A/O
3885

COMPUTER PROCESSING OF MEDIEVAL SLAVIC MANUSCRIPTS

PROCEEDINGS

FIRST INTERNATIONAL CONFERENCE

24 – 28 July, 1995, Blagoevgrad, Bulgaria

"Prof. Marin Drinov" Academic Publishing House

КОМПЬЮТЕРНА ОБРАБОТКА НА
СРЕДНОБЕЛКОВИ СЛАВЯНСКИ
РЪКОПИСИ

ДОКЛАДИ

ТРЕТА МЕЖДУНАРОДНА КОНФЕРЕНЦИЯ
24 – 25 юли 1995
Благоевград, България

COMPUTER PROCESSING OF MEDIEVAL SLAVIC MANUSCRIPTS

PROCEEDINGS

FIRST INTERNATIONAL CONFERENCE
24 – 28 July, 1995, Blagoevgrad, Bulgaria

КОМПЮТЪРНА ОБРАБОТКА НА СРЕДНОВЕКОВНИ СЛАВЯНСКИ РЪКОПИСИ

ДОКЛАДИ

ПЪРВА МЕЖДУНАРОДНА КОНФЕРЕНЦИЯ

24 – 28 юли, 1995, Благоевград, България

Академично издателство "Марин Дринов"

София

1995

681.39:091=81"04/14"+091=81"04/14":681.39
B/6398

This conference was supported by grants from the International Research and Exchanges Board (IREX), with funds provided by the U. S. Department of State (Title VIII) and the John D. and Catherine T. MacArthur Foundation, by the Joint Committee on Eastern Europe of the American Council of Learned Societies (ACLS), and by the Soros Center for the Arts - Sofia. Hospitable local organizer of the conference was American University in Bulgaria, Blagoevgrad.
The book of proceedings is sponsored by grant from IREX.

None of these organizations is responsible for the views expressed. Editors of this book would also like to acknowledge the generous support of the Bulgarian Academy of Sciences, University of Sofia and the University of Pittsburgh, PA, USA.

ПРОВЕРКА 2003
0 3885
БАН ЦЕНТРАЛНА БИБЛИОТЕКА ДАР
Б 2573/96

© David J. Birnbaum, Andrej T. Bojadžiev, Milena P. Dobreva,

Anisava L. Miltenova (editors)

© The copyrights of the proceedings belong to their authors

ISBN 954-430-417-7

CONTENTS

Foreword.....9

Session: **STATE OF THE ART**

Birnbaum, David. How Slavic Philologists Should Use Computers.....19

Bader, Winfried. Bible and Computer: A Brief History of the Last Ten Years (abstract)

.....29

Paskaleva, Elena, Milena Dobreva. New Tools for Old Language: Computer Processing of Old Bulgarian Texts.....30

Session: **ENCODING PROBLEMS**

Birnbaum, David. Informational and Presentational Units in Early Cyrillic Writing.....41

Grünberg, Karsten. Transcription Rules for Old Church Slavonic Writing.....50

Kempgen, Sebastian. Complex Script-Systems on Today's Personal Computer.....68

Session: **TEXT PROCESSING: PROBLEMS, METHODS AND TOOLS**

MacRobert, Mary. A TEI View of MS Pec 68.....79

Azarova, Irina. Processing Slavonic Bible Texts in the Russian Bible Society (Sankt-Petersburgh branch).....89

Молдован, Александр. О создании фонда письменных источников русского языка XI—XII вв. в Институте русского языка РАН.....94

Session: **MANUSCRIPT COLLATION**

Veder, William. Transcription and Edition (abstract).....96

Bakker, Michael. Computer Collation of Manuscript Transcription (abstract).....96

Ivanova, Tanja, Nina Šojleva, Irina Bolcheva. Advanced Searching.....97

Session: **FONTS AND IMAGE PROCESSING**

Majstoroska, Mimoza. A Church Slavonic Alphabet for Reprinting Old Manuscripts Using a Microcomputer.....103

