custom page layouts for reading the texts of the corpus.

In our project, only OD's diary required split-merge stage processing, due to the complexity of its text structure described earlier. After a conversion of transcription Word documents to TEI XML with OxGarage, the data was reorganized to create six files (three for the original and three for the copy) corresponding to three separate "works" we identified earlier.

At the next stage, a series of transformations was applied to obtain a TEI-conformant XML file ready for tokenization. As a matter of fact, text-processing software like Microsoft Word or LibreOffice Writer may create artificial divisions between text segments with identical markup. It may also introduce segments with additional formatting information (such

as font size) as a result of copy and paste or other editing actions. These segments and divisions are invisible to users but they may cause errors in tokenization and element indexing. Therefore, we had to remove superfluous segments and merge artificially divided adjacent elements before any further processing. Once the merger of adjacent identical elements is complete, it is possible to convert project-specific styles and special characters into proper TEI tags. Any styles not recognized as TEI tags are rendered as <hi rend="styleName"> by OxGarage. Their conversion to proper TEI is pretty straightforward with XSLT. Transforming special characters into XML tags requires a sequence of templates that parse text nodes with regular expressions.

Tokenization is operated by a Groovy script internal to the TXM platform. With this script, the user may adjust the list of word-separating characters and pre-tokenize some segments that require complex processing. The user can also provide a list of elements that should not be tokenized (for instance, editorial notes).

The tokenized text is once again submitted to a sequence of XSLT transformations. At this stage it is possible to merge or split some tokens and to add token-level annotations to facilitate corpus queries. In our case, an attribute "error" was introduced for various types of errors marked up in the Word document. Another attribute, "crochets", re-introduces some special characters to the word form (such as brackets for additions and deletions) in order to allow queries on the word-internal markup.

After the third series of XSLT transformations, the properly tokenized text is submitted to TreeTagger (Schmid 1994) for automatic part-of-speech tagging and lemmatization. The POS tags and lemmas provided by Tree-Tagger, as well as all other word-level annotations, are finally recorded in TEI-TXM XML format. The TXM extension for TEI provides the possibility to create an unlimited number of annotations at the word level thanks to <txm:ana> element.

At the next stage, importing a corpus, TXM produces indexes for the CQP search engine and generates edition pages. These can be created either by a default Groovy script or by a custom XSLT stylesheet. In our case, XSLT customization was necessary in order to visualize various levels of markup.

Figure 1. Screenshot of a page transcription from OD's diary

8. Reading and analysing the corpus with TXM

In Figure 1 we provide a screenshot of a page from OD's diary. The page number is given in brackets, as the page is not actually numbered in the source document. The normalized date and the call number of the source document are displayed in gray in the beginning of each entry. In the transcription that follows one can see a place name in green and the date in red, the errors highlighted in yellow, and the correction in the source document highlighted in blue. Additional comments are provided in footnotes. When the mouse cursor is placed over a word, a tooltip with corresponding annotations is displayed.

Figure 2 represents a screenshot from AC's diary containing an illustration followed by a table with its description. In this corpus image, descriptions and the transcription are both searchable, but it is of course possible to work on a subcorpus of pure transcriptions.
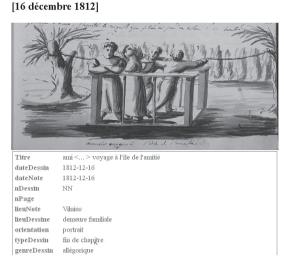
[16 décembre 1812]

Titre	ami <… > voyage à l'île de l'amitié
dateDessin	1812-12-16
dateNote	1812-12-16
nDessin	NN
nPage	
lieuNote	Vilnius
lieuDessine	demeure familiale
orientation	portrait
typeDessin	fin de chapitre
genreDessin	allégorique

Figure 2. Screenshot of an illustration and its description from AC's diary

A screenshot of the transcription of AC's diary is given in Figure 3. Background colour is used to indicate the semantic structure of the record that was marked up with paragraph styles in the initial transcription. The yellow background corresponds to a preamble, blue to an action and green to retrospection. As in OD's diary, personal names are displayed in blue font colour and corrections are highlighted in blue. Illegible segments are represented by "<…>".
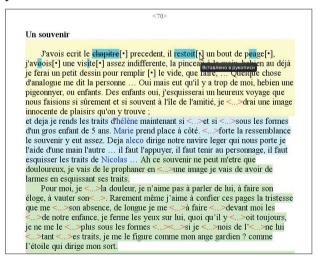

Figure 3. Screenshot of a page transcription from AC's diary

In addition to visualising the edition, TXM provides a wide range of tools for qualitative and quantitative research on the corpus. For instance, the *index* command allows you to obtain a list of personal names in the corpus with the following query: *<persname> []+ </persname>*. There are 7041 occurrences of 1595 different forms of personal names in the copy of OD's diary. Another query used with the *progression* command produces a chart with a curve increasing with every occurrence of a Russian word (*<foreign_lang="ru"> []+ </foreign>*). The steeper the curve, the denser is the use of the searched term over time. As one can clearly see in Figure 4, the density of Russian words increases considerably in 1845 and the following years. The distance between the vertical lines representing years corresponds to the volume of writing. We can see that this volume increases dramatically in 1844 and decreases progressively later.
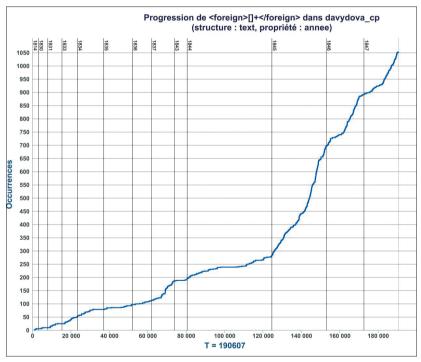

Figure 4. Progression chart of Russian words in OD diary

The full range of research tools provided by TXM is provided in the TXM User Manual and various tutorials available at the TXM project website (http://textometrie.org).

9. Conclusion

This project on Francophone diaries has provided an occasion to set up a workflow for creating a complex digital edition by an international team and to test the opportunities that TXM provides for working with such kind of corpora. The workflow based on transcription and pre-annotation with Microsoft Word and a series of XSLT transformations incorporated in the TXM import process has been efficient, although it is less suitable for complex cases of embedded annotations (e.g., manuscript errors in proper names).

The digital editions AC's and OD's diaries are available online on the TXM demo portal[5]; the access codes to read and search theses resources are provided upon request. Work on their proof-reading and refining is still going on. Once the text of the editions is stabilized, open public access to them will be provided.

References

Borodikhin & Dolgushin 2019: Borodikhin, A. Yu., Dolgushin, D. V. "Dnevniki O. I. Orlovoj-Davydovoj kak chast' Tihomirovskij kollekcii GPNTB SO RAN." *Quaestio Rossica*, Forthcoming (article accepted).

Debrenne 2016: Debrenne, M. "Sopostavitel'nyj deviatologicheskij analiz perepisannyh dnevnikov O. I. Davydovoj i pervichnyh tekstov." *Vestnik NSU. Series: Linguistics and Intercultural Communication* 14-3 (2016), 59–75.

Debrenne 2017a: Debrenne, M. "Sozdanie sistemy razmetki francuzskogo teksta dnevnikov O. I. Davydovoj dlja avtomaticheskoj obrabotki teksta." *Vestnik NSU. Series: Linguistics and Intercultural Communication* 15-1 (2017), 34–40.

Debrenne 2017b: Debrenne, M. "The French Language in the Diaries of Olga Davydova." In Strien-Chardonneau, M. van & Kok Escalle, M.-Ch. (eds.), *French as Language of Intimacy in the Modern Age. Le français, langue de l'intime à l'époque moderne et contemporaine.* Amsterdam: Amsterdam University Press, 2017.

Debrenne 2018: Debrenne, M. "Sopostavitel'naja tipologija oshibok i ispravlenij v originalah i kopijah dnevnikov O. I. Orlovoj-Davydovoj". *Vestnik NSU. Series: Linguistics and Intercultural Communication* 16-2, 127–143 (DOI: 10.25205/1818-7935-2018-16-2-127-143).

---

[5] http://portal.textometrie.org/demo/?command=page&path=/JournauxFrancophones.

Egorov 2018: Egorov, O. G. *Russkij literaturnyj dnevnik XIX veka. Istorija i teorija zhanra. Issledovanie*. 2nd edition. Moscow: Flinta, 2018.

Engel' & Perper 1966: Engel', S. T., Perper M. I. (eds.) *Dnevnik Aleksandra Chicherina, 1812–1813*, Moscow: Nauka, 1966.

Gretchnaia & Viollet 2008: Gretchanaia, E., Viollet, C. *"Si tu lis jamais ce journal…": Diaristes russes francophones, 1780–1854*, Paris: CNRS-Editions, 2008.

Gretchnaia et al. 2012: Gretchanaia, E., Stroev, A., Viollet, C. (dir.) *La francophonie européenne aux XVIIIe–XIXe siècles. Perspectives littéraires, historiques et culturelles*. Bruxelles, Bern, Berlin, Frankfurt am Main, New York, Oxford, Wien: Peter Lang, 2012.

Heiden 2010: Heiden, S. "The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme." in Otoguro, R., Ishikawa, K., Umemoto, H., Yoshimotoand, K. and Harada, Y. (eds.) *24th Pacific Asia Conference on Language, Information and Computation*. Tokyo: Institute for Digital Enhancement of Cognitive Development, Waseda University, 2010, 389–398 (https://halshs.archives-ouvertes.fr/halshs-00549764).

Mischenko 2001: Mischenko, T. K. "A. V. Chicherin – uchastnik voennyh dejstvij pod Malojaroslavcem: iz opyta raboty s podlinnym dnevnikom A.V. Chicherina." in *Otechestvennaja vojna 1812 goda v Kaluzhskoj gubernii i rossijskoj provincii. Materialy nauchnoj konferencii*. Malojaroslavec, 2001, 10–20.

Panina 2016: Panina, N. "Dnevnik Aleksandra Chicherina 1812–1813 gg.: Avtobiograficheskij zamysel v kontekste frankofonii avtora". In Dmitrieva, E.E., Lebedeva, O.B., Stroev, A.F. (eds.) *Sibirsko-francuzskij dialog XVII-XX vekov i literaturnoe osvoenie Sibiri: materialy Mezhdunarodnogo nauchnogo seminara, Tomsk, 11-15 iyunya 2015 / In-t mirovoj literatury im. A. M. Gor'kogo Rossijskoj akad. nauk,* Moscow: IMLI RAN, 2016, 202–216.

Panina 2019: Panina, N. "Francuzskij original dnevnika Aleksandra Chicherina 1812–1813 gg." *Tekst. Kniga. Knigoizdanie*. Forthcoming (article accepted).

Schmid 1994: Schmid, H. "Probabilistic Part-of-Speech Tagging Using Decision Trees." in *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, 1994. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger

About the authors

Alexei Lavrentiev, PhD, research engineer at the Institut d'Histoire des Représentations et des Idées dans les Modernités, CNRS & ENS de Lyon, Lyon, France, alexei.lavrentev@ens-lyon.fr.

Michèle Debrenne, Dr Sc, professor at the Institute for the Humanities, Novosibirsk State University, Novosibirsk, Russia, micheledebrenne@gmail.com.

Nina Panina, PhD, associated professor in the Department of History, Culture and Art of Novosibirsk State University, fellow researcher at Novosibirsk State University of Architecture, Design and Arts, Novosibirsk, Russia, panina@mmedia.nsu.ru.

Dmitry Dolgushin, PhD, associated professor in the Department of 19th- and 20th-century Literature and Department of History of Culture, Novosibirsk State University, Novosibirsk, Russia, d_dolgushin@mail.ru.

Andrey Borodikhin, PhD, Head of the Department of Rare Books and Manuscripts, State Public Scientific Technological Library of the Siberian Branch of the Russian Academy of Sciences.

Релюме

В статье представлен проект подготовки электронного издания и компьютерного корпуса франкоязычных дневников двух русских дворян – Александра Васильевича Чичерина (1793–1813) и Ольги Ивановны Орловой-Давыдовой (1814–1876). Оригинальный французский текст этих дневников никогда не публиковался, несмотря на значительную ценность их материала для исследований в области лингвистики, литературоведения, истории и истории культуры. Дневник А. В. Чичерина содержит большое число авторских рисунков, которые тесно переплетаются с текстом записей. Дневник О. И. Орловой-Давыдовой содержит ценные сведения о повседневной жизни русской аристократии, о ряде исторических событий, а также штрихи к портрету личности ряда исторических фигур. Оба дневника являются богатым источником данных для исследования русско-французского взаимовлияния в области языка и культуры. Технологический процесс проекта состоит в первичной транскрипции а программе Microsoft Word с использованием стилей и ряда специальных символов (микросинтаксис). Транскрипция автоматически конвертируется в формат TEI XML и импортируется на платформу TXM для анализа и публикации корпуса

(http://textometrie.org). Прототипы изданий обоих дневников доступны не демонстрационном портале платформы TXM (http://portal.textometrie.org/demo/?command=page&path=/JournauxFrancophones).

APPROACHES TO ELECTRONIC PROCESSING OF ARCHIVAL MASS SOURCES: REFORMATTING OR CREATING DIGITAL RESOURCES1

Lyudmila Mazur and Oleg Gorbachev

Abstract: There are two approaches commonly applied in research to digitize primary data sources: problem-orientated and source-orientated. The first one is more economical, efficient and flexible. What we get, however, is a single-use resource. The second approach is aimed at designing a system that would contain the complete set of authentic data. Thus, the system would be suitable for different purposes, but it would take much more time to construct.

This paper analyzes a case of two sets of primary data and the problems faced when designing and creating a digital multi-purpose resource system. The first data set includes primary forms of the All-Russia Communist Party Census of 1922 and 1927. The second set comprises family budget records taken during budget surveys of 1929 and 1963. Both sources have a complex hierarchical structure, difficult to adjust for data normalization. Finally, the records taken in different years characterize the dynamics for specific social groups, which presents us with the challenging task of organizing this information and solving the problem of incompatibility between the structure of the records. All the above-mentioned characteristics can lead to the loss of primary data when converting them into a database.

When creating a digital resource system, it is essential to develop the right digitization strategy, one that is able to compensate for data loss and able to support a system suitable for a variety of purposes and adjustable to the individual needs of each user.

Keywords: Reformatting, digital resources, digital archive, database, budget surveys, Communist Party Census

Today there is a common understanding (more or less) of how to digitize ancient manuscripts: a complete copy is usually required, with all the

---

original features of the document, but there is no universally accepted approach to documents of the 20th century. They have not yet been covered by the noble patina of the past. And very often at first sight they contain an excessive amount of information. What if the volume of the information in the document significantly exceeds the immediate needs of the researcher?

The formation of the digital information environment is one of the trends in the development of the discipline of history, which, along with other humanities, is undergoing a digital change and faces a lot of longstanding unresolved problems. We'd like to begin with outlining the main trends of the development of history as a discipline:

– *thematically*, from the middle of the 20th century, there has been a transition from the history of events to the study of processes and history of everyday life, mentality, memory, etc.

– *methodologically,* the zone of interdisciplinarity is expanding, which affects not only the conceptual apparatus but also the formulation of research issues and the use of methods from related sciences, especially sociology, economics, statistics, geography, and so on.

– *from the point of view of the information and resource potential*, there is an expansion of the base of available sources for historical research to include mass sources, primarily ego-documents, nominative sources, periodicals and so on.

– *technologically,* there is a total computerization of historical research practices, because today the computer is indispensable in a historian's work, although the scope of its use can vary significantly: from searching for information and composing text to using the capabilities of computer technologies at all stages of the research process, including the processing and transformation of historical data with the help of a variety of standard or specialized program packages – databases, GIS, media and so on. As a significant part of the professional community of historians begins to actively use digital resources, there comes a wide recognition of the fact that digital orientation in the expanding world of electronic resources becomes part of the "historian's craft", according to Leonid Borodkin (Borodkin).

An integral element of all these directions is *the creation of digital resources* (in Russian, *digitalizatsia*). It should be distinguished from *digital reformatting*, which means transferring information from physical to digital

media, such as scanning a document. Reformatting is now realized in the archival sector. In Russia, according to the Program for the Informatization of Archives, by 2020, 20% of all documentary arrays should be reformatted. In the process of reformatting, there is no change in the structure of information: it simply acquires an electronic form, convenient for use in digital format. This approach can be correlated with the third industrial revolution, which lasted from 1969 to 2010.

But the creation of an electronic analogue of a paper document does not fundamentally change the information environment and only facilitates remote access to archival documents. This procedure is appropriate when working with unique texts. If we are dealing with complexes of mass nominative sources, then an intermediate step in the process of information transfer is its formalization and transformation into data structures available for additional quantitative or spatial processing.

The creation of digital resources is understood more widely – this is the process of creating a new information product in digital form. For example, the creation of databases on census materials, electronic maps, 3D-reconstructions and so on. In this case, the information of the primary source is transformed into a new form with the inevitable change in the initial information potential (its expansion or contraction). The key advantage of such work is the creation of a new innovative product, with new functional and consumer properties. And if *reformatting* is primarily aimed at improving existing traditional research practices, *the creation of digital resources* allows us to bring them to a new level of interdisciplinary interaction and use. On this basis, the creation of digital resources can be considered a phenomenon of the fourth industrial revolution.

The Russian specificity of the process reveals two main trends:

- in the archival sphere, most digital transformation initiatives are aimed at *reformatting*;

- in the discipline of history, primarily in digital history, the creation of new information products prevail. However, the possibilities of using them are limited, since they presuppose the creation of information products intended for a specific task. As a rule, the secondary use of these products is difficult or impossible.

The most important task of developing the infrastructure of historical research is the creation of open information resources, which are the result

of *creating digital resources*. Requirements for the development of resources of this kind include the following criteria:

1. *Source-orientation*. The creation of a digital resource assumes the measurement, formalization, normalization of primary materials of mass historical sources (individual forms of population censuses, metric records, *revizskiye skazki*, and other variants of population registers), which contain mainly personal (nominative) information.

2. *Completeness*. The resource includes the entire array of surviving documents or records.

3. *Efficiency*, i.e. resource uses a data format that provides storage, import, integration of non-recurring resources and their collective use.

4. *Openness*. The resource is available to the scientific community.

The ideology of creating source-oriented resources available online was developed in the concept of the "big database". It is built on the observance of certain standards for the description of data and the use of such storage formats that ensure their viability under conditions of rapid change of hardware and software (Gorbachev).

The implementation of international projects for the creation of big databases (IPUMS-USA, NAPP, Mosaic, the Vienna Database on European Family History, church books, the Demographic Data Base at Umeå University, etc.) demonstrates the high scientific potential of this scientific trend, which contributes to the formation of a new research paradigm that supplants the national frameworks for research practices and creates opportunities for cross-national research.

There are two main strategies for creating digital resources of primary mass sources: a *research-oriented* and *source-oriented approach*. The former implements the principle of selecting information from a source in accordance with the question being studied. It is most often used because it is time-saving, operative and adaptive, considers the research requests and hypotheses. But in the end, a one-time resource is created.

The latter is aimed at maximally complete and authentic representation of source information in the system being designed. In this sense, it corresponds to the tasks of multi-purpose use, but it is more labor-intensive, especially in the case of information-intensive and complexly organized sources. These are, for example, the primary forms of the All-Russian Communist Party Census of 1922, which include 59 questions, many of which contain

sub-questions. After normalization, the number of questions in the form has grown to 120.

Another example of a complex structured source is the primary forms of budgetary surveys of peasant households, whose information potential reaches several thousand features (about 5,000). Implementing a source-oriented approach to them is almost impossible: loss of information is inevitable. What to do in this case?

Thus, one of the crucial tasks of creating a multipurpose information resource is the development of a strategy for creating digital resources that would allow to ensure a mode of multipurpose use of the resource with compensation of information losses and flexible adaptation to individual requests of different users.

As an approach to creating digital resources, we offer a combination of a database and an electronic archive of document scans. This strategy was tested using the example of the two sets of sources mentioned above: the first complex includes the primary forms of the All-Russian CP Census of 1922 and 1927, and the second complex are the household budgets, preserved as a result of budgetary surveys in 1928/29 and 1963.

Both sources are characterized, first, by a complex hierarchical structure, poorly adapted to normalize the data; secondly, they contain a large amount of primary data describing the object (party member or peasant household). Thirdly, they represent a dynamic complex, including descriptions of the object from different years. All these features cause loss of primary information when converted to a database.

Let us dwell on their characteristics.

I. When designing the Information system "All-Russian CP census 1922–1923", the main question was how to implement the project: to follow the structure and features of a source or to create an information resource according to the concrete research task? (Taller) The second option is in a sense simpler and more efficient since the problem of formalizing the source information is also solved, taking into account the requirements and tasks formulated by the researcher. However, the choice was made in favor of *a source-oriented model* that allows the reuse of data for different research purposes by maximizing the full display of information contained in a complex of sources (Gutnov & Pereverten'). Source orientation manifests itself in the structure of the information array, which includes a database with

structured but not formalized questionnaire information and electronic copies of documents (questionnaires) in JPEG format. Electronic copies are tied to specific records – i.e. the user, working with the archive, can study the information entered into the database and simultaneously view the electronic copy of the questionnaire by checking or specifying information.

The need to include electronic copies of documents in the archive is justified by the following considerations:

1. *The needs of control over the correctness of data entry*. This control can additionally be carried out by the user, comparing information in the database with a copy of the document;

2. *The possibilities of reorganizing the database*. Despite the declared principle of completeness in the development of the structure of the database, it was impossible to structure and reflect all the information available in the questionnaires. Something is inevitably lost or does not fit into the database format, but can be interesting to the user. In this case, it is necessary to create conditions for the user to "refine" the database for their tasks;

3. *Visualization of the source allows us to clarify its features*: the appearance, the nature of the records, additional remarks on the questionnaire, the correctness of its design – this is especially important in cases of illegible handwriting of the registrar.

The type of this electronic resource can be defined as *prosopographical* since it has the characteristic features caused by a complex of sources. For example, the database contains personal data of communists not only at the time of the census but also in dynamics concerning labor and party activities and military service. Generalization of this information makes it possible to compile a collective biography of the members of the Ural organizations of the Russian Communist Party (Bolsheviks) of the 1920s.[2]

The information system operates in two modes: search and browsing. In the first case, the user can search for a name or the name of the locality to find information about a person. The second mode is working with filters to sample records according to a given criterion (age, nationality, education, social status, etc.) to use statistical methods to obtain indicators for both creating a collective social portrait of the communists of the Urals from

---

[2] See the interpretation of prosopography in the framework of historical informatics in: Yumasheva.

1922 to 1924 and for the study of individual groups of communists. Filters can be used not only in database mode; tables can be imported into a more flexible and user-friendly format (for example, Excel) and be processed.

The database "All-Russian Census of Members of the RCP(b): Ekaterinburg Province, 1922" consists of seven tables in accordance with the main thematic blocks of Form A: general information (14 fields), education (18 fields), social origin (9 fields), labor activity (12 fields), party activities (15 fields), revolutionary activity (17 fields) and military service (27 fields). The tables are linked through the key fields "Sequence number" and "Last name, first name, middle name." The size of the database is 12,000 records, plus about 10,000 for the set from 1924. The database is posted on the website of the project "Early Soviet Society As a Social Project: Ideas, Implementation Mechanisms, Design Results", currently test mode. A copy of the database was provided to the Public Organizations Documentation Center of the Sverdlovsk Region for use as a search tool.

II. Another resource, the creation of which began in 2018, includes information on primary forms of budgetary surveys of peasant farms from 1928 to 29 and in 1963.

Methods for organizing and carrying out budgetary surveys were developed in the Russian Empire at the end of the 19th century, thanks to the activity of zemstvo (a type of regional representative authority). After the Bolshevik Revolution of 1917, these investigations were renewed and originally acquired the form of a periodic analysis, carried out in cluster surveys, covering mainly peasants' households.

In 1921 and 1922, about 500 households were surveyed; in 1922 and 1923, 4,000. Between 1923 and 1930, the Central Statistics Office surveyed from 8,000 to 20,000 peasant households annually (Istoriya obsledovaniy…). The scope of these surveys reached its peak just before the beginning of the collectivization campaign. In 1928 and 1929, in the USSR, budget surveys covered about 20,000 households (Bokarev).

Peasant households were selected using the cluster method: at first, specialists specified the number of households for each region; then this region was divided into production areas (specializing in crop growing, cattle farming, timber production, fishery, pasture farming and so on). This approach gave statisticians the full picture of the life of peasant households and enabled them to study the impact of macroeconomic factors on family budgets.

Beginning in 1932, a permanent network of households was formed that included around 0.01% of the USSR's population (about 6,500 families). From 1969 to 1987, budget surveys covered 62,000 households (Istoria obsledovaniy…). Since the 1970s, in many parts of the country, including Sverdlovsk region, the surveys started to focus on families of *sovkhoz (state farms)* workers, which replaced *kolkhoz* families.

The data from the budgetary survey was aggregated and was found to contain the following sections: common household data with sizes and composition; working hours; income obtained from work (in a collective farm, at an enterprise or in an institution); turnover of products in a household; expenses on acquisition of industrial goods, transport, housing, household services, taxes and debts; and food structure.

The most complete data is the *primary* forms of budgetary surveys, which were completed for each farm.

In the State Archives of the Sverdlovsk Region, 325 budgets were found for 1928–1929, including several thousand data points, as well as 221 "Control Writing Books of the Statistician" for 1963, which allows for a comparative analysis. Meanwhile, the budgetary inspections allow us to reconstruct the life of a household not only demographically but also in terms of its budget and consumption.

Based on the comparable indicators of 1928–29 and 1963, a dynamic database is being formed to study the evolution of the peasant family in the Urals from the 1920s to 1960s.

The database being created implements a strategy that is oriented towards creating a *research*-oriented database but with the possibility of modifying it for secondary use. The basic structure of the table includes 43 fields, most fully revealing the size, composition, type of family and factors influencing its demographic characteristics. They can be divided into the following blocks:

Demographic, such as the household head's profile (age, sex, nationality, literacy, affiliation with public organizations), data on the family size and structure, the age of household members, the number of minor children and the demographic family type

Economic, such as the year when the household was created; the land, cattle and other livestock; farming practices such as growing grain and vegetable crops and haymaking; crafts; transport and living conditions

Budget-related, such as income and expenditures, consumption patterns and valuable possessions

The database reflects approximately 4% of the information stored in the primary sources. Therefore, in parallel with the database, an e-archive of electronic copies of archival documents is formed. Each record of the database is associated with the corresponding file in the e-archive. Creating an open resource, we are striving to ensure that the user, relying on the e-archive, could make corrections or additions to the database while performing research.

Returning to the requirements for creating an open information resource based on historical mass sources, especially nominative sources, it should be noted that the requirements are fully feasible only when working with sources that include a relatively small set of data. If the source has a more complex structure and contains a large amount of primary information, it is possible to create a database including part of the primary source, with the ability for users to correct and enhance the data.

References

Bokarev, Yu. P. *Byudzhetnye obsledovaniya krest'yanskikh khozyaystv 20-kh godov kak istoricheskiy istochnik (Budget surveys of peasant farms of the 1920s as a historical source)*. Moscow : Nauka, 1981.

Borodkin, L. I. "Digitalizatsiya, vizualizatsiya, reprezentatsiya, analitika?" (Digitalization, visualization, representation, analytics?), *Informatsionnyy byulleten' assotsiatsii 'Istoriya i komp'yuter'* 44 (2015), 3–8.

Digitalizatsia (Digitalization) // Entsiklopedicheskiy slovar' SMI [Electronic resource]. URL: https://smi.academic. https://smi.academic.ru/ 74/%D0% 94%D0%B8%D0% B3%D0%B8%D1%82%D0%B0%D0%BB%D0%B8 %D0 %B7%D0%B0%D1%86% D0%B8% D1%8F (addressed 03.03.2019).

Istoriya obsledovaniy byudzhetov domashnikh khozyaystv (History of household budget surveys) [Electronic resource]. URL: http://stavstat.gks.ru/ wps/wcm/connect/rosstat_ts/stavstat/resources/f233f100415db7128daaef-367ccd0f13/История+обследований+бюджетов+домашних+хозяйств.pdf. (addressed 03.03.2019).

Gorbachev, O. V. "*Istoriko-demograficheskie bazy dannykh v kontekste evropeyskikh komparativnykh issledovaniy: proekt EHPS-NET*" (Historical and demographic databases in the context of European comparative research: the EHPS-NET project), Istoricheskaya informatika. Informatsionnye tekhno-

logii i matematicheskie metody v istoricheskikh issledovaniyakh i obrazovanii. 2014. No. 2-3 (8-9). Pp. 4–9.

Gutnov, D. A. and Pereverten', V. A. "Rossiyskie istoriki XVIII – nachala XX vv.: proekt i informatsionnaya Sistema" (Russian historians of the 18th - early 20th centuries: project and information system). in *Krug idey: novoe v istoricheskoy informatike*. Moscow, 1994, 49–50.

Thaller, M. "Chto takoe 'istochniko-orientirovannaya obrabotka dannykh'; chto takoe 'istoricheskaya informatika'" (What is the "source-oriented data processing"; what is "historical informatics"). in *Istoriya i komp'yuter: novye informatsionnye tekhnologii v istoricheskikh issledovaniyakh i obrazovanii*. Goettingen, 1993, 5–18.

Yumasheva, Yu. Yu. "Istoriografiya prosopografii" (Historiography of prosopography). *Izvestiya Ural'skogo gosudarstvennogo universiteta.* Seriya 2. Gumanitarnye nauki. 10/39/ (2005), 95–127.

About the authors:

Mazur Lyudmila N., Doctor of Historical Sciences and Professor at the Ural Federal University in Ekaterinburg (Russia), the Director of International Demographic Unit. *E-mail:* Lmaz@mail.ru

Gorbachev Oleg V., Doctor of Historical Sciences and Professor at the Ural Federal University in Ekaterinburg (Russia). E-mail: og_06@mail.ru

Резюме
Существует два подхода, обычно применяемых в исследовательской практике для оцифровки первичных источников данных: проблемно-ориентированный и источнико-ориентированный. Первый более экономичный, эффективный и гибкий. Однако в этом случае мы получаем одноразовый ресурс. Второй подход направлен на разработку системы, содержащей полный набор достоверных данных. Таким образом, систему можно будет использовать для различных исследовательских целей, но для ее создания потребуется гораздо больше времени.

В этой статье анализируются два комплекса первичных данных и проблемы, с которыми мы столкнулись в ходе проектирования и создания многоцелевого цифрового ресурса. Первый комплекс включает в себя первичные формы Всероссийских партийных переписей 1922 и 1927 гг. Второй содержит данные семейных бюджетов, собранные в ходе бюджетных обследований 1929 и 1963 гг. Оба источника имеют сложную иерархическую структуру, затрудняющую нормализацию. Записи, сделанные в

разные годы, позволяют проследить динамику для конкретных социальных групп, но для этого необходимо решить сложную задачу организации информации с целью достижения совместимости структур первичных данных. Существует опасность потери первичной информации при преобразовании в единую базу данных.

При создании системы цифровых ресурсов важно разработать правильную стратегию оцифровки, которая может компенсировать потерю данных и создать систему, которая подходит для различных целей и может быть адаптирована к индивидуальным потребностям каждого пользователя.

EXPERIENCES IN THE ANALYSIS OF MANUSCRIPTS AT CIMA

Heinz Miklas

Abstract: Since the 1970s, we have witnessed steady growth of non-invasive analytical methods that can be applied in the field of cultural heritage and are summarized under the term *analytical archaeography*. A wide variety of these methods are used by the Center for Image and Material Analysis in Cultural Heritage (CIMA) in Vienna, whose work currently focuses on ancient manuscripts. In particular, poorly conserved sources and sources with special ornamentation are regularly analyzed by multi-spectral imaging (MSI) and image processing, on the one hand, and by XRF, UV-Vis, rFTIR and Raman spectroscopy, on the other. If possible, DNA analysis of parchment is also performed. After a brief overview, this article reports some experiences using these techniques of computer vision. In the end, the author discusses the problem of open access to valuable source data in connection with their long-term storage. In offering one solution to this problem, he recommends a new type of source publication with the addition of a flash drive with multispectral images and their description, as well as a description of the chemical (and possibly DNA) composition of the monument.

Keywords: CIMA, manuscript analysis, analytic archaeography, computer vision, material analysis

1. Introduction

Scholars of the humanities have always dreamed of getting closer to the natural sciences and advancing towards verifiable results by incorporating scientific methods. From the middle of the nineteenth century – to name only August Schleicher with his *Stammbaum-Theory* and the Neo-Grammarians with their *sound laws* – such attempts had a considerable impact on the progress of linguistic research, advancing the field more than any other school before them. What lagged behind was the development of technical support. Thus, when I began my career as a researcher, the typewriter and the magnifying glass were about the only technical means which a philologist could rely on in his source work. Microfilm had just been developed

based on analogue photography, and a few libraries disposed of ultraviolet lamps for deciphering palimpsests.

Only from the late 1970s did things begin to develop rapidly: first the triumphal march of the personal computer, then the invention of digital photography, and soon the Internet joined in as a revolutionary development.

From then on we have observed two major branches of development: on the one hand the broad river of constantly growing software for description, editing and textual analysis of written heritage, which we subsume under the term *digital humanities*, and on the other a slowly broadening stream of hard- and software innovations for the visualization, conservation and investigation of the supporting material as well as the content handwritten or printed on it. For the latter, a proper name has not been found yet; but since it is urgently needed, I suggest the term *analytic archaeography*[1], meaning an interdisciplinary field in which natural scientists and humanities scholars work together on problems in codicology, epigraphy, numismatics and other areas. The two terms – digital humanities and analytic archaeography – must neither be fused nor confused, as the former concerns the application of the computer as the only instrument, while in the latter the computer is just part of the set of instruments, the majority of which function with radiation – such as cameras, scanners, computer tomographs or spectrometers. Thus, archaeography comprises both archaeoscopy and archaeometry and has in this sense already been used in archaeology (Kleijn 1977; Binford 2001, 205). It is left to discussion whether other methods such as microbiological DNA or stable-isotope analyses should also be included in the term. It might be safe to do so. In any case, all instruments and subareas mentioned above are necessary for contemporary source work, and most of them were discussed at the 2018 El'Manuscript Conference. In this way, written heritage studies are getting more and more complex and can only be reasonably executed either trans- or interdisciplinarily, depending on the subject.

---

[1] The term has already been used in Malina 1977, but in a different sense: "As distinct from archaeological theory, archaeography (p. 4) is defined as the methods for formalized description in the processing of archaeological material. This new term refers to the use of computers to aid in the objective handling of large amounts of archaeological data. Thus, archaeography is simply concerned with the tools necessary for description and classification." (Hodder 2010).

2. The CIMA project

Interdisciplinarity and analytic archaeography are the two keywords which characterize the Centre of Image and Material Analysis in Cultural Heritage in Vienna (CIMA).[2] It was founded in 2014 with a grant from the Austrian Federal Ministry of Science and Research as an inter-university research institution of four institutes from three universities: the Department of Slavonic Studies and the Department of Byzantine and Modern Greek Studies at the University of Vienna, the Computer Vision Lab of the TU Vienna and the Institute of Science and Technology in Art at the Academy of Fine Arts Vienna. For specific tasks we co-operate with colleagues from other Austrian and foreign institutions. For instance, our Internet editions have been handled by our Izhevsk colleagues under the lead of Victor A. Baranov,[3] and in our new Glagolitica project[4] microbiologists from the University of Natural Resources and Life Sciences in Vienna[5] and a restoration-conservation expert from Krems University[6] take part in order to broaden our expertise.

As is often the case, the impetus for interdisciplinarity arose from need. In the late 1990s difficulties in deciphering the new Slavonic finds at St. Catherine's Monastery in the Sinai caused me get in touch with the codicological laboratory of the Russian National Library in St. Petersburg under Denis O. Tsypkin, who had been the first to use multispectral imaging on Slavic material[7]. It was the time when the Archimedes Project started in the USA,[8] and somewhat later Dieter Harlfinger was to launch the European palimpsest project "Rinascimento virtuale."[9] Tsypkin and I made various

---

[2] http://www.cima.or.at, https://hrsm.cvl.tuwien.ac.at/.

[3] See http://mns.udsu.ru/mns/portal.main?p1=55&p_lid=2&p_sid=1 and the articles of V. A. Baranov, R. M. Gnutikov and O. V. Zuga in this volume.

[4] The Austrian Science Foundation (FWF) project no. P29892 (https://cvl.tuwien.ac.at/project/the-origin-of-the-glagolitic-old-church-slavonic-manuscripts/), in the framework of which this article was written.

[5] See https://www.boku.ac.at/en/bioconversion/partner/#c72521 and the presentation of Guadalupe Piñar in *El'Manuscript 2018* (Programme).

[6] See https://www.donau-uni.ac.at/en/department/bauenumwelt/team/11426/index.php and the article by Patricia Engel in this volume.

[7] After the preliminary work of D. Èrastov, cf. Miklas & Rapp 2015, 4, 11.

[8] See Easton, Knox, Christens-Barry 2003, and http://www.archimedespalimpsest.org/.

[9] http://www.rinascimentovirtuale.eu/.

tests and then worked out a concept for the establishment of a technical laboratory for the analysis of written records, which we proposed to the Austrian Academy of Sciences. Even if this plan could not be realized for a lack of funding, it represented the nucleus of long-lasting cooperation in various projects supported by the Austrian Science Fund[10]. The foundation of CIMA was the logical consequence of these endeavours, which in its turn forms the basis for current projects.[11]

When you tackle such a venture, you have to have a vision and you must develop a working philosophy. Our vision was outlined on the 2004 EVA conference (Miklas 2004) as a scenario in which the investigation of written heritage will be:

- done almost exclusively based on images
- enabled or improved by special recording-methods
- more exhaustive, precise, objective and less time consuming via automated image analysis
- executed with a set of tools which can be applied also by humanities scholars after a learning process

Even if our aims and tools have already exceeded these starting limits, our philosophy has not changed, so we still follow the same principles:

(1) non-destructiveness
(2) rapidity and economy of analyses
(3) transportability of the apparatusses
(4) robustness and reliability of the instruments

3. Fields of work

Now, let us turn to practice. I will shortly characterize the work done by the three core groups and then go into some detail when describing our experiences.

Philology. Philologists propose the objects of investigation, supervise the work, discuss solutions for the technical tasks with the scientists and execute the philological analysis of the sources.