Mitrevski, Ljupčo. Fund, Fonts, and Character Set of Church Slavic Graphemes.....114

Stoyanov, Ivelin. Optical Character Recognition of Historical Documents.....119

Session: **PROBLEMS IN RUNNING A CONCRETE PROJECT:**

Computer-Supported Processing of Medieval Slavic Manuscripts

Miltenova, Anisava. Computer Assisted Analysis of the Macrostructure and Typology of Medieval Slavic Miscellanies.....129

Bojadžiev, Andrej. Paleographic and Orthographic Features in the Description of Slavic Manuscripts.....140

Session: **APPLICATIONS OF DATA BASE MANAGEMENT SYSTEMS**

Jankoska, Stana, Dragan Mihajlov, Ljubomir Josifovski. Database of Slavonic Manuscripts in Macedonia.....147

Ribarova, Zdenka, Kirill Ribarov. Computer Processing of Old Church Slavonic Manuscripts: Results and Prospects.....153

Lampe, Reinhard. Создание единного сводного каталога славянского письменного наследия.....159

Лабынцев, Юрий. Международный исследовательский проэкт "Slavia Orthodoxa et Slavia Romana: Взаимодействие славянских миров: Духовная культура Подляшья".....168

Щавинская, Любовь. Основные задачи и принципы построения банка данных Международной славяноведческой программы "История книжной культуры Подляшья".....177

Session: **CASE STUDIES. TYPES OF MANUSCRIPTS**

Camuglia, Monia. The Psalter: From the Oral to the Informatic Tradition.....185

Vakareliyska, Cynthia. Medieval Slavic Menologies On Line.....197

Guergova, Emilia. Computational Analysis of Hymnographic Compendia.....207

Šniter, Marija. A System for Encoding Euchological (mainly non-calendar) Texts within the Parameters of the *IST* Computer Program.....215

Salter, Francis. Some Suggestions Arising from a Personal Computer Analysis of the Svetostefansko Hrisovulje.....228

Session: MULTIMEDIA

Bratanova, Velina, Elena Kotseva, Dragomir Petrov. The St. Sophia Project. A Stage in its Realization and the Prospects for its Completion.....238

Session: KLEIO

Тяжелникова, Виктория. Русская версия источнико-ориентированного программного обеспечения КЛИО: новые возможности обработки источников.....257

Tikhonov, Vladimir. The Problems of Categorisation in Content Analysis.....267

Юшин, Игорь. Источнико ориентированный подход и проблемы социальной истории.....277

Angusheva, Adelina. Computer Investigation of Medieval Prognostic Books by Means of KLEIO.....291

Dimitrova, Margaret. Loanwords in the New York Missal and the KLEIO Computer Program.....298

SUPPLEMENT*

Янакиев, Мирослав. Компютърните технологии и текстологията.....315

Klasinc, Peter Pavel, Vinko Kurent. First Results of Testing the Computer Recognition of Manuscrips.....323

Xristova, Borjana. The First Bulgarian CD-ROM: Bulgarian Letters in the Context of Balkan Literature, Tetraevangelia and Qurans.....332

* Papers were included in the program of the conference, but had not been read out because of the absence of their authors.

Foreword

I. COMPUTER-SUPPORTED RESEARCH AND TEACHING in the humanities has been growing at an increasing pace over the past two decades, with new methods for using computers to increase productivity in these areas, and new types of software, appearing continuously. Unfortunately, the field of Slavic studies (and especially medieval Slavic studies) has lagged well behind other areas of the humanities in applying modern information technology to improve research, scholarship, and teaching. The absence of well-developed computational standards and resources in Slavic studies results largely from peculiarities of our discipline, which have made it impossible simply to plug Slavic data into systems developed for studying other cultures. As this conference has demonstrated, these difficulties are not insurmountable.