---

[10]   https://cvl.tuwien.ac.at/project/the-sinaitic-glagolitic-sacramentary-euchologium-fragments/; https://cvl.tuwien.ac.at/project/the-enigma-of-the-sinaitic-glagolitic-tradition-2/; https://cvl.tuwien.ac.at/project/cima/.
[11]   https://cvl.tuwien.ac.at/project/the-origin-of-the-glagolitic-old-church-slavonic-manuscripts/.

Computer Vision. Here we focus on three major areas – imaging, image processing and automated document analysis. Most important for the examination of manuscripts, which are often preserved in bad condition or contain palimpsests, has been multispectral imaging (MSI) and the enhancement of its results by image processing.

As the name indicates, MSI means taking photographs in a number of narrow spectral bands, which can then be combined in various constellations and further developed. For this purpose a mobile acquisition system, which can be surrounded by a black tent, has been acquired for photographing manuscripts. It comprises six major elements and presently works with 11 bands reaching from ultraviolet (UV) to near infrared (365-940 nm). For the lowest UV illumination, a fluorescence image is acquired in addition to the reflectography image. Contrary to our American EMEL colleagues[12] who are working with more images, but only one camera, we usually apply two – an achromatic PhaseOne for MSI and a normal chromatic or RGB camera – to produce both true-color pictures and color-separated UV-fluorescence images. This is important for three reasons: first, in order to provide a "natural" impression of the object for scholars working with the pictures; second, to ease the production of true-color editions, and third, in order to bridge minor gaps in the chosen MSI range by the wide and uninterrupted range of the RGB camera under broad-band illumination. The spectral range of the two cameras is sufficient for average objects written in metal (such as iron-gall) inks; for manuscripts written with carbon ink or darkened by various influences, infrared (IR) radiation up to about 1100 nm is required. Therefore, in such instances we use a special IR camera with the beautiful name "Osiris."[13]

Most colleagues doing MSI have so far concentrated their efforts on the production of readable images of text. This is only natural, as the text is the heart of the manuscript. But there are many other important items which are usually the topic of research for codicologists and restoration-conservation specialists. In order to describe them properly, so far mainly oblique- and transmissive-light images have been used, on the one hand, and microscopes or microscopic images on the other. Now, after acquiring our 60-megapixel

---

[12] http://emel-library.org/; http://sinaipalimpsests.org/technologies.

[13] From Opus Instruments/U.K., with a spectral response of 900–1700 nm and a resolution of 4096x4096 px.

camera we decided that the most economic means to combine the outcomes of these techniques could lie in the construction of a "light dome" in which this high-resolution camera might deliver 3D images of almost microscopic clarity due to the object being so close to the camera. Of course, this is what we call "a pious wish", but last year our colleague Simon Brenner already developed a first prototype which yields remarkable results, not only for codicology but also for the decipherment of hidden texts.[14] Presently, he is trying to convert it into a fold up-version which can be easily carried, so we can use the dome regularly in our imaging sessions.

Let us now turn to the processing part.[15] First, in order to avoid misplacements and aberrations, the images have to be aligned to each other or "registrated."[16] Then dimensionality-reduction techniques divide different layers or segment parts of interest. Finally, mere visual methods can be added such as pseudo-colouring in order to further enhance the contrast or the like.

Especially in the first project our computer scientists invested much energy in the development of graphemic tools which were combined in a Toolbox for Manuscript Research.[17] Now they are concentrating on subjects such as scribe identification and OCR software.[18] An important preprocessing step for such high-level image processing applications is binarization (Hollaus, Diem, Sablatnig 2018).

Chemical Material Analysis. The determination of the material composition of a manuscript is based on elemental analysis and spectroscopic methods and concerns the support, inks, pigments, dyes and binding media used for the production of written sources. After starting with a self-built X-ray fluorescence (XRF) instrument in the first project (Desnica & Schreiner 2006), CIMA now applies, depending on the object and the situation, two to four complementary spectroscopic methods in order to analyze the chemical elements and compounds: the compulsory X-ray fluorescence analysis, the

---

[14] With a focus on manuscripts see Brenner 2018a; with a focus on epigraphics Brenner, Zambanini, Sablatnig 2018.

[15] For a recent overview see Arsene, Church, Dickinson 2018; for the following see also Miklas, Brenner, Sablatnig 2017, and Knox 2018.

[16] See i.a. Diem & Sablatnig 2008, and Diem, Lettner, Sablatnig 2008.

[17] See i.a. Kleber, Sablatnig et al. 2008, Diem & Sablatnig 2008, Diem, Lettner, Sablatnig 2008, Gau, Vill et al. 2009.

[18] See, e.g., Fiel & Sablatnig 2013, Fiel, Kleber et al. 2017, and Hollaus, Diem, Sablatnig 2014.

Fourier Transform InfraRed spectroscopy in the reflection mode (rFTIR), and, in addition, either Raman or UV-VIS spectroscopy.[19] The Elio XRF device can also be used for mapping, which shows more clearly the elemental distribution in a certain area, but is also more time- and space-consuming.

These analyses serve three coherent aims: first and foremost they lead (paradigmatically) to the detection of the elements and compositions; secondly, by comparing the data they help (syntagmatically) to determine relationships between objects and contribute to their dating and localization; and thirdly, they can explain certain phenomena we observe on the objects, such as parchment degradations or materials applied during conservation-restoration treatments.

In this way, about 90 important objects from 24 sites and 12 traditions have already been partly or fully imaged, of which some 60 have also been chemically examined – not only manuscripts but also epigraphy on stone, leather etc.

4. Experiences in Computer Vision

After this short overview we will treat some interesting cases in connection with MSI and post-processing. There can be no doubt that MSI has widely improved over the past fifteen years and has been of extreme value when badly preserved or palimpsested sources are concerned. Even so, it still shows certain limitations and drawbacks, which have to be surmounted.

In certain cases the application of MSI alone almost yields satisfactory results. This is true of the Cyrillic palimpsest in the Dragotin codex of the Bulgarian National Library, whose images Fabian Hollaus had to enhance only by contrast-stretching and false-color treatment after combining and registrating the RGB with the UV-reflectography image.[20] In other instances, such as the Sinaitic Liturgiarium 5N, several post-processing techniques were needed, some of which were developed by our computer-vision specialists.[21] This gradual improvement took years and led to three major stages

---

[19] See, e.g., Frühmann, Cappa et al. 2018a.

[20] See the description Hollaus, Brenner, Miklas 2018 and the images attached to Hristova-Shomova 2018.

[21] See i.a. Diem & Sablatnig 2008, Lettner, Sablatnig 2009, and Hollaus, Gau, Sablatnig 2013.

of readability. As a consequence, we had to decipher the text three times until we received a final result for most pages![22]

From this we can draw two important conclusions. First, even if the mentioned images were still made with comparably low-resolution cameras,[23] they contain a huge amount of information which presently can only be fully used after applying a series of complicated post-processing techniques. And second, in order to do so, safe long-term storage of the image material is absolutely essential. Many producers of beautiful images reflect too little on this topic, and you come across photographers who store their material in non-adequate formats such as JPEG even though TIFF is the most common used among MSI specialists. But even TIFF will not do because the conversion into this format causes a loss of information, and certain software products can only be applied to the raw data produced in the imaging process! Therefore, the raw data are more valuable than any converted or further developed images. Unfortunately, to the best of my knowledge, only one hardware producer has made attempts to create a common standard for the preservation of their format, namely Adobe with its Digital Negatives (DNG).[24] If others do not take this on as well, this could prove fatal one day!

The most challenging object to photograph were the Glagolitic Medical Folia from the Sinai. When working with them in the library, we did not guess that these three tiny bifolia hide a multi-layer palimpsest! Being under time pressure, we made only two series of color and UV images, which at first sight did not exhibit any palimpsest traces either. Only when we had improved them, blown them up on the screen and turned them in various directions, we gradually noticed traces of three different rulings indicating three under-text layers[25]. When we followed the lines systematically, Dana Hürner and I

---

[22] Cf. Miklas & Rapp 2015, 11 (fig. 2).

[23] I.e., a normal RGB Nikon D2Xs camera (resolution 4288x2848 px) and an achromatic Hamamatsu MSI-camera (resolution 4000x2672 px) with a spectral response of 330–1000 nm.

[24] https://en.wikipedia.org/wiki/Digital_Negative. A sophisticated user system is being developed by a group associated to the EMEL team and called IIIF Image API 3.0 ALPHA DRAFT, see https://iiif.io/api/image/3.0/#table-of-contents: "This document describes an image delivery API defined by the International Image Interoperability Framework (IIIF, pronounced "Triple-Eye-Eff") Consortium. The IIIF Image API specifies a web service that returns an image in response to a standard HTTP or HTTPS request".

[25] See the image in Miklas et al. 2012, 301.

discovered first a whole phrase in Latin, then a few single Glagolitic letters (Miklas et al. 2012, 124-125). Yet, the third underlayer remained invisible.

This last layer we discovered only a few years later on the basis of a new MSI series that our American EMEL colleagues made in the framework of their Sinai Palimpsests project.[26] Surprisingly, our old images and the new ones behaved somehow complementarily to each other: both exhibited two underlayers each, but none of them all three. While both showed the Glagolitic underlayer, instead of the Latin layer,[27] the new images disclosed traces of a Greek or rather Cyrillic undertext!

Furthermore and unexpectedly, I discovered on them remnants of miniatures covering the two pages where we had found the Latin characters. Finally, during an additional look at the old pictures I noticed that I could have discovered the drawings already earlier; what made it impossible was the obstacle that in order to see letters, I had enlarged the images on the computer-screen, which prevented me from discerning large entities!

Beyond this practical conclusion we must ask why the UV series made with a 12 megapixel RGB camera and the highly sophisticated MSI series executed with a 50 MP achromatic camera[28] behaved differently? Several explanations are possible – for instance, a different amount of noise or different lighting conditions. But whatever the reason for this outcome may be, they raise a number of general questions.

The first concerns the number of bands needed to yield the best possible results. The answer to this question has yet to be found. Of course, we have learned that UV and IR bands are absolutely necessary, but the number and choice of bands between these extremes has so far relied on personal impressions rather than on scientific evaluation. Therefore, we have started developing means to combine our MSI results with those deriving from our spectroscopic measurements in order to draw conclusions from their comparison.[29]

---

[26] http://sinaipalimpsests.org/. For the following see Miklas 2018 (the full version is forthcoming).

[27] On very few images of the three series only the first letter *S* can be guessed when comparing them with the old images.

[28] See the description and image under http://sinaipalimpsests.org/technologies.

[29] See his presentation "Towards a unified Database for Multispectral Images and Spectroscopic Material Analysis of Manuscripts", held at the Intern. Congress "Visual Heritage", Vienna, Nov. 11, 2018.

Finally, since the results of punctual spectroscopy might be insufficient, we also apply hyperspectral scanning, which works with a much higher number of bands (Frühmann, Cappa et al. 2018b). But hyperspectral imaging has two major disadvantages: it produces an enormous amount of data which are difficult to store and handle, and the relevant instruments are still very expensive. Affordable types usually start at 400 nm and go up to 800 nm – a spectral range which is too narrow to cover all source requirements.

As is well known, the spectral range is also crucial with MSI cameras insofar as, especially in the IR range, the usual glass lenses cause distortions that blur the imaging result. While this has often been taken as an unavoidable mishap, our colleague Simon Brenner has evaluated the distortions of our new MSI camera and developed a program for their mathematical correction (Brenner 2018b). As a consequence, we now get pictures which are homogenously in focus in all ranges!

Although the number of images to be analyzed can be significantly reduced by post-processing, the current rate of analysis still places a heavy burden on scholars involved in deciphering the pictures. Just imagine you have to compare several series of some thirty different pictures in order to decipher one and the same page! This happened to us especially when working on the Sinaitic Liturgiarium mentioned above.

Processing. Meanwhile, there are numerous ways to "enhance" pictures, i.e., to convert the raw material of the imaging process into the best possible image for reading – automated and manual, supervised and unsupervised, statistic and mere artistic. When asked which method is the best, at least for a given case, some colleagues confess that they just work their way through according to the "trial and error" principle, and few really understand the effect of PCA, ICA, LDA etc.[30] Easily understandable, however, is our LDA method based on ruling analysis (Hollaus, Gau, Sablatnig 2013) because – as a rule, not without exception – where there is ruling there is (or was) also text, as we have just observed. Equally simply comprehended is the effect of false colouring. But in other instances we can be confronted with the obstacle that computer vision sometimes produces good

---

[30] Only recently have scientists started systematically to compare and evaluate the different approaches, see Arsene, Church, Dickinson 2018, 80-96 (ch. 4-5).

practical results while the theory runs behind. Yet, without theory progress will always remain limited.

5. Open access and long-time storage

Let us close with the open-access availability of valuable source files in connection with long-term storage. Recently, we have developed a plan for new manuscript editions and analyses which foresees two major improvements: first, one or more flat memory sticks would be attached to the relevant book cover, in which color and MSI series of the source are stored. This would not only preserve them in the best possible manner but also be of extreme value for further studies. And second, in addition to the printed text or philological study of the source, a description of the images and their production would be given, as well as the results of its chemical and, if possible, microbiological analyses.[31] In this way a full, lasting account of all the specifics of the given source would be provided. A first example of such a new edition has just appeared in Sofia and concerns the early 12th-century Cyrillic Menaion palimpsest in the already mentioned Dragotin Apostolus of the Bulgarian National Library (No. 880) (Hristova-Shomova 2018).

Further experiences related to the above outlined topics were presented during the conference in the workshop of S. Brenner, B. Frühmann, F. Cappa, H. Miklas, M. Schreiner and W. Vetter[32].

References

Arsene, Church, Dickinson 2018: Arsene, C. T. C., Church, St., M. Dickinson. "High Performance Software in Multidimensional Reduction Methods for Image Processing with Application to Ancient Manuscripts." *Manuscript Cultures* 11, 2018, 73–96.

Binford 2001: Binford, L. R. *Constructing Frames of Reference. An Analytical Method for Archaeological Theory Building Using Ethnographic and Environmental Data Sets*. Berkeley, Los Angeles, London: Univ. of California Press.

---

[31] See Hollaus, Brenner, Miklas 2018 on the former, and Frühmann, Vetter, Schreiner 2018 on the latter topic.

[32] Cf. *El'Manuscript 2018* (Programme); see also the contributions of C. Römer et al. and P. Engel in this volume.

Brenner 2018b: Brenner, S. "Error correction in Multispectral Imaging – A blackbox approach." in *R-CHIVE Conference*, Univ. of Rochester (NY), June 8, 2018 (https://r-chive.com/2018-2/).

Brenner 2018a: Brenner, S. "Photometric Stereo for Palimpsest Analysis." in Rapp, C. et al. (eds.): *New Light on Old Manuscripts: Recent Advances in Palimpsest Studies.* Intern. Conf. Vienna, 25–27 April 2018. Abstracts. Vienna: AAS, 15/22.

Brenner, Zambanini, Sablatnig 2018: Brenner S., Zambanini S., R. Sablatnig. "An Investigation of Optimal Light Source Setups for Photometric Stereo Reconstruction of Historical Coins." in Sablatnig, R.; M. Wimmer (eds.). *Eurographics Workshop on Graphics and Cultural Heritage*. Vienna, 203-206 (http://diglib.eg.org/handle/10.2312/gch20181362).

Desnica & Schreiner 2006: Desnica, V., M. Schreiner. "A LabVIEW-controlled portable x-ray fluorescence spectrometer for the analysis of art objects." in *X-Ray Spectrometry* 35, 2006, no. 5, 280–286. (https://doi.org/10.1002/xrs.906).

Diem & Sablatnig 2008: Diem, M., R. Sablatnig. "Registration of Ancient Manuscript Images Using Local Descriptors." in *Proc. of the 14th Intern. Conf. on Virtual Systems and Multi Media Dedicated to Cultural Heritage*, 2008, 188–192.

Diem, Lettner, Sablatnig 2008: Diem M., Lettner M., Sablatnig R., "Registration of Manuscript Images using Rotation Invariant Features." in Pers J. (ed.). *Proc. of Computer Vision Winter Workshop 2008*. Moravske Toplice, Slovenia, 2008, 107–114.

Easton, Knox, Christens-Barry 2003: Easton, R. L., Knox, K. T., W. Christens-Barry. "Multispectral Imaging of the Archimedes Palimpsest." in *32nd Applied Image Pattern Recognition Workshop*, Washington, DC: IEEE Computer Society, 111–118.

El'Manuscript 2018: Baranov, V., Engel, P., Fuchsbauer, J., H. Miklas (eds.). *El'Manuscript 2018. 7th Intern. Conf. on Textual Heritage and Information Technologies, Vienna and Krems, Austria, 14-18 September 2018. Abstracts, Participants, Programme.* Vienna: Univ. of Vienna; Krems: Danube Univ. for Continuing Education, Aug. 2018.

Fiel & Sablatnig 2013: Fiel, S., R. Sablatnig. "Writer Identification and Writer Retrieval using the Fisher Vector on Visual Vocabularies." in *Proc. of 12th Intern. Conf. on Document Analysis and Recognition (ICDAR 2013)*." Washington DC, Aug. 23-29, 2013, 145–149.

Fiel, Kleber et al. 2017: Fiel, S., Kleber, F., Diem, M., Christlein, V., Louloudis, G., Stamatopoulos, N., V. Gatos. "Competition on Historical Document Writer Identification (Historical-WI)." in *14th Intern. Conf. on Document*

Analysis and Recognition (ICDAR), Kyoto, 2017, 1377–1382 (doi: 10.1109/ICDAR.2017.225).

Frühmann, Cappa et al. 2018a: Frühmann, B., Cappa, F., Vetter, W., M. Schreiner. Three Complementary Non-invasive Methods Applied to Historical Manuscripts." – *Manuscript Cultures* 11, 2018, 97–108.

Frühmann, Cappa et al. 2018b: Frühmann, B., Cappa, F., Vetter, W., Schreiner, M., Father Petrus. "Multianalytical Approach for the Analysis of the Codices Millenarius Maior and Millenarius Minor in Kremsmuenster Abbey." *Heritage Science* 6, 2018, 10 (https://doi.org/10.1186/s40494-018-0176-3).

Frühmann, Cappa et al. 2018c: Frühmann, B., Cappa, F., Vetter, W., Schreiner, M., Miklas, H., Rapp, C., Hollaus, F., R. Sablatnig. "Interdisziplinäre Studien mittelalterlicher Handschriften im CIMA-Projekt/Interdisciplinary Studies of Medieval Manuscripts in the CIMA Project. Review." *Restauratorenblätter* 35, 2018, 7–30.

Frühmann, Vetter, Schreiner 2018: Frühmann, B., Vetter, W., M. Schreiner. "Spectroscopic Material Investigations on the Palimpsest-Codex No. 880 of the Bulgarian National Library 'SS. Cyril and Methodius'." in Hristova-Šomova, I.: *Dragotin Minej. Bălgarski răkopis ot načaloto na XII v., zapazen kato palimpsest v Dragotinija apostol (NBKM 880)*. Sofija: Univ. izd. 'Sv. Kliment Ochridski', 2018, 743–751.

Gau, Vill et al. 2009: Gau, M., Vill, M. C., Kleber, F., Diem, M., Miklas, H., R. Sablatnig. "A Toolbox for Manuscript Analysis." in *Proc. of Conf. on Computer Applications and Quantitative Methods in Archaeology (CAA). Making History Interactive*. Williamsburg (VA): Archaeopress, 2009, 86–92.

Hodder 2014: Hodder, I. Review of Malina 1977. *Proc. of the Prehistoric Society*, vol. 4, 2014, 158 (https://doi.org/10.1017/S0079497X00027274.

Hollaus, Brenner, Miklas 2018: Hollaus, F., Brenner, S., H. Miklas. "The Images of the Dragotin Palimpsest Provided by the Centre of Image and Material Analysis in Cultural Heritage." Hristova-Šomova, I.: *Dragotin Minej. Bălgarski răkopis ot načaloto na XII v., zapazen kato palimpsest v Dragotinija apostol (NBKM 880)*. Sofija: Univ. izd. 'Sv. Kliment Ochridski', 2018, 737–742.

Hollaus, Diem, Sablatnig 2014: Hollaus, F., Diem, M., R. Sablatnig. "Improving OCR Accuracy by Applying Enhancement Techniques on Multispectral Images." in *Proc. of the 22nd Intern. Conf. on Pattern Recognition (ICPR 2014)*. IEEE 2014, 3080–3085.

Hollaus, Diem, Sablatnig 2018: Hollaus, F., Diem, M., R. Sablatnig. "MultiSpectral Image Binarization using GMMs." in *16th Intern. Conf. on Frontiers in Handwriting Recognition (ICFHR)*, Aug. 5-8, Niagara Falls, USA, 2018, 570–575 (https://doi.org/10.1109/ICFHR-2018.2018.00105).

Hollaus, Gau, Sablatnig 2013: Hollaus, F., Gau, M., R. Sablatnig. "Enhancement of Multispectral Images of Degraded Documents by Employing Spatial Information." in *12ᵗʰ Intern. Conf. on Document Analysis and Recognition (ICDAR)*, Washington DC 2013, 145–149 (doi:10.1109/ICDAR.2013.36).

Hristova-Shomova 2018: Christova-Šomova, I.: *Dragotin Minej. Bălgarski răkopis ot načaloto na XII v., zapazen kato palimpsest v Dragotinija apostol (NBKM 880).* Naučen redaktor Ch. Miklas. Redaktor na grăckite tekstove Iv. Petrov. Sofija: Univ. izd. 'Sv. Kliment Ochridski', 2018.

Kleber F., Sablatnig R., Gau M., H. Miklas. "Ancient Document Analysis based on Text Line Extraction" – *19ᵗʰ IAPR Inter. Conf. on Pattern Recognition (ICPR'08).*" Tampa (FL), Dec. 2008 (https://ieeexplore.ieee.org/document/4761530).

Klejn 1977: Klejn, L. S. "A Panorama of Theoretical Archaeology." *Current Anthropology* 18, 1977, no. 1, 1–42 (https://www.jstor.org/stable/2741222).

Knox 2018: Knox, K. T. "Image Processing Software for the Recovery of Erased or Damaged Texts." *Manuscript Cultures* 11, 63–72.

Lettner, M., Sablatnig, R. 2009: Lettner & Sablatnig 2009: "Spatial and Spectral Based Segmentation of Text in Multispectral Images of Ancient Documents." in *10ᵗʰ Intern. Conf. on Document Analysis and Recognition (ICDAR).* Barcelona, 2009, 813–817 (http://www.cvc.uab.es/icdar2009/papers/3725a813.pdf)

Malina 1977: Malina, J. *System of Analytical Archaeography* (Studie Archeologického ústavu Českoslov. akad. věd v Brně 2). Praha: Academia, 1977.

Miklas 2004: Miklas, H. "Analysis of Traditional Written Sources with the Aid of Modern Technologies." in *EVA 2004 Moscow Conference*, 29 Nov–04 Dec 2004 (http://conf.evarussia.ru/eva2004/eng/reports/report_211.html).

Miklas 2018: Miklas, H. "'Excavating' the Slavonic Palimpsests in the New Sinaitic Finds." in Rapp, C. et al. (eds.): *New Light on Old Manuscripts: Recent Advances in Palimpsest Studies.* Intern. Conf. Vienna, 25–27 April 2018. Abstracts*.* Vienna: AAS, 2018, 12/22.

Miklas, Brenner, Sablatnig 2017: Miklas H., Brenner S., R. Sablatnig. Multispectral photography for the digital restoration of ancient manuscripts: devices, methods and practical aspects of Historical computer science. *Istoričeskaja informatika* 3, 2017, 116–134 (http://e-notabene.ru/istinf/article_23697.html).

Miklas & Rapp 2015: Miklas, H., C. Rapp. "The Centre of Image and Material Analysis in Cultural Heritage (CIMA) in Vienna and its Present Activities. Part I: Introduction and Philology." *Palaeobulgarica* 39, 2015, no. 1, 3–17.

Vill, Gau et al. 2008: Vill M., Gau M., Miklas H., R. Sablatnig. "Static Stroke Decomposition of Glagolitic Characters" in Sablatnig R., Hemsley J., Kammerer P., Zolda E., J. Stockinger (eds.). *Proc. of 2ⁿᵈ EVA 2008 Vienna Conf., Digital Cultural Heritage – Essential for Tourism.* Vienna (Austria), 2008, 95–102.

About the author

Heinz Miklas is a retired professor of the Department of Slavonic Studies, Vienna University, 1090 Vienna, Austria (Heinz.Miklas@univie.ac.at). He is one of the founders of CIMA (https://cima.or.at/) and currently leads the Austrian Science Fund project "The Origin of the Old Church Slavonic-Glagolitic Manuscripts."

Резюме

С 1970-х годов мы стали свидетелями неуклонного роста неинвазивных аналитических методов, которые могут быть использованы в исследовании культурного наследия и обобщены под термином «аналитическая археография». Некоторые из этих методов используются Центром анализа изображений и материалов в области культурного наследия (CIMA) в Вене, который в настоящее время занимается древними рукописями. В частности, специально плохо сохраняющиеся памятники и памятники с замечательным украшением анализируются, с одной стороны, с помощью мультиспектральной съемки (MSI) и обработки изображений, а с другой – с помощью рентгеново-флуоресцентной ультра-фиолетовой (XRF), Фурьер-трансформационной инфракрасной (rFTIR) и Рамановской спектроскопии. Если возможно, пергамент дополнительно исследуется ДНК. После краткого введеня в статье рассматриваются важные опыты использования мультиспектральной визуализации. В конце автор рассуждает о проблеме открытого доступа к ценным исходным данным в связи с их длительным хранением. Как одно из решений он рекомендует новый тип издания памятников с добавлением съемной памяти (флешки) с мультиспектральными снимками и их описанием, а также описанием химического (и возможно метагеномного) состава памятника.

THE FIVE *W*S OF THE OLD CHURCH SLAVONIC CODEX ZOGRAPHENSIS: RECENT STUDIES, FUTURE TASKS[1]

Yavor Miltenov

Abstract: Ever since the famous Glagolitic Old Church Slavonic Codex Zographensis was discovered some 175 years ago, its features have been subject to comprehensive research. Naturally, they have also become a major source for the reconstruction of Old Church Slavonic grammar and for research into various aspects of Slavonic written culture from the 9th to 11th centuries. However, the critical question remains – whether the accumulation of this large amount of scholarly literature leads to answers to the five classical questions to be posed to any such historical source: who, what, when, where and why? The present article intends to gather the information available in one place and to critically revisit the different opinions. The focus will be on studies conducted over the last thirty years which offer new approaches and new answers.

Keywords: Old Church Slavonic, Codex Zographensis, Glagolitic, scribe, Cyrillic, glosses.

1. It is well-known that the scholarly editions of and the extensive research on early Glagolitic manuscripts in the late 19th and in the beginning of the 20th century have led to a very important breakthrough in the study of the Old Church Slavonic language and the history of written Slavic culture. Is it necessary today, after a hundred and fifty years of abounding scholarly work in the field, to go back to these extensively researched monuments? The answer is undoubtedly yes. This becomes obvious when we review the arguments behind the well-known basic information in textbooks and encyclopedias. Let us take, for example, the Codex Zographensis (Zogr). Based on the same linguistic, orthographic and paleographic features, scholars have expressed three different views about its regional provenance and

three notions as to what is its relative dating. Although some of them are now widely accepted as fundamental while others were set aside, it is clear that the features allowing such a variety of interpretations cannot be called or treated as unambiguous.

Apart from the need to refine the methodology for addressing such issues, there are new, better conditions for new research. Let us remember, for example, that the lower text of the added palimpsest folia in the manuscript has not yet been fully deciphered, and that there has been no detailed analysis of the inks used. In this regard, we can rely on the project *The Origin of the Glagolitic Old Church Slavonic Manuscripts*, supported by the Austrian Science Fund,[2] which aims to study Glagolitic sources employing philological and historical methods, in combination with relevant current technologies, allowing multispectral, spectroscopic and microbiological analyses. Yet another development is the publication of a digital edition of Zogr, freely accessible on the website of the Russian National Library.[3] Until recently, most scholars had to suffice with the print edition of V. Jagić (1879) and with a microfilm copy of the manuscript of poor quality. Thanks to the high-resolution images we can now check the conclusions of previous research, re-examine the paratexts (that were not included in Jagić's edition) and explore in more detail the decoration and other elements which are much more easily observable now.

2. In the framework of the above-mentioned project, this paper aims to reassess some lesser-known facts about Zogr which refer to the five classical questions of any such investigation: who, what, when, where and why. I will try to combine my personal observations with the achievements of studies conducted over the last thirty years, which offer new approaches and new hypotheses.

---

[2] Three interconnected research projects concerning the Glagolitic tradition have been carried out in Vienna under the leadership of Heinz Miklas and funded by the Austrian Science Fund FWF: The Sinaitic Sacramentary (Euchologium) Fragments (2007–2010), The Enigma of the Sinaitic Glagolitic Tradition (2011–2014), The Origin of the Glagolitic Old Church Slavonic Manuscripts (2017–2020).

[3] http://expositions.nlr.ru/ex_manus/Zograph_Gospel/_Project/page_Manuscripts.php?izo=D2D92E28-51F6-4085-B3D7-B2AAB8DA9BDD.

It is well-known that the original manuscript was supplemented with later additions, including 17 palimpsest folia in Glagolitic,[4] a Cyrillic synaxarium and menologion at the end, notes, emendations and instructions in the margins. Here I will focus on the earliest part of the manuscript, while leaving the discussion concerning the evidence of its later history to be considered elsewhere.

3. Zogr is an extensively researched monument, and its characteristics are familiar to anyone working in the relevant field. For that reason I will not venture to repeat its typical catalogue description. Instead, I will present some of these data as a source of information about the creation process of the manuscript.

What did the original manuscript initially include?

First of all, undoubtedly, the text of the four Gospels, each preceded by a table of contents.

In 2009 Zhanna Levshina, on the basis of a first quire-analysis by her late husband Vyacheslav Zagrebin,[5] made public the latter's hypothesis that miniatures of the evangelists introduced every Gospel. Indeed, if we count carefully the lost and present folia in the quaternions which contain the table of contents and the beginning of the Gospels, it appears that there was one leaf each without text (Zagrebin, Levshina 2009). Although it seems obvious that they were cut out in modern times, we do not know whether they have been preserved, and if so, in which private or state collection.

And last, the usual liturgical apparatus, including Eusebian tables, synaxarion and menologion, may have been added at the end. No traces of it survived except the marking of the Eusebian canons and Ammonian sections throughout the manuscript executed by the (main) scribe himself.

Zogr was written on a thin, white, smooth, high-quality parchment, in hanging Glagolitic script, which preserves the old round style of the letters.

---

[4] The palimpsest folia 41–57 substitute lost quaternions of the original manuscript; both the upper and the lower text-layer are written in Glagolitic, and just like the overtext the undertext must have contained Mt 16:19–24:20. It seems plausible that the latter was considered to be of inferior quality and was washed off shortly after it had been written in order to be renewed by a more experienced scribe. This assumption is supported by the assumption that the overtext exhibits more archaic orthographic features than the undertext.

[5] In connection with the preparation of a facsimile edition initiated by H. Miklas, in which V. Zagrebin, L. Moszyński and Z. Hauptová participated.

This means that it derives from a monastery which could afford excellent writing materials and, furthermore, a monastery which was inhabited by monks accustomed to writing and reading Glagolitic texts.

To judge from the relatively small format of the manuscript and its moderate but stylish decoration, Zogr may not have been meant to function as a large, representative codex. Nevertheless, it still had to meet certain standards. This becomes more evident when we compare the original part with the later insertions in which either the handwriting or both the material and the ductus reveal a lack of experience and thus point to a completely different background and lower quality level **(figs. 1-3)**.

Fig. 1. Codex Zographensis, f. 168, main scribe

Fig. 2. Codex Zographensis, f. 182, main scribe and second hand (three or four lines at the bottom)

Fig. 3. Codex Zographensis, f. 54, later Glagolitic addition.

4. The scribe was sufficiently prepared to convey his Glagolitic anti-graph without introducing significant changes to the text. His ductus shows indisputable experience in the use of the Glagolitic script. Nevertheless, his usual writing routine seems to have instead been Cyrillic.

The first evidence which reveals this assumption can be found in his marginal markings of the Ammonian sections and Eusebian canons. L. Mo-szyński (1967) identified a large number of inaccuracies in the numerical designations deriving from the influence of the Cyrillic tradition. According to Moszyński, the Glagolitic antigraph may have featured certain mistakes which were repeated by the scribe of Zogr; however, he himself was defi-nitely responsible for many more.
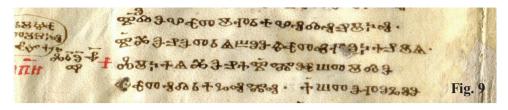
Secondly, the scribe of Zogr added marginal glosses concerning words or phrases in the main text (Miltenov 2019). Most of them are Cyrillic and written with the same ink. There are also three notes in Glagolitic (ff. 189v, 190r, 192v) which reveal the same handwriting (see **fig. 4-11**):

Fig. 4

Fig. 5

Fig. 6

Fig. 7

Fig. 8

Fig. 4-11. Codex Zographensis, marginal glosses and comments (ff. 6v, 70v, 89v, 128r, 132r, 189v, 190r, 283r).

Fig. 9

Fig. 10

Fig. 11

folio	Gloss in the margin:	Word or phrase in the text:
6v	трьхть	ⰡⰓⰀⰑⰎⰕⰏⰅ (конъдратъ)
70v	воеводѣ	ⰃⰎⰡⰅⰂⰓⰀⰐⰅⰏⰅ (ігемономъ)
82v	[.л/м]оужь: [.]ънъиа	ⰀⰟⰁⰕⰘⰕⰟⰁⰀ (ѣвиафарѣ)
89v	кънаꙃь сънъмишта	ⰕⰠⰟⰁⰅⰓⰟⰍⰁⰅⰄⰁⰕ (архисѵнагога)
100r	противьне	ⰈⰅⰏⰅⰓⰅ (сотоно)
108v	[.]лѣпьца слѣпъ	ⰂⰕⰟⰏⰅⰆⰅⰓⰅⰆ ⰈⰅⰄⰓⰅ (вартимеи слѣпъ)
128r	корицꙗ и б(!)[.]сѣкꙗ вони[.]	ⰕⰠⰅⰆⰕⰟⰍⰅⰉ (ароматъи)
131r	начальнꙑи і др[ь]жьвьниу[.]	ⰈⰁⰕⰂⰅⰓⰅⰆ ⰍⰅⰅⰕⰁⰆⰅ (славънъи теофиле) ?
131v	переми[.] авиана	ⰄⰟⰓⰅⰂⰟⰓⰟⰒⰅ ⰟⰠⰁⰄⰟⰞ ⰕⰂⰟⰕⰟⰓⰅ (дьневьнъиꙗ урѣдъи авиана)

folio	Gloss in the margin:	Word or phrase in the text:
132r	есикера	(ꙗ)ⷳⷳⷳ (к(ва)са)
189v	[ⷪ]ⷠⷮ ꙁⷷ[ⷨ]ⷮⷪⷮ ꙁ[ⷨ][ⷦ]ⷷⷪⱈ ([т]ри ѵа[с]ти скъ[л]аза)	(драгъма)
190r	ⷠⷠⷶ[.] ⷮ[.] ⷫⷶ[.] (дрѣ[.] на[.] пл[.])	(рожьць)
192v	[.]ⷪⷶ[.]ⷷ ([.]ти[.]є)	(мамонинꙑ)
269v	[к]овъ	(патѫ)
283r	сѫдиште	(приторъ)

Apparently, the scribe was confident and learned enough not only to be a copyist, but also to take over the role of an editor. The glosses document the scribe's attempt to interpret and/or offer a replacement for unclear Greek loanwords. Were they chosen as result of a comparison with other Gospel copies? Some examples can be explained in this way, but in other cases the glosses seem the product of the scribe's own initiative. After consulting the recent critical editions of the Gospels of Matthew (Alekseev et al. 2005) and John (Alekseev et al. 1998), further studies as well as the Old Slavonic dictionaries, it appears that the rare word скълаѕь is not attested in any Gospel text of Luke 15, 7. The same applies to корица in Mark 16, 1. Кънаѕь сънъмишта instead of архисѵнагога in Mark 5,35 at first sight seems like an East Bulgarian feature; however, it cannot be found in other witnesses. A possible explanation for proposing тръхътъ as a replacement for конъдратъ in the text of Matthew 5, 26 could be that the scribe of Zogr was influenced by the same episode in Lk 12, 59, where we find the only occurrence of this lexeme in the Gospel texts. Some variants also agree with a few later testimonies; e.g., the change of воевода for ꙉемонъ can be testified only in the Miroslav and Karpino gospels, the explanation of пата with ковъ occurs only in the Dobromir gospel and in Theophylact's commentaries on the gospels, and сѫдиште instead of приторъ is not attested prior to the London gospel.

These marginal notes in Cyrillic and Glagolitic are among the earliest examples of editorial work in a written Slavonic source ever. Other early and typologically similar cases have recently been explored by C. M. MacRobert (2014) and are attested in the Sinai Psalter and the Psalter of Demetrius. As in Zogr, alternatives to Slavic translations of Greek words, explanatory or interpretative comments, and Slavic equivalents for Greek loanwords also occur in these manuscripts. Such direct evidence confirms the assumption that one of the main ways of editing in the Old Bulgarian period consisted exactly in the glossing of specific words. Although these medieval scribes nourished the idea of a sacred text, this was no obstacle for those equipped with the relevant schooling, experience and self-esteem to try to improve the text in the margins with their own linguistic suggestions. At one of the following stages of transmission these suggestions could become incorporated into the text; then other scribes, either on their own initiative or under the supervision and discretion of a preceptor, applied the same method of introducing into the copy new glosses or direct replacements in the text.

Fig. 12

Fig. 12-14. Codex Zographensis, initials, ff. 77r, 131r, 225r.

Fig. 13

Fig. 14

5. In addition to being both copyist and editor, the creator of Zogr most probably took care of the decoration as well. In this direction point the same brown-orange, brick-colored ink that is used for writing marginal instructions (e.g., once for the gloss воеводѣ on f. 70v), capital letters and the coloring of initials. The most characteristic feature of Zogr's decoration represent the bows of the geometric initials, shaped as an animal head with an eye and a prolonged muzzle **(figs. 12–14)**. Recently, E. Moussakova (2016) showed that similarly shaped bows can also be found in the Codices Assemanianus and Marianus, the Glagolita Clozianus, the Boyana Palimpsest, and the Euchologium Sinaiticum (for some examples see **figs. 15–29**). These comparative observations
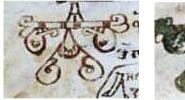

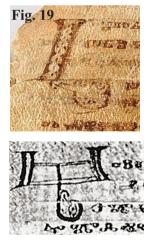**Fig. 15-16. Codex Assemanianus, initials, ff. 121v, 146r.**

Fig. 15

Fig. 16

Fig. 19

Fig. 17

Fig. 18

Fig. 20

Fig. 17-18. Euchologium Sinaiticum, 1/N, initials, ff. 1r, 28v.

Fig. 19-20. Clozianus, initials, ff. 3r (Insbruck), 12v (Trento).

Fig. 21

Fig. 22

Fig. 23

Fig. 24

Fig. 25

Fig. 26
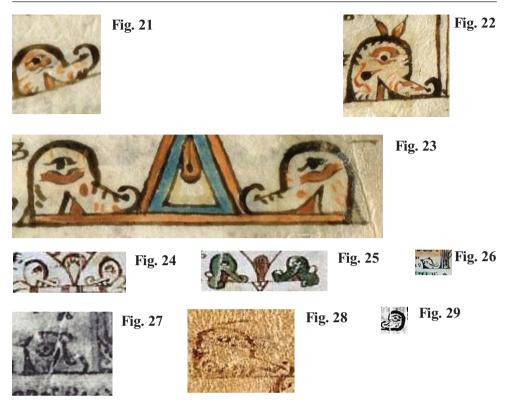

Fig. 27

Fig. 28

Fig. 29

Fig. 21-29. Dragon head element from initials in Zographensis, Assemanianus, Euchologium Sinaiticum, Clozianus.

show that the scribe of Zogr followed a previously established Glagolitic tradition in ornamentation. And this feature places the manuscript together with Glagolitic sources which otherwise could hardly be grouped together.

6. Another interesting though not so popular piece of evidence from Zogr represents its traces of liturgical markings and instructions **(figs. 30–32)**. Here I am not referring to the later Cyrillic apparatus but to the marking of readings inserted in the margins or in the text itself, either by the hand of the scribe or by another hand soon after the completion of the gospel copy (see Moszyński 1985, Pentkovskiy 1998: 27-30). Drawing upon such traces of lectionary instructions A. Pentkovskiy concluded that the Zogr's antigraph might have been a tetraevangelion intended for liturgical use. This hypothesis looks solid, but it has to be verified by a thorough palaeographic study in order to find out which of these notes were inserted by the scribe himself and which were added later.

Fig. 30

Fig. 31

Fig. 32

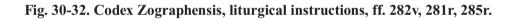**Fig. 30-32. Codex Zographensis, liturgical instructions, ff. 282v, 281r, 285r.**

7. In 2016, the Bulgarian scholar Dimitar Iliev reported that during his work on cataloging the old printed books of the Zograph monastery he came across a handwritten note in Greek,[6] according to which in 1786 the precious Glagolitic manuscript was in the church of the Zograph metochion in the Provlaka region (Iliev 2016). According to a plausible hypothesis the manuscript appeared in the monastery at the end of the 13th century, but it clearly was not created there. According to the Athonite documents, which have been recently edited, reedited and analyzed by Kiril Pavlikianov (Pavlikyanov 2005: 17–30, 188–196, Pavlikyanov 2011: 86–90), until the mid-11th century, and possibly still in the early 12th century, Zograph was an

---

[6] Discovered in the Θεῖον καὶ ἱερὸν εὐαγγέλιον, printed by Nikolaos Glykys in Venice in 1671 and kept in the Zograph monastery library under No. 19023.

average Greek monastery. In the mid-12th century, the Bulgarian presence may have been significant and even dominant, but the first Greek documents in which the monastery was called "the monastery of the Bulgarians", date back to 1276 and 1279 (a *Prostagma* of Emperor Andronikos II and a *Praktikon*).

For these reasons the origin of Zogr remains a subject of debate. The least supported hypothesis claims that the manuscript originated in the Rhodope region. This assumption is based on a single argument, namely the change of the big into the small *yer* preceding a syllable with *yat*. The most widespread and established hypothesis argues that Zogr was copied in "the Ochrid Literary School" or "the area of Macedonia", respectively, or is of "West Bulgarian" provenance (which by and large refers to the same linguo-geographic region). This opinion is again supported by one single phonetic feature: the inconsistent vocalization of the big *yer*, attested in about 50 examples across the entire manuscript. Most of them concern the suffix -ъв- in the noun-forms смокы, црькы, любы, the adverb тъкъмо, the forms of the pronoun къжьдо. Sometimes the vocalization occurs in the endings of masculine nouns before an enclitic demonstrative pronoun.[7]

Over the past 50 years, Bulgarian scholars such as I. Galabov (Galabov 1968: 148, 1980: 62–63), B. Velcheva (Velcheva, Todorov 1993: 16, Velcheva 2012: 15) and T. Slavova (Slavova 1989), have also hypothesized that Zogr might have been copied in Eastern Bulgaria. According to this assumption, first, the aforementioned phonetic features – the mergers of big into small *yer* preceding a syllable with *yat* and of big *yer* into о – were characteristic for the ancient Moesian dialects, too. Second, in the text of

---

[7] Examples of -ъв- from *ū include смоковьницѫ Mk 11:13, Lk 13:6, 21:29, смоковьници Lk 3:7, смоковьницеѫ Lk 1:49, 1:51, смоковьница Mt 24:32, Mk 13:28, црьковь Mt 26:18, 27:40, Mk 11:16, 14:58, Lk 19:45, 21:38, Io 2:19, 8:2, црˊковь Mt 26:61, Lk 1:9, црковь Lk 18:10, црѣковънаѣ Mt 27:51, Lk 23:45, црѣковънъимъ Lk 22:52, любовь Lk 11:42 and Io 13:35. See also the forms of the root скѫдьль- (скѫдоль Lk 5:19, скѫдольницѣ Mk 14:13, Lk 22:10), the adverb тъкъмо (токмо Mt 5:47, 10:42, токˊмо Lk 10:22), and the various forms of къжьдо (кождо Mt 25:15, 26:22, Io 16:32, кожьдо Lk 2:3, 13:15, Io 6:7). Sometimes vocalization occurs in the endings of masculine nouns before an enclitic demonstrative pronoun (родо-сь Mt 11:16, 24:34, Mk 8:12, 13:3, Lk 11:29, образо-сь Mk 12:16, рабо-тъ Lk 12:43, позоро-сь Lk 23:48, праздьнико-сь Io 7:8, народо-сь Io 7:49). Other instances of о < ъ include волиѣ Io 13:5, оусохъшѫ Mk 11:20, трѣхоть Lk 12:59, тоижде Lk 2:8 and ст҃ои Lk 1:72.

Zogr a few secondary readings occur which are considered to be specifical-ly Eastern Bulgarian.[8] Third, Zogr reflects a Glagolitic orthography which can be associated with the East Bulgarian centre Preslav. Its features include the lack of specific letters for /št/ and /dz/[9], the appearance of the (new) differentiation of yotated and non-yotated small *yus*, the new distribution of letters for /i/, the letter *izhitsa*, and the introduction of new diacritical marks.

To summarize, two of the hypotheses about the origin of Zogr are based on a single phonetic feature, while the third uses the same arguments in order to claim a different (speaking in Bulgarian terms: opposite!) place of origin, backing them up with other linguistic (mainly graphemic) evi-dence. The main obstacle to the use of such features to determine the origin of a particular manuscript lies in well-known but often neglected limita-tions – most significantly, a scribe's native dialect and the places where he receives his education and/or where he copies a particular manuscript need not necessarily agree. Furthermore, morphological and lexical features may also stem from an antigraph or protograph composed in a different region. Consequently, the linguistic preferences of the scribe who copied Zogr are most likely reflected in the glosses he left behind, and it is these which need thorough further study.

8. So far, the dating of the earliest Slavic manuscripts has relied on the more or less subjective assessment of scholars – how far the orthographic and linguistic features of the specific manuscript deviate from the "ideal" Old Church Slavonic grammar, and how far its paleography deviates from a presumed ancient ideal, on the one hand, and from the similarly subjective chronology of the development of the relevant changes, on the other. It has to be recognized that such a method is speculative by its very nature (Lunt 1986: 112–113).

The most common dating of Zogr in the scholarly literature is in the area of the 10th–11th centuries. Bold suggestions speak of the 10th century, while more cautious scholars prefer the 11th century. The early dating of this

---

[8] E.g. акъі instead of the original ѣко, великъ instead of велии, дохъторъ instead of възглавница, жидъ instead of июдѣи, кладазь instead of стоуденьць, лоучити са instead of ключити са, лѣвъ instead of шоуи, масло instead of олѣи, оставлениѥ instead of отъпоуштениѥ and отъданиѥ, пастоухъ instead of пастырь, понѣвица instead of плаштаница, съборъ instead of съньмь and цвѣтъ instead of кринъ.

[9] The same applies to the Cyrillic glosses mentioned above.

and other manuscripts derives from a number of newly discovered Bulgarian epigraphic monuments from the 10th century. Their research led a number of scholars to the conclusion that the established chronology of orthographic norms and linguistic processes in 10th-century Bulgaria needs to be re-evaluated, as many features previously considered as Middle Bulgarian turned out to be common and widespread already during the Old Bulgarian period.[10]

Recently, the most significant attempt at improving the principles for dating the earliest Slavic manuscripts has been made by Heinz Miklas and his doctoral student Dana Hürner. In the framework of the above-mentioned Vienna projects they developed a graphemic approach for the spatio-temporal classification of Glagolitic manuscripts.[11] In brief, in order to determine the chronology and the region of origin (redaction) of manuscripts, Miklas and his team use twenty graphemic classes, divided into more than 50 features such as the number and specifics of *yers* and graphemes for nasal vowels, the type of *yery*-, *o*-, *u*-, and *i-graphs*, the rendering of **tj*, presence or absence of the *l*-epentheticum, etc. After collection of the relevant data they are converted into numbers that, according to a chronological and regional (redactional) key, are used to calculate a timeframe and regional distribution (redaction) in which the text may have been written. By this method the calculations for Zogr indicate that the most likely period of its creation lies between 987 and 1006. Compared to the calculations of other early Glagolitic manuscripts this result means that Zogr was copied at nearly the same time as the Rila Folia, Codex Assemanianus, and Clozianus, while the Codex Marianus, for example, is to be dated approximately 30 years later.

9. Recent scholarship has shown that there are still a number of features and aspects of Zogr which ought to be subjected to further research. Established scientific concepts and the basic principles for classifying the earliest manuscripts require reassessment and refinement based on a complex and multidisciplinary approach. The greater accessibility of the sources and the development and use of modern technologies for their study will undoubtedly unearth further important details concerning the Slavic written culture of the 10th–11th centuries.

---

[10] See for example Velcheva, Todorov 1993, Totomanova 2014, and the critical point of view, e.g., in Florya, Turilov, Ivanov 2000: 107–121.

[11] Described in detail in Miklas, Hürner 2015; see also Hürner 2010, Miklas 2017.

References

Alekseev et al. 1998: *Evangelie ot Ioanna v slavyanskoy traditsii*. Izd. podg. A.A. Alekseev, A.A. Pichhadze, M.B. Babitskaya i dr. Sankt-Peterburg, 1998.

Alekseev et al. 2005: *Evangelie ot Matfeya v slavyanskoy traditsii*. Izd. podg. A.A. Alekseev, A.A. Pichhadze, M.B. Babitskaya i dr. Sankt-Peterburg, 2005.

Florya, Turilov, Ivanov 2000: Florya B. N., A. A. Turilov, S. A. Ivanov. *Sud'by kirillo-mefodievskoy traditsii posle Kirilla i Mefodiya*. Sankt-Peterburg, 2000.

Galabov 1968: Galabov, Iv. "Ranni shkoli na stariya balgarski knizhoven ezik." *Balgarski ezik*, 18/2-3 (1968), 141–148.

Galabov 1980: Galabov, Iv. *Starobalgarski ezik s uvod v slavyanskoto ezikoznanie.* Sofia, 1980.

Hürner 2010: Hürner, D. *Paläographie und Graphematik der westkirchenslavisch-glagolitischen Kiever und Prager Blätter im Kontext der sinaitischen Tradition unter besonderer Berücksichtigung der Einträge des Demetrius Sinaiticus.* PhD-diss. Vienna univ. Wien, 2010. (http://othes.univie.ac.at/13004/)

Iliev 2016: Iliev, D. "New Information about the Provenance of the Glagolitic Codex Zographensis from an Unpublished Marginal Note (Zographou Monastery Library Inv. No. 19023)." in *Proceedings of the 23rd International Congress of Byzantine Studies. Belgrade, 22–27 August 2016. Thematic Sessions of Free Communications.* Belgrade, 2016, 24.

Jagić 1879: Jagić, V. *Quattuor evangeliorum codex glagoliticus olim Zographensis nunc Petropolitanus.* Berolinum, 1879 (repr. Graz, 1954).

Lunt 1986: Lunt, H. "On Old Church Slavonic Gospel Texts." *Byzantinobulgarica* 8 (1986), 112–121.

MacRobert 2014: MacRobert, C. M. "On the Headings and Marginal Notes in the two Glagolitic Psalter Manuscripts in St. Catherine's Monastery on Mount Sinai." in *Philology Broad and Deep: In Memoriam Horace Gray Lunt.* Ed. Michael S. Flier, David J. Birnbaum, Cynthia M. Vakareliyska. Bloomington, IN, 2014, 175–185.

Miklas 2017: Miklas, H. "The Graphemic Coordinates of the Old Church Slavonic Glagolitic Horologion Fragment from Sinai." *Palaeobulgarica* 61 (2017), 3, 3–18.

Miklas, Hürner 2015: Miklas, H., D. Hürner. "A Numerical Approach for Dating and Localising Glagolitic-Old Church Slavonic Documents on Graphemic Grounds." in *I. A. Boduen de Kurtene i mirovaya lingvistika. Mezhdunarodnaya konferenciya V Boduenovskie chteniya*. Kazan. feder. universitet, 12-15 okt. 2015 g. *Trudy i materialy.* Kazan', 2015, 14–23.

Miltenov 2019: Miltenov, Ya. "Tri plasta na leksikalno redaktirane v Zografskoto evangelie." *Slověne*. International Journal of Slavic Studies 1 (2019), in print.

Moszyński 1967. Moszyński, L. "Kanony Euzebiusza w głagolskim rękopisie kodeksu Zografskiego." *Slovo* 25-26 (1976), 77–119.

Moszyński 1985. Moszyński, L. "Zografskie i Mariańskie tetraewangelia jako księgi liturgiczne. Część I: Etap Głagolski." *Polata Knigopisnaia* 14–15 (1985), 76–99.

Moussakova 2016. Moussakova, E. "The Dragon/Snake Motif in the Illuminated Old Glagolitic Manuscripts." *Slovo* 66 (2016), 139–163.

Pavlikianov 2011. Pavlikianov C. *The Early Years of the Bulgarian Athonite Monastery of Zographou (980-1279) and its Byzantine Archive.* Sofia, 2011.

Pavlikyanov 2005. Pavlikyanov, K. *Istoriya na balgarskiya svetogorski manastir Zograf ot 980 do 1804 g.* Sofiya, 2005.

Pentkovskiy 1998. Pentkovskiy A. M. "Lektsionarii i chetveroevangeliya v vizantiyskoy i slavyanskoy liturgicheskih traditsiyah." in *Evangelie ot Ioanna v slavyanskoy traditsii.* Sankt-Peterburg, 1998. Prilozhenie 1, 3–54.

Slavova 1989. Slavova T. "Kam lokalizatsiyata na Zografskoto evangelie – starobalgarski pametnik ot X-XI v.." *Palaeobulgarica* 13 (1989), 1, 33–38.

Totomanova 2014. Totomanova, A.-M. "Edinstvoto na balgarskiya knizhoven i neknizhoven ezik prez X–XI v." in Popkonstantinov, K., A.-M. Totomanova. *Epohata na balgarskiya tsar Samuil. Ezik i pismenost*. Sofia, 2014, 79–173.

Velcheva 2012. Velcheva, B. "Suprasalskiyat sbornik i vremeto na tsar Petar." in *Preotkrivane: Suprasalski sbornik, starobalgarski pametnik ot X vek*. Sofia, 2012, 13–16.

Velcheva, Todorov 1993. Velcheva, B., A. Todorov. "Za nova hronologiya na rannite slavyanski nedatirani rakopisi." in *Slavyanska filologiya*, t. 21, Sofia 1993, 14–21.

Zagrebin, Levshina 2009. Zagrebin V. M., Zh. L. Levshina. "O predpolagaemyh miniatyurah Zografskogo evangeliya." in *Hrizograf*. Vyp. 3. *Srednevekovye knizhnye tsentry: Mestnye traditsii i mezhregional'nye svyazi*. Trudy mezhdunarodnoy nauchnoy konferentsii, Moskva, 5–7 sent. 2005 g. Moskva, 2009, 478–491.

About the author

Yavor Miltenov (yavor.miltenov@gmail.com) is Associated Professor at the Institute for Bulgarian Language of the Bulgarian Academy of Sciences. He has published two books (*The Erotapokriseis of Pseudo-Caesarius in*

the Slavonic Manuscript Tradition and *Zlatostruy: An Old Slavonic Homiletic Collection Commissioned by the Bulgarian Tsar Symeon. Text-Critical and Source Study*) and over 40 articles on the history of Medieval Slavic texts and manuscripts and on their linguistic peculiarities and role from a literary and cultural point of view.

Резюме

С тех пор как был открыт знаменитый древнеславянский Codex Zographensis (Зографское евангелие) 175 лет назад, он стал предметом всесторонних исследований. Одновременно он является одним из самых важных источников для реконструкции древнеславянской грамматики и изучения различных аспектов славянской письменной культуры IX-XI вв. Тем не менее, остается вопрос: приводит ли накопление такого большого количества научной литературы к ответам на пять классических вопросов, которые нужно задать к любому важному источнику: кто – что – когда – где – почему? Статья содержит обзор возможных ответов, а также критический пересмотр различных мнений. Основное внимание уделяется исследованиям, проведенным за последние тридцать лет, которые предлагают новые подходы, решения и гипотезы.

LINGUISTIC ANALYTICAL COMMENTS MADE IN THE "MANUSCRIPT" HISTORICAL CORPUS (ON THE APOSTOLOS COLLECTION)

Mariia Novak

Abstract: This paper focuses on analytical comments to erroneous writings in the digital edition of the Tolstovskii Apostolus (14th century) published on the "Manuscript" portal containing Old Slavonic and Old Russian medieval written sources. The codex is characterized by a large number of various errors and peculiar conjectures, which require linguistic commentary embedded into the "OldEd" specialized editor and visualized on the "Manuscript" website. This paper aims to classify errors in the codex and describe their causes. The study demonstrates that there are complicated cases when insignificant graphic substitutions lead to false syntactic coherence or, vice versa, when variations of the grammatical standard seem to be spelling defects. These cases require commentary primarily to prevent misinterpretation. At the lexis level, the comments refer to the incorrect interpretation of the Greek source, conjectures, and exotic inclusions. The author also discusses more trivial cases of dittography, grapheme metathesis, syllable omission, and anticipating the vowel in the next syllable. The evaluation of context correctness is possible owing to the mainstream tradition of the Book of Acts and the epistles, particularly of the Greek versions.

Key words: historical corpus, comments, errors, grammar, lexis

Introduction

Currently, text corpora of medieval written sources are becoming a significant part of the Digital Humanities. Among them, the steadily developing "Manuscript" portal (http://manuscripts.ru/) is quite a remarkable phenomenon. This historical corpus offers an impressive number of analytical tools for any thorough diachronic linguistic research (Baranov et al. 2004; Baranov 2007; Baranov 2010; Baranov 2015). In particular, there is a possibility to mark and comment on misspellings, slips of the pen, and various non-standard features of a text via the "OldEd" specialized editor, as demonstrated in (Baranov et al. 2014).

In 2017, a joint research team from Izhevsk and Kazan (including the author of this paper) completed an internet edition of two Old Russian manuscripts containing the Book of Acts and New Testament epistles: Apostolus Christinopolitanus (a commented text from the 12th century) and Tolstovskii Apostolus (a continuous text from the 14th century). The final results of the project were a comprehensive analysis of orthography, phonetics, grammar, and lexis of the sources, and their marked-up transcription represented on the "Manuscript" website (machine-readable copies of each are available at http://manuscripts.ru/mns/portal.main?p1=61 and http://manuscripts.ru/mns/portal.main?p1=64).

Aim and Methodology

This paper focuses on the Old Russian Tolstovskii Apostolus from the 14th century, which is characterized by a large number of various scribal lapses and peculiar conjectures which require explanations. There are different types of erroneous writings in this old codex: more trivial cases (such as haplography, dittography, and syllable omission) and extraordinary ones, which attract the most research interest. Commenting on them, we extend the basic linguistic research as far as analyzing not only simple graphical cases as mentioned above but also complicated ones when insignificant graphic substitutions may create new text structures.

We aim to demonstrate various types of errors which make an impression of correctness and therefore need linguistic or exegetical comments.

What exactly is to be commented? We do not comment on variations that form the grammatical "image" of the manuscript (for instance, uncoordinated forms of participles). Since they reflect the trends of language development and their penetration into the bookish church language, they do not represent any deviations. On the contrary, unclear contexts or clauses with the illusion of correct interpretation are commented on foremost, for we consider it essential to warn users about possible misinterpretation. The evaluation of context correctness is possible owing to the mainstream tradition of Acts and the epistles, particularly of the Greek versions.

Simpler lapses are also commented on. The most common commentary formula is "так! должно быть / вместо" = "sic! it should be / instead of …" The commenting procedure is available within the "OldEd" specialized editor with further visualization on the website, for instance:
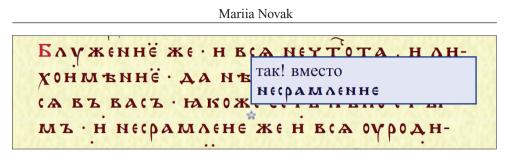
так! вместо
несрамленне

Figure 1. The example of comment visualization, f. 29.1.2[1]

Complicated contexts sometimes require more extensive comments, as we see below while discussing the most intriguing cases.

Facts and Discussion

1. False syntax coherence.

а) болѣ во близъ смр̃ти · нъ б̃ъ помилова · и не того же точью нъ и мен͛ ∴ Да не *печална* прииму (f. 31.1.2; Philippians 2:27) 'he was ill, near to death. But God had mercy on him, and not only on him but on me also, lest I should not get him sad' instead of …да *не печаль на печаль* прииму '…lest I should have sorrow upon sorrow' (cf. in Greek ἵνα **μὴ λύπην ἐπὶ λύπην** σχῶ; the English translation here and below follows the English Standard Version[2]). The second печаль 'sorrow' is omitted in Tolstovskii Apostolus, but this shortened collocation, however, does not seem incoherent. The wider context explains that. The phrase refers to Apostle Paul's disciple Epaphroditus, who had been ill and then got better, for "God had mercy on him" not to aggrieve Paul. Thus, it is entirely possible to relate the form печална 'sorrowful' to Epaphroditus and not to Paul himself.

b) Ктому воды не пии *ни* вина мало испиваи (f. 39.2.2; 1 Timothy 5:23) 'No longer drink only water, neither drink wine, even a little' instead of …*но* вина мало испиваи '…but use a little wine' (cf. in Greek …ἀλλ' οἴνῳ ὀλίγῳ χρῶ).

Here, the substitution of one grapheme gives the opposite meaning to the whole context, which remains syntactically correct. This lapse might be provoked by the vicinity of several и-symbols, both in the previous and following words.

c) There is a similar semantic shift in the following context:

И стго дха раздѣлениѐ по воли ёму своёи *но* англ́мъ бо повину вселеную (f. 44.2.1; Hebrews 2:4-5) 'and by gifts of the Holy Spirit distributed according to his will. But it was to angels that God subjected the world' instead of … *ни* англ́мъ бо повину вселеную '…But it was not to angels…' (cf. in Greek καὶ πνεύματος ἁγίου μερισμοῖς, κατὰ τὴν αὐτοῦ θέλησιν. **Οὐ** γὰρ ἀγγέλοις ὑπέταξεν).

2. False grammatical change.

a) *Иже* бо зависть и ревениѐ · ту нестроёниѐ 'who is jealousy and selfish ambition, there will be disorder' (f. 52.1.2; James 3:16) instead of *Идеже* бо зависть и ревениѐ… 'Where jealousy and selfish ambition exist…' (cf. in Greek Ὅπου γὰρ ζῆλος καὶ ἐριθεία). The form иже occurs owing to a syllable omission. It looks like a masculine pronoun which in standard Old Slavonic grammar could not cohere with the feminine зависть and neuter ревениѐ. However, researchers may have the illusion of observing a new grammatical phenomenon, namely the generalization the gender of pronouns.

b) да ёгоже ради ѻклеветаю въ (f. 54.2.2; 1 Peter 3:16) 'so that, when you are slandered,' cf. in Greek ἵνα, ἐν ᾧ καταλαλοῦσιν ὑμῶν.

This case demonstrates how a variation of the grammatical standard could look like a spelling defect. One can interpret ѻклеветаю as a 1Sg. present form 'I slander,' while it also can be a variant of 3Pl. present form ѻклеветаютъ 'they slander.' The Greek text contains 3Pl. present form (and this is the reason to consider the English passive form "you are slandered" as a failure, though inevitable). The existence of such present forms in Old Russian sources was reliably proved in (Žolobov 2016a; Zholobov 2016b).

3. False lexical coherence.

a) Испроси *веселиѐ* ѻбрѣсти бу иаковлю[3] (f. 67.2.1, Acts 7:46) 'asked to find the mirth for the God of Jacob' instead of Испроси *селениѐ* ѻбрѣсти бу иаковлю 'asked to find a dwelling place for the God of Jacob' (cf. in Greek

---

[3] We use here and below ꙗ as a substitute of the Old Russian "iotated a" because of its absence in the font recommended by the journal.

ἠτήσατο εὐρεῖν **σκήνωμα** τῷ θεῷ Ἰακώβ). This context refers to King David; thus the shift to веселиѥ 'mirth' does not contradict its sense since the biblical tradition attributes to King David the Book of Psalms, where the concept of joy and mirth is essential (Hatch & Redpath 1897-1906).

b) There is a frequent substitution of participle forms with the въш-base by ones with the въш-base, for instance: хⷭъ же пришедъ старѣишина стⷧескъ въшшации благⷣти (f. 47.2.1; Hebrews 9:11) 'Christ appeared as a high priest of the highest good things' instead of *въшшации* благⷣти 'the good things to come' (cf., in the majority of Greek versions, τῶν μελλόντων ἀγαθῶν). The altered context might not seem incoherent to the codex scribe, who probably was able to relate the въш-base with the verb въꙑсити. Its semantics corresponds with the exalted tone of the whole epistle. The form въшшации has something in common both with a participle and a comparative adjective but is neither.

c) всь иже ѿ хⷭъ ѥстьственъ грѣха ѥсть (f. 57.1.2; the chapter-list to John's First Catholic Epistle Catholic Epistle) 'every Christian naturally belongs to the sin' instead of всь иже ѿ хⷭъ ѥсть свенъ грѣха ѥсть 'every Christian is beyond sin' (cf. in Greek ἐκτός ἁμαρτίας (Migne 1864: 685). It is possible that the codex scribe did not comprehend the bookish word свенъ 'outside, beyond' and made a conjecture by inserting an extra letter. This step also changed the meaning to its opposite.

d) *Старица* избранъ господꙑни (f. 59.2.1; 2 John 1) 'The old lady to the elect lady' instead of *Старьць* избранъ господꙑни 'The elder to the elect lady' (cf. in Greek Ὁ πρεσβύτερος). Possibly, it is not a random but a deliberate substitution which reflects a peculiar contextual interpretation since it is not a simple lapse altering one grapheme.

4. Exotic lexical inclusions.

The codex we discuss also contains a certain amount of transliterated Greek inclusions, which also receive comments. They represent various parts of speech.

a) *лифинъ* забꙑтиѥ вземъ (f. 56.1.1; 2 Peter 1:9) 'having forgotten' (cf. in Greek λήθην λαβὼν, literally 'got oblivious'). The accusative form λήθην is translated as забꙑтиѥ, but the transliteration лифинъ precedes the translation.

b) смирениѥ оумоу *комвосфъ* оукрасите (f. 55.2.1; 1 Peter 5:5) 'Clothe yourselves with humility' (cf. in Greek τὴν ταπεινοφροσύνην ἐγκομβώσασθε).

Here, the incomplete (with several letters omitted) transliteration precedes the translation again. The Greek aorist imperative form ἐγκομβώσασθε undergoing the transliteration presents the verb ἐγκομβόομαι which means 'clothe oneself, get dressed.' It is absent in the Septuagint (Schmoller 1994: 140); its Slavonic interpretation as оукрасите 'decorate, adorn' is also unconventional and reflects the idea of moral and aesthetic values unified, so it is characteristic of a holistic Christian worldview.

c) и кли ѳнонусъ (f. 54.1.1; 1 Peter 2:1) 'and envy' (cf. in Greek καὶ φθόνους). Firstly, there is a similarity of the Cyrillic transliteration and Slavonic translation (и кли : καὶ 'and') that is also similar to the previous cases. Secondly, we observe the form ѳнонусъ which accurately transliterates the Greek accusative φθόνους 'envy,' but no translation follows it. All these facts point to a particular contact with Greek sources which might have taken place at the level of the text prototype, with the results entering later codices.

5. Simple lapses.

Here we refer to cases of dittography (ꙗвлаёмоёмоё, f. 29.2.1, instead of ꙗвлаёмоё), grapheme metathesis (алодокиꙗ, f. 34.1.1, instead of лаодикиꙗ), syllable omission (мелхисекову, f. 46.2.1, instead of мелхиседекову; ницьца, f. 41.2.1, instead of женицьца), and anticipating the vowel in the next syllable (ӧбѣщинии, f. 49.1.1, instead of ӧбѣщании). They all receive appropriate comments.

6. Complicated lapses.

Some lapses have a less transparent character, for instance:
имуще пища и ӧдежа се нꙑ гнѣётъ (f. 40.1.1, 1 Timothy 6:8) instead of … имуще пища и ӧдежа се нꙑ годъ ёстъ 'if we have food and clothing, with these we will be content' (cf. in Greek ἔχοντες δὲ διατροφὰς καὶ σκεπάσματα τούτοις ἀρκεσθησόμεθα; the latter form presents the verb ἀρκέω, which in passive voice means 'to be content'). Such an unobvious deformation might occur step by step in different predecessors of the Tolstovskii codex.

Conclusions

In this paper, we focused on different types of erroneous writings existing in the Tolstovskii Apostolus from the 14th century. Its machine-readable representation on the "Manuscript" portal allows the insertion of an analytical commentary through the "OldEd" editor tools. We aimed to de-

fine which phenomena are worth such commenting. The study led to the following conclusion: there are contexts which create the illusion of syntactical correctness and coherency. They require commenting first to prevent misinterpreting and a false impression of grammatical changes taking place in the codex. The Greek text gives the basis for the verification of Slavonic contexts. Besides, we demonstrated more trivial cases which also needed linguistic commentaries. It is difficult to confirm whether discussed errors occurred directly in the Tolstovskii Apostolus or in any of its prototypes (protograph or antigraph).

References

Baranov et al. 2004: Baranov V., Votintsev A., Gnutikov R., Mironov A., Oshchepkov S., Romanenko V. "Old Slavic Manuscript Heritage: Electronic Publications and Full-Text Databases." In Hemsley, J. (ed.): *EVA 2004 London (Electronic Imaging, the Visual Arts Conference & Beyond). Conference Proceedings*. London, University College London; Institute of Archaeology, 2004, 11.1-11.8.

Baranov 2007: Baranov, V. "The Ideology and Technology of Creating Online Full-Text Digital Collections of Ancient and Medieval Slavonic Manuscripts." In: *International Conference on Applied Natural Sciences*. Trnava, 2007, 199-207.

Baranov 2010: Baranov, V. "Machine-Readable Linguistic Internet Resources as a Basis for Historical-Philological Studies." *Journal of Applied Mathematics, Statistics and Informatics* 6(2), 63–79.

Baranov et al. 2014: Baranov V., Anikina R., Gnutikov R. "A Model of Transcription Markup for Manuscripts in the "Manuscript" system: Haplography, Dittography, Inserts, Omissions and Comments." In Baranov, V., Zheliazkova, V., Lavrentiev, A. (eds.): *Pismenoto nasledstvo i informatsionnite tehnologii: materiali ot V mezhdunarodna nauch. konf. (Varna, 15–20 septemvri 2014 g.)*. Sofiia, Izhevsk, 2014, 155–161.

Baranov 2015: Baranov, V. "Istoricheskii korpus kak tsel' i instrument korpusnoi paleoslavistiki." *Scripta & e-Scripta: The Journal of Interdisciplinary Medieval Studies* 14-15 (2015), 39-62.

Hatch & Redpath 1897-1906: Hatch, E., Redpath H. *A Concordance to the Septuagint and Other Greek Versions of the Old Testament (including the Apocryphal Books) in Three Volumes*. Oxford: Clarendon Press, 1897–1906.

Migne 1864: Migne, J.-P. *Patrologiae cursus completus. Series graeca*. T. 85. Paris, 1864.

Schmoller 1994: Schmoller, A. *Handkonkordanz zum griechischen Neuen Testament. 8. Auflage*. Stuttgart: Deutsche Bibelgesellschaft, 1994.

Zholobov 2016a: Zholobov, O. *Drevnerusskii glagol: k istorii prezensa*. Kazan', 2016.

Žolobov 2016b: Žolobov, O. "Present Tense Forms Variability in the Paroemiarion Zacharianum d. 1271 (to the Parchment Internet-Edition)" *Zeitschrift für Slawistik* 61/2 (2016), 305–321.

About the author:

Mariia Novak, Doctor hab. of Philology, V.V. Vinogradov Russian Language Institute of the Russian Academy of Sciences, mariaonovak@gmail.com.

Резюме

В статье обсуждаются виды лингвистических комментариев к электронной публикации Толстовского списка Апостола XIV в., осуществленной объединенным исследовательским коллективом из Казани и Ижевска в 2015-2017 гг. (http://manuscripts.ru/mns/portal.main?p1=64). Данный список Апостола характеризуется большим количеством разнородных ошибок и своеобразных конъектур, которые часто нуждаются в пояснениях. Специализированный редактор "OldEd", в котором осуществляется создание машиночитаемых копий древнеславянских источников, дает возможность встраивать в текст комментарии к различного рода графическим неточностям, ошибкам писца либо намеренным изменениям текста, с последующей их визуализацией на сайте. Цель данной статьи – выявить феномены Толстовского Апостола, нуждающиеся в комментариях, и определить степень их важности. Комментирование Толстовского списка затрагивает как простые графические случаи (диттография, метатеза графем, предвосхищение графемы в соседнем слоге, пропуск слога), так и более сложные ситуации, когда незначительные графические замены могут создавать впечатление ложной связности текста на уровне синтаксиса либо, наоборот, варианты грамматической нормы представляются графическими дефектами. Например: воды не пии *ни* вина мало испиваи (л. 39.2.2; 1-е послание Тимофею, 5:23) вместо …*но* вина мало испиваи; И стго дха раздѣлениѐ по воли ѐму своѐи *но* англмъ бо повину вселеную (л. 44.2.1; посланием Евреям, 2:4-5) вместо …*ни* англмъ бо повину вселеную; *Иже* бо зависть и ревениѐ · ту нестроѐниѐ (л. 52.1.2; послание Иакова, 3:16) вместо *Идеже* бо зависть и ревениѐ; да ѐгоже ради Ѡклеветаю въ (л. 54.2.2; 1-е послание Петра, 3:16 – в последнем случае форма

ѻклеветаю вполне может быть "нулевым" вариантом формы 3 лица множественного числа презенса). На уровне лексики комментируются случаи ложной связности текста, сочетания перевода и транслитерации, неверной интерпретации греческого первоисточника, конъектуры, например: Испроси веселиё ѻбрѣсти бу иаковлю (л. 67.2.1, Деяния апостолов, 7:46) вместо Испроси селение ѻбрѣсти бу иаковлю; всь иже ѻ хѣ ёстьственъ грѣха ёсть (л. 57.1.2; предисловие к первому посланию Иоанна) вместо всь иже ѻ хѣ ёсть свенъ грѣха ёсть; Старица избранѣ господъини (л. 59.2.1; 2-е послание Иоанна, 1) вместо Старьць избранѣ господъини (в двух последних случаях возможна целенаправленная конъектура); лифинъ забъітиё вземъ (л. 56.1.1; 2-е послание Петра, 1:9 – в этом случае перевод сочетается с транслитерацией).

Случаи грамматической и лексической ложной связности в первую очередь нуждаются в лингвистических комментариях, поскольку последние позволяют предотвратить неверное понимание контекста либо ложное впечатление об отражении грамматической ситуации в рукописи, которое может возникнуть у исследователей – пользователей исторического корпуса "Манускрипт". Основанием для верификации чтений в Толстовском списке является греческий первоисточник.

Также в статье приводятся более простые случаи диттографии, предвосхищения графемы в соседнем слоге, метатезы графем, пропуска слога: иавлаёмоёмоё, л. 29.2.1, вместо иавлаёмоё, алодокина, л. 34.1.1, вместо лаодикина, мелхисекову, л. 46.2.1, вместо мелхиседекову, ницьца, л. 41.2.1, вместо жениьцьца, ѻбѣщинии, л. 49.1.1, вместо ѻбѣщании. Эти и подобные неточности, выявленные в Толстовском Апостоле, также снабжаются соответствующими комментариями в системе "Манускрипт".

Вопрос о происхождении комментируемых ошибок и конъектур выходит за рамки статьи, поскольку имеет текстологический характер. Обсуждаемые явления могли возникнуть как в самой рукописи XIV века, так и в ее антиграфе либо протографе.