The difficulties in question can be addressed and resolved through the active collaboration of Slavic philologists and information science specialists who understand the requirements of humanities computing. While there is a real danger that the field of Medieval Slavic studies will continue to remain isolated from modern research and teaching methods, the present conference has demonstrated that an opportunity exists to bridge the gap between Slavic studies and developments elsewhere in humanities computing. The papers included in this volume provide evidence of substantial results both in standardizing the computational treatment of Slavic materials and in generating significant products suitable for distribution to researchers and students. Because the manuscript sources with which we work are basic tools in both the social sciences and the humanities, the results of our efforts ultimately will lead to increased access to research and teaching materials in both of these broad disciplinary areas.

One aspect of the explosive growth of computer-supported research and teaching in the humanities, aided particularly by the development of hypermedia technology, has been the collection of large text corpora for research purposes, including such significant, widely-accessible, large-scale projects as the Greek *Thesaurus Linguæ Graecæ* and *Perseus Project.* Some of these developments are described by Linda W. Helgerson in "CD-ROM and Scholarly Research in the Humanities" (*Computers*

and the Humanities, vol. 22 [1988], 111–16) and in *The Humanities Computing Yearbook 1989–90*, ed. I. Lancashire, Oxford: Clarendon, 1991 (passim). However, as is noted above, the field of Slavic studies, and especially medieval Slavic studies, lags well behind other areas of the humanities in the application of modern information technologies to research, scholarship, and education. This is readily apparent from the modest "Slavic Languages" section of the aforementioned *Humanities Computing Yearbook* (by David J. Birnbaum and Harry Gaylord, pp. 300–29). While Greek scholars, who also have to deal with alphabetic peculiarities, have produced the major projects mentioned above, there are no comparable resources in medieval Slavic studies. This disparity can be attributed to at least four sets of reasons:

A. Lack of International Coordination and Cooperation

For political and economic reasons, there has long been a lack of information in the West about research activities in Central and Eastern Europe, and vice versa. There has even been a lack of coordination and cooperation within Central and Eastern Europe, so that while researchers in Bulgaria have begun to use computers for word processing, collecting bibliographies, etc. over the past several years, similar work done in Macedonia, Russia, the Czech Republic, Poland, and elsewhere has been carried out on different computer platforms, using different word processors, different fonts, and, most important, different text architectures and character and glyph encoding systems. A principal goal of this conference has been to facilitate coordinated and compatible resource development by scholars in all participating countries.

B. Encoding and Rendering (Display and Printing) of Medieval Slavic Texts

Popular and widely available computer tools for word processing, collecting text corpora, collating critical editions, generating bibliographies, and other activities involving text processing are normally designed first for an English-language market, and then sometimes expanded to deal with other Latin-alphabet languages, and only then sometimes expanded to address non-Latin text processing. Computer manufacturers are understandably reluctant to develop special products for academic "niche" markets that do not present a financially significant pool of potential customers. While some

developers have produced limited support for the computational processing of modern Slavic languages written in Cyrillic (primarily modern Russian), these products are often incompatible with one another, resulting in, for example, the generation of text files that cannot be exchanged among users of different hardware, operating systems, application software, or fonts. While the problems facing those who use Cyrillic, and particularly those who use medieval Cyrillic, are more complex than is the case with ASCII-only systems, there is no inherent technical reason why electronic versions of medieval Slavic texts should not be as standardized, uniform, and interchangeable as English-language ASCII texts. This issue, too, was a central theme of the Blagoevgrad conference.

C. Difficulties in Using Modern Educational and Research Software

The computer processing of images, which is widely used in modern educational and research tools, was not commonly available in Central and Eastern European countries for economic reasons. As a result, established libraries of texts, images, and audio and visual aids in computer form, which could be included in hypermedia databases, have begun to appear only relatively recently. As the economies of these countries stabilize and the costs of hardware drop, this situation should begin to remedy itself, and many of the contributions to this volume describe both on-going research (such as innovative optical character recognition programs from Macedonia and Bulgaria) and the development of significant databases that incorporate image files (such as the multimedia *St. Sofia* CD-ROM).

D. Specific Features of Early Slavic Writing

One difference between the practice of classical and Slavic philology is that classical scholars most often work with normalized texts and concentrate on major editorial variation, with fine paleographic and orthographic vagaries