ЦИФРОВОЙ ИССЛЕДОВАТЕЛЬСКИЙ КОМПЛЕКС РУКОПИСНЫХ И СТАРОПЕЧАТНЫХ КНИГ[1]

Починская Ирина Викторовна, Палкин Александр Сергеевич

Аннотация. В статье речь идет о цифровом исследовательском комплексе рукописных и старопечатных книг, созданном Лабораторией археографических исследований Уральского федерального университета с целью облегчения доступа специалистов к ее фонду памятников кириллической письменности. Этот фонд сложился в результате экспедиций по местам компактного проживания старообрядцев, которые не только сохранили до наших дней средневековую православную книжность, но и на ее базе создали широкий пласт демократической литературы. Книжный фонд ЛАИ представляет собой ценнейший корпус исторических источников. Цифровой исследовательский комплекс рукописных и старопечатных книг лаборатории позволит ученым работать с фондом разными способами, выбор которых зависит от цели исследователя. Доступ к нему осуществляется через сайт лаборатории в разделе «Древлехранилище». Комплекс состоит из трех частей. Первая часть – это описи фондов, которые представлены двумя подразделами: общего списка 27 территориальных собраний фонда и электронного поиска по фонду. Вторая часть – база данных, которая состоит из полных научных описаний старопечатных книг и рукописей. Схема описания была разработана коллективом лаборатории. Для доступа к описаниям используется веб-интерфейс. Третья часть комплекса – полнотекстовые цифровые копии старопечатных книг и рукописей. В качестве хостинг-площадки для публикации цифровых копий был выбран Яндекс. Диск. В настоящее время на сайт выставлены описи фондов и 68 копий памятников письменности. В завершающей стадии находится работа над электронной базой данных, которая будет завершена к концу 2019 г.

Keywords: digital research complex, database, Cyrillic books, manuscripts, old printed books.

Цифровой исследовательский комплекс рукописных и старопечатных книг создан Лабораторией археографических исследований Уральского федерального университета (Екатеринбург) с целью облегчения доступа специалистов к ее фонду памятников кириллической письменности.

Лаборатория существует 45 лет, она была создана для организации и проведения археографических экспедиций в уральском регионе, для хранения и изучения полученных в экспедициях памятников древнерусской книжно-рукописной традиции. Средой бытования этой традиции в Новое и Новейшее время стало русское старообрядчество – религиозно-политическое движение, возникшее во второй половине XVII в. Археографические экспедиции ЛАИ проходили по местам компактного проживания старообрядцев.

Сохраняя основы традиционной книжной культуры, старообрядчество дало новый импульс развитию демократической литературы и сформировало огромнейший комплекс не только сохраненных ими памятников средневековья, но и оригинальных сочинений, отразивших как общероссийские проблемы существования самого масштабного оппозиционного движения России, так и его региональные особенности.

Фонд ЛАИ УрФУ включает в себя литургическую литературу, многочисленные сборники самого разного характера, образцы позднего русского летописания, памятники историко-географического и естественнонаучного характера, оригинальные полемические, исторические, эсхатологические сочинения старообрядцев, певческие рукописи, сборники духовных стихов. Коллекция старопечатных книг дает возможность проследить историю развития восточно-русского кириллического книгопечатания от его истоков до начала XX в., а также отражает основные этапы развития западно-русского книгопечатания. Многие книги имеют автографы видных деятелей российской истории и культуры. В составе фонда ЛАИ УрФУ, наряду с памятниками литературы, содержатся архивные документы по истории старообрядчества (личные архивы) (Починская 1999).

Книжный фонд лаборатории состоит из 27 территориальных коллекций. Его особенность заключается в том, что, будучи сформированным в последней четверти XX в., он отражает, во-первых, современное состояние древнерусской книжно-рукописной традиции огромного региона, распростершегося от Северного Урала до Казахстана, от сама-

ро-саратовского Поволжья до Тюмени, во-вторых, позволяет проследить исторические процессы, протекавшие в социально-политической и духовной жизни не только данной территории, но и России в целом. Он представляет собой ценнейший корпус исторических источников по проблемам конфликтов на религиозной почве и поисков путей их решения. Фонд дает возможность значительно расширить представления о взаимоотношениях между государством, официальной церковью и оппозиционным им старообрядчеством, а также внутри старообрядчества. С использованием фонда рукописных и старопечатных книг ЛАИ написана не одна сотня исследовательских работ и защищен ряд диссертаций, но его научный потенциал не исчерпан.

Сегодня в ЛАИ работает 11 человек, которые занимаются научно-исследовательской работой, основной источниковой базой которой является свой книжный фонд, обеспечивают сохранность и введение в научный оборот этого фонда. Одним из направлений деятельности коллектива ЛАИ в настоящее время стало решение задачи создания удобной цифровой среды для обеспечения удаленного доступа исследователей к памятникам письменности.

Лаборатория археографических исследований принадлежит к числу книгохранилищ, которые первыми в России еще в 1980-х гг. начали работу над созданием электронных баз данных памятников письменности. Созданная нами база данных получила название «Рукописные и старопечатные книги Урала» (РИСК) и была предназначена для внутреннего пользования. После многолетнего апробирования возникла мысль о размещении базы в Internet на сайте лаборатории. Для этого была создана ее новая версия. Также было принято решение объединит базу РИСК с описями фонда, сопровожденными поисковой системой, и полнотекстовыми электронными копиями книг. Таким образом сложился цифровой исследовательский комплекс рукописных и старопечатных книг лаборатории, позволяющий ученым работать с фондом разными способами, выбор которых зависит от цели исследователя. Доступ к цифровому исследовательскому комплексу ЛАИ осуществляется через сайт лаборатории (lai-urgi.urfu.ru) в разделе «Древлехранилище».

Первая часть комплекса. Описи фондов представлены двумя подразделами в разделе «Древлехранилище». В подразделе «Список территориальных собраний» можно получить общее представление о

каждом из 27 территориальных собраний, ознакомиться с кратким описанием коллекции. Второй способ работы с описями – это электронный поиск по фонду, который позволяет искать интересующую информацию из описи отдельно взятого территориального собрания или по всему книжному собранию в целом. По умолчанию поиск работает в режиме совпадения фрагмента текста. Так, по запросу *старо* поиск выдаст результаты со словами «старообрядец», «старовер» и их словоформы. Если нужно сузить результаты поисковой выдачи, запрос необходимо делать в кавычках: так, по запросу *«старообрядец»* поиск выдаст результаты, где есть это слово, проигнорировав однокоренные слова и словоформы. Особенно актуально использование кавычек при поиске материалов, датированных определенным веком. По запросу *XVI* будут выданы результаты, включающие XVI, XVII, XVIII, в то время как по запросу *«XVI»* результаты будут относиться только к XVI столетию.

Работа с описями позволяет сделать общий обзор фонда: понять географическое и конфессиональное происхождение памятников письменности, определить их специфику (иллюстрированные, музыкальные и т. д.), выделить тематические группы. Такой вариант работы удобен для первоначального ознакомления с фондами или для быстрого поиска материалов, относящиеся к одному виду книжных памятников (например, псалтыри).

Вторая часть комплекса. База данных РИСК состоит из полных научных описаний старопечатных книг и рукописей[2]. Схема описания была разработана коллективом лаборатории (Починская 2010). Для доступа к описаниям используется веб-интерфейс, который в данный момент находится в финальной стадии разработки и проходит внутреннее тестирование, его открытие планируется в 2019 г.

База данных РИСК состоит из основного массива информации, с которым связаны вспомогательные («Орнаментика печатных книг», «Орнаментика рукописей», «Шрифты», «Указатель имен в записях») (рис. 1). Такая структура позволяет находить элементы одного типа и связывать их с записями в основном информационном массиве (см. рис. 2). На рисунке показано, например, что шрифт 1 присутствует в книге A и B, а

---

[2] База данных «Рукописные и старопечатные книги Урала» получила свидетельство о государственной регистрации Федеральной службы по интеллектуальной собственности Российской Федерации № 2018620531.

шрифт 2 только в книге В. Поиск работает и в другую сторону, когда можно найти все книги, напечатанные определенным шрифтом.

Рис. 1. Схема устройства базы данных «Рукописные и старопечатные книги Урала».
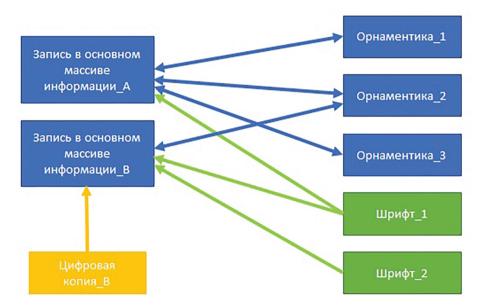

Рис. 2. Связи элементов в базе данных «Рукописные и старопечатные книги Урала».

Большое внимание мы уделяем удобству представления результатов поиска, которые должны иметь привычный исследователю вид археографического описания, а не таблицы (или таблиц), восприятие которой затруднено вследствие большого количества вспомогательных и служебных полей, необходимых для машинного поиска. Ниже приведены примеры фрагментов основного массива информации, экспортированные в Excel (рис. 3 и 4), и пример вывода одного из результатов поиска (рис. 5).

№	Deposit_city	Deposit_Place	Collection	Col_Number	Invent_Number	Prtint_manuscript	Author	Title	Place_creation	Typograp by_scriptorium	Publisher	Printer_scriptor	Date	Century_begin	Century_end	White_date	Year_begin	Month_begin	Day_begin	Year_end	Month_end	Day_end	Format	Tech_parameters
726	Екатеринб ург	ЛАИ УрФУ	V (Курганск ое собр.)	3п	2292	печать	0	Сборник	Почаев	0	0	0	1787 (7295)	18	18	0	1787	0	0	1787	0	0	2	формат: 2, 330 х 210 8.6 А.6 – Т.6 [У].6 Ф.6 – V.6 АА.6 – VV.6 ААА.6 – VVV.6 АААА.6 – СХСХСХС).6 ЩЩЩЩ.6+1 = л.: [1–6, 1–17], 18–23, [24–54], 55–56, [57–111], 112–113, [114–117], 118–658, 657, 660–705, 7052, 706–738, 737, 740–789, 800–881 = 887 л., А–V = 40 букв, А–Ц = 25 букв формат полоса набора 234 х 127 высота 10 строк: 86 строк на с. 25 колонтитулы печать в две краски перепечатано с изд.: М., 29.06.1647(01.03–29.06.7155). Зернова, IV, 200.
830	Екатеринб ург	ЛАИ УрФУ	V (Курганск ое собр.)	9п	2455	печать	0	Службы и житие Николая Чудотвор ца	[Москва]	[Печатный двор]	0	0	[20.04.164 1 (01.04–20 .04.7149)]	17	17	0	1641	4	20	1641	4	20	4	формат: 4, 192 х 132 [1] 41.8 – 30.8 31.4 = л.: [1 пуст., 1–3, 1 пуст., 1–4], 5–38, [39, 40], 41–55, 57–64, 64–235 [236–243] = 243 л. формат полоса набора 150 х 95 высота 10 строк: 89/90 строк на с. 15 печать в две краски

Рис. 3. Фрагмент базы данных, экспортированной в формат Excel

Décor	Illustrated	Contents	Preservation	Notes_Stamps	Bibliography	NB	Year_recieve	Region_recieve	Description_Author	Description_year
Украшения: заставок 3 с 3 досок на л.: 1 втор. сч. (Поч., 239), 616 (Зерн., 253), 875 (Поч., 240); концовок 8 с 6 досок на л.: 6 об. перв. сч. (Поч., 259), 392 (Поч., 266), 398 об. (Поч., 255), 447 (Поч., 257), 452 (Поч., 257), 536 об. (Поч., 256), 550 об. (Поч., 269), 871 (Поч., 258); инициалов 7 с 5 досок на л.: 1 втор. сч (Поч., 280), 6 (Поч., 290), 7 об. (Поч., 279), 9 об (Поч., 286), 18 (Поч., 286), 30 об. (Поч., 290), 875 (Поч., 286); рамки на полях: 1 втор. сч (Поч., 274), 616 (Поч., 272).	нет	Состав: [Оглавление, л. 1–6 об. перв. сч.]; слова отцов церкви, л. 1 втор. сч–874 об.; послесловие с выходными данными: «А богодохновенная книга напечатана с переводу во всем согласно ни в чем разньствует, напечатана же в лето 7295 в начале книги 2 ф. л.»	Сохранность: сохранившиеся л. 18–23, 55, 56, 112, 113, 118–881. Переплет: доски, обтянутые кожей с тиснением, две застежки; в типографии Почаевской в славу и хвалу Вседержителю Богу и всем святым, л. 875–881 об.	Записи: скорописью XX в. шариковой ручкой, «Сия книга Моисеева Иосифа Федоровича благословлена с Котовым Харитоном Филипповичем в 1953 году».	Вознесенский, 316; Починская, 68.	0	1983	Курганская	Починская И.В.	2017
Украшения: заставок 4 с 3 досок на л.: [1 втор.] сч. (Зерн., 352), 40 (Зерн., 352), 642 (Зерн., 271), [242 (Зерн., 257)]	нет	Состав: [оглавление, л. [1–3]; служба Николе, л. 1–4] 5–641; «Житие и жизнь и отчасти сказание о чюдесех иже во святых отца нашего Николы, архиепископа мирликийских чюдотворца», л. 642–241 об.; [послесловие с выходными сведениями, л. 242–243 об.]	Сохранность: сохранившиеся л.: 5–38, 41–64, 64–231, 233, 234, 235, л.: 1–4, 39, 232, 237–239 – рукописные на бумаге втор. пол. XVIII–нач. XIX в., почерки те же. Переплет: доски, обтянутые кожей с тиснением; две застежки, одна утрачена; корешок реставрирован кожей.	Записи: л. 5, скорописью кон. XVIII в. 1 почерком: «Колуга град»; л. 5 об., тем же почерком: «Город Орел»; л. 13–14, скорописью XVIII в. 2 почерка: Синьбирскою; л. 6, тем же почерком: «Город Орел»; л. 37, скорописью XVIII в. 3 почерком: «Поликанперцист Фед...»; л. 38 об. тем же почерком: «Туда раекой пшеничной муки сто пуд рука моа 1748»; л. 50, 51, образвки записей тем же почерком; л. 61, скорописью кон. XVIII в. 1 почерком служебного характера; л. 63, скорописью XVIII в. 3 почерком, типа проба пера: «Каи ...» (повторяется 3 раза); л. 641 об., скорописью XVIII в. 3 почерком, полустав, полустертые записи; л. 642–69, скорописью XVIII в. 4 почерка: «Сия глаголемая книга (?)»; л. 79, скорописью XVIII в. 3 почерком: «Кашцеляристу (повторяется)»; л. 81, 83, 84, тем же почерком: «Каи ... (неразборчиво)» дьяк вологодский дьячек березовский; л. 153, 154, 170–172, проба пера, л. 201, 213 об.–214, 235, скорописью XVIII в. 1 почерком служебного характера, л. 236, скорописью XVIII в. 5 почерка: «...Книга глаголемая [житие] Николы чюдотворца»; внутр. сторона нижней крышки переплета, скорописью XVIII в. 6 почерком (?): «Сия книга тобольского мещанина Афанасья Макарова продана Бронниковской волости деревни Зориапшовой Матфею сыну Зориапшову»; ниже, скорописью XVIII в. 7 почерком: «Читал деревни Абрамовой Иван Игнатьев и прочитал до конца его чудеса и ж[итие] святителя Николы».	Зернова, 155.	0	1988	Курганская	Починская И.В.	2017

Рис. 4. Фрагмент базы данных, экспортированной в формат Excel (продолжение)

Пример поисковой выдачи

V.8п/2450.

ТРИОДЬ ПОСТНАЯ. М., печ. Андроник Тимофеев Невежа, 08.11.1589 (20.12.7096–08.11.7098). 2º, 280 × 182. [1]⁸–[58]⁸ [59]⁶ = л.: 1–128, 126, 130–139, 141, 141–157, 156, 159–423, 414, 425–470 = 470 л.; формат полосы набора 228 × 121; высота 10 строк = 88/89, строк на странице 24, на л. 461 – 25 строк (в каталоге Зерновой ошибочно указано во всей книге 18 строк); на л. 3 – фолиация «6», исправленная в типографии на «3» красной краской; печать в две краски. Вариант набора 1-го вида (см.: Зернова, 10; Лукьяненко 1993, 99).

Украшения:

заставок 16 с 9 досок на л.: 1 (Зерн., 127), 11 (Зерн., 128), 18 (Зерн., 129), 31 (Зерн., 131), 65 (Зерн., 129), 79 об. (Зерн., 1-е состояние), 164 (Зерн., 127), 215 (Зерн., 128), 256 (Зерн., 132), 309 об. (Зерн., 133), 386 (Зерн., 134), 420 (Зерн., 134), 423 (Зерн., 133), 441 (Зерн., 130), 459 (Зерн., 135), 469 об. (Зерн., 135);

инициалов 2 с 2 досок на л.: 441 (Зерн., 109а), 469 (Зерн., 117).

Состав: службы с недели мытаря и фарисея по шестую неделю поста, л. 1–419 об.; «Подобает ведати, когда поются троичны дневнии, на осмь гласов, поются во святый великий пост и в прочая посты, егда поется аллилуйя», л. 420–422 об.; «Седалны, ото осмогласника, яже поем на первых стихологиях, святыя четыредесятница», л. 423–440; «Месяца априля, в первый день. Житие и жизнь преподобныя матери нашея Марии Египетьския, описано Софронием, патриархом Иерусалимьским», л. 441–459; «Повесть полезна, от древняго писания сложена, воспоминание являющи преславно бывшаго чюдеси егда персы и варвары, царьствующий град облегоша бранию, иже и погибоша божиим судом, и искушени бывше. Град же неврежен быв, молитвами пречистыя госпожа нашея Богородица и оттоле молебное благодарение поется, неседалное день тои именуя», л. 459–468 об.; послесловие и выходные сведения, л. 469 об.–470 об.

Библиография: Зернова, 10; Лукьяненко, 99.

Сохранность: сохранившиеся л.: 2, 3, 6–18, 25–27, 31, 33–136, 138, 139, 141, 141, 142–157, 156, 159–316, 318–423, 414, 425–440, 459–468; блок разбит, листы выпадают, потрепаны; л. 1, 4, 5 – рукописные, кон. XVII– нач. XVIII в. (?) (филигрань «Герб Амстердама»; полуустав, черные и красные чернила, на 1-м л. примитивная заставка).

Переплет: доски, обтянутые кожей с тиснением, на нижней крышке кожа не сохранилась; две застежки, одна утрачена; блок оторван от переплета.

Запись: л. 63, скорописью XVII в.: «Сия книга Пелымского города».

Работа с подобного рода описаниями позволяет исследователям получить максимально полную информацию о формальных характеристиках того или иного книжного памятника.

Третья часть комплекса – полнотекстовые цифровые копии старопечатных книг и рукописей. Цифровые копии создаются на специализированном программно-аппаратном комплексе «РуСкан 2А3». В основе аппаратной части лежит колыбель, позволяющая раскрывать книгу на 90 градусов. Сверху разворот книги придавливается стеклом, что позволяет устанавливать одинаковое фокусное расстояние во время сканирования всей книги и получать цифровые копии одинакового размера, упростить обработку изображений и их дальнейшую публикацию (рис. 5).

Рис. 5. Сканер РуСкан 2А3.

Начало систематической работы по оцифровке фондов лаборатории было положено в 2012 г. в связи с началом проекта «Книжное наследие уральских старообрядцев» в рамках Endangered Archives Program (Программа по сохранению архивов, находящихся в опасности) Британской библиотеки. Технические параметры сканирования применялись в соответствии с требованиями Программы. После завершения этого проекта в 2013 г. сканирование книг было продолжено в соответствии с теми же стандартами. В настоящее время оцифровано 68 старопечатных книг и рукописей, включающих наиболее редкие и ценные памятники, а также часть коллекции XVI в. (см.: https://lai-urgi.urfu.ru/ru/drevlekhranilishche/ehlektronnye-kopii-knig/).

Необходимо сказать несколько слов по поводу технических параметров оцифровки. Сканирование осуществлялось посредством двух

камер Canon EOS 600D с объективами Canon 15-85 mm. Свет белый, одинаковый за счет использования темных штор сканирующего устройства. Баланс белого установлен вручную. Разрешение снимков: 18 Мпикс. Экспозиция для съемки листов: iso: 400; выдержка: 1/100 с.; диафрагма: 8.0; все листы снимались с цветовой (стандарты RGB и CMYK) и размерной шкалой.

Отдельной задачей была съемка филиграней на сканере. Для этого необходимо было разработать тонкий подсвет, который мог бы помещаться под листом книги, прижатым стеклом, не нагреваться и не повреждать памятник. В ходе ряда экспериментов была разработана следующая технология. Съемка осуществлялась теми же камерами, что и съемка листов, но при выключенном свете сканера. Iso было установлено на 400 ед., экспопара выдержка-диафрагма устанавливалась в зависимости от типа и плотности бумаги и подбирались опытным путем для каждой книги отдельно.

Для подсветки филиграней во время съемки использовался подсвет оригинальной конструкции, по эксплуатационным параметрам практически не уступающий американским аналогам, а по себестоимости – дешевле примерно в 15 раз (в ценах и валютных курсах 2012–2013 гг.). Пример работы подсвета изображен на рис. 6.

Невозможность снимать книжные переплеты в сканере из-за опасности повреждения оборудования выступающими металлическими и деревянными частями переплетов заставила проводить их съемку фотокамерой Canon EOS 7D, закрепленной на штативе, и объективами Tamron 24-70 mm, Tamron 70-200 mm. Съемка осуществлялась на белом фоне, параметры экспозиции выбирались индивидуально под условия съемки.

После съемки изображения проходили процесс обработки, состоящий из следующих этапов. Обрезались лишние поля по периметру снимков в специализированном программном обеспечении, входящем в комплект поставки сканера. Далее проходила визуальная проверка наличия всех листов в цифровой копии памятника. После этого удалялись «черные» листы, появлявшиеся в результате съемки филиграней (особенностью сканера является съемка разворота, разделенного на 2 файла; при съемке филиграней один из файлов оказывался черным). Далее осуществлялась конвертация в формат pdf для публикации на сайте. Важным моментом является и то, что исходники изображений сохра-

Рис. 6. Пример отснятой филиграни. Библия, Острог, 1581, печатник Иван Федоров
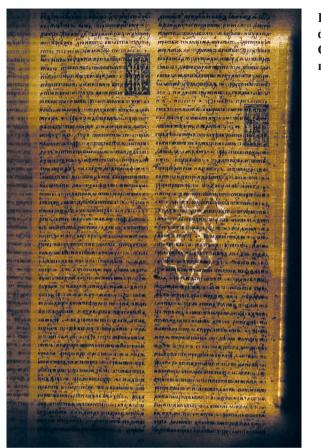
няются отдельно, это позволяет при необходимости исправить ошибки, выявленные после публикации цифровой копии памятника, а также сохранять цифровые копии в более высоком качестве, чем это позволяет сделать формат pdf.

Хостинг-площадкой для публикации цифровых копий был выбран Яндекс.Диск вследствие более дружественного интерфейса, высокой скорости работы и меньшей вероятности блокировки, чем у зарубежных сервисов. В будущем планируется улучшить качество выложенных изображений посредством переезда на более удобные и профессиональные платформы.

Размещение на сайте цифрового исследовательского комплекса в полном объеме позволит сделать уникальный источниковый корпус Ла-

боратории археографических исследований доступным для всех специалистов в области истории, литературы и культуры России.

Список литературы

Починская 1999: Починская, И. "Уральский археографический центр и его коллекция старопечатных и рукописных книг кириллической традиции" *Solanus* 13, 60–79.

Починская 2010: Починская, И. "Описание рукописных и старопечатных книг." В Починская, И., Мангилев, П., Ануфриева Н.: *Рукописные и старопечатные книги кириллической традиции: изучение, описание, хранение: учебное пособие.* Екатеринбург: Издательство Уральского университета, 419–455.

Авторы

Починская Ирина Викторовна, доктор исторических наук, заведующий Лабораторией археографических исследований Уральского федерального университета, Екатеринбург, Россия, e-mail: poirvi12@gmail.com

Палкин Александр Сергеевич, кандидат исторических наук, старший научный сотрудник Лаборатории археографических исследований Уральского федерального университета, Екатеринбург, Россия, e-mail: alexander.palkin87@gmail.com

Summary

The Laboratory of Archaeographical Studies (LAI) at Ural Federal University has created a digital research resource with the aim of easing specialist access to its collection of Cyrillic literary artefacts. This collection was the result of expeditions to locations with large numbers of Old Believers. While maintaining the foundations of traditional book culture, Old Belief also gave new impetus to the further development of folk literature. They also created a large collection of medieval artefacts and original pieces that reflect both the Russia-wide problems of the country's largest dissent movement and its regional particularities.

The LAI's collection contains liturgical literature, numerous anthologies of a highly diverse character, examples of late Russian chronicles, geographical and scientific texts, original polemical, historical and eschatological Old Believer essays, medieval Russian music sheets, and collections of religious

poetry. The collection of old printed books provides the opportunity to trace the development of eastern Russian Cyrillic book printing and its sources prior to the 20th century: it also reflects the developmental stages of western Russian book printing. Many of the books contain autographs from prominent actors in Russian history and culture.

Alongside literary artefacts, LAI's collection contains documents on the history of Old Belief taken from personal archives. Thus, the collection is an extremely valuable set of historical sources on religious conflicts and the search for means to resolve them. It allows researchers to significantly widen conceptions about connections between the state, the official church, and Old Belief: it also provides insight into intra-Old Believer relationships. Hundreds of books have been written and a range of theses defended while using the LAI's manuscript and book collection: however, its academic potential has yet to be exhausted.

Currently, one of the main areas of the LAI's activity is the creation of an appropriate digital method to guarantee remote access to the aforementioned artefacts. With this aim in mind, a digital research resource containing the Laboratory's manuscripts and old printed books has been prepared in order to allow academics to work with the collection in a variety of different ways depending on their research aims. One can access this resource through the Laboratory's website (lai-urgi.urfu.ru), in the section entitled 'Depository'. The resource consists of three parts.

The first part contains lists of fonds, which are themselves divided into two subsections. In the subsection 'List of Territorial Sets', one can find general information and descriptions about each of the collection's 27 sets. The second way of working with the lists is an electronic search, which allows one to look for interesting information either in one individual territorial set or across the entirety of the collection.

The second part contains a database consisting of detailed descriptions of the old printed books and manuscripts. The descriptions were prepared by the Laboratory's team. To access these descriptions, a web interface will be made available: at the moment, this is in the final stages of development and is undergoing internal testing. We plan to open it to the public at the end of 2019.

The third part consists of fully digitised copies of old printed books and manuscripts. The digital copies were created on the special device 'RuSkan 2A3'. Systematic work on the digitisation of the Laboratory's collection was begun in 2012 in connection with the project 'The Textual Legacy of the Ural Old Believers', which was conducted within the framework of the British Library's Endangered Archives Program. After this project concluded in 2013,

scanning continued in accordance with the same standards. Currently, 68 old printed books and manuscripts have been digitised, including the rarest and most valuable artefacts and a considerable part of the collection dating from the 16th century. Yandex.Disk serves as the hosting site for publishing the digitised copies.

ON PART-OF-SPEECH ATTRIBUTION AND GRAMMATICAL TAGGING OF OLD RUSSIAN *КРИВО* (KRIVO) AND *ПРАВО* (PRAVO)

Anna Ptentsova

Abstract: The paper discusses the identity of ancient Russian lexemes кри-во and право as part of speech; it is shown that their unambiguous assignment to the category of adverbs is controversial in a large number of cases and is incorrect in at least one of them; it is suggested that in a number of contexts for such words there is an indelible grammatical ambivalence of an adverb / noun (at least for native speakers of the modern language), and therefore we can only assume a greater or lesser degree of probability of one or another part of speech ; It is concluded that the corpus mark-up must necessarily reflect this ambivalence.

Keywords: Old Russian, semantics, part-of-speech attribution, grammatical tagging, corpus linguistics.

The part-of-speech attribution of some Old Russian words ending in -o seems to be difficult in certain cases. I'll discuss two of them: криво and право. The research is based primarily on the Old Russian subcorpus of the Russian National Corpus (RNC), as well as historical dictionaries, tagged texts from the collection of Old Russian manuscripts (*Rukopisnye pamyatniki Drevnej Rusi*), and also a number of offline sources.

Let us consider the following examples from the RNC:

(1) да аще кто ѿ руси или ѿ грекъ створи <u>криво</u> да оправлаеть то

'If any of the Russians or Greeks broke the rules ("done awry"), let them correct it' (Rus'–Byzantine Treaty (945), Laurentian Chronicle, f. 12r).

(2) иде же <u>криво</u> братие исправивъше ѵьтѣте

'Where [it is written] wrong, brothers, correct and read' (Izbornik 1076, ff. 275v – 276r, a note by scribe Ioann).

In both cases криво is tagged as an adverb in the RNC, and the same decision is made for (1) in *Rukopisnye pamyatniki Drevney Rusi*. The latter

contains no information on (2), because Izbornik 1076 is not present in this collection. However, the dictionary accompanying the 2009 publication of Izbornik 1076 also lists криво as an adverb. The historical dictionaries of 11th–14th and 11th–17th century Old Russian (Slovar' 11th–14th; Slovar' 11th–17th), as well as *Materialy dl'a slovar'a drevnerusskogo yazyka* ("Materials for the Dictionary of Old Russian") by Izmail Sreznevskiy (Sreznevskiy 1958), attribute криво exclusively to adverbs.

In Old Russian, the adverb криво carried two meanings: "incorrectly" and "unfairly". These meanings will be discussed further. In Modern Russian the word also has the meaning 'askew'. The earliest occurrence of криво in this sense, however, appears only in the 17th century. Compare an example from the Old Russian subcorpus of the RNC:

Карабль (sic!) *наш был мало грузен и шел таково* <u>*криво, что пушки*</u> <…> *чертили по воде*

'Our ship was under-loaded and sailed so askew that the cannons <…> stroked the water' (Travel of the stolnik P.A. Tolstoy to Europe, 1697–1699).

Returning to examples (1) and (2) above, we must note that treating криво as an adverb is not the only possible approach. Due to homonymy, криво may also be a neuter short adjective, or a noun similar to добро 'good' or зъло 'evil'. The (latter) substantival usage is, in fact, probable in the contexts considered. In (1), let us replace криво with зъло, which is a semantically legitimate substitution:

(1') да аще кто ѿ руси или ѿ грекъ створи <u>зъло</u> да оправлаеть то

'If any of the Russians or Greeks broke the rules ("done evil"), let them correct it'.

In this case, the lexeme зъло is more likely to be interpreted as a noun. Compare to similar examples from the RNC:

(3) не мозите сего створити <u>зла</u> ни оубиваите игора

'Do not dare to do this evil, do not kill Igor' (Kiev Chronicle, 1147, f. 129r).

(4) не тако бо бѣ казалъ имъ цесарь немечьскыи и папа римьскыи ιако же си зло оучиниша цѣрюграᴬ

'What the emperor and the Pope have ordered them [to do] was not to cause harm to Tsargrad' (Novgorod First Chronicle, 1204, f. 68v).

In both (3) and (4) the word зъло is marked as a noun. In (3), it is in genitive case and does not match the adverbial form, but this decision also

appears correct in the case of (4), where the noun and adverbial forms are identical.

We should note that the (Sreznevskiy 1958), as well as the (Slovar' 11th–17th), describes зъло as a noun, while the adverbial meaning is only attributed to the word зълѣ. On the contrary, the (Slovar' 11th–14th) acknowledges the part-of-speech homonymy of зъло and зълѣ, like all dictionaries do for the pair криво – кривѣ. Consider one of the examples of зъло as an adverb from this dictionary:

(5) да не въмѣнить имъ г҃ь грѣха ѥже помыслиша зло о мнѣ

'Let not God deem it a sin that they thought evil of me' (Tolkovaya Paleya 1406, f. 117r).

In this context, however, the word зъло can easily be interpreted as a noun, and this appears to be a more appropriate attribution. Therefore, in case we favor the substantive interpretation of зъло in створи зъло 'do evil', it is natural to adhere to this choice for криво in an analogous construction.

Let us consider other examples of криво, the majority of which come from the RNC. They make up a continuum of marginal cases: in some, криво (like зъло in (5)) can be interpreted as either an adverb or a noun. In others, the attribution inclines more naturally towards the adverb. First, an ambiguous example from RNC:

(6) а не на мнѣ та кровь боудеть но на виноватомъ но на томъ кто боудеть криво оучинилъ

'This blood shall not be on me, but on the guilty, on the one who does wrong' (Volhynian Chronicle, 1289, f. 930r).

Here the RNC marks криво as a neuter single adjective. *Rukopisnye pamyatniki* has an additional note, "serves as a noun". The substantive interpretation seems justified, especially considering the semantically close construction зло оучиниша (4), where зло is marked as a noun. However, if we consider криво a noun in криво оучинилъ, it is obvious that the same choice should be made in створи криво.

Consider also another case of part-of-speech ambiguity in криво, the famous note by scribe Putyata in Menaion XI:

(7) поутата пьсалъ . даѥ криво да исправите а не кльните

'Putyata wrote (this). If (it is) wrong, correct (it), but do not curse' (Menaion XI, f. 135r, cited according to (Karskiy 1901: 301)).

Let us turn to the contexts where криво is unambiguously an adverb: example (8) from the RNC, and example (9) from the (Slovar' 11th–14th):

(8) гоудьцю ладацю гоусли своꙗ и филипъ сѣда и рече криво ладиши инако было

'When the guslar tuned his gusli, Philip said: "Tuning (it) wrong, it has been (tuned) differently"' (Ptchela, end of 14th century, f. 49v).

In (8), a parallel adverb инако 'differently' influences the adverbial interpretation of криво.

(9) а ꙋто взато криво а то по исправѣ отъдати

'Whatever was taken wrongly, return after closing the case' (a letter missive, Moscow, 1390).

Криво must be an adverb in (9), because a noun would require genitive marking.

Now let us consider one more example from the RNC, where криво may be an adverb, a neuter short adjective, or a noun. The same ambiguity applies to право:

(10) реклъ бо еси ѡже ꙗ саду вь киеву то ꙗ тебе надѣлю ⟨...⟩ нынѣ же ты сѣлъ еси право ли криво ли надѣли же мене

'You said: if I reign in Kiev, I will vest you <…> Now it is you who is in Kiev, rightly or wrongly; vest me' (Kiev Chronicle, 1174, f. 204v).

For право, consider also the well-known margin note in the 11th century copy of *13 Sermons by Gregory the Theologian*, where, once again, the choice between noun and adverb is unobvious:

(11) ꙋьгьле кривага главо пиши право

'Goldfinch, you fool, write correct things / correctly' (cited according to (Karskiy 1901: 303)).

The question arises: in which contexts are криво and право unambiguously treated as adverbs, and in which contexts the attribution is not clear-cut? Evidently, the ambiguity arises in the examples with transitive verbs; such verbs must have an accusative object, and the action they signify should be naturally evaluated on the scale of good to bad or right to wrong. The verbs сътворити 'do', оуꙋинити 'commit', взати 'take', мыслити 'think', and пьсати 'write' all fall into this category.

Let us consider another well-known example that helps to resolve the discussed ambiguity in at least a number of cases. It is a margin note by scribe Domka in the 11th century Novgorod Menaion:

(12) дажь въ на поа обраψе <u>криво</u> а вьсе ⟨…⟩

'Should (anyone) conducting a sermon by them (these books) find a mistake… (fragment ends)'

(Novgorod Menaion 1095-1097, September, f. 176r, cited according to (Obnorskiy, Barkhudarov 1999: 37)).

What would be the preferred interpretation in this case? Let us assume, by the logic of corpus and dictionary markup, that криво stands in for an adverb, as in examples (8), (9), and possibly (10). If so, криво describes an action as 'containing a mistake or certain unfairness'. Note that we encounter this meaning in modern spoken Russian:

(13) *Они как-то <u>криво</u> договорились и не смогли встретиться*

'They made an awry agreement and failed to meet'.

(14) *Договор был составлен <u>криво</u>, и теперь ему отказываются возмещать убытки*

'The contract was drawn up incorrectly, and now they refuse to compensate his losses'.

The old meaning appears to be broader than the modern, because the latter lacks the component of 'unfairness'. In principle, however, they are semantically close. The fact that in Modern Russian криво is an adverb may vote in favor of adverbial interpretation of the old криво. The intuition that such криво is relatively new in the language is, of course, false.

Returning to example (12), we will discuss the construction дажь ⟨…⟩ обраψе криво. As a side note, the word обраψе 'will find' is a present form without the standard suffix –ть, which is a feature of the Old Novgorod dialect. Криво, in both Old and Modern Russian, is a floating scope adverb. According to (Boguslavskiy 1996), the scope of a lexeme is the fragment of the syntactic structure, such that the components of this fragment – word meanings, constructions and intonation – fill one semantic slot of the given lexeme. Adverbs in Modern Russian, as suggested in (Filipenko 1998), may have either "floating" or "fixed" scope, depending on whether the adverb may relate to a single component in a predicate, or to many such components. To provide an example (taken from Filipenko), adverbs like *тихо* 'quietly', *медленно* 'slowly', or *ночью* 'at night' have "fixed" scope, because they characterize a single component of the situation (manner or time), and also because they are used with a limited number of predicates. On the contrary,

the adverbs like *легкомысленно* 'carelessly' and *правильно* 'correctly' have "floating" scope.

E. Rudnitskaya in (Rudnitskaya 1991) contrasts two uses of *легкомысленно*. In the first use the adverb may be defined as follows: 'in the course of an action, the actor does not think (enough) about the consequences caused by the manner of their action'. Such meaning is exhibited in e.g. *отвечать легкомысленно* 'reply carelessly' or *поступать легкомысленно* 'act carelessly'; note that the adverb assesses the quality of an action, but not the fact of performing that action. In the second use, on the contrary, the adverb describes the very fact of taking an action, which Rudnitskaya defines as follows: 'when causing a situation, the actor does not think (enough) about the consequences of this situation'. We encounter this meaning in the so-called sentential uses, like *Он легкомысленно согласился* 'He carelessly agreed', which we may paraphrase as *То, что он согласился, было легкомысленным* 'The fact that he agreed was careless'. Similarly to the above examples, the phrase *Он правильно выступил* lit. 'he rightly spoke' may be interpreted in two ways: either as 'What he said during his talk was right', or as 'Speaking was the right thing to do'.

Considering this, we attribute both modern and old криво to floating-scope adverbs, because they can modify a broad range of predicates and characterize various aspects of the situation. The modern *криво договориться о встрече* means 'mistakenly (or imprecisely) agree on the time and place (for a meeting)', while *криво составить документ* means 'commit an error in wording when designing a document', etc. Similarly, in the old лдити, взти, надѣлити криво the adverb describes different aspects of the situation depending on the predicate type (which may be numerous). Криво лдити means 'tune (the instrument) incorrectly', взти ⟨надѣлити⟩ криво – 'to take <to vest> illegally'. Note that, unlike *легкомысленно*, modern *криво* may not refer to a phrase as a whole, and apparently, this holds for the old криво as well.

Let us turn once more to дажь ⟨…⟩ обраще криво. Should we translate it into Modern Russian, we would not be able to use modern *криво*, unlike in the expressions лдити, взти, надѣлити криво, and many others. While the collocations *криво ладить* 'tune incorrectly', *криво взять* 'take wrongly', and *криво наделить* 'vest illegally' are legitimate in Modern Russian, *криво найдет* 'will find wrongly' is nonsensical. What are the reasons for

this discrepancy? Recall that in the above Modern Russian examples *криво* means 'making a mistake in the process of action', and the same or slightly broader meaning is implied in most of the Old Russian examples, but not in обраще криво. Обраще криво does not imply a mistake in the action of acquiring (обрѣсти), but on the contrary, this action may be considered an attempt to correct the mistake.

What, then, is the meaning of *криво* in relation to обрѣсти? Adhering to our theory of *криво* as an adverb, we must expand its floating scope in the case of Old Russian. Evidently, in our example it modifies the object (the mistake in the text), while the scope of modern *криво* does not extend past the predicate. In other words, we suppose that криво обрѣсти in Old Russian means 'to find an object that contains a mistake', but this is equivalent to treating *криво* as a noun. Indeed, modifying the original assumption removes unnecessary complexity in our logic.

In accordance with the argumentation above, we should also select the substantive rather than adverbial interpretation for *криво* in examples (2) and (7), and for *право* in example (11). In (8) and (9) *криво* should be interpreted as an adverb. In the remaining contexts, we recognize unresolvable ambiguity.

Compare also some cases of part-of-speech ambiguity in *право*:

(15) с҃тополкъ же сматеса оумомъ . рекии еда се <u>право</u> боудеть или лжа

'Sv'atoslav got excited and said: will be this correct / the truth or false?' (Povest' Vremennykh Let, Hypatian codex, 1097, f. 88v).

(16) рече има ꙗнь то вамъ <u>право</u> молвать б҃ѣ ваши

'Yan' told them [sorcerers]: your gods tell you correctly / the truth' (Povest' Vremennykh Let, Hypatian codex, 1071, f. 66r).

(17) повѣжь брате мои право своею ли волею сн҃ъ твои сѣлъ в берестьи ци ли твоимъ повелениемь

'Tell me, my brother, correctly / the truth: if your sun took a throne at Berestye by his own decision or by yours?' (Volhynian Chronicle, 1289).

In all three cases *право* is interpreted as an adverb in the RNC, but it could be also interpreted as a noun (notice for (15) that the noun лжа is connected with право by the conjunction или 'or'). The same problem we have for the construction право ти молвлю 'I tell you the truth' which is a sort of an invariable expression:

(18) рече емоу кн҃же право ти молвлю стерези са хотать ти ꙗти

'He said to him: oh prince, I tell you the truth, be careful, they want to catch you' (The Kiev Chronicle, 1161, f. 183v).

Unresolvable ambiguity like this should be reflected in the RNC, as well as in other grammatically annotated resources. One possible solution would be to introduce double part-of-speech marking for these cases.

Electronic Resources

Old Russian manuscripts: www.lrc-lib.ru

Russian National Corpus: www.ruscorpora.ru

References

Boguslavskiy 1996: Boguslavskiy, I. M. *Sfera deystviya leksič'eskix yedinits*. Moscow: Škola "Yazyki russkoy kultury", 1996.

Izbornik 2009: *Izbornik 1076 goda*. Moscow: Yazyki slavyanskoy kultury, 2009.

Karskiy 1901: Karskiy, E. M. *Oč'erk slavyanskoy kirillovskoy paleografii*. Warsaw: Tipografiya varšavskogo uč'ebnogo okruga, 1901.

Obnorskiy, Barxudarov 1999: Obnorskiy, S. P., Barxudarov, S. G. *Xrestomatiya po istorii russkogo yazyka*. Part I. Moscow: Aspekt-press, 1999.

Rudnitskaya 1991: Rudnitskaya, E. L. "Slovarnaya statya nareč'iya lekgomyslenno." *Semiotika i informatika* 32, 156–159.

Slovar' 11th–14th : *Slovar' drevnerusskogo yazyka XI – XIV vv*. Vol. 4. Moscow: Russkiy yazyk, 1991.

Slovar' 11th–17th: *Slovar' drevnerusskogo yazyka XI – XVII vv*. Vol. 8. Moscow: Nauka, 1981.

Sreznevskiy 1958: Sreznevskiy, I. I. *Materialy dlya slovarya drevnerusskogo yazyka*. Vol. I. Moscow: Gosudarstvennoye izdatel'stvo inostrannyx i natsional'nyx slovarey, 1958.

Filipenko 1998: Filipenko, M. V. "Ob adverbialax s plavayušč'ey i fiksirovannoy sferoy deystviya." *Semiotika i informatika* 36, 120–140.

About the author

Anna Ptentsova (anna.ptentsova@gmail.com);

Lomonosov Moscow State University, Faculty of Philology, senior lecturer. Scientific interests: Old Russian, Modern Russian, semantics, lexicology.

Резюме

В работе обсуждается частеречная принадлежность древнерусских лексем криво и право; показывается, что однозначное отнесение их к разряду наречий спорно в большом числе случаев и неверно как минимум в одном из них; выдвигается предположение о том, что в ряде контекстов для подобных слов на -о существует неснимаемая (по крайней мере, для носителей современного языка) грамматическая амбивалентность наречие / существительное, в связи с чем можно лишь предполагать бо́льшую или меньшую степень вероятности той или другой части речи; делается вывод о том, что корпусная разметка обязательно должна отражать эту амбивалентность.

RECOVERING A 16TH-CENTURY OTTOMAN DOCUMENT DAMAGED BY SPILLED INK

C. Römer, S. Brenner, F. Cappa, B. Frühmann,
E.-G. Hammerschmid, M. Schreiner

Abstract: An Ottoman firman of Suleiman the Magnificent (1520-1566), dated August 13, 1557, and kept at the Vienna Haus-, Hof- und Staatsarchiv (HHStA, Turcica 13, Konv.2 [1557 VII–XII], f. 33–39, Petritsch 1991: no. 351), was badly damaged by ink spilled over it in 1860. Therefore, its text has not been completely deciphered. As an attempt to recover the text under the spilled ink, Multispectral Imaging (MSI) of the document was performed by the Computer Vision Lab at TU Vienna and the Institute of Science and Technology in Art (ISTA) of the Academy of Fine Arts Vienna, where the following investigations were performed:

1. IR reflectography has made the damaged text fully readable.

2. By applying X-ray Fluorescence analysis (XRF) and Raman spectroscopy the original as well as the spilled inks could be identified.

3. Additionally, the pigments of the *tughra* and the "gold dust" used for drying the ink were analysed.

Keywords: Ottoman documents, damaged documents, Multispectral Imaging (MSI), IR reflectography, X-ray fluorescence analysis, Raman spectroscopy

An Ottoman firman dated August 13, 1557, and kept at the Vienna Haus-, Hof- und Staatsarchiv (HHStA, Turcica 13, Konv. 2, (1557 VII–XII) f. 33–39, UR TUK 75; Petritsch 1991: 127–128, no. 351) is an especially interesting document, not so much due to its contents, but in view of its condition. In it, Sultan Suleiman the Magnificent (r. 1520–1566) complains vis-à-vis King Ferdinand I of Habsburg (1503–1564) about Habsburg incursions into the Ottoman territory of Hungary. Incursions of this kind were a consequence of the unstable conditions along the Ottoman-Habsburg borderline. The Habsburg-Ottoman peace treaty of 1547 and all subsequent treaties stipulated joint commissions to investigate infringements of the peace from both sides. In the document, Suleiman also gives his permission for Ferdinand's ambassadors Franz Zay (1498–1570) and Anton Veranc-

sics (1504–1573) to return home, ordering the destruction of the fortress at Szigetvár (Hungary) as a condition for renewing the peace treaty. The third ambassador, Ogier Ghislain de Busbeck (1522–1592), is supposed to stay in Constantinople.[1]

This document was published in 1983 in a volume comprising 36 orders of Suleiman directed to emperors Charles V, Ferdinand I, and Maximilian II. (Schaendlinger & Römer 1983: 54–56, no. 21). The original document's length is 171 cm and its width 40 cm. It has a golden *tughra* (the name of the sultan written and often beautifully painted by a special clerk at the Ottoman court who was called *nişancı*). In some instances the original "gold dust" (*rîk*) used for drying the ink can still be seen. This firman was badly damaged when in 1860 Walter Friedrich Adolf Behrnauer (1827–1890), librarian at the Vienna Hofbibliothek, inadvertently spilled a pot of ink over it (Ill. 1[2]).

Behrnauer came from Saxony. His academic teacher was Heinrich Leberecht Fleischer (1801–1888), a renowned orientalist. Behrnauer taught Ottoman Turkish and Ottoman literature as well as Ottoman diplomatic and palaeography in Vienna between 1852 and 1861[3]. Subsequently, he became an *Assistent* at the Königliche Öffentliche Bibliothek in Dresden[4]. Behrnauer was highly esteemed by another German orientalist, Wilhelm Theodor Ahlwardt (1828–1909, Littmann 1953: 112), who heavily criticized Hammer-Purgstall's (1774–1856) philological competence, stating that there was not a single valiant Arabist in Austria except Behrnauer who had suffered from attacks by Hammer and his "satellites".[5]

The incident at the archive caused a turmoil of bureaucratic correspondence, especially as Behrnauer not only had reacted by lifting the document and holding it vertically so that the ink spread downward, but according to

---

[1] On these ambassadors, see, e.g., Schaendlinger 1984: 194.

[2] The copyright of all illustrations, which are stored in a CD, as well as of Table 1 and Figure 1 is with © NTK 2018 Univ.-Prof. Dipl.-Ing. Dr. M. Schreiner. Neither this CD nor any part may be reproduced, stored or transmitted in any form or by any means, electronical or mechanical, including photocopying, microfilming, and recording, or by any information storage and retrieval system, without permission in writing from the publisher.

[3] See Herzog 2010: 32, Bihl 2009: 24–25.

[4] See Haan 1875: 15.

[5] See Höflechner et al. 2018: 210.

an archivist's report[6] had handed back the document without telling anybody. Behrnauer himself seems to have explained his behaviour as resulting from a nervous disease, which was used by the *Obersthofmeisteramt* as a reason for denying him access to the archive at least until his recovery. Even then he would not be left alone with any document.[7] Behrnauer seems to have left Vienna in 1861, thus ending his career at the Hofbibliothek. The last time that he taught palaeography and reading Ottoman documents of a higher level at Vienna University appears to have been the summer term of 1861. One may wonder where he would have had access to material from at this date.

Another consequence of the Behrnauer incident, apparently, was a new regulation that working in the archive would only be allowed with the use of pencils instead of pens and ink. Today, this rule is respected in archives and manuscript collections around the world.

Let us now turn to the analysis of the document.

First of all, we wish to sincerely thank the director of the Haus-, Hof- und Staatsarchiv, Mag. Thomas Just, who immediately gave his permission to have the document moved from the archive in order to carry out the photographic documentation and material analytical investigations. Our thanks also go to Mag. Kathrin Kininger from the Haus-, Hof- und Staatsarchiv, who was responsible for the transport of the document and was present throughout the measurements.

As a first step, the document was brought to the Computer Vision Lab of the TU Wien (https://cvl.tuwien.ac.at/), where Multispectral Imaging (Hollaus 2013, Sablatnig 2014) was carried out by Simon Brenner. The document was imaged under 11 narrow-band illuminations with peak wavelengths between 365 and 940 nm using an achromatic middle-format camera. Additionally, true-colour and UV-fluorescence images were taken with a conventional DSLR camera. As expected, the best results were achieved in the IR region, especially at 940 nm (Ills. 1 and 2). As the covered text

---

[6] The archivist's name was Carl Rohenauer (HHStA Kurrentakten, Zl. 170/1860, report of 24 September 1860).

[7] The whole documentation of this incident is to be found under HHStA Kurrentakten, Zl. 170/1860 and 207/1860 (see Petritsch 1991: 128). For a transliteration of the final decision about Behrnauer, see the Appendix. Unfortunately, for reasons of space, it is impossible to transliterate all the documents here.

already was fairly readable in those images, no further processing was necessary.

Subsequently, the document was brought to the Institute of Science and Technology in Art (ISTA, http://www.ntk.akbild.ac.at/) at the Academy of Fine Arts Vienna and analysed by the other co-authors of this publication.[8] Here, the following non-invasive material analytical techniques were applied:[9]

1. X-ray Fluorescence analysis (XRF)
2. RAMAN spectroscopy
3. optical microscopy
4. IR reflectography

All these investigations must and can be carried out in air and are accepted to be non-destructive and non-invasive, although a low amount of energy is emitted into the object (heat in Raman spectroscopy and X-ray radiation in XRF, causing free radicals and hence oxidation or hydrolysis).

1) X-ray Fluorescence analysis (XRF)

This method is used to determine the chemical elements of materials. It works on the basis of photoelectric absorption, where primary X-ray radiation of an X-ray tube is absorbed by the material, removing electrons from the inner atomic shells. This interaction creates for a short period (circa 1 ns = 10^{-9} sec) an "excited state" and an empty slot in the electron shell of the sample atoms. With relaxation (i.e., returning to the lower energy state)

---

[8] Analysis was conducted within the framework of the HRSM-project "Analysis and Conservation of Cultural Heritage – Modern Imaging and Material Analysis Methods for the Visualization, Documentation and Classification of Historical Written Material (Manuscripts)" centred at CIMA (Centre of Image and Material Analysis in Cultural Heritage, https://hrsm.cvl.tuwien.ac.at/). The main emphasis of CIMA is to extend and strengthen multidisciplinary cooperation and to promote research at the intersection of natural sciences and humanities. Comparable initiatives can be found at the University of Hamburg (Centre for the Study of Manuscript Cultures, http://www.manuscript-cultures.uni-hamburg.de, as well as the Network Comparative Oriental Manuscript Studies – COMSt, https://www.aai.uni-hamburg.de/en/comst.html), and in California (i.e., the computer vision activities of The Early Manuscripts Electronic Library – EMEL, http://emel-library.org/). CIMA's activities have so far been focused on the investigation of medieval manuscripts (including palimpsests), charters, and illuminated historic maps from the 8th to the 14th centuries, which in many cases are poorly preserved and in a fragmentary state. Further information is provided in the contribution of H. Miklas in this volume.

[9] Cf. Frühmann et al. 2018.

secondary X-ray radiation is emitted. This fluorescence radiation has an energy/wavelength typical for chemical elements present in the material.

For the analysis of the Ottoman firman an ELIO-type spectrometer developed by XGLab[10] was used for material analysis of art works. Its measuring head comprises a 4 Watt rhodium X-ray tube, a silicon drift chamber detector of 25 mm², and two laser pointers. The primary X-ray beam and hence the area of analysis has a diameter of approximately 1 mm.

In general, the inks and pigments of a manuscript (text and illumination) can be deduced from the elements determined – e.g., the detection of sulphur (S) and mercury (Hg) in a red area definitely indicates the presence of the pigment of vermilion (HgS). Another example would be the detection of iron (Fe) in the ink – a clear indication for the application of iron gall ink, which is an Fe(III) bis-gallate complex. Due to the chemical structure of this compound, the ratio of iron to carbon (Fe:C) is equal to 1:14, but by XRF only the detection of Fe is possible, since in air elements with an atomic number below that of sulphur – such as carbon, oxygen, sodium, magnesium, and aluminium – cannot be detected (Ill. 3).

For these reasons, Raman spectroscopy was applied for the analysis of the Ottoman firman as a valuable complementary technique to XRF, as it is also non-destructive and can be carried out in air.

2) RAMAN spectroscopy

Raman spectroscopy is based on the interaction with matter of high-intensity monochromatic electromagnetic radiation ranging from the UV to visible and IR regions. Materials such as pigments, dyes, and inks are characterized by diverse scattering processes based on their molecular structures. Therefore, specific information related to the chemical composition can be obtained using Raman spectroscopy.

Usually, a Raman system consists of five main components:
• laser source for stimulating the molecular vibrations
• integrated microscope with illumination system
• dispersion unit (spectrometer) for separating the Raman scattering of the material
• detector – usually a CCD plate
• computer for controlling the instrument and evaluating the results

---

[10] See XGLab S.R.L., Milano, Italy. http://www.xglab.it.

In our case, the measurements were carried out with a portable confocal Raman spectrometer (Pro-Raman L-Dual-G[11]; Ill. 4), equipped with 2 Diode Lasers for the excitation: 785 nm (~350 mW) and 532 nm (~50 mW) with narrow line-widths of 2.0 cm^{-1} and 1.5 cm^{-1}, respectively. The instrument is based on a two dimensional CCD array detector, which is temperature regulated (-60°C). The integrated microscope is equipped with a 1.3 Mpixel CMOS camera with in-line LED illumination. Analyses were performed using the 785 nm excitation laser with 50x LWD (long working distance) objective lens. The spectra were evaluated by comparison to the ISTA reference database. In contrast to XRF, where the domain of analysis is approximately in the range of several mm, the area of analysis for the Raman is in the range of several hundred microns, which in some cases hinders a clear comparison of the results obtained.

3) Optical microscopy

Additionally, our document was also investigated by applying a Wild M 650 stereo microscope at a magnification of about 30-40x, starting with 6x, in order to determine the golden colours of the *tughra* as well as of the so-called gold dust (*rîk*), which was used for drying the ink. According to Redhouse 1890: 999a, *altun rîki* was powdered gold leaf used upon important documents. According to Kütükoğlu 1994: 44, gold colour was made of gold leaf, which was mixed with gum Arabic or honey and water. After an hour it was sieved and dried over the fire. Before use, gelatine was added to make it fluid and smooth again. As gelatine dries fast, gold ink could not be made beforehand but had to be prepared when needed. This kind of ink was brushed onto the pen with a brush. Texts written with gold ink were polished with a small stone to make them shine. Another recipe was to let tin boil with water, first adding powdered saffron, then gum Arabic and continuing to stir until it cools. *Rîk* looked like a fine pink dust, but there were also varieties that shone like stars. Some were more expensive than others. Kütükoğlu does not say anything on the composition of *rîk*.

4) IR reflectography

Art objects and especially archival materials are usually photographed in the visible (ca. 400-780 nm), UV (ca. 320-400 nm) and IR (ca. 700-1100

---

[11] See Enwave Optronics, Irvine, CA, USA, http://www.tsi.com/Raman-spectrometers/.

nm) ranges. UV radiation is already absorbed by the uppermost layers of material and used for UV fluorescence photography in order to visualize traces of bleached ink. IR radiation, however, can penetrate colour or ink layers. In contrast to conventional IR photography using standard silicon sensors, for IR reflectography InGaAs detectors are applied, which are characterized by a sensitivity of up to approximately 1700 nm. For this study an Osiris camera of Opus Instruments (www.opusinstruments.com) was available with wavelengths between 900 and 1700 nm and a resolution of 4096×4096 pixels (Ill. 5).

Results

1. Inks containing carbon absorb light far into the IR region, whereas iron gall inks do not and become "transparent". Thus, as regards readability, a combined study of Simon Brenner's 940 nm photograph and those taken with the Osiris camera mentioned in the previous paragraph resulted in the recovery of most of the hitherto invisible or unreadable words of the document. In the 1980s, the text covered by Behrnauer's ink had already partly become decipherable by holding the document against a normal lamp. On the basis of the new pictures, nearly all the missing words could be deciphered. As a consequence of ink corrosion through Behrnauer's ink, some words were thought to be lost forever. However, Ottoman documents use a highly stereotyped formulaic language. As the surroundings of these lacunae are better visible now, they can be reconstructed. In what follows, a list of the corrections and additions is presented[12]:

   • line 7[13]: instead of *ḥudūdından olub* read *ḥudūdında olub* "are **at the** border of…"

   • line 8: instead of *ḳalᶜe-i mezbūreye ilticā ve istīnād alub* read *ḳalᶜe-i mezbūreye ilticā ve istīnād **eyleyüb***

   • line 8: *mādāmki mezbūr ḳalᶜe ḥāli üzre ṭu[ra] … [ṭara?]fları żabṭ olınmayub reᶜāyāyı [ve] fuḳarāyı rencīde ėdüb ḳavāᶜid-i ᶜahd kemāliyle muḥkem u mü'ekked olmaḳ* "As long as the aforementioned fortress **stays** as it is …. **its surroundings [?] cannot be checked, they molest the sub-**

---

[12] The new readings are rendered with bold characters. Translations will only be added if the meaning changes compared to Schaendlinger – Römer 1983.

[13] The line numbering is done on the basis of the whole document. Illustration 1 shows lines 7–11 and the place of issue.

jects and the poor, and it is practically impossible that the rules of the **treaty be completely firm** and confirmed".

• last word of line 8 read *olmaḳ*

• line 9 read *ve aradan refᶜ ėdesiz* "**and that you make it disappear**"

• line 10 read *eglendürmeyüb muᶜtemedün ᶜaleyh ė*[*lçi?*]/*ā*[*dem ? gö*] *nderesiz* "You shall not wait but **send an ambassador / a man you trust**".

The word after *kesim ḫuṣūṣı* remains unclear, although the ink is well visible. But in its upper part there is a piece of destroyed and restored paper.

• The place of issue at the bottom left reads *bi-maḳām Ḳosṭantiniyye el-maḥrūse* "at the residence of Constantinople **the well-protected**".

2. For XRF and Raman spectroscopy, several domains (measurement points – MP) were chosen, which can be seen in illustration 6. This was done in order to characterise the support paper, the black ink and the "gold" dust. Both, the basic material and the inks showed varying amounts of sulphur, chlorine, potassium, calcium, and iron, in some areas also titanium. None of them except iron are relevant for identifying the inks.
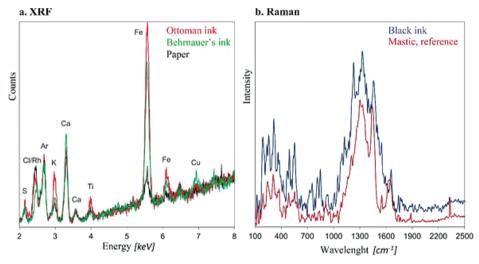
The brownish-black ink of the text contains more iron and potassium than the paper (MP 2), whereas the spilled ink (MPs 3 and 4) has a different composition. A higher intensity of copper (Cu) and calcium (Ca), but definitely a lower amount of iron (Fe), was detected in comparison to the support material. Therefore, the conclusion can be drawn that the chemical composition of the two inks is different.

MP 11, belonging to the part of the text that was not damaged, does not show any difference to the composition of the paper. This suggests that the ink was made of soot. Raman spectroscopy would be able to differentiate between both, iron gall ink and soot ink, but unfortunately the document was restored by applying a layer of resin, most probably mastic, thus making further analysis impossible.

Conclusion

The following figure (fig. 1) shows the results of the X-ray fluorescence analysis and Raman spectroscopy: (a) spectrum of XRF: red = Ottoman ink; green = Behrnauer's ink; black = paper. The Ottoman ink contains K and Fe, whereas in Behrnauer's ink small amounts of Ca, Fe, and Cu were determined. (b) Raman spectrum of the measuring point of the ink (blue)

and reference spectrum of mastic (red). By the treatment of the surface the information of the ink is covered.

XRF of Süleyman's *Tughra* has corroborated the idea that it contained gold (ills. 7 and 8); no gold could be found in the so-called gold dust (ills. 9 and 10). These measurement points only show silicon (Si), potassium (K), and iron (Fe) as well as small amounts of titanium (Ti). Raman spectroscopy identified a band of silicon, but the material applied for the restoration treatment prohibits further investigation. A higher amount of iron does not necessarily mean that the ink is an iron gall ink. More measurements would be needed to clarify the questions that have come up during these investigations.

Figure 1: The spectrum of X-ray fluorescence analysis reveals higher intensities of iron (Fe) in the black ink in comparison to Behrnauer's ink, and the Raman spectrum confirms the application of mastics during restoration.
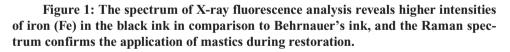
Table 1: Summary of the XRF and Raman results.

MP	Material/Measuring place	Analysis[14]	Interpretation
1	brownish basis material	S, Cl, K, Ca, (Ti), Fe	
2	black ink, text	(S), K, **Fe**, (Pb)	iron gall ink possible
3	brown stain	Ca, Ti, **Fe**, (Cu)	iron gall ink possible
4	brown stain	Ca, **Fe**, (Cu)	iron gall ink possible

---

[14] With the measuring points, only those elements are indicated in comparison with the basis material that were detected additionally or in larger quantities.

5	paper	S, Cl, K, Ca, (Ti), Fe	less intense than MP1
6	"gold dust" (*rîk*)	**Si**, **K**, Ti, **Fe**	
7	"gold dust" (*rîk*)	not different from MP 1	
8	"gold dust" (*rîk*)	Si, K, Ti, Fe	
9	"gold dust" (*rîk*)	**Si**, **K**, Ti, **Fe**, (Pb)	
10	"gold dust" (*rîk*)	**Si**, **K**, Ti, **Fe**, (Pb)	
11	black ink, text	not different from MP 1	carbon black?
12	golden ink, *Tughra*	**Au**	

Appendix

HHStA Kurrentakten, HHStA Zl. 170/1860.

12629/K

Die von der Direktion des Haus- Hof- und Staats Archives mit Bericht vom 24ten d. M.[15] angezeigte, durch den Amanuensis der K.K. Hofbibliothek dr. Adolph Behrnauer geschehene höchst bedauerliche Verletzung einer im Kais. Haus- Hof- und Staatsarchive aufbewahrten türkischen Staats Urkunde ist zur Kenntniß des kais. Obersthofmeister Amtes gebracht und von letzterem die Hofbibliothek angewiesen worden, sein bey diesem unliebsamen Vorfalle bewiesenes, jedenfalls höchst tadelnswerthes Benehmen mit allem Nachdrucke vorzuhalten und demselben über hierortige Aufforderung zu bedeuten, daß ihm die Benützung des Staats Archives für einige Zeit und wenigstens inso lange er von seiner angeblichen Krankheit nicht vollkommen hergestellt seyn wird verweigert sey, und auch später nur unter Beobachtung gewißer Vorsichtsmaßregeln wieder zugestanden werden wird.

Die Direction wird unter Rückschluß des in Frage stehenden Dokuments hievon zur Darnachachtung und mit dem Auftrage verständiget, fürderhin zum Schutze wertvoller seltener Dokumente die vorgeschlagene Vorsichts Maßregel in Anwendung zu bringen, darin bestehend, daß statt der Tintenfäßer, Bleistifte denjenigen verabreicht werden, die Abschriften oder Auszüge von derley Urkunden nehmen wollen.

So unangenehm auch das Ministerium durch den obigen Vorfall berührt worden, so kann dasselbe es doch nur billigen, daß hievon unverzüglich die

---

[15] The document mentioned here is HHStA Kurrentakten, HHStA Zl. 170/1860, 12926 K 184 860.

Anzeige anher erstattet wurde, und auch das Kais. Oberst-Hofmeister Amt hat die ihm vom Kais. Ministerium des Äußern gemacht Mittheilung – mit Hinblick auf die Gefahr, welcher die zugänglichen Schätze der Hofbibliothek ausgesetzt seyn könnten – dankbar anerkannt.

Wien den 24 Oktober 1860

In Abwesenheit des Ministers des Äußern, der Unterstaats Sekretär:

B Koller[16]

An die Direction des Kais. Haus- Hof- und Staats Archives.

References

Bihl 2010: Bihl, W. *Orientalistik an der Universität Wien. Forschungen zwischen Maghreb und Ost- und Südasien. Die Professoren und Dozenten*. Wien – Köln – Weimar: Böhlau, 2010.

Frühmann et al. 2018: Frühmann, B., Cappa, F., Hammerschmid, E., Schreiner, M. *Untersuchungsbericht Nr. 2018/9. Materialanalytische Untersuchungen an einer Osmanischen Urkunde Ur TUK75 des Österreichischen Staatsarchivs, mittels Röntgenfluoreszenzanalyse (FA), Raman-Spektroskopie, Licht-Mikroskopie und Infrarot-Reflektografie durchgeführt im Rahmen des HRSM Kooperationsprojektes „Erschließung und Erhaltung von Kulturgut"*, unpublished, dated Vienna, March 2018 and directed to Kathrin Kininger of the Haus-, Hof- und Staatsarchiv, Vienna, 2018.

Haan 1875: Haan W. *Sächsisches Schriftsteller-Lexicon*. Leipzig: Robert Schaefer's Verlag. https://de.wikisource.org/wiki/Haan:Walter_Friedrich_Adolph_Behrnauer (last accessed 23 February 2019).

HHStA, http://archivinformationssystem.gv.at/archivplansuche.aspx?ID=1521550, last accessed 23 February 2019.

Höflechner et al. 2018: Höflechner W., Wagner, A., G. Koitz-Arko (eds.). *Joseph von Hammer-Purgstall. Briefe, Erinnerungen, Materialien.* Version 2. Publikationen aus dem Archiv der Universität Graz (Quellen zur geschichtlichen Landeskunde der Steiermark, 30). (https://hdl.handle.net/11471/559.10, last accessed 23 February 2019), 46.

---

[16] Baron Koller was entrusted with this office on December 1, 1859 (see Haus, Hof- und Staatsarchiv, http://archivinformationssystem.gv.at/archivplansuche.aspx?ID=1521550, last accessed 23 February 2019).

Hollaus 2013: Hollaus, F.; Gau, M.; Sablatnig, R. "Enhancement of Multispectral Images of Degraded Documents by Employing Spatial Information." in *12th International Conference on Document Analysis and Recognition (ICDAR)* (2013), 145–149 (doi:10.1109/ICDAR.2013.36), 2013.

Herzog 2010: Herzog, Chr. "Notes on the Development of Turkish and Oriental Studies in the German Speaking Lands." in *Türkiye Araştırmaları Literatür Dergisi* 8.15 (2010), 7–76.

Janssens 2004: Janssens, K. "X-ray based methods of analysis." in Janssens, K. & Van Grieken, R. eds. *Comprehensive Analytical Chemistry*, vol. XLII. Amsterdam: Elsevier B.V. 2004, 129–226.

Kütükoğlu 1994: Kütükoğlu, M. S. *Osmanlı Belgelerinin Dili: (Diplomatik)*, Kubbealtı Neşriyatı 35. İstanbul: Kubbealtı, 2004.

Littmann 1953: Littmann E. "Ahlwardt, Wilhelm Theodor" – *Neue Deutsche Biographie* 1, 112 [Online-Version]; URL: https://www.deutsche-biographie.de/pnd11625744X.html#ndbcontent (last accessed 23 February 2018).

Müller 1967: Müller, R. O. *Spektrochemische Analysen mit Röntgenfluoreszenz*. München: R. Oldenburg. 1967

Petritsch 1991: Petritsch, E. D. *Regesten der osmanischen Dokumente im Österreichischen Staatsarchiv*. Mitteilungen des Österreichischen Staatsarchivs, Ergänzungsband 10/1. Wien: Österreichisches Staatsarchiv, 1991.

Redhouse 1890: Redhouse, Sir J. W. *A Turkish and English lexicon: shewing in English the significations of the Turkish terms*. Constantinople: The American Mission, 1890.

Sablatnig 2015: Sablatnig, R.; Hollaus, F.; Čamba, A. "The Centre of Image and Material Analysis in Cultural Heritage (CIMA) in Vienna and its Present Activities. Part II: Computer Vision." –*Palaeobulgarica* Vol. 39.1 (2015), 18–28.

Schaendlinger & Römer 1983: Schaendlinger, A. C. *Die Schreiben Süleymans des Prächtigen an Karl V., Ferdinand I. und Maximilian II. aus dem Haus-, Hof- und Staatsarchiv zu Wien*. Unter Mitarbeit von Claudia Römer. (Denkschriften der Österreichischen Akademie der Wissenschaften, phil.-hist. Kl., Bd. 163.) Wien: Verlag der Österreichischen Akademie der Wissenschaften, 1983.

Schaendlinger 1984: Schaendlinger, A. C. "Die osmanisch-habsburgische Diplomatie in der ersten Hälfte des 16. Jahrhunderts." *Osmanlı Araştırmaları / The Journal of Ottoman Studies* 4, 181–196.

Smith & Dent 2005: Smith, E.; Dent, G. *Modern Raman Spectroscopy*. A practical approach. Wiley, 2005.

About the Authors

Claudia Römer (claudia.roemer@univie.ac.at) is Associate Professor (1992) at the Department of Oriental Studies, University of Vienna. She completed her PhD in Turkish Studies (minor: Arabic Studies) in 1980. Among her fields of study are Ottoman documents, Ottoman social and economic history, Ottoman historical grammar, and Ottoman syntax.

Simon Brenner (sbrenner@cvl.tuwien.ac.at) is a research assistant and PhD student at the Computer Vision Lab, TU Wien.

Federica Cappa (f.cappa@akbild.ac.at) completed her PhD in 2014 at the University of Vienna, Faculty of Organic Chemistry. Since then she has been employed as Project Staff at the ISTA and a member of CIMA.

Bernadette Frühmann (B.Fruehmann@akbild.ac.at) is a trained physicist, did her PhD at the TU-Wien and is Assistant Professor at the ISTA. She is responsible for the non-destructive X-ray Fluorescence (XRF) analysis within the CIMA projects.

Ernst Hammerschmid (ernst.hammerschmid@me.com) is Senior Lecturer at the Academy of Fine Arts Vienna for photographic techniques and documentation of art works and takes care of the photographic documentation in the CIMA projects.

Manfred Schreiner (m.schreiner@akbild.ac.at) is Full Professor and Head of the Institute of Science and Technology in Art (ISTA), Academy of Fine Arts Vienna. He studied chemistry at the TU Wien, where he completed his PhD in materials science in 1979. After postdoc positions in the USA and Germany, he became full professor for Colour Science and Materials Science in Art at the Academy of Fine Arts Vienna. Since 2013 he has coordinated material analytical investigations by non-invasive techniques within CIMA.

Summary

An Ottoman firman of Suleiman the Magnificent (1520-1566), dated August 13, 1557, and kept at the Vienna Haus-, Hof- und Staatsarchiv (HHStA, Turcica 13, Konv.2 [1557 VII-XII], f. 33–39, Petritsch 1991: no. 351), was badly damaged by ink spilled over it by Walter Friedrich Adolf Behrnauer, librarian at the Vienna Hofbibliothek in 1860. Behrnauer had made things even worse by folding the document and handing it back to the clerk without telling anybody what had happened. The severe measures taken against him at the archive when this damage was found out are discussed, and the German text of the final document from the Vienna Haus-, Hof- und Staatsarchiv that was to close the "Behrnauer case" is given as an appendix to this article (HHStA Kurrentakten, HHStA Zl. 170/1860). Due to the damage, the text of the document could not be completely deciphered when it was part of a publication (transliteration, translation, and facsimile) of thirty-six Ottoman firmans of Suleiman the Magnificent from the Haus-, Hof- und Staatsarchiv in Vienna (Schaendlinger & Römer 1983). Some words, however, had become visible when a lamp was held at the back of the document.

In an attempt to recover all the text covered by the spilt ink, Multispectral Imaging (MSI) of the document was performed in the Computer Vision Lab, TU Vienna; additionally, Infrared reflectography was done at the Institute of Science and Technology in Art (ISTA) of the Academy of Fine Arts Vienna:

1. As expected, multispectral imaging with wavelengths ranging from 365–940 nm showed the best results in the IR range at 940 nm. Similar results were achieved by the IR reflectography done at ISTA and executed with an Osiris camera of Opus Instruments (www.opusinstruments.com) with wavelengths between 900 and 1700 nm. Both techniques have made the damaged text nearly fully readable.

2. By using X-ray Fluorescence (XRF) analysis, the original as well as the spilled inks could be identified and their composition determined. The analyses showed two distinct inks, with the undamaged Ottoman ink in the upper part of the document being identifiable as soot ink, and the possibility of iron gall ink for Behrnauer's ink as well as the Ottoman ink in the damaged parts of the document.

3. Raman spectroscopy did not show clear results, as the document was restored by applying a layer of resin, probably mastic, which covers the inks and obstructs further analysis.

4. Additionally, the pigments of the *tughra* and the composition of the "gold dust" used to dry the ink were analysed. It turned out that the *tughra* was painted with gold, whereas the "gold dust" did not contain any gold. Moreover, both the *tughra* and the "gold dust" were characterized by allying a stereo microscope in order to show the different structures of the materials.

ILLUSTRATIONS
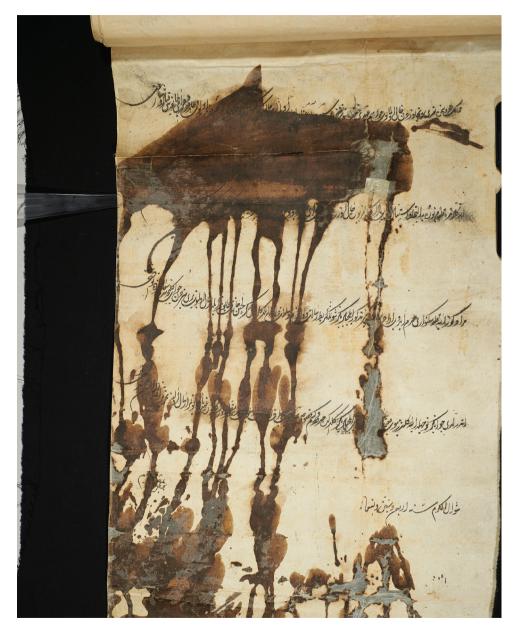
Illustration 1: TUK75-1_D4___white2

Illustration 2: TUK75-1_P1_940nm.

Illustration 3: The instrument ELIO for XRF analysis

Illustration 4: Portable confocal Raman spectrometer (Pro-Raman L-Dual-G) positioned above the Ottoman firman

Illustration 5: Image of IR-reflectography
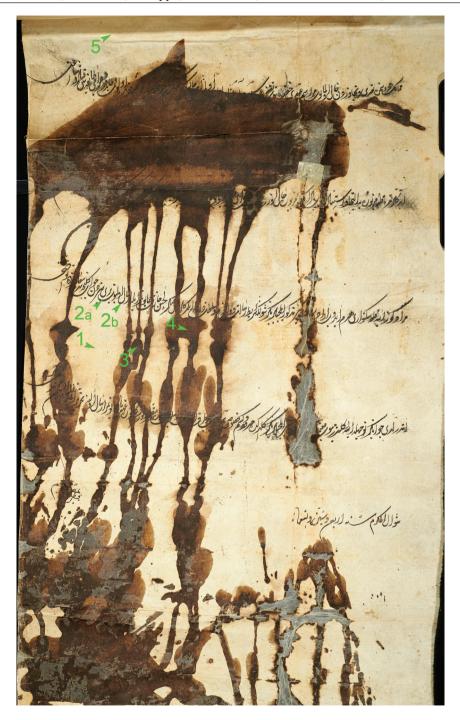
Illustration 6: The document with the areas for XRF measurements

Illustration 7: Süleyman's Tughra, Au, stereo microscope 40x, upside down

Illustration 8: Area of the Raman measurements for the Tughra

Illustration 9: "Gold dust" (rîk), stereo microscope 40x, upside down

Illustration 10: Area of the Raman measurements for the "gold dust" (rîk).

CONSERVATION TREATMENT OF MANUSCRIPTS AND CODICOLOGY: QUESTIONS AND CONSIDERATIONS

Angeliki Stassinou and Penelope Banou

Abstract: Codicology is the study of the physical structure of the book, a field that promotes a better understanding of book structures and their method of construction, which can also provide information about the place of origin, provenance and subsequent history of a book. In many cases conservation treatments can highlight the elements of a manuscript, but sometimes interventions intending to restore a manuscript to its original appearance unfortunately result in obscuring original elements significant to the study of the manuscript. For example, there are cases of codices restored by creating new bindings based on a hypothesis of the binding characteristics of a certain era and origin of production.

Should we consider a new approach to the conservation of manuscripts? Is it necessary for the codices to be restored in their original form? Should this be regarded as scientifically safe or ethical? How can we ensure that the researchers have access to the original material or recorded information? Should this information be in open access to the public?

Meticulous documentation of the binding structure, the parts of the codex, and their condition must precede any form of conservation treatment. Should manuscript catalogues or databases include more detailed descriptions, graphs or photos to enrich the core of the documentation?

Keywords: conservation, digitization, codicology, diplomatic, manuscript, archaeology of the book.

Introduction

In recent decades there has been a demonstrated interest in book archaeology and codicology, with numerous manuscripts included in digitization projects worldwide and available on the websites of libraries, archives and relevant collections. This tendency has provoked various discussions over the need for enriching the description of manuscripts in printed and digital catalogues.

The main purpose of this paper is to convey how book and paper conservators could complement research on codicology and book archaeology and could contribute to the value of manuscripts so they can be appreciated and promoted as cultural heritage objects. In this context, a number of questions and considerations are established about the effect of conservation treatments to the study of the original features of the manuscripts and the need for a new approach in matters of intervention.

Codicology as a discipline deals with the study of the handmade book as an archaeological object but also collects information about the date, place of origin, provenance and subsequent history of a book. Codicology mainly focuses on the physical aspects of codices and manuscript books: it examines the book's size, materials (such as the substrate, the writing media, the painting and drawing media), preparation of the writing surface, structure of the bookblock, sewing, bookbinding and rebinding, ex-libris inscriptions, and so forth, all of which could have changed over time and place. In so doing, it often determines characteristics specific to various production centers (scriptoria) and libraries (The Oxford Dictionary of Byzantium 1991).

So, the discipline of codicology aims to develop a better understanding of the structures and methods of construction of books, and to record the historical development of the physical production of manuscripts through time, since such structures vary according to time and place. Recent codicological studies put increasing emphasis on the book as a material vector of culture, unlike traditional palaeography, which tends to study the book in a cultural vacuum, focusing on the history of the scripts and the handwriting, as well as the archaeology of the manuscript folio, the textual and visual elements of the various layers of its central and marginal space (The Oxford Dictionary of Byzantium 1991).

Certainly, the text is the most significant part of a manuscript. However, the book has to be studied holistically as an entity, and the material structure that bears the text should not be ignored or underestimated. The binding characteristics and decorative elements are often related to the context of the manuscript, while the materials, genre and style of a manuscript create the whole picture of what it represents. Without a doubt the style of a manuscript is influenced by many factors – religious, social, financial, matters of taste and the purpose or intention of use. Manuscripts also bear significant intrinsic information and invaluable "fingerprints" of the crafts-

manship, fashion and technological, intellectual and scientific achievements of the era and place of origin. All of these automatically characterise manuscripts as cultural objects (Jokilehto 2005).

The "contribution" of book conservators

Conservators often have the fortune and privilege to put their hands on rare and valuable codices and manuscripts – representative samples of an era, a religion, a region, a kingdom, a political or military administration. Before treatment, conservators take notes of identification details and possible signs, stamps, notes or features, but they mainly have to record all the physical and structural elements of the authentic manuscript, its original characteristics and its condition. In many cases, conservators are the first to observe and study the authentic features in detail and to examine and investigate every single part of the original manuscript, which in many instances has not previously undergone any interventions. Depending on the condition of a manuscript or the application of certain practices, like local or total dismantling, during treatment, conservators can access and study parts of the manuscript which are not normally visible to readers or researchers in a library. Definitely, these parts – e.g., the binding thread and the sewing supports, lacing of the boards, materials and sizes of the spine linings, the endbands, the material of the boards, etc. – are not easily discerned after conservation treatments or the restoration of the manuscript.

Photos, sketches and graphical representation of the structure or the method of construction of every part, plus the attachment or the format of the several elements of the manuscripts, are performed by the conservators and are included in the condition reports of the manuscripts. Additionally, conservators in collaboration with conservation scientists and specialists perform non-destructive examination and use analytical methods for identification of the material and mapping of the extent of damage so as to determine the appropriate conservation treatment. Application of these techniques can reveal information, changes or parts not easily visible.

The kind of information that conservators can contribute to the research field of codicology or the archaeology of the book can involve:

● the writing substrate – i.e., whether it is papyrus, parchment or paper. If it is parchment, the type of animal skin and origin, method of processing and manufacture, surface preparation and possible coatings and relative

disposition of the hair and flesh side are recorded. If it is paper, the method of manufacture and relevant characteristics, weight and size of paper, colour, surface characteristics (texture or burnishing), watermark and possible provenance are recorded.

• the number of leaves used in a quire and the collation and variations within them

• the types of inks and colours (pigments and binding media)

• the manner of pricking and ruling, and whether these processes were conducted before or after the leaves were folded, one or more leaves at a time, or with the aid of a template

• the format of the book

• illumination, including the technique and methodology of application

• how a book was sewn and bound (type of binding structure and special elements such as materials and styles used for sewing and lacing, covers made of different kinds of leather, parchment, embroidery or precious metals and enamels, fastenings (clasps or ties) and furnishings (corner pieces, bosses, and other appendages), end leaves, embossed decoration, the method of construction and materials.

• endbands (the type of endband, the materials used for the primary and secondary support and the colour and thickness of the threads used)

• material of boards (wood, paper, etc.), method of attachment, bevelling of the edges, lacing paths, spine linings and their materials and dimensions

• rubrication, illumination and edge decoration

• the results of technical examination (such as photography with raking and transmitted light or microphotography, including applications in the non-visible range of the spectrum like multispectral imaging, UV reflectance and UV-induced fluorescence photography) and image and statistical processing, which provide significant information about the materials and the creation of a manuscript which are not evident with optical observation in normal light (e.g., the surface characteristics, preparation marks for the text, undertexts in palimpsests, faded inscriptions, preliminary drawings, etc.)

• the results of scientific analysis for the inks, pigments and binding media, painting layers, the preparation of the substrate and the illuminations or gilding.

Without a doubt, all the aforementioned provide significant information and shed considerable light on the original materials and structure of the manuscript and on the methods employed for its construction – data useful for research of manuscripts for scientists of other disciplines. Nevertheless, all this information is rarely accessible to researchers and, without a standard format for description, is difficult to search for.

Questions and considerations related to conservation treatments

Although data gathered by the conservators in a standardized form could contribute to the research of codicologists, there are matters arising from any conservation treatments applied in the past: Do conservation treatments affect the authenticity of a manuscript? Do conservation treatments, and in particular restoration, preserve all the original elements and intrinsic information of a codex? Could conservation treatments influence the intellectual and structural integrity of a manuscript? Could they also impede future research? Is it possible to perform scientific analysis and study of the materials with safe results after conservation treatment?

It is common knowledge that the extent of intervention of conservation treatments varies, depending on the approach adopted by a conservation studio, dictated by the curators of an institution or imposed by the taste of the owner of the manuscript. In many cases conservation treatments highlight the elements of a manuscript. On the other hand, when a manuscript is completely or partially restored, the original elements can be obscured and the style of the book may be altered. Isn't this crucial for the study of and research on the manuscript? Would the outcome of extended conservation treatment in certain cases be regarded as a counterfeit, with misleading aspects?

Reference to real studies justifies the aforementioned questioning but also raises more considerations. There are cases of codices where only the damaged body of the book block was saved without any remnants of the cover, the boards or the binding, or of any evidence of other particular elements, which were later completely restored, based on hypothesis, following the binding characteristics of the era and origin of production of the codex. Is this scientifically safe? Will codicologists or researchers of the future be misled by interventions like these? Is it ethically safe to apply extensive renovations when existing evidence is limited?

In other cases, manuscripts have been rebound following the fashion of the time, usually in a completely different manner than that of the era in which the codex was produced, the place of origin or the style used for such types of manuscripts. This kind of intervention has been justified as a practice, which is easily discernible, is not misleading and enables safe handling and adequate capture for digitization. From another point of view, could one suggest that rebinding in a different style interferes with the cultural interpretation of a manuscript.

Another matter that causes dilemmas and discussion is what treatment approach should be followed for a manuscript where parts of the original binding structure have been preserved. For example, in this case, the hand-written text block sewn on double hemp supports is still partially attached to the wooden boards in the lacing paths, with remnants of the spine lining exposed between the sewing supports and the leather covers adhered to the boards. There are also traces of the thread found in the fold of the quires, remnants of the end bands attached to the boards, etc. The embossed decoration on the leather cover, traces and indications of previous metal decoration on the cover and holes, remnants of the leather anchorage and metallic nails for the attachment of the clasps can be discerned.

Ideally, the purpose of conservation treatment should be to preserve and reinforce the binding structure without risking the integrity of the manuscript. In this way, our treatments do not interfere with the interpretation of a manuscript as a cultural object or with the study of the original elements.

In case the sewing is completely dismantled, should we resew the gatherings following the original sewing pattern? Should the endbands be restored? Should we restore the binding structure using the original parts (e.g., the wooden boards or the leather cover)? For the latter, could existing traces and features be removed, altered or hidden due to our interventions? What could be the effect of conservation treatment and materials application?

Ultimately, what is the aim and purpose of conservation and preservation? The protection, treatment and care of manuscripts. But all practices should ensure unobstructed study and research, encourage dialogue, fostering mutual respect, appreciation and understanding of the object. This paper does not intend to suggest a certain approach, but we wish to share our thoughts and concerns with the professional community of conservators and other scientists involved in the study of manuscripts. However, many

colleagues would agree that the current approach suggests that the best thing could be the storage of a manuscript in a box in optimum conditions, without the application of extended interventions (Clarkson 1999, Szirmai, 1989). In certain cases, the application of a limp paper or vellum binding that is simple, strong, durable, flexible and easily removable, and which cannot damage the textblock, could also be an arguable suggestion (Clarkson 1975, Szirmai, 1991).

Description of manuscripts in the era of digitization

Nowadays, access can be easily provided via digitized surrogates, but how can we ensure access to the physical aspects of the manuscript? A plethora of information concerning the physical characteristics of manuscripts is gathered by conservators but usually remains under limited access by only the conservation staff, often unknown even to colleagues of other disciplines working in the same library, archive or institution. These data might also be useful for researchers of the archaeology of the book and codicology, but, sadly, they become known only when included in articles published by conservators.

At the same time, looking at published catalogues of manuscript collections found in libraries all over the world, it is observed that although certain features of the manuscripts from a paleographical point of view are usually described, little attention is paid to the physical characteristics of the manuscript, with only brief description of the materials, dimensions, foliation, and collation. The same approach is adopted for the physical descriptions of digitized manuscripts currently displayed in databases available in the websites of generally recognized libraries, archives and institutions. Only in the case of exhibition catalogues are manuscripts described in a more detailed way.

Without a doubt, digitization provides easy access to manuscripts and documents of historical, artistic, cultural or generally significant value, from different periods of time, now held in institutions all over the world. Digital technology has opened a totally new perspective. Manuscripts are easily accessible from the place of our convenience, at any time of the day, even when the library, archive or museum is closed.

Libraries and archives are creating digital surrogates from their existing resources. Ambitious projects seek not merely to locate, catalogue and preserve manuscripts but also to enhance access, spread awareness and en-

courage their use for educational and research purposes. Should the data gathered by the conservators be included in the core of the physical description in the manuscript catalogue and relevant databases? The matter at stake is what purpose this could serve.

One of the main aims of digitization is to reduce the handling and use of fragile or heavily used original material and create a "back up" copy for endangered material. As Flavio Marzo remarks, "To justify restriction in handling objects, we need to improve the online metadata and the amount of information available with the surrogates"(Marzo 2018).

At the same time access to cultural-heritage material should not be provided simply as fast-track access to commercial or social web pages. We often get the impression that digitization acculturates the public to a shallow approach and undermines the value of the original object in people's mind. Access to cultural material should attract the attention of the viewer in an enlightening or even educating way. In this context, it could be suggested that information gathered during preservation and conservation projects should be included in the description record of digitized collections of manuscripts and be available in open access not only to researchers but also to the public, possibly through links so as to avoid extended descriptions.

To conclude, we believe that sharing the data and changing our attitude on treatment could benefit scientific research and could catalyze public appreciation and respect for manuscripts and codices as cultural heritage objects, with a constructive gain for future generations.

Bibliography

Clarkson 1975: Clarkson, C. "Limp Vellum Binding and Its potential As a Conservation Type Structure for the Rebinding of Early Printed Books :A Break with 19th and 20th Century Rebinding Attitudes and Practices." in *International Council of Museums Committee for Conservation, 4th Triennial Meeting*, *Venice,* 51/3, 1975, pp. 1-15.

Clarkson 1999: Clarkson, C. "Minimum Intervention in Treatment of Books." in *Pre-Prints, 9th IADA-Congress, Copenhagen, International Association of Book and Paper Conservators,* pp. 89–96.

Jokilehto 2005: Jokilehto, J. *Definition of cultural heritage, reference in documents in history*. ICCROM Working Group "Heritage and Society", January 2005.

Marzo 2018 : Marzo, F. "Digitising books as objects: The invisible made visible." in *The British Library Collection Care Blog, https://blogs.bl.uk/collectioncare/2018/02/digitising-books-as-objects-the-invisible-made-visible.html*

Szirmai 1989: Szirmai, J. A. "Stop Destroying Ancient Bindings." *The Abbey Newsletter, vl. 13, nr5*, Sept 1989, reprinted with permission from *Gazette du Livre Médievale*, No. 13, Fall 1988, pp. 7-9 (editorial offices at CNRS-CEMAT, 156, avenue Parmentier, F-75010 Paris, France).

Szirmai 1991: Szirmai, J. A. "Old bookbinding techniques and their significance for book restoration." in *Preprints, 7th International Congress of IADA,* Uppsala, 26-30 August 1991.

The Oxford Dictionary of Byzantium, vol. 1, Kazhdan Alexander (ed.) Oxford University Press, Inc., 1991.

About the authors
Angeliki Stassinou
e-mail address: astassinou@yahoo.com
Angeliki Stassinou graduated from the Department of Conservation of Antiquities and Works of Art in the Technological Educational Institute (T.E.I.) of Athens in 1995 and specialized in book and paper conservation after her Master of Arts at Camberwell College of Arts, The London Institute, London (1996). Since 2004, she has been a permanent member of staff in the Conservation Department of the General State Archives of Greece, where she designed and set up the conservation studio and works on library and archive material. She had been a lecturer of the Department of Conservation (T.E.I of Athens) for a decade, has published papers on archive and limp bindings and participated in several European projects.
Penelope Banou
e-mail address: pbanou@yahoo.gr
Penelope Banou belongs to the permanent staff of the Conservation Department of the General State Archives in Athens since 2008. She graduated from the Department of Conservation of Antiquities and Works of Art in the Technological Educational Institute (T.E.I.) of Athens (1996) and specialized in conservation of works of art on paper after her postgraduate studies, a Master of Arts in Conservation of Fine Art at the Northumbria University in UK (1998). Ever since, her professional activities include participation in

preservation and conservation projects of works of art on paper and archival material collections belonging to public and private collections, while she is involved in education (as a lecturer in the Department of Conservation, T.E.I of Athens) and in research programs, with several publications.

OCS BIBLICAL LANGUAGE IN
THE ERA OF THE NEW TECHNOLOGIES

Anna-Maria Totomanova

Abstract: This paper describes the history and potential of the Histdict system (http://histdict.uni-sofia.bg/) that was created for producing an online historical dictionary of Bulgarian. Within the last two years Histdict system tools were used for producing other types of online historical dictionaries: an electronic dictionary of the language of Patriarch Euthymius and an online terminological dictionary of the language of John the Exarch. The problem of integrating the new entries into the *Historical Dictionary of Bulgarian* is also addressed.

Key words: OCS, Slavia Orthodoxa, Diachronic Corpus of Bulgarian Language, search engine, specialized software for the Historical Dictionary of Bulgarian, morphological annotation, specialized dictionaries

The history of the Eastern Orthodox community known as *Slavia Orthodoxa* started with the translation of the Gospels made by St. Cyril and the creation of a new *lingua sacra* based on the linguistic features of the Bulgarian Slavonic dialect of Thessaloniki, where the brothers Cyril and Methodius were born and grew up. After the end of the Moravian mission (863-886) the Cyrillo-Methodian language was resurrected and developed as an official liturgical language on Bulgarian soil and very soon spread out among the Eastern and Southern Slavs that converted to Christianity between the 9th and the 12th centuries. In the next centuries Old Bulgarian (known also as Old Church Slavonic or "OCS") functioned as a common literary/bookish language of all Orthodox Slavs and of Romanian populations in Medieval Walachia and Moldova up to modern times, developing two regional recensions – Russian and Serbian. This is the language of OCS bibles that served as a model for the entire literary production of Slavia Orthodoxa in the Middle Ages. The late Russian recension known as Church Slavonic survived as a language of the printed liturgical books and in the 18thand 19th centuries played a crucial role in shaping the modern languages of Russians,

Bulgarians and Serbs. Even today there are writers and poets who compose their works in Church Slavonic.

Given the great importance of Old Bulgarian to both the history of the Slavonic Bible and to the cultural and religious history of modern Russians, Bulgarians and Serbs, ten years ago we decided to facilitate the work of researchers by introducing modern digital technologies in such a conservative field as diachronic linguistics and designed our first project entitled "ICT Tools for Historical Linguistic Studies", which was funded by the European Social Fund as BG051PO001-3.3-04-0011 in 2009. Our goal was twofold:

• to speed up data collecting from the Slavonic medieval books created between the 10th and 18th centuries and accelerate further data processing

• to make diachronic linguistics more attractive for young people born in the Computer Age for whom computers are a part of their natural habitat

In fact we planned to design and construct a system that will enable us to produce a web-based *Historical Dictionary of Bulgarian*. The goal was rather ambitious since ICT approaches had not been applied to OCS before. Therefore starting from scratch we faced a lot of problems that we were not able to foresee. As you can figure out, we solved them using the old trial-and-error method. Very soon however we realized that in order to pursue and achieve our objectives we'll need additional time and funding, and so the first project was followed by three other projects following on the first:

• **BG051PO001-4.3.04-0004** E-Medievalia (Electronic Resources for Distant Learning in Medieval Studies) (2012-2014)

• **BG051PO001-3.3.06-0024** Informatics, Grammar, Lexicography (2012-2015)

• **BG05M2OP001-2-009-0005** Modern Palaeoslavonic and Medieval Studies (2017-2018)

Detailed information about all the projects can be obtained on the web portal *Cyrillomethodiana* http://cyrillomethodiana.uni-sofia.bg.

1. *Cyrillomethodiana web portal*

My contribution will dwell on both our achievements and failures as well as on our plans for the future.

Now almost ten years after the beginning of the first project we have at our disposal a unique electronic system for digital processing of Medieval and Early Modern Bulgarian texts of the 10th to 18th centuries.

2. *Histdict System*

Though the system was designed for producing a web-based *Historical Dictionary of Bulgarian,* in fact it allows for and facilitates the studying, adequate presentation and popularization of the most important part of Bul-

garian cultural and historical heritage: Bulgarian language and mediaeval literature in general.

The system includes a ***Diachronic Corpus of the Bulgarian Language*** and the respective electronic tools for processing linguistic data. The corpus contains more than 150 texts of proven Bulgarian origin from different genres of the 10th to 18th centuries and has its own specific software that enables textual and palaeographic annotation. Given the fact that Bulgarian literature transmitted the Byzantine cultural and literary model to the other Orthodox nations in our part of Europe, the corpus contains both translated and original Medieval Bulgarian texts. The texts are digitally typed, and all of them (excluding the works of St. Kliment Ohridski) are reproduced in the orthography in which they survived (Bulgarian, Serbian or Russian).The corpus also includes Early Modern Bulgarian texts – mostly Damaskini and other compilations as well as some non-literary texts such as scribal notes, inscriptions and juridical documents. Back in 2011 we had only 75 texts, whereas now their number has doubled, and we continue uploading new texts. Some of them are provided from colleagues abroad who also use the corpus. One of the largest texts uploaded onto the system is the so-called Chronograph of the Archive that includes the oldest text of the Octateuch and Kingdoms translated in Bulgaria during the reign of Simeon the Great (893-927). Each text is introduced by the respective archaeographic data (title, source, dating, orthography, author etc.).

3. *Diachronic Corpus*

For compiling the corpus we needed not only the respective software but also specialized Unicode (UTF-8) OCS fonts to make the texts readable online. In order to avoid the retyping of already keyboarded texts the new fonts were installed in a converter that allowed for the conversion of the texts to the new fonts. Now we use the third version of the converter, which includes 3 specialized fonts with different design: *CyrillicaBulgarian10U, Cyrillica-Ochrid10U* and *CyrillicaOldStyleU.* The last one is meant for typing Early Modern Bulgarian texts. In the beginning we were able to convert only the fonts of the families *Cyrillica Bulgarian, Cyrillica Ohrid* and *Cyrillica Shafarik*, developed by *Synthesis Soft* company and widely spread among the Palaeoslavonic research community, to which gradually other OCS and Ancient Greek fonts were added: the Italian *PopRetkov;* the Unicode font *BukyVede*, which is also based on Synthesis Soft design; *TimesGreekClassic* and *TimesGreekOld*, also developed by the same company, converted now into Palatino; as well as all varieties of the modern font *TimesCyrillic* in *Times New Roman.* Our OCS fonts and the converter spread throughout Europe since they were compatible with the editorial software programs, and a series of books and periodicals in both Bulgaria and abroad are using them. Some colleagues of ours created a special blog for them.

Under the first project we managed to digitize 11 000 entries of the Old Church Slavonic Dictionary, compiled and published by IBL for the Bulgarian Academy of Sciences because we wanted to use it as a basis for the Historical dictionary of Bulgarian. The last 500 entries were digitized under the second project since they do not comply with the entire format of the other entries or contain different types of errors and had to be processed one by one. Thus now we have a full electronic version of this dictionary that is available online and contains the lexis of classical OCS manuscripts of the 10 and 11th centuries.

4. *Digitized Dictionary*

The structured XML document that included all entries was used to produce a reversed Greek-Old Bulgarian Dictionary, which is also available online and is widely used by colleagues who deal with the metaphrastic practices of men of letters writing in Old Bulgarian. The reversed dictionary was elaborated under the E-Medievalia project that produced an interactive teaching-learning platform in Medieval Studies. The platform contains 24 interactive courses for both under- and postgraduate students in the areas

of philology, philosophy, theology, history and arts, among them courses on Old Bulgarian (OCS), Old Bulgarian literature and the history of Bulgarian, which are available also in English.

5. *Reversed Greek-Old Bulgarian dictionary*

The design of the historical-dictionary software turned out to be the biggest challenge for the project team and especially for our ICT specialists since, as mentioned above, we decided to edit and upgrade the digitized version of the Old Bulgarian Dictionary, complementing it with new entries and new meanings in the old entries. Neither linguists nor ICT specialists were

of the Corpus. The temporary solution came from our main software specialist, who digitized the OCS dictionary and developed the software for the historical dictionary. The search engine is installed in the Histdict system, and one can search across all its entries (from the two dictionaries, the corpus and the chronograph) or from just one of them. The search engine shows results with snippets of context; to find a result together with the respective full context one has to use the browser internal search functionalities.

7. *Search engine in histdict.uni-sofia.bg*

A virtual keyboard helps to enter the string to search for.

8. *Virtual Keyboard*

In the meantime we continued our work on elaborating a grammatical dictionary of Medieval Bulgarian (OCS) in order to produce tools for morphological annotation of the texts. The first step in this direction was the compilation of the tagset of OCS, which contains 2 200 tags and describes the complicated grammar of OCS.

9. *Tagset of OCS*

Pp1lp	Pronoun, personal, first person, locative case, plural
Pp2op	Pronoun, personal, second person, nominative case, plural
Pp2gp	Pronoun, personal, second person, genitive case, plural
Pp2dpl	Pronoun, personal, second person, dative case, plural, full form
Pp2dpt	Pronoun, personal, second person, dative case, plural, schort form
Pp2ap	Pronoun, personal, second person, accusative case, plural
Pp2ip	Pronoun, personal, second person, instrumental case, plural
Pp2lp	Pronoun, personal, second person, locative case, plural

2 лице, дв. ч.

Pp1ou	Pronoun, personal, first person, nominative case, dual
Pp1au	Pronoun, personal, first person, accusative case, dual
Pp1Gu	Pronoun, personal, first person, genitive/ locative case, dual
Pp1Dul	Pronoun, personal, first person, dative/ instrumental case, dual, full form
Pp1dut	Pronoun, personal, first person, dative case, dual, schort form
Pp2Ou	Pronoun, personal, second person, nominative/ accusative case, dual
Pp2Gu	Pronoun, personal, second person, genitive/ locative case, dual
Pp2Dul	Pronoun, personal, second person, dative/ instrumental case, dual, full form
Pp2dut	Pronoun, personal, second person, dative case, dual, schort form

III.2. Възвратно лично местоимение

Ppxas	Pronoun, personal, reflexive, accusative case, singular
Ppxgs	Pronoun, personal, reflexive, genitive case, singular
Ppxdsl	Pronoun, personal, reflexive, dative case, singular, full form
Ppxdst	Pronoun, personal, reflexive, dative case, singular, schort form
Ppxis	Pronoun, personal, reflexive, instrumental case, singular
Ppxdsl	Pronoun, personal, reflexive, locative case, singular

III.3. ПОКАЗАТЕЛНИ МЕСТОИМЕНИЯ
Мъжки род

Pdmos	Pronoun, demonstrative, masculine, nominative case, singular
Pdmgs	Pronoun, demonstrative, masculine, genitive case, singular
Pdmas	Pronoun, demonstrative, masculine, accusative case, singular

The tagset was followed by a grammatical dictionary that includes paradigms of the inflectable words taking into account different phonetic and orthographic variations. The paradigms (rules) were ascribed to all inflectable words in the historical dictionary, and as a final step the grammatical dictionary was installed into it. Now by clicking the sign + located next to the dictionary entry title, the user can see the whole paradigm of the respective word.

10. *Grammatical Dictionary*

Based on the grammatical dictionary, our ICT specialists created a prototype of the morphological annotator, which is also an open-access tool on our site.

11. *Morphological annotator*

paradigms depending on the verb type and so far have produced the participial forms of the verbs of the first two OCS conjugations (i.e., 894 verbs with over 3 million forms). It is up to our ICT specialists now to find out how to include them into the dictionary and the search engine. The grammatical dictionary upgrade made us realize the need to edit the lemmas and the grammatical definitions in the historical dictionary, which seems to turn into a constant process.

Our running project explores the potential of electronic tools and resources for producing specialized web-based dictionaries: a terminological dictionary of the Old Bulgarian language and, as a tribute to one of the most prominent Bulgarian spiritual leaders, a dictionary of the writings of St. Patriarch Euthymius of Tarnovo.

Since the specialized dictionary software seemed too complicated to the new project participants, especially PhD and postdoc students who had no experience working on the previous projects, we decided to compile the new dictionary entries in a Microsoft Word environment and include them into respective specialized dictionaries through a converter. For this purpose each working group elaborated a special entry format to be recognized by the converter. The converter for St. Patriarch Euthymius' dictionary turned out to be a great finding because it enables us to convert more than 400 dictionary entries in few seconds.

12. *Dictionary Entries Converter*

The first part of St. Patriarch Euthymius' dictionary (from ⰰ to ⱀ) is already available online, and soon a paper version will follow. The production of the second part/volume will be supported by the National Research Programme "Cultural Heritage, National Memory and Social Development", funded by the Bulgarian Ministry of Education and Science.

The converted dictionary entries allow for online redaction using the specialized software originally developed for the historical dictionary.

13. *Patriarch Euthymius' dictionary online*

For compiling the dictionary entries, an alphabetical list of all word forms in Patriarch Euthymius' writings contained in the corpus was produced electronically. The list shows the title of the opus and the exact place the forms occur and allows a quick work lemmatization.

A similar list for the terminological dictionary was also made available, but given the complexity of the task the working group decided to limit the linguistic material and explore the terminology of John the Exarch's *Hexaemeron* and *Theology*. As a result this working group identified around 1000 terms to be included into the specialized dictionary. The terminological dictionary is also available online and will have a paper copy as well.

14. *The terminological dictionary online*

The lexical material from Patriarch Euthymius' dictionary entries has already been integrated into the historical dictionary. This process helped us to reveal the flaws of the specialized software we produced for the historical

dictionary of Bulgarian. Our decision to focus on adapting the dictionary entries converted for the historical dictionary as well will optimize and facilitate the production of new entries.

15. *New material in the historical dictionary*

Inspired by the success of the reversed Greek-Bulgarian dictionary, some project team members dedicated themselves to the production of a reversed Greek-Church Slavonic dictionary based on the Church Slavonic dictionary by Atanasiy Bonchev that contains a larger lexical corpus (over 24 000 lexemes) than the Old Bulgarian dictionary (Bonchev 2012). Finan-

cial support was provided by the Zograph Heritage Foundation that also funded the paper edition of the dictionary (Christov 2019). As agreed, a year after its publication an electronic version will be uploaded into the Histdict system.

As you can see we continue to expand the functionalities of the system in order to produce different type of dictionaries as well as to present and popularize Bulgarian cultural heritage – Bulgarian language and religious culture. Within the last five years since the system became operational, in our department for Cyril and Methodius studies alone, 5 PhD theses on different topics were successfully defended and a series of books on medieval history and language published. Colleagues from both Bulgaria and abroad are using the corpus and the dictionaries and support us in our pioneering work. We obtained positive results on an educational level as well: for the last five years, due to the E-Medievalia platform, the number of students choosing the state exam in the history of the Bulgarian language increased five times. Last year almost 80% of the graduates selected this field.

The Histdict system resources have also been recently included in a new effort, Research Infrastructure on Religious Studies (ReiReS) – an EU funded project under Horizon 2020 that is considered to be a preparation for inclusion in the EU roadmap in the near future.

References

Bonchev 2012: Арх. Бончев, А. *Речник на църковнославянския език*. Т. I -А-О, С., 2002, Т. II П-Я, С. 2012.

Miklosich 1876: Miklosich, F. *Die christliche Terminologie der Slavischen Sprachen: Eine sprachgeschichtliche Untersuchung von Franz Miklosich. Denkschriften der kaiserlichen Akademie der Wissenschaften*. Philosophisch-Historische Klasse. Band 24. Wien, 1876.

Christov 2019: *Гръцко-църковнославянски речник*. Съставен от Иван Христов въз основа на *Речника на църковнославянския език* от архимандрит д-р Атанасий Бончев. Редактор А. Тотоманова, София 2019, ISBN 978-954-770-451-0.

About the author:

Anna-Maria Totomanova is professor of the history of the Bulgarian language at Sofia University. Her research interests focus on the fields of historical phonetics, morphology and morphonology, lexicology and lexicography. She has published and studied several important Medieval Bulgarian texts such as the *Vatican Palimpsest*, the *Synodikon of Tzar Boril*, the Slavonic version of the so-called *Chronicle of George Synkellos* and *Stanislav's Menaion Reader.* During the last decade she has been involved in a series of projects aiming at introducing ICT into the field of medieval Slavonic studies.

Abstract

The paper describes the history of producing the unique Histdict system, which was designed for producing an online historical dictionary of the Bulgarian language. The system was structured and equipped with the respective electronic tools during the last decade, with the support of several projects funded by European Structural Funds. Now the Histdict system includes a diachronic corpus of Bulgarian language (10th–18th centuries), the digitized dictionary of Old Bulgarian (OCS), a historical dictionary of Bulgarian and two specialized historical dictionaries – the dictionary of the language of Patriarch Euthymius of Tarnovo (14th century) and a terminological dictionary based on John the Exarch's opera (10th century). The corpus contains more than 150 keyboarded medieval texts of proven Bulgarian origin. Each text is preceded by metadata including title, provenance, author, dating, orthography etc. The software of the corpus allows for introducing different types of comments. With few exceptions all texts are reproduced in their original orthography (Old or Middle Bulgarian, Russian and Serbian). The historical dictionary entries follow the structure of the Old Bulgarian dictionary and its software. This allows for editing the entries of the Old Bulgarian dictionary through introducing new meanings and examples and creating new entries, but it proved to be rather complicated. That is why the new project designed and produced a converter that enables the researchers to work in a Microsoft Word environment, sticking to a certain formatting. The converter was tested during the production of the two specialized dictionaries, and based on that the project participants decided to use the converter for producing and editing the entries for the historical dictionary. The latter will require additional testing and upgrading of this unexpected but valuable tool. The online dictionaries of Patriarch Euthymius' language and of John the Exarch's terminology are an attempt to expand the function-

alities of Histdict system and face new research challenges. In April 2019 the paper versions of the two specialized dictionaries were produced and made available to specialists in Palaeoslavonic and Medieval studies. The diachronic corpus and historical dictionary are constantly developing resources, and their completion with new texts and entries is an ongoing process. A grammatical dictionary of Old Bulgarian (OCS) is installed into the historical dictionary, and based on that the system now is equipped with a semi-automatic morphological annotator that recognizes grammatical features of all inflectable words in the historical dictionaries. Recently the grammatical dictionary was complemented by the tags and forms of the uninflectable words, and the testing of an automatic annotator has already begun. The Histdict system is equipped with a virtual keyboard and a search engine that allow for exploring all its resources apart from the Old Bulgarian dictionary. Though the system was meant to serve the production of a historical dictionary, now it is a complex and sophisticated tool for studying and publishing medieval Slavonic texts. The system turned out to be an excellent resource for presenting Bulgarian written heritage to a broader audience.

ANALYSIS OF BINDING STRUCTURE
AND PAPER IN VALVASOR'S GLORY

Jedert Vodopivec Tomažič

Abstract

Die Ehre dess Hertzogthums Crain ("The Glory of the Duchy of Cariola"), printed in Nüremberg in 1689, is the masterpiece of the polymath Johann Weichard Valvasor. This encyclopaedic work is one of the pillars of Slovenian history and culture, appreciated by many disciplines, from the natural and social sciences to the arts and humanities. Although Valvasor's work was the object of extensive research, some aspects remain unanalysed, among them the bookbinding structure as well as the paper of this important book.

Fourteen copies of volume III of *Die Ehre dess Hertzogthums Crain* held in Slovene and Croatian institutions, were selected and examined both in terms of binding and paper. More in-depth analyses were carried out on Valvasor's personal copy, held in the Metropolitan Library Zagreb, and on a damaged volume held in the library of the Slovenian Academy of Science and Arts in Ljubljana.

Key words: Valvasor, book, *Die Ehre dess Hertzogthums Crain*, paper, binding, structure.

1. Introduction

Die Ehre dess Hertzogthums Crain ("The Glory of the Duchy of Carniola"), hereinafter *Die Ehre*, is the masterpiece of the polymath Johann Weichard Valvasor.[1] This encyclopaedic work is one of the pillars of Slovenian history and culture, appreciated by many disciplines, from the natural and social sciences to the arts and humanities (Golec 2013). Although Valvasor's *Die Ehre* was the object of extensive research, some aspects remain unanalysed, among them the technical background as well as the material and physical structure of this important book. There are no archival sources clarifying the details regarding the printing and publication process of *Die Ehre*, so other means of acquiring this information would be necessary.

---

[1] Johann Weichard Valvasor, *May 28, 1641, Ljubljana Slovenia, †September 1693, Krško, Slovenia.

Such an opportunity arose when a badly damaged copy of volume III[2] was sent to the Book and Paper Conservation Centre of the Archives of the Republic of Slovenia (BPCC ARS).

A conservation intervention usually offers an opportunity to examine in great detail the structure and material of the object being treated. A comparison of the physical characteristics of other existing copies is important before undertaking any conservation-restoration intervention on a printed book, particularly in relation to evaluating appropriate treatments for the binding. On this particular occasion the decision was accepted that in addition to the conservation treatment, which is fully documented and published both in written and photographic form,[3] a survey of other known copies of the work would be undertaken for the purposes of comparison along with some scientific analyses which are in details presented elsewhere (Vodopivec et al 2014).

As the conservation treatment was to be carried out on one copy of volume III of *Die Ehre*, our examination focused on other copies of the same volume. Since volume III contains more illustrations than the other volumes, it was always of particular interest to the general public. Due to extensive use, volume III suffered the most damage, and this fact made the physical structure easily visible.

In total 14 copies of volume III, printed in 1689, were selected and examined. More in-depth analysis was carried out on the conserved volume and on the copy from Valvasor's personal library, today kept at the Metropolitan Library in Zagreb (M 11249 T3).[4] This research and the related discussion and conclusions are presented here.

---

[2] Milko Kos' Historical Institute of the Scientific Research Centre of the Slovenian Academy of Science and Arts in Ljubljana (Library, SAZU II 11161/3 H).

[3] Archives of Book and Paper Conservation Centre, Archives of the Republic of Slovenia, no.: 11/68.

[4] In 1690, Aleksandar Ignacije Mikulić, the bishop of Zagreb, bought Valvasor's library and print collection (probably through Pavao Ritter Vitezović as an intermediary). The collection has been preserved in its entirety in the Metropolitan Library in Zagreb. Since 2010, the Library and the Archbishop's archives have been a part of the Croatian State Archives in Zagreb.

2. Scope of the Research

The research focused on the structural elements of the original binding in order to identify which elements of the binding were original and to establish whether a uniform type of binding had existed. This information would then determine to what extent the conservation treatment should follow the type of binding in the preserved original copies as opposed to that of the rebound book, if at all. The research also included the structure of the textblock[5] and some properties of the paper used for printing the textblock. Papers on which the supplements were printed and papers used for endleaves were not a subject of this examination.[6]

Selection of the Copies of Volume III for the Survey

A search of the Co-operative Online Bibliographic System and Services (COBISS) database showed that at least one copy of the first edition of *Die Ehre* is kept in 22 libraries in Slovenia[7] and, in addition, according to Majda Oražem Stele's records from 1989, copies are held in 76 European and American libraries. (Oražem Stele 1990, 254–256)

For this study, 14 copies kept in Slovenia and Zagreb[8] were examined. They are presented in table 1 and listed below:

1. Bogenšperk, Castle Institute**,** JZB No 56[9]
2. Ljubljana, Archives of the Republic of Slovenia, Library, AS, P 1015
3. Ljubljana, Franciscan Monastery Ljubljana, FSLJ 10 al.[10]
4. Ljubljana, National Museum of Slovenia, NMS 1625/3
5. Ljubljana, National Museum of Slovenia, NMS 1610/3

---

[5] http://www.ligatus.org.uk/lob/concept/1663.

[6] The examination and research of papers used for endleaves and the printing of full-page supplements would be a complex and extensive task and as such are beyond the scope of this research project and are omitted from this research.

[7] http://www.cobiss.si/scripts/cobiss?ukaz=getid&lani=en.

[8] *Die Ehre* from Valvasor's private library now kept by the Metropolitan Library in Zagreb.

[9] Donated in 1993 by Vanja Lokar from Trieste**.**

[10] In her doctoral thesis *Knjižna zbirka p. Žige Škerpina (1689–1755) v ljubljanskem frančiškanskem samostanu* (= Žiga Škerpin's book collection in the Franciscan Monastery, Ljubljana), Sonja Svoljšak states that in 1735, the then guardian Mariašič arranged for the binding of the books purchased by that year (including *Die Ehre*); the chronicler, however, does not mention any names of the bookbinders active in Ljubljana at the time, 2007, 52 and 169.

6. Ljubljana, National and University Library, NUK R II. 6632/3[11]

7. Ljubljana, National and University Library, NUK R II. 361 643/3

8. Ljubljana, National and University Library, NUK R 117455/3

9. Ljubljana, Slovenian Academy of Science and Arts, Library, SAZU TR, II. 713 TR[12]

10. Ljubljana, Slovenian Academy of Science and Arts, Library, SAZU U – II 6462 U

11. Ljubljana, Slovenian Academy of Science and Arts, Library, SAZU H – II 11161/3 H[13]

12. Ptuj, Ivan Potrč Library Ptuj, KIP N-16214D[14]

13. Zagreb, Croatian Academy of Science and Arts, Library, HAZU R-910[15]

14. Zagreb, Metropolitan Library Zagreb, MKZg M 11249 T3[16]

Working Method

All selected copies were examined in relation to the structure of both the binding and the textblock paper. The results are presented in table 1.

The damaged copy and the Valvasor's personal copy are described in more detail in later sections.

---

[11] The copy is from the Lyceum Library in Ljubljana.

[12] The copy is from the Robida's collection (Adolf Robida, active as writer, translator and critic, *June 14, 1885, Ljubljana Slovenia, †February 29, 1928, Ljubljana, Slovenia).

[13] The conservation treatment was carried out on this copy by the CCR ARS (No. 11–68).

[14] The copy originates from Dornava castle library.

[15] The copy from the collection of Dr Ivan Kukuljević Sakcinski (1816–1889), a prominent Croatian historian and politician.

[16] The Valvasor's personal copy with his ex libris.

Table 1: Comparison of copies of volume III of *Die Ehre dess Hertzogthums Crain* ("The Glory of the Duchy of Carniola"), 1689 Nurmberg, W.E. Endter

Institution Library call number Date of examination Provenance Date of the binding Photo	Size (mm) **A:** binding **B:** front page **C:** layout on page chapter IX, p.44 **D:** thickness of paper	Structure, specific features: **TB:** texblock **Damage grade:** I (very good), II (good), III (bad), IV (very bad)
1 **Bogenšperk Castle Institute** **No. 56** 7 May 2013 *From the collection of Vanja Lokar, Trieste* *Rebinding, probably late 19th century*	A: 330 × 210 × 95 B: 320 × 208 C: 291 × 172 D:0.114	*Covering material:* tanned skin on spine, marbled paper sides *Covering decoration:* blind tooling and gilt decoration *Board material:* paste-laminate board *Endbands:* no *Endleaves::* not original (contemporary with the last intervention on the binding) *Edge decoration:* bluish *No. of sewing supports:* 5, double cords *Opening:* flat (nearly180°) *Watermark in TB paper:* see Fig. 2b *Feel of paper:* slightly soft *Sound of paper:* intense *TB damage grade: II (wear, tears, staining)* *Binding damage grade: III (torn, spine folds)* All four volumes bound in the same manner, in tanned leather on spine and corners, marbled paper sides

Institution	Size (mm)	Structure, specific features:
Library call number		
Date of examination	**A:** binding	**TB:** texblock
	B: front page	
Provenance	**C:** layout on page chapter IX, p.44	**Damage grade:**
Date of the binding		I (very good), II (good), III (bad), IV (very bad)
Photo	**D:** thickness of paper	
2	A: 340 × 220 × 100	*Covering material:* tanned skin
Ljubljana, ARS, P 1015, 4°		*Covering decoration:* slightly sprinkled leather
	B: 333 × 205	*Board material:* paste laminate boards
22 December 2011		*Endbands:* yellow-white, stuck-on endband
and 7 August 2013	C: 291 × 174	*Endleaves:* contemporary with the binding
		Edge decoration: slightly sprinkled edges (orange and brown)
Binding contemporary	D: 0.126	*No. of sewing supports:* 5, double cords
with the textblock		*Spine:* slightly rounded, 5 raised bands
		Opening: flat (nearly 180°)
		Watermark in TB paper: see Fig. 2b, visible in 20 leaves
		Feel of paper: compact and soft
		Sound of paper: not intense
		TB damage grade: II-III (wear, tears, minor foxing),
		TB supplements: wear, tears
		Binding damage grade: I-II (wear)
		All four volumes bound in the same manner, in tanned leather

Institution Library call number	Size (mm)	Structure, specific features:
	A: binding	**TB:** texblock
Date of examination	**B:** front page	
	C: layout on page chapter IX,	**Damage grade:**
Provenance	p.44	I (very good), II (good), III (bad), IV (very bad)
Date of the binding		
	D: thickness of paper	
Photo		
3	**A:** 338 × 208 × 130-135	*Covering material:* alum-tawed skin
Ljubljana FSLj. 10a 1		*Covering decoration:* blind stamped
		Board material: wood
26 February 2013	**B:** 326 × 200	*Clasps:* 2 fastenings (on the fore-edge)
7 August 2013		*Endbands:* blue-white, stuck-on endband
	C: 291x 173	*Endleaves:* contemporary with binding, watermark and orientation of
From the collection of		chain lines different from textblock
Luka Vodnigk		*Edge decoration:* sprinkled in red-brown colour
	D: 0.100	*No. of sewing supports:* 5, double cords
Binding from mid 18th		*Spine:* slightly rounded
century		*Opening:* restricted as vols. III and IV are bound together
(see note 15)		*Watermark in TB paper:* see Fig. 2b
		Feel of paper: compact (stiff)
		Sound of paper: intense
		TB damage grade: I(no stains, no foxing)
		Binding damage grade: I
		Bound in two volumes, binding not contemporary with TB **Škerpin** type of binding (mid 18th century) in alum-tawed skin

Institution Library call number *Date of examination* *Provenance Date of the binding* *Photo*	Size (mm)	Structure, specific features:
	A: binding **B:** front page **C:** layout on page chapter IX, p.44 **D:** thickness of paper	**TB:** texblock **Damage grade:** I (very good), II (good), III (bad), IV (very bad)
4 **Ljubljana, NMS 1625/3** 19 December 2011 From Augustin monastery in Ljubljana, 1731 *Binding probably contemporary with the textblock*	A: 330 × 230 × 80-95 B: 323 × 188 C: 291 × 173 D: 0.120	*Covering material:* alum-tawed skin *Covering decoration:* decorative blind tooling *Board material:* wood *Endbands:* brown-white, stuck-on endband *Endleaves:* 19th century marbled paper *Edge decoration:* blue *No. of sewing supports:* 5, double cords *Opening:* restricted (book supports necessary) *Watermark in TB paper:* see Fig. 2b; well visible in the title page *Feel of paper:* uniform *Sound of paper:* partly intense *TB damage grade:* I *Binding damage grade:* I-II Binding probably contemporary with TB, all four volumes bound in the same manner

Institution Library call number *Date of examination* *Provenance* *Date of the binding* *Photo*	Size (mm) **A:** binding **B:** front page **C:** layout on page chapter IX, p.44 **D:** thickness of paper	Structure, specific features: **TB:** texblock **Damage grade:** I (very good), II (good), III (bad), IV (very bad)
5 **Ljubljana, NMS 1610/5, 6** 19 December 2011 Binding probably late 19th century	A: 344 × 217 × 65 B: 330 × 208 C: 291 × 172 D: 0.125	*Covering material:* grey bluish paper *Covering decoration:* none *Board material:* paste laminate board *Endbands:* not preserved *Endleaves:* not original *Edge decoration:* not decorated *No. of sewing supports:* 3, single recessed cords *Spine:* tubular, slightly rounded *Opening:* flat (180°) *Watermark in TB paper:* see Fig. 2b *Feel of paper:* raggy *Sound of paper:* not intense *TB damage grade:* III (wear, tears, staining, foxing) *Binding damage grade:* III (wear, tears) All volumes bound in the same manner, binding not contemporary with TB

Institution **Library call number** *Date of examination* *Provenance* *Date of the binding* *Photo*	Size (mm) **A:** binding **B:** front page **C:** layout on page chapter IX, p.44 **D:** thickness of paper	Structure, specific features: **TB:** texblock **Damage grade:** I (very good), II (good), III (bad), IV (very bad)
6 **Ljubljana, NUK, R II 6632/3** 24 July 2013 *From Lyceum Library in Ljubljana* *Binding probably contemporary with the textblock*	A: 338 × 205 × 95 -110 B: 321 × 202 C: 291 × 172 D: 0.063	*Covering material:* alum-tawed skin *Covering decoration:* composition of blind tooling over the whole cover *Board material:* wood *Clasps:* 2 fastenings, rigid part (metal catchplate) preserved, flexible part: strap cut off *Endbands* green-white, stuck-on endband *Endleaves:* original *Edge decoration:* greenish *No. of sewing supports:* 4, double cords *Spine:* slightly rounded *Opening:* flat (nearly180°) *Watermark:* see Fig. 2b; clearly visible on the title page *Feel of paper:* very soft very used *Sound of paper:* not intense *TB damage grade:* II-III (wear, tears, interventions in the past) *Binding damage grade: I* All four volumes bound in the same manner in alum-tawed skin

Institution Library call number Date of examination Provenance Date of the binding Photo	Size (mm) A: binding B: front page C: layout on page chapter IX, p.44 D: thickness of paper	Structure, specific features: TB: texblock Damage grade: I (very good), II (good), III (bad), IV (very bad)
7 Ljubljana, NUK, R II 361 643/3 2 July 2013 Binding contemporary with the textblock	A: 339 × 205 × 78-83 B: 330 × 205 C: 291 × 172 D: 0.112	Covering material: tanned skin (sheep) Covering decoration: blind tooling on spine and gilt decoration Board material: paste-laminate board Endbands: natural white, stuck-on endband Endleaves: original Edge decoration: sprinkled in brownish red colour No. of sewing supports: 5, double cords Spine: slightly rounded Opening: flat (nearly 180°) Watermark in TB paper: see Fig. 2b Feel of paper: firm, not soft Sound of paper: intense TB damage grade: I (almost no foxing) Binding damage grade: I All four volumes bound in the same manner, in tanned skin

Institution Library call number / Date of examination / Provenance / Date of the binding / Photo	Size (mm) A: binding / B: front page / C: layout on page chapter IX, p.44 / D: thickness of paper	Structure, specific features: TB: texblock / Damage grade:
8 **Ljubljana, NUK, R117 455/3**	A: 332 × 207 × 85	**TB:** texblock
		Damage grade: I (very good), II (good), III (bad), IV (very bad)
	B: 331 × 200	*Covering material:* black thin leather
2 July 2013		*Covering decoration:* gilt decoration on spine and inner part of the edges
	C: 291 × 172	*Board material:* probably paste laminate board
		Endbands: stuck-on endband
Rebinding, probably beginning of the 19th century	D: 0.126	*Endleaves:* highly decorated
		Edge decoration: green
		No. of sewing supports: 4, probably single cords
		Spine: straight, tubular (hollow back), decoratively marked bands
		Opening: flat (nearly180°)
		Watermark in TB paper: see Fig. 2b
		Feel of paper: soft, thin
		Sound of paper: not intense
		TB damage grade: II (some tears, foxing on some leaves)
		Binding damage grade: I
		All four volumes rebound in the same manner, in thin black leather.

Institution Library call number *Date of examination* *Provenance* *Date of the binding* *Photo*	Size (mm) **A:** binding **B:** front page **C:** layout on page chapter IX, p.44 **D:** thickness of paper	Structure, specific features: **TB:** texblock **Damage grade:** I (very good), II (good), III (bad), IV (very bad)
9 **Ljubljana, SAZU II. 713 TR** 20 December 2011 29 July 2013 *From Adolf Robida collection* *Binding contemporary with the textblock*	A: 344 × 210 × 80 B: 330 × 230 C: 291 × 172 D: 0.111	*Covering material:* smooth tanned skin *Covering decoration:* central ornament in blind impression on front and back cover *Board material:* pasteboard *Endbands:* stuck-on endband, green surface coloured thread, *Endleaves:* original *Edge decoration:* not decorated *No. of sewing supports:* 5, double cords *Spine:* slightly rounded *Opening:* flat (nearly 180°) *TB paper:* fine, homogenous *Watermark in TB paper:* see Fig. 2b *Feel of paper:* homogenous, not soft *Sound of paper:* intense *TB damage grade: I (tears in certain places, foxing only sole leaves)* *Binding damage grade: I-II (tailband broken)* All four volumes bound in the same manner, in tanned skin

Institution Library call number / Date of examination / Provenance / Date of the binding / Photo	Size (mm)	Structure, specific features:
	A: binding **B:** front page **C:** layout on page chapter IX, p.44 **D:** thickness of paper	**TB:** texblock **Damage grade:** I (very good), II (good), III (bad), IV (very bad)
10 **Ljubljana, SAZU II 6462/U** 29 July 2013 *Binding probably contemporary with the textblock*	A. 330 × 210 × 90 B: 320 × 200 C: 291 × 172 D: 0.117	*Covering material:* parchment, contemporary *Covering decoration:* none *Board material:* paste-laminate board *Endbands:* ochre-white, original (stuck-on endband) *Endleaves:* original *Edge decoration:* greenish *No. of sewing supports:* 4, double cords *Spine:* slightly rounded, bands visible but not prominent *Opening:* flat (nearly180°) *Watermark in TB paper:* see Fig. 2b, visible on 10 leaves *Feel of paper:* slightly soft *Sound of paper:* not very intense *TB damage grade: I-II (foxing throughout the TB, earlier repairs on some leaves)* *Binding damage grade: I-II (repairs on the spine)* All four volumes bound in the same manner, in parchment

Institution Library call number Date of examination Provenance Date of the binding Photo	Size (mm) A: binding B: front page C: layout on page chapter IX, p.44 D: thickness of paper	Structure, specific features: TB: texblock Damage grade: I (very good), II (good), III (bad), IV (very bad)
11 Ljubljana, SAZU 1116l/3 H Conserved copy 24 November 2011 until 30 January 2012 Rebinding probably end of 19ᵗʰ century, additional interventions mid 20ᵗʰ century	A: 330 × 205 × 98 B: 320-315 C: 291 × 174 D: 0.118	Covering material: tanned skin; paper Covering decoration: fields, made with blind tooling on the spine Board material: paste-laminate board Endbands not preserved Endleaves: contemporary with last intervention on the binding Edge decoration: edges sprinkled (colour now faded) No. of sewing supports: 4, double cords Spine: heavily glued Opening: flat (nearly180°) Watermarks in TB paper: see Fig. 2b Feel of paper: soft Sound of paper: not intense TB: incomplete, 585 printed leaves, 11 supplements TB damage grade: IV (wear, tears, staining, foxing, abundance of glue adhesive tapes) Binding damage grade: IV (broken spine joints, spine heavily glued) 4 volumes rebound in different manners, part 2 in contemporary tanned skin binding

Institution / Library call number / Date of examination / Provenance / Date of the binding / Photo	Size (mm)	Structure, specific features:
	A: binding **B:** front page **C:** layout on page chapter IX, p.44 **D:** thickness of paper	**TB:** texblock **Damage grade:** I (very good), II (good), III (bad), IV (very bad)
12 **Ptuj, Ivan Potrč Library, N-16214D** 20 June 2013 *From Dornava Castle library, 1919* *Binding contemporary with the textblock* 	A: 343 × 220 × 73 B: 335 × 212 C: 291 × 172 D: 0.080	*Covering material:* tanned skin *Covering decoration:* titling in gold/ gilt decoration? on the spine *Board material:* paste-laminate board *Endbands:* not preserved *Endleaves:* not original (vol. I), marbled paper (vols. II-IV) *Edge decoration:* very slightly sprinkled edges *No. of sewing supports:* 6, single cords *Spine:* slightly deformed - concave *Opening:* flat (nearly180°) *Watermark in TB paper:* see Fig. 2b *Feel of paper:* smooth, thin *Sound of paper:* intense *TB damage grade:* III (wear, tears) *Binding damage grade:* II-III (wear, tears) All four volumes bound in the same manner, in tanned leather.

Institution Library call number *Date of examination* *Provenance* *Date of the binding* *Photo*	Size (mm)	Structure, specific features:
	A: binding **B:** front page **C:** layout on page chapter IX, p.44 **D:** thickness of paper	**TB:** texblock **Damage grade:** I (very good), II (good), III (bad), IV (very bad)
13 **Zagreb, HAZU-Bib, R-910** 16 July 2013 *From the collection of Dr Ivan Kukuljević Sakcinski* *Binding contemporary with the textblock*	A: 331 × 203 × 88-104 B: 325 × 203 C: 291 × 172 D: 0.118	*Covering material:* parchment *Covering decoration:* dark brown/black central stamp with coat-of-arms *Board material:* paste-laminate board *Endbands:* blue-white, stuck-on endband *Endleaves:* contemporary with the binding *Edge decoration:* blue *No. of sewing supports:* 5, parchment tapes *Spine:* tubular (hollow) *Opening:* flat (nearly 180°) *Watermark in TB paper:* see Fig. 2b *Feel of paper:* fine, smooth *Sound of paper:* intense *TB damage grade:* I *Binding damage grade:* I Vol. III: from another set, bound in parchment Vols. I, II, and IV: bound in tanned leather

Institution Library call number	Size (mm)	Structure, specific features:
Date of examination	**A:** binding	**TB:** texblock
	B: front page	
Provenance	**C:** layout on page chapter IX, p.44	**Damage grade:**
Date of the binding		I (very good), II (good), III (bad), IV (very bad)
	D: thickness of paper	
Photo		
14 **Zagreb, Metropolitan Library, M 11249 T3**	A: 380 × 245 × 110	*Covering material:* tanned skin
		Covering decoration: sprinkled
	B: 372 × 245	*Board material* paste-laminate board
		Endbands: orange-white, stuck-on endband
4 July 2012 and 16 July 2013	C: 291 × 172	*Endleaves:* original
		Edge decoration: edges sprinkled in red and blue
		No. of sewing supports: 5, double cords
		Spine: slightly rounded
Personal copy of J. W. Valvasor with his ex libris	D: 0.168	*Opening:* flat (nearly180°)
		Watermark in TB paper: different than in other copies, best visible ANH + coat-of-arms, visible in 81 leaves
Binding contemporary with the textblock		*Feel of paper:* very diverse (throughout TB yes), from smooth to rough
		Sound of paper: intense
		TB damage grade: I (minor tears in certain places)
		Binding damage grade: I (locally some bruises)
		All four volumes bound in the same manner

Measurements were carried out with a ruler and a flexible measure with a millimetre scale. The paper thickness was measured using *Cordik®* and *Mitutoyo®* micrometers. The feel of the paper was established by touching it with the fingers; the sound of the paper was established by listening to the paper as it was rattled. The homogeneity and the fibrillization/fibrillation of the paper were examined visually and with a hand lens[17] with a magnifying power of 30. The traces of the impression of the paper-making mould and the presence of watermarks were examined in transmitted light. All examinations were carried out by the same person and in room climate conditions.

The structural elements of the binding that were examined were:
– The size of both the textblock and the binding
– covering: material and decoration
– boards: material
– endleaves: material, date i.e. contemporary with the textblock or from a later rebinding
– endband: structure, material
– sewing supports: number, material
– the form of the spine
– the opening of the textblock
– edge decoration
– damage: type and grade
– specific features, provenance
The textblock paper was examined in relation to the following elements:
– the size of the leaf and printed layout
– the thickness of the paper
– the feel of the paper
– the sound of the paper
– watermark: mould impression, watermark and its position in the quire and on the sheet
– damage: type and grade
– specific features

---

[17] A portable microscope Light Scope 1174 Eschenbach Optik.

Examination of the Damaged Copy (SAZU II 11161/3 H)

Both the binding and the textblock of this volume were badly mechanically damaged (figs. 1b and 1c). In addition, the textblock was incomplete; the first seven printed leaves were missing and had been replaced by a manuscript copy of the text consisting of 24 pages (12 leaves). (fig. 1a).

Figs 1a-1c: Condition of copy of Die Ehre SAZU II 11161/3 H before the conservation treatment

Textblock Structure

The textblock (316–320 × 200 × 98 mm) contained 12 handwritten leaves,[18] 557 printed leaves (organised in 98 quires (3 + 95), the first three being quaternions and the rest (95) ternions or incomplete ternions (quires 4 and 61) and all 9 full-page supplements, almost all of them badly damaged (figs. 1b and 1c). The size of the printed surface is 291 × 172 mm, which is the same as in all examined copies.

---

[18] These manuscript leaves were used to replace the missing first 7 pages of the book.

The text block paper did not rattle.[19]

On all the leaves of the textblock, an impression of the papermaking mould, as well as chain lines and laid lines, are visible to the naked eye. The traces of laid lines run parallel to the text on all leaves, except for the appendices, while the traces of chain lines are always perpendicular to the printed text.

On a small number of leaves, a watermark is visible but hardly readable. It is composed of two parts; the lower one is formed as a shield, ending in a fleur-de-lys and divided in three sections, containing letters A and D. The upper part shows a triangular object, similar to a stonemason's drill. This watermark and the paper-mill is at the moment not identified, but it is most likely to be of German provenance.

Fig. 2a: The watermark in the textbook paper is visible only on a small number of leaves and even there the visibility is generally poor

Fig. 2b: Drawing of watermark in textbook paper

Judging from the impression of the laid and chain lines, the position of the watermarks and the number of sheets in a quire, it was established that

---

[19] Rattle: the sound the paper makes when the paper is moved.

the leaves of the textblock were printed in folio size (2º),[20] i.e., four pages per one sheet[21] of paper (Tschudin 2012, 19).

The paper on which full-page supplements are printed is not of the same quality as the paper on which the text is printed (figs. 1b and 1c). They differ in terms of weight, thickness and watermarks. As these papers would require further comparisons and research, they were omitted from the present analysis.

The Binding

The book's boards and covers were partly replaced, probably, based on an analysis of the materials used, in the mid-20th century. No written report or similar documentation was available. The textblock was trimmed by at least 10 mm in height, and new endleaves were added at the front and back. The textblock was not complete at the time of this rebinding. The first 7 leaves at the beginning were missing and had been replaced by the 12 leaves (24 pages) of manuscript (fig. 1a).[22]

The only originally preserved component is the sewing: the textblock was sewn with linen thread on four double supports made of hemp cord with a Z-twist. The supports were broken in the middle. Both endbands were missing.

The textblock was sewn all-along in the first and last two quires and on and off (multi-section sewing) in other quires. The initial stitch (a kettle stitch) was begun 20 mm from the head (the top edge) of the section (K). The first support (1) was attached 60 mm from the head, the second (2) 123 mm, the third (3) 183 mm, the fourth (4) 243 mm and the final stitch (k) 296 mm (all measurements are from the head edge of the textblock – see the diagram below). The end (tail edge) of the section was 318 mm from the top edge of the textbook.

Head	K	1	2	3	4	k	Tail
0	20	60	123	183	243	296	318

---

[20] Philip G., *A new Introduction to Bibliography, Book production: The Hand-Press Period 1500–1800* (Oxford: Oxford University Press 1972), 78–109.

[21] The sheet of paper on which the printing plate was impressed.

[22] The manuscript copy was made with iron gall ink on industrially made paper. The text of the first 14 printed pages of Part 3 of *Die Ehre* was copied.

The endleaves are not original, but rather date from the mid-20ᵗʰ century repair. The pastedown on the front cover has survived but is badly damaged. The front free endleaf that belongs to it has been lost, as has the entire back pastedown. As a result, it has not been possible to determine with accuracy the endleaf format.

The textblock was sewn on four raised bands. There were no original spine linings. Bookbinding gauze, added during the partial rebinding, was pasted all along the length of the spine. Gelatine glue and subsequently synthetic glue (PVAc) had been applied to the spine, making it very rigid.

The book boards measure 330 × 205 mm and are made of pasteboard.

The covering is not completely original. The leather spine and corners, decorated with blind impressions of linear stamps, may be a part of the original leather covering. The rest of the board is covered with a red-brown decorative paper that was added when the volume was partially rebound (fig. 14).

Valvasor's Personal Copy from the Metropolitan Library in Zagreb (M 11249 T3)

The copy marked M 11249 T3, held by the Metropolitan Library in Zagreb (MKZg), is of particular importance because it looks quite different than the other examined copies. Since it contains Valvasor's ex libris, it is thought to have once been Valvasor's personal property. The copy was examined twice in the Metropolitan Library reading room (on 4 July 2012 and 16 July 2013) and is described in detail below (figs. 3a-3d).

Fig. 3a   Fig 3b

Figs 3a-3d: Condition of the copy of Die Ehre vol. III, from The Metropolitan Library Zagreb

Fig. 3c   **Fig 3d**

This copy differs considerably from the other examined copies. The material structure is schematically presented in table 1. The findings are as follows:

The external dimensions of the book (380 × 245 × 110 mm) and of the binding (371 × 245 mm) are considerably greater than those of the other examined copies. The book differs from the average by approximately 50 mm in height and by approximately 40 mm in width.

All four volumes of this copy of *Die Ehre* are of the same size and bound in the same way. The books were sewn on five double bands. The covers are made of pasteboard and covered by sprinkled tanned leather with some gilt decoration on the spine. The spine is slightly rounded. (figs.s 3a and 3b)

The spine of volume III is partly damaged at the head and foot (table 1, no. 14). The surface of the leather covering is partly worn (fig. 3a).

The textblock has sprinkled edges in red and blue. The endbands are original and sewn with orange and white thread. (fig. 3b)

The endleaves, both the pastedown and the free endleaf, are original – that is, contemporary with the binding (fig. 3c).

The textblock is very well preserved (fig. 3d).

The measurements of the paper thickness showed that the paper of the MKZg copy was 0.164 mm thick, which is considerably thicker than in other examined copies. The leafing sound (rattle) is quite intense.

68 leaves of the textbook bear visible watermarks that all differ from the watermark visible in the paper of the textblock of the other copies examined. The watermark bears the acronym ANH and a coat-of-arms are clearly visible in transmitted light (Fig. 4a). This watermark was identified in the

full-page supplement of Turjak and on the textblock leaves; it is clearly visible in volume IX, leaf 111–112, and volume XI, leaf 651–652.[23]

Fig. 4a

Fig. 4b

Fig. 4c

Figs 4a-4c: Paper in the textblock of Die Ehre, vol. III, from The Metropolitan Library, Zagreb

Examination with and without a magnifying glass showed that individual leaves of paper in this copy vary considerably in terms of thickness,[24]

[23] The copy, M 11249 T3, held by The Metropolitan Library in Zagreb.

[24] E.g., volume IX, leaves 209–210 or volume XI, sheets 147–148.

roughness[25] and their level of rust-coloured impurities.[26] The textblock paper, however, does not show any signs of ageing, mechanical damage or chemical damage. The fibrillation is rather non-homogeneous in the majority of leaves. In transmitted light, spots with piles of fibres are often visible, while the structure of the paper shows traces of drops of water that fell on the newly made paper sheet,[27] or the fingerprints of the paper maker.[28] (Figs 4b and 4c) The non-homogeneity and the presence of impurities clearly indicate that the paper was of a lower price range.

3. Discussion and Results

The visual examination of paper revealed that 13 of the examined copies of *Die Ehre* contain the same watermark in the textblock paper (fig. 2a). In the majority of cases, the watermark was detected on less than a third of leaves (table 1), and on many leaves it was difficult to perceive even in transmitted light. It is often obscured by print and in some cases also by the darkening (browning) of leaves caused by impurities. The traces of chain and laid lines from the papermaking mould are clearly visible on all sheets in all the copies. Although several different watermarks in the MKZg copy (from Valvasor's library) were identified, they did not include the above-mentioned watermark.

On the basis of the position of the watermarks in the paper of the 13 examined copies and the traces of impression of the pattern of the papermaking mould, it is clear that the paper sheets in the textblock were printed in folio size (4 pages to one sheet of paper). However, during the conservation treatment of volume III of *Die Ehre* (SAZU II 11161/3 H), specifically the pulling the textblock, it was found that the quires were assembled in ternions, which is a form usually found in books of the period.

The *measurements of the paper thickness* showed the differences in thickness among the examined copies (fig. 1). The average difference is approximately 0.12 mm. The thickness of the textblock leaves is significantly greater in the copy from Metropolitan Library, amounting to 0.168 mm on

---

[25] E.g., volume X, leaves 253–254.
[26] E.g., volume IX, leaves 113–114 or leaves 167–168.
[27] E.g., volume IX, leaves 11–12 and 41–42.
[28] E.g., volume XI, leaves 153–154.

average, as well as in the copy kept by NUK (RII 6632/3), where the difference is 0.063 mm.

The past practice in the area of today's Germany (in contrast to other areas) was to print on paper not treated with gelatine, as printing on unsized paper was easier and less expensive. Older German bookbinding manuals indicate that sizing was undertaken by bookbinders before the book was bound (probably at the customers' request, as this was additional work that would have increased the cost).[29] This information offer an explanation as to why the feel and sound of the papers in the examined copies are so different (i.e., sized versus un-sized). A final confirmation of this theory, however, would certainly require additional analyses.

The size of the *printed layout* is the same in all fourteen copies, which is to be expected as the printing form was not changed. With respect to the size of individual books, Reisp states that the external dimensions of individual volumes are 330 × 205 mm. The examination of the copies led to the conclusion that the copies with their cotemporary binding slightly differ in size and indicate that they were not all trimmed to the same size. The size of the MKZg copy from the Metropolitan Library in Zagreb is significantly greater, being considerably taller and wider than average.

An *examination of the spines* of the majority of the books still having a contemporary binding has shown that the textblocks were sewn on four or five double linen cords. The most common board material was pasteboard covered by tanned leather. There are eight such examples. In three cases (NUK R II. 6632/3, NMS 1625/3 and NMS 1625/3), the boards are of wood covered with alum-tawed skin decorated with blind stamp impressions. The binding on the copy kept by FSLJ was made later, as it belongs to a set of Škerpin's bindings (mid-18th century). The examined copies include two with parchment covers (SAZU U – II 11162 U and HAZU R-9). The binding of the copy kept by HAZU differs from the others, being a limp parchment binding.

These findings indicate that the binding of the book was not uniform across copies. The practice (in the Germanic part of Europe) in the 16th and

---

[29] Information regarding gelatine sizing in the area of today's Germany in comparison with the practice in other lands was provided by Nicholas Pickwood.

17th century was that books were bound after the purchase of the textblock.[30] This is partly related to the fact that already bound books were more expensive to export due to increased weight and tax, so it was cheaper to transport unbound textblocks. The owners or commissioners determined the type of binding in accordance with their own wishes and prices.

On the basis of visual examination of the damaged paper in the analysed textblocks of *Die Ehre*, it was found that the most common damage is mechanical damage and staining (foxing) (table 1). Mechanical damage in the copies is directly related to the intensity of use of individual copies and the mechanical properties of the selected paper. Browning (foxing) is not dependent on how the book has been handled. It is probably a result of impurities that ended up in the paper, mostly during the printing procedure (fig. 1b). Its intensity could also be a result of the environmental conditions in which the individual copies have been stored. According to Pickwoad's[31] and Orti's (Orti 2016) findings, this phenomenon is most prominent in the paper of books printed in the Germanic part of Europe. Our investigation showed that the appearance of browning is more prominent in those copies of *Die Ehre* where the sound of the paper is not intense. This phenomenon is very likely to be related to the absence of a surface gelatine treatment, but unforunatly this could not be proved with FTIR mesurments.[32]

All four volumes of *Die Ehre* in the *copy kept by the Metropolitan Library in Zagreb* are bound in boards covered with sprinkled tanned leather. Valvasor's ex libris is found on the front pastedown. The minor mechanical damage visible on the binding is as a result of handling. The textblock paper does not show any signs of ageing or of mechanical or chemical damage. The leafing sound is intense. A detailed examination of the textblock as regards transmitted and direct light showed that various papers had been used. They differ in terms of thickness, smoothness or roughness, non-homogeneity, fibrillization/fibrillation, irregularities in the manufacture and presence of impurities. This leads to the conclusion that the paper used for printing

---

[30] Rosenfeld, H. 'Buchpreis, Antiquariatspreis und Einbandpreis im 16. und 17. Jahrhundert', in: *Gutenberg Jahrbuch*, 33 (1958), 358–363.

[31] Nicholas Pickwoad, personal correspondence, 2014.

[32] Kavkler, K. *Poročilo naravoslovnih preiskav v knjigi Slave vojvodine Kranjske, 3. del, Nürnberg 1689* (internal report, Ljubljana: ZVKDS Restoration Centre, June 2013).

was of a lower quality. Watermarks detected in this copy differ completely from the watermarks present in the paper of other examined copies. Judging from the position of the paper mould and the position of the watermark, the book was printed in folio size.

Taking into consideration the properties of the textblock papers that have been noted, it is clear that the paper used for printing this copy is completely different from the papers on which other examined copies were printed. In view of this fact there is a possibility that this copy is a "trial" impression before printing. The confirmation of this hypothesis (assumption) would require further evidence.

The analysis of the selected paper sheets showed that papers in the same quire may differ in their basic physical characteristics, too. The paper sheets differ particularly in terms of weight and thickness, which is understandable given that they were handmade.

The comparison of the examined copies of volume III of *Die Ehre* showed that the full-page supplements are always located in the same position in the textbook, but not always inserted in the textblock by the same method.

As regards binding, all examined copies with their contemporary bindings were bound in four volumes; even these contemporary bindings, however, differ from one another.

The early binding of the copy of *Die Ehre* (SAZU II 11161/3 H) assigned for conservation to the BPCC ARS was not preserved. In view of this fact and the damage, it was decided to sew the textblock anew, following the technique and materials detected in the contemporary bindings. During the process of conservation, it was decided to preserve the mid-20th century form (covering) it had when it was sent for conservation. All those materials found in and on the cover before the treatment were retained in their original positions: the red paper covering, the leather spine fragments and the corners and pasteboards.

4. Conclusion

Although Valvasor's work has been extensively researched, certain aspects remain unanalysed. These include the technical background and the material and physical structure of his works, such as the textblock structure, the paper and the binding. A conservation treatment provides an exceptional

opportunity to examine these characteristics in detail. Such a possibility arose in the process of the conservation of a badly damaged copy of volume III of *Die Ehre des Hertzogthums Crain* (SAZU II 11161/3 H).

The book in question is a printed work, and several copies in much better condition are available. According to published data, a copy of *Die Ehre* is held in 22 institutions in Slovenia, and quite a few copies are kept outside of Slovenia as well. Our survey comprised twelve copies from Slovenia and two copies from Croatia. All 14 copies were examined to compare their bindings, the textblock structure and the paper used.

The research covered only the paper used for textblock printing, not the paper on which supplements were printed or the paper used for endleaves. The textblock paper was examined in order to attempt to establish the different types of paper used for printing *Die Ehre* and the format of the printing.

The examination and analysis of the paper in the damaged copy and the comparison with the papers in the 12 examined textblocks led to the finding that all the textblocks were printed on paper with the same watermark, which, judging by its physical and optical properties, is of a lower quality type. The textblock paper in individual examined copies was also found to differ in terms of feel and sound. These differences are likely to be the result of gelatine sizing carried out in German lands by bookbinders after printing. This procedure was probably not performed regularly because it was time-consuming and increased the cost of the binding. It presumably was done at the request of the commissioner of the binding. Further clarification of all questions related to the selection of paper for textbook printing would require additional analyses.

The structure of the quires, the positioning of the watermarks, the size of the paper leaves and the impression of the laid and chain lines led to the conclusion that all copies were printed in folio size.

Valvasor's personal copy kept by the Metropolitan Library in Zagreb was found to differ from other examined copies in terms of the size and type of paper used. This is considerably different from the paper of all the other examined copies, although it also is of lower quality.

Scientific analysis showed that papers in the same quire may also differ in their basic and physical characteristics, especially in terms of grammage, thickness and bulk.

Before beginning the conservation intervention, it was discussed whether, given that every conservation intervention reduces the authenticity of the original item, it was actually necessary to intervene on such an important book. The book in question is a printed work, not a unique manuscript, and it had been previously repaired. In addition, copies of the text that have been very well preserved are available for use.

After a thorough examination of the condition of the textblock and binding, it was decided to undertake a comprehensive conservation treatment that would include unbinding, surface and wet cleaning, sizing and repair of all textbook leaves and re-binding. Given the abundance of stains of various kinds and the water-soluble glue in adhesive tapes, wet-cleaning of the paper would have been the only effective cleaning method.

The comparative analysis of physical and optical properties of papers before and after washing and sizing showed that the changes of some properties appear after washing. After sizing, the values mostly resemble those before washing. The paper is more voluminous, more porous, less opaque and with a slightly increased surface roughness, while the brightness and colour remain almost the same. All these changes were expected.

The examination of the bindings of several copies of *Die Ehre* was undertaken in an attempt to identify the original elements of the binding in the preserved copies and to establish whether a uniform type of binding had existed. The results of the analysis of the bindings led to the conclusion that the manner of binding was not completely uniform. It was decided in rebinding the conserved copy to follow the preserved sewing structure and to retain the covering materials added during its partial rebinding in the mid-20th century. Although *Die Ehre des Hertzogthums Crain* is a printed book, the physical appearance of each preserved copy is different, and each is a source of data valuable for the fields of paper, printing and binding, as well as for the technical history of the book.

Bibliography

Gaskel 1976: Gaskel, P. *A new Introduction to Bibliography, Book production: The Hand-Press Period 1500–1800* (Oxford: Oxford University Press 1972), 78–109.

Golec 2013: Golec, B. "Valvasorjev izvor, družina in mladost – stare neznanke v novi luči (1. del)." *Kronika,* 61 (2013), no. 1, 6.

Oražem-Stele 1990: Oražem-Stele, M. "Dela Valvasorja leta 1989 v sre-dnjeevropskih knjižnicah in njegov sodobnik pisatelj Franz von Wützenstein." In *Valvasorjev zbornik* (Ljubljana: Slovenska akademija znanosti in umetnosti, 254–256.

Orti 2016: Orti, G. "The quality of the paper used in printed books of the 15th and 16th centuries in the Episcopal Seminary's library of Padova." In *Book and paper conservation II*, (Ljubljana: Arhiv Republike Slovenije, 40–47.

Tschudin 2012: Tschudin, F. P. *La carta. Storia, materiali, tecniche,* Roma, Passariano: Centro di catalogazione e restauro dei beni culturali, 2012), 19.

Vodopivec Tomažič 2014: Vodopivec Tomažič, J. "Slava Vojvodine Kran-jske 1689, materialna podoba in konservatorsko-restavratorski poseg na 3. delu." in *Studia Valvasoriana*, ed. by J. Weis, Ljubljana: Zavod Dežela Kranjs-ka, 2014, 131–171.

Vodopivec Tomažič, 2016: Vodopivec Tomažič, J. "Analyses and conser-vation of Valvasor's Glory,1689, Vol. III." In *Book and paper conservation II*, Ljubljana, Arhiv Republike Slovenije, 2016, 68–107.

Vodopivec Tomažič, 2017: Vodopivec Tomažič, J. "Valvasor's Glory of the Duchy of Carniola, 1689, Vol III: Comparative analysis of binding structure and paper." *Bookbindings Theoretical Approaches and Practical Solutions,* Bibliologia 45, Turnhout, BREPOLS, 2017, 227–251.

Weis 2014: Weis, J. (ed.). *Studia Valvasoriana*, Ljubljana: Zavod Dežela Kranjska, 2014.

About the author

Jedert Vodopivec Tomažič, Prof. Dr., Universty of Ljubljana & Alma Mater Europaea ECM

Head of Conservation Department, Archives of the Republic of Slovenia

Zvezdarska 1, 1000 Ljubljana, Slovenia

jedert.vodopivec@gov.si, jedrt.Vodopivec@gmail.com

She obtained her BSc and MSc degree in Chemistry Science and PhD degree in Conservation Science at the University of Ljubljana. After a short period working in a research laboratory, she joined the Conservation De-partment of the Archives of the Republic of Slovenia in 1980 as a chem-ist and book and paper conservator working on archival, library and other graphical materials. Since 1990 she has also been involved in professional training and since 1997 as lecturer of preservation of written heritage at the

University of Ljubljana and Alma Mater Europaea ECM. From 1996 to 2006 she was a full member of the Preservation Committee of the Internacional Council on Archives (ICA/CPTE), from 2003 to 2006 was active in the EU COST Action G8-WG3 'Degradation processes, corrosion and weathering' as coordinator for organic materials of the WG3 'Non-destructive Analysis and Testing of Museum Objects'. Vodopivec has been the author of national and international papers, organiser of conferences, seminars and workshops on the conservation of cultural heritage on paper and parchment and has collaborated in research projects dealing with conservation of archival, library and graphic works on paper.

Zusammenfassung

Der Beitrag präsentiert einen Vergleich von 14 Kopien des dritten Bandes von Valvasor's "Die Ehre des Herzogthums Crain". Unter ihnen ist auch Valvsor's persönliche Kopie, die sich aktuell in der Metropolitan Bibliothek in Zagreb befindet. Das betreffende Buch ist eines der Säulen slowenischer Geschichte und Kultur, eine enzyklopädische Arbeit, die von allen Wissenschaftsdisziplinen wertgeschätzt wird. Das Buch ist in 15 Kapitel unterteilt und üblicherweise in 4 Bänden mit insgesamt 3532 Seiten gebunden. Da Band 3 der an Illustrationen reichhaltigste ist, handelt es sich um den am meisten genutzten und folglich auch meistbeschädigsten Band. Die Kopien des dritten Bandes wurden untersucht, um ihre Bindungen, Buchblock und das verwendete Papier zu vergleichen.

ORTHOGRAPHIC VARIANT OR SCRIBAL ERROR? THE CASE OF THE LATE EIGHTEENTH-CENTURY SLOVENIAN POLJANE MANUSCRIPT

Andrejka Žejn

Abstract: Dating from the late eighteenth century, the Poljane manuscript is one of the more recent and more important discoveries in the ongoing systematic collection of lesser known and unknown manuscript texts of old Slovenian literature, and as such is an interesting subject of methodologically diverse analyses. The text is an example of the baroque tradition of describing the life and, in particular, the sufferings of Jesus Christ, and is specifically related to the work of the famous German baroque writer Martin of Cochem. In the tradition of scholarly critical editions of Slovenian manuscripts, the so-called critical transcript has become established. While also being an interpretation of the manuscript, the transcript is intended to bring the text closer to the contemporary reader by means of modern spelling, as well as providing a basis for further language analysis. When interpreting and analysing the language of texts from distant periods, we must take into account the language situation and the level of language development so that individual text renderings can be appropriately categorised as normal or acceptable variants, scribal errors, or renderings that cannot be unambiguously defined as normal or incorrect. Taking the Poljane manuscript as an example, we show the factors that must be taken into account in a critical assessment, while also presenting the editorial solutions adopted in the scholarly digital edition as suggested by the TEI Guidelines.

Keywords: Poljane manuscript, eighteenth-century Slovenian, scribal errors, language variation

Introduction

For more than a decade, the systematic collection and analysis[1] of lesser known and unknown manuscript texts, especially texts from the seventeenth

---

[1] A representative result of the collecting of manuscripts is the online *Register of Early Modern Slovenian Manuscripts (Register of Baroque and Enlightenment Slovenian Manuscripts, NRSS)*, available at http://ezb.ijs.si/fedora/get/nrss:nrss/VIEWENG/ (see Ogrin et al. 2013). The register contains freely accessible digital facsimiles of already known and newly discovered manuscripts equipped with codicological and palaeographic data, as well as known or established facts about authorship, time and period of creation, content,

and eighteenth centuries, has been in progress at the ZRC SAZU Institute for Slovenian Literature and Literary Studies.[2] One of the most significant discoveries of this research is the so-called Poljane manuscript from the late eighteenth century. This 700-page volume offers ample material for further research in the history of language, literary stylistics, literary history, cultural history and a number of other research areas. Such research should supplement the current knowledge of Slovenian literary and cultural history while enabling the Poljane manuscript to be positioned with respect to the stages of the evolution of Slovenian language, literature and culture.

For this purpose, a digital scholarly critical edition of the manuscript is being prepared (see Žejn et al. 2017). Scholarly critical editions of this and other manuscripts[3] represent material that opens new perspectives on the development of Slovenian literature and encourages the research of later periods that have not yet received the deserved attention.[4]

Taking the example of the Poljane manuscript, the present paper focuses on one of the questions of textual criticism related to historical linguistics and dialectology that contributes to a holistic analysis of a manuscript. The known facts concerning the circumstances of the creation of the text will be presented, and it is on the basis of these facts that the text must be observed and studied. Referring to specific examples from the Poljane manuscript, certain dilemmas will be highlighted, along with solutions that serve

---

classification, etc. The register currently contains over 120 manuscripts and is still being expanded with newly discovered manuscripts.

[2] This research began as part of the *Register of Baroque and Enlightenment Slovenian Manuscripts: Computer-Assisted Scholarly Publications and Analyses (2008-2011)* and its follow up *Slovenian Literature in Unknown Manuscripts during the Reformation and Romanticism: Computer-Assisted Analyses and Scientific Publications (2013-2016)*, both initiated and led by Dr Matija Ogrin.

[3] Scholarly digital editions, like the repository, are freely available at the eZISS digital library/*Scholarly Digital Editions of Slovenian Literature* (http://nl.ijs.si/e-zrc/). Scholarly critical editions in printed book form have been published since 2017 as part of the *Works of Old Slovenian Literature Collection* (lead editor Matija Ogrin).

[4] Thus, the Poljane manuscript, as a meditative narrative text, will be included in the corpus of texts in the project *The Beginnings of Slovenian Narrative Prose*, conducted at the ZRC SAZU Institute for Slovenian Literature and Literary Studies. The project, whose project leader is the author of the present paper, is funded by the Slovenian Research Agency. The issue we are dealing with in this paper is a preparatory step in the analysis of the text with regard to narrative elements and connections with later canonised narrative texts.

to avoid inadequate interpretations of a manuscript text. There will also be a presentation of the TEI tag sets used to adequately represent the solutions to dilemmas in a scholarly digital edition.

Basic data on the Poljane manuscript

The Poljane manuscript was discovered in Poljane above the town of Škofja Loka, Slovenia, in 2009, hence its editorial name. It is held by the National and University Library in Ljubljana, and the facsimile edition and description of the manuscript are freely accessible at the *Register of Early Modern Slovenian Manuscripts*, library call no. Ms 023, at http://ezb.ijs.si/fedora/get/nrss:nrss_ms_023/VIEW/. We can conclude that it was written by an unknown author around 1800, most likely in the nearby Capuchin monastery in Škofja Loka. More than 700 pages of the folio volume have been preserved, but the initial and concluding pages are missing, as are several intermediate pages. In terms of content and style, the manuscript is associated with one of the adaptations of the monumental baroque work *The Life of Jesus* by the German Capuchin friar Martin of Cochem, who was active in the second half of the seventeenth and the early eighteenth century. The manuscript is the second most extensive and the second oldest preserved text of Cochemian adaptations in Slovenian.[5] In terms of content, it is divided into two parts that are disproportionate in scope: the first part gives an account of Jesus' life (birth, childhood, deeds and miracles) on 87 pages, while the second part, comprising 577 pages, depicts the Passion of Christ from the moment of his arrival in Jerusalem to the washing and anointment of his dead body.[6] Throughout the manuscript, the course of events is highlighted and clarified by testimonies and interpretations from the works of church fathers, and by medieval mystical revelations that extensively complement the evangelical reports. The Passion section is character-

---

[5] Five manuscript adaptations in Slovenian have been discovered to date. For one of them, only a report has been preserved, while another, an adaptation of Cochem's *The Life of Jesus*, was published in the years 1868-1973 as a printed book in Slovenian. In the book edition, the manuscripts are not mentioned, but from certain renderings we can conclude that the translator also relied on the manuscript tradition.

[6] Though the concluding pages of the manuscript are missing, we can conclude that this section's "99th law" contained the scene of laying Jesus in his grave, while the "100th law", which is missing in its entirety, most likely contained a description of the resurrection as the conclusion and climax of the whole event.

ised by detailed and vivid descriptions of what Jesus and Mary experienced in the days prior to Jesus' death. (Ogrin 2011: 394-395)

Critical transcript of the manuscript

By adopting the TEI Guidelines, the planned scholarly digital edition will contain, as its basis, the digital facsimile edition of the manuscript, including a diplomatic and critical transcript. In addition to these basic elements of scholarly critical editions of Slovenian manuscripts, the edition will feature a differential glossary and a number of accompanying studies. Critical transcription means rendering the original text in the Bohorič alphabet (Sln. *bohoričica*) in the form of Gaj's alphabet (Sln. *gajica*), while also taking into account the rules of modern orthography, especially with regard to capitalisation and punctuation, as well as a modern conceptualisation of word boundaries. The Bohorič alphabet, in which the manuscript was written, was in use almost from the beginnings of the Slovenian standard language in the age of Protestantism. On the basis of the use of the Bohorič alphabet in Slovenian Protestant books and, above all, the conclusions of the audit committee of the first translation of the *Bible* into Slovenian from 1581, the Bohorič alphabet was codified by Adam Bohorič in his grammar *Articae Horulae Succisivae* from 1584.[7] In the 1830s, another Latin alphabet, known as Gaj's Latin alphabet, started to be used for Slovenian. It first appeared in print in 1838, but the Bohorič alphabet continued to be used as a second alphabet until 1848.[8] Today, Gaj's Latin alphabet is the only established script for the Slovenian language. Even with just its critical transcript, a text can be made more accessible to the contemporary reader while at the same time representing a basis for further textual analysis and, to a degree, interpretation of the text.

In a critical transcript of the text or manuscript with inconsistent spelling and a not-yet-codified (standard) language variant – both of which are characteristic of the Poljane manuscript – there is often a dilemma about

---

[7] The first book in Slovenian was written by Primož Trubar in the Gothic script (*Catechismus and Abecedarium*, 1550), but later, in the second edition of 1555, it was transcribed by the author into the Humanistic script and partly amended on the basis of the German model. Nevertheless, the script was named after its founder, Adam Bohorič.

[8] In the meantime, there were some attempts to introduce other scripts, but they were unsuccessful in the long run.

how to read and understand a certain letter and, consequently, a word that is unknown, incomprehensible or at least uncommon in terms of contemporary linguistic experience. When the critical transcription of a text is undertaken, the question usually arises: *Is the individual reading the intended word, and thus one of the valid or acceptable historical possibilities, or is it a scribal error*? In order to determine the most objective response to this question, we must investigate in as much detail as possible the various circumstances surrounding the creation of the manuscript text. The main factors that can or should be taken into account are: the alphabet of the text, the Slovenian standard language and dialect situation at the time of the production of the manuscript text, and its transmission.

The Bohorič alphabet and Gaj's Latin alphabet – similar but different types of script

In the Bohorič alphabet and Gaj's Latin alphabet, the individual sounds in Slovenian are mostly represented by the same letters, so determining the relationship between the two writing systems, which forms the basis for a critical transcript, is especially problematic with respect to the "non-matching" letters, in particular, the letters for the (alveolar) sibilants, the [i] and [j] sounds with their clusters, as well as for the [u] and [v] sounds. For some (alveolar) sibilants, digraphs are used that are unknown to Gaj's Latin alphabet. Although the Bohorič alphabet was codified in a grammar as early as in the sixteenth century, inconsistencies in the rendering of certain sounds continued to appear centuries later. By inconsistencies, we do not only mean differences in the rendering of the same sound in different texts, but differences in rendering the same sound by the same author in the same text. Such differences can only partly be attributed to the phonemic environment or the tradition of the phonetic transcription of a particular word. Inconsistencies can also be the result of the conscious or unconscious impact of writing practices in German and Latin texts. An example of the inconsistency found in the manuscript is the [s] sound, which is rendered in as many as five different ways in the Bohorič alphabet (henceforth: BoA): as *s* (e.g., *postava* [law] critically transcribed in Gaj's Latin alphabet (henceforth: GLA) as *postava*, or *gospud* [lord], in GLA *gospud*), *sʃ* (e.g., *usʃakter* [each, every], in GLA *vsakter; popisʃat* [describe], in GLA *popisat*)*; ʃs* (*popiʃsan*, in GLA *popisan; ceʃsarju*, in GLA *cesarju*)*; ʃ* (*uʃmilene* [mercy], in GLA *usmilene;*

boſ [barefoot], in GLA *bos*)*;* or **ss** (e.g., *bessedi* [word], in GLA *besedi*). Or, from another perspective: in the manuscript*,* the digraph **sh** in BoA represents either the [š] or [ž] sound, e.g., for [š]: *starshi* [parents], in GLA *starši; nimash* [(you) do not have], in GLA *nimaš*; for [ž]: *sheno* [wife], in GLA *ženo; boshja* [divine], in GLA *božja*. The transcription of the digraph **sh** as **ž** or **š** in GLA is problematic when it is found in a word unknown in contemporary standard Slovenian, and we need to decide upon the appropriate transcription into GLA. In so doing, Slovenian historical, dialectological and etymological dictionaries can be helpful, as well as German historical and dialectological dictionaries, such as in: *rishtenga* critically transcribed as *rištenga* in the sense "tool, aid" from Ger. *Ausrüstung*; *shirm* critically transcribed as *širm* [shield, defence] from Ger. *Schirm; po***sh***aubat* critically transcribed as *požavbati* [to anoint] with a root in Ger. *Salbe*.

Despite the described inconsistencies in spelling, we can get a relatively clear picture of the typical correspondences between BoA and GLA when transcribing the former into the latter. In addition to consistent practices, deviations are identified when a sound in BoA is rendered with a specific letter or several different letters with multiple occurrences, while exceptional (single) or rare (few) renderings are also present. An example of this is the rendering of the [č] sound: in the manuscript, this sound is in principle rendered with the digraph *zh*; however, in the text as a whole, there are five renderings of this sound with a letter *z* in four different words: *zisto* "čist" [completely], *zes* "čez" [over/across], *temuz* "temuc" [but, however] (occurring twice) and *konzan* "končan" [finished]. On first glance, these renderings appear to be scribal errors in which the second part of the digraph *zh* has been accidentally omitted. Nevertheless, the digraph's double occurrence in the same word is a cause for some doubt. It is precisely in relation to these renderings of the [č] sound that a key connection has been established between the manuscript and its base text, i.e., an older Slovenian manuscript that served as an older model text for the Poljane manuscript.[9] In fact, the Poljane manuscript is one of the few manuscripts in Slovenian whose base text is known: a manuscript with the editorial title *Jesus' Life in*

---

[9] Among other things, comparison of the two texts shows that, in addition to the preserved base text, the author of the Poljane manuscript had access to another source, which he used in part to complement his text. (Ogrin & Žejn 2016: 131)

a Hundred Laws (see http://ezb.ijs.si/fedora/get/nrss:nrss_ms_028/VIEW/), which dates back to the middle of the eighteenth century and of which only a tenth of the original text has been preserved. A text-critical comparison of both texts using software tools (see Ogrin & Žejn 2016) showed that few linguistic changes had been made; significantly, more differences were found in the rendering of sounds. Analysis of the rendering of the [č] sound in the base text shows that it is rendered there either as **zh**, the same as in the Poljane manuscript, or as **z**. The latter rendering is less common but still relatively frequent. It can be concluded that, in the base text, it is the [č] sound that is typically rendered by *z*, such as in compounds and derivatives of *srce* [heart], which is the result of velars changing to palatals before the Proto-Slavic j (e.g., *serzne* from the base text is rendered in the manuscript as *serzhne*), *resnica* [truth] (e.g., *reſniznu* from the base text is rendered as *resnizhnu* in the manuscript) and *konec* [end] (*pokonzanh* from the base text corresponds to *pokonzhanh* in the manuscript). Furthermore, this rendering is frequent in the word forms of *mrlič* [dead body] (*Merliza* in the base text is *Merlizha* in the manuscript and *merliča* in GLA) and, sporadically, in certain other words. The author of the manuscript clearly used a single digraph *zh* for both renderings of the [č] sound but preserved *z* in the four words listed above. If we acknowledge these facts rather than attributing departures from the norm to scribal errors, we can speak of the influence of the base text on the more recent renderings.

Linguistic evolution: The case of variant forms

There is another type of variation in the manuscript that is unrelated to the question of alphabet. Variant readings of the same word are common, giving rise to a dilemma about the correct or most appropriate reading of, for example, a word meaning "whole/all": the nom./gen. sg. neutral gender: *zelu* (in GLA *celu*), *zelo* (in GLA *celo*) and *zel* (in GLA *cel*); or "society": *drushina* (in GLA *drušina*) and *drushna* (in GLA *drušna*); or "a hundred": *stou*, *sto* and *stu* (in GLA *stou, sto* and *stu*). The legitimacy of these variants can be verified on the basis of a knowledge of the standard norm of Slovenian and the dialect situation at the time of the compilation of the manuscript.

The situation regarding Slovenian (standard) language in the late eighteenth century could itself be defined as "variant-based". To begin with, standard Slovenian was not artificial in its origin but evolved from a spoken

dialect, though by omitting the more specific dialectal features (Orel 2003: 551). In this regard, mention should be made of the diverse structuring of Slovenian dialects into seven dialect groups, further subdivided into numerous dialects and speeches.[10] The basis of the standard language formulated by the Protestants in the sixteenth century was the language of the central Slovenian regions, albeit characterised phonetically and morphologically by that of the Lower Carniolan and Notranjska regions. The lexicon, in particular, was influenced by other languages, primarily German, which for centuries and in several respects was the dominant language in the area where Slovenian was spoken (Orožen 1996: 105). At the beginning of the second half of the eighteenth century, individual so-called regional varieties of standard language began to evolve, resting on and drawing from other Slovenian dialects, with the process peaking at the very end of the eighteenth century. Thus, four regional varieties of standard language existed in the territories where Slovenian was spoken from the mid eighteenth to the mid nineteenth century: alongside the Carinthian and the Eastern Styrian and Prekmurje varieties, the most important and normatively most accomplished was the Upper Carniolan variety, which was based on the sixteenth-century tradition (Orožen 1996: 122). Nevertheless, it must be pointed out that no individual variety of the Slovenian standard language was firmly codified by the end of the eighteenth century.

At the end of the eighteenth century, the Upper Carniolan variety of the Slovenian standard language was used in the region where the manuscript originated. This standard language variety found its codification in *Kranjska gramatika*, a grammar by Marko Pohlin, which was first published in 1768 and comprehensively edited in 1783. Notably, however, the norm was determined by some important printed publications, the most relevant for the Slovenian standard language variety of the time being the translation of the Bible published in the period from 1786 to 1802 by Jurij Japelj and his translation team. This was the first so-called Catholic translation of the Bible, following that by Protestant Jurij Dalmatin in the sixteenth century, which remained the model text for the Catholic translation despite the di-

---

[10] According to the dialectological chart of the Slovenian language, over thirty dialects and speeches are classified within the seven dialect groups (Carinthian, Styrian, Lower Carniolan, Pannonian, Primorska and Rovte).

vergence in religious outlooks. At the same time, the region where the manuscript was written belongs to the Western parts of the Rovte dialect group. The latter was formed as a blend of the languages of the Slavs who partly inhabited this area as early as the sixth century, and of the Slovenian and German colonists who began to populate these lands from the eleventh and especially the twelfth century. The Western branch of this dialect group is characterised by the influence of the neighbouring Upper Carniolan dialect (Ramovš 1935: 48-49, 99, Logar 1996: 407).

Examining the manuscript in the light of the above facts about the standard language norm from the end of the eighteenth century and the dialect region in which it originated, we can establish that the variant readings of the same word are a consequence of a set of influences on language that, at the time, were acceptable. A more detailed analysis of both the phonological and lexical characteristics of the language of the manuscript shows that forms typical of the Protestant standard language tradition are used in the text, followed first by forms typical of the Upper Carniolan variety of the Slovenian standard language, and then by a number of Rovte and/or Upper Carniola dialectal manifestations, which in principle were unacceptable in the standard language of the period. On the basis of these findings, the legitimate variation of the manuscript's language was defined as its main characteristic (Žejn 2016: 421-422).

Returning to the variant readings listed at the beginning of this section, it is possible, on the basis of the history of the Slovenian standard language and dialects, to conclude that the *celu* variant can be attributed to the influence of the Protestant standard language tradition, the *celo* variant to that of the then Upper Carniolan standard language variant, and the *cel* variant to a reflection of the Rovte (as well as the neighbouring Upper Carniolan) dialect group, whose main characteristic is strong vowel reduction. Similarly, the readings *drušina* [society] signifies the following of the then standard language norm, while the reduced variant *drušna* is a result of dialectal influence. The reading *človik* [man] observes the older tradition of the Protestant standard language, while *človek* is a variant introduced by Japelj in his Catholic translation of the Bible for the Upper Carniolan variety of the standard language and is also characteristic of today's standard language. The *stu* and *sto* [a hundred] variant readings again represent the opposition between the Protestant traditions of the standard language and the central

Slovenian variant from the end of the eighteenth century, while the *stou* variant can be attributed to a reading in a base text. The latter is a rare example of phonetic evolution that, in comparison with the earlier manuscript, the author of the manuscript changed to *stu* or *sto* in some instances but kept it in its original form in others.

It is not the purpose of a critical transcript to harmonise the language of a manuscript text with the currently recognised standard language norm but rather to preserve all of the occurring variants so that, on the basis of a critical transcript, we can observe and describe the language, thus significantly advancing our knowledge of the history of the Slovenian standard language.

Types of scribal error

In the light of the above, linguistic analysis helps us to determine which of the forms of a language variant were possible and acceptable in a specific period of time, thus providing us with a basis for the identification of potential scribal errors. The described dilemmas encountered in the linguistic analysis of the manuscript do not, of course, imply that it contains no errors. In fact, the following types of scribal error were identified in the manuscript:

– Missing letters in words at the level of sounds/graphs: although vowel reduction and consonant assimilation are a very common and characteristic phenomenon of the Rovte spoken language or dialect, the missing consonants in *srit* instead of, correctly, *strit*, in GLA *strit* ("storiti" [to do]); *Shentjash* instead of *Shentjansh*, in GLA *Šentjanž* ("Sveti Janez" [St John]); and *pershi* instead of *pershli*, in GLA *peršli* (past part. masc. pl. of the verb "priti" [to come]) should be attributed to scribal errors. This decision is supported by the fact that the remaining occurrences of these words and their derivatives are rendered without the missing letter.

– The occurrence of a wrong letter for which it can be inferred that it is not a result of the linguistic evolution of Slovenian, e.g., *posravi* instead of *postavi* (in GLA *postavi*) as the 3rd ps. sg. form of the verb "postaviti" [to place] – as opposed to *posravi* (in GLA *pozravi*) as the 3rd ps. sg. of the verb "pozdraviti" [to say hello] – which is the result of the zd > z dialectal assimilation.

– Repeated (redundant) word-central syllables, such as *kervavavu* (the same in GLA) as the neutral gender of the adjective "krvav" [bloody] instead of *kervavu*; or *zheszhesnaturalski* (in GLA *čezčeznaturalski*) "nadnar-

aven" [supernatural] instead of *zhesnaturalski* (in GLA *čeznaturalski*). All of the variants listed here as correct occur more than once in the text.

– The use of the wrong word, which is unambiguous from the context, e.g., *Kaiphas* (in GLA *Kajfas*) instead of *Christus* (in GLA *Kristus*).

– Words that were obviously omitted by mistake, most frequently the auxiliary "to be".

Questionable readings

A special case is represented by readings that, when encountered for the first time in the text, might be classified as errors given the absence of an explanation grounded in historical linguistics but for which subsequent readings reveal that it is not a case of a single occurrence but a minimum of two occurrences. One such example is the aforementioned reading of the [č] sound, for which an explanation was found by comparing it with a base text. This prompts us to consider whether such readings can really be regarded as candidates for scribal errors or whether we should allow the possibility that, despite their being exceptional in character, such readings are motivated by reasons unknown to us. Among such cases are:

– three readings of *pra* (3rd ps. sg. of the verb "praviti" [to say]) with the omission of the final *v* – alongside other occurrences of *prau* (in GLA *prav*)

– the readings of *ozhes* (in GLA *očeš*) and *nozhes* (in GLA *nočeš*) (2nd ps. sg. of the verb "(ne) hoteti" [to (not) want]), where *h* was omitted from the digraph in the coda

In the scholarly digital edition, conclusions will be drawn on the acceptability of certain variant readings, as well as on scribal errors and readings that could only conditionally be classified as such. In the concluding section of the paper, we present solutions for the planned scholarly digital edition based on the technological and editorial solutions proposed in the *Text Encoding Initiative (TEI) Guidelines*.

The scholarly digital edition and adopted TEI elements

The scholarly digital edition will be included in the eZISS digital library, whose critical editions each contain a cover page in Slovenian and English, the entire facsimile edition, the original text in the TEI XML transcript, and a web presentation in HTML (Erjavec & Ogrin 2009: 126). The

display in HTML also shows authorial and editorial interventions, thus enabling a parallel viewing of the individual transcripts and the facsimile edition (Erjavec 2005: 62).

In order to implement solutions for the above dilemmas in the scholarly digital edition, we have adhered to the *TEI Guidelines*, from which we have selected elements for annotating errors and questionable readings, while variant readings that can be attributed to the diversity of acceptable instantiations resulting from language evolution are not specifically marked. Nonetheless, a detailed analysis of special linguistic features and their origin will be given in one of the accompanying studies, providing an explanation for the variation in the reading of sounds and words.

1. Readings/variants that were found to be almost certainly, or with the highest degree of certainty, scribal errors are annotated with a TEI <sic> element followed by an editorial correction of the reading marked with the <corr> element, both of which are embedded within the <choice > element. In this way, the choice of elements allows us to preserve the original written form while also providing information on which reading/variant would be correct according to editorial judgment. For example:

<choice>
      <sic>srit</sic>
      <corr>strit</corr>
    </choice>

2. In addition, the dilemmas are specifically marked where it was not possible to say definitively and unequivocally whether it was a case of scribal error or one of the possible variants, for example:

<choice>
      <orig>jogne</orig>
      <reg>jagne</reg>
    </choice>

Thus, the reader of the scholarly digital edition will be informed in two ways about the relationship between the manuscript and the editorial interventions in the text: by the display of the editorial interventions, and by the

possibility of a parallel display of the facsimile and its diplomatic and critical transcript. However, more detailed analyses and mutual comparisons of other previously unknown and unexplored manuscripts will be needed in order to provide definitive answers to unresolved dilemmas and to identify possible modifications of the process of determining specific scribal errors.

List of manuscripts

Register of Early Modern Slovenian Manuscripts. Ms 023. *Poljanski rokopis. Jezusovo življenje v sto postavah*. [*Poljane Manuscript. The Life of Jesus in a Hundred Laws*] URL: http://ezb.ijs.si/fedora/get/nrss:nrss_ms_023/VIEW/ [11. 2. 2019]

Register of Early Modern Slovenian Manuscripts. Ms 028. *Jezusovo življenje v sto postavah*. [*The Life of Jesus in a Hundred Laws*] URL: http://ezb.ijs.si/fedora/get/nrss:nrss_ms_028/VIEW/ [11. 2. 2019]

References

Erjavec & Ogrin 2009: Erjavec, T., Ogrin, M.: "Elektronske znanstvenokritične izdaje slovenskega slovstva eZISS: metode zapisa in izdaje." In Stabej, M. (ed.): *Infrastruktura slovenščine in slovenistike*. Ljubljana: Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik, 2009, 123–128.

Erjavec 2005: Erjavec, T. "Elektronske znanstvenokritične izdaje slovenskega slovstva: standardi in izzivi." In Ogrin, Matija (ed.): *Znanstvene izdaje in elektronski medij.* Ljubljana: Založba ZRC, ZRC SAZU, 2005, 51–70.

Logar 1996: Logar, T. *Dialektološke in zgodovinske razprave.* Ljubljana: ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša, 1996.

Ogrin & Žejn 2016: Ogrin, M., Žejn, A. "Strojno podprta kolacija slovenskih rokopisnih besedil: variantna mesta v luči računalniških algoritmov in vizualizacij." In Erjavec, T., Fišer, D. (ed.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika, 29. september–1. oktober 2016.* Ljubljana: Znanstvena založba Filozofske fakultete, 2016, 125–132.

Ogrin 2011: Ogrin, M. "Neznani rokopisi slovenskega slovstva 17. in 18. stoletja." *Slavistična revija* 59/4 (2011), [386]–399.

Ogrin et al. 2013: Ogrin, M., Javoršek, J. J., Erjavec, T.: "A register of early modern Slovenian manuscripts." *Journal of the Text Encoding Initiative* 4 (2013), 1–13.

Orel 2003: Orel, I. "Slovenski pisni jezik nekdaj in danes – med izročilom in govorom." In Vidovič-Muha, A. (ed.): *Slovenski knjižni jezik – aktualna*

vprašanja in zgodovinske izkušnje. Ljubljana: Center za slovenščino kot drugi/ tuji jezik pri Oddelku za slovenistiko FF, 2003, 551–562.

Orožen 1996: Orožen, M. *Poglavja iz zgodovine slovenskega knjižnega jezika. Od Brižinskih spomenikov do Kopitarja.* Ljubljana: Filozofska fakulteta.

Ramovš 1935: Ramovš, F. *Historična gramatika slovenskega jezika 7, Dialekti.* Ljubljana: Učiteljska tiskarna, 1996.

Žejn et al. 2017: Žejn, A., Erjavec, T., Ogrin, M.: "Poljanski rokopis. Baročno besedilno izročilo o Jezusovem življenju v elektronski izdaji." In Bjelčevič, A., Ogrin, M., Perenič, U. *Rokopisi slovenskega slovstva od srednjega veka do moderne.* Ljubljana: Filozofska fakulteta, Oddelek za slovenistiko, Center za slovenščino kot drugi in tuji jezik. 2017, 111–118.

Žejn 2016: Žejn, A.: "Jezik *Poljanskega rokopisa*: jezik rokopisa na prehodu iz 18. v 19. stoletje." *Slavistična revija* 64/4 (2016), 407–425.

About the author
Andrejka Žejn, andrejka.zejn@zrc-sazu.si, is a researcher at the ZRC SAZU Institute for Slovenian Literature and Literary Studies in Ljubljana, Slovenia. She received her PhD in Carinthian dialectology from the University of Ljubljana. As well as carrying out research on the beginnings of Slovenian narrative prose, she is involved in the publication of the Institute's scholarly critical editions of Slovenian manuscripts from the eighteenth century.

Summary
The discovery of numerous manuscript texts in Slovenian in recent years has provided important material for furthering and re-examining our knowledge of the history of Slovenian language, literature and culture. For a complex analysis of new discoveries, a multidisciplinary approach, including historical linguistics and dialectology, is required, which should help resolve dilemmas about whether a particular reading in the manuscript text is an acceptable, legitimate historical variant or a scribal error. Using one of the most prominent recent discoveries for Slovenian language, literature and history, the Poljane manuscript, as an example, we demonstrate the factors that must be taken into account when deciding whether a particular reading is correct or incorrect. This assessment refers to the so-called critical transcript of a manuscript, an integral part of the scholarly critical edition of the manuscript, which, among other things, involves a transcription of the text written in the historical Bohorič alphabet using contemporary spelling (Gaj's Latin alphabet). In addition to the

alphabet, an important role is played by the history of the Slovenian standard language and the dialectal situation, as well as the relationship between the manuscript in question and its base text.

An analysis of the readings of individual sounds in the Bohorič alphabet in relation to Gaj's Latin alphabet gives a relatively clear picture of the relationship between the two spellings, despite the diverse readings of certain sounds, especially (alveolar) sibilants. However, readings can be found that deviate and, on the face of it, appear to be errors. One such example is the case of several exceptions in the reading of the [č] sound, which we can conclude are due to the influence of a base text that served as a model for the manuscript.

In the manuscript, many variants of the same word occur that cannot be associated with the alphabet used. These variants can be accounted for by an analysis of the development of the Slovenian standard language and by observation of the characteristics of the dialect spoken in the manuscript's place of origin. Based on the tradition of the mid-sixteenth-century standard language, the Slovenian standard language in the period of the *Manuscript's* compilation evolved in four varieties, all of which were subject to dialectal influences. The area to which the emergence of the manuscript is ascribed was characterised by the Upper Carniolan variety of standard Slovenian, which, to a higher degree than other varieties, followed the existing standard norm. At the same time, it was the area where the Rovte dialect was spoken under the influence of the neighbouring Upper Carniolan dialect. Consequently, the following can be found in the manuscript: 1) variant readings typical of the standard language from the sixteenth century onwards; 2) variant readings that deviate from tradition and that observe the then Upper Carniolan variety of standard Slovenian; and 3) variant readings emerging under the influence of varieties of dialectal evolution that were not acceptable for the standard language. The critical transcript refrains from correcting the numerous variant readings, irrespective of whether they are defined as part of the standard language or a dialect.

A detailed linguistic analysis also serves as the basis for determining one-off readings that can be defined unequivocally as scribal errors, e.g., an omitted letter that cannot be attributed to vowel reduction or consonant assimilation, a repeated syllable, or a semantically incorrect word. In addition to linguistically acceptable varieties and scribal errors, the third type is readings that are exceptional with respect to the dominant practice but that, despite initially appearing to be scribal errors, actually occur more than once. The question therefore arises as to whether it is a case of genuine scribal errors, or whether a satisfactory justification for these occurrences could potentially be found on the basis of a knowledge of several texts from the same period.

In the scholarly digital edition, the readings defined as scribal errors and those for which we are unable to answer unambiguously whether they are scribal errors or legitimate variants are annotated by means of TEI elements. In this way, and with the possibility of comparing the facsimile edition with the diplomatic and critical transcripts, dilemmas arising with regard to individual readings are presented along with their possible solutions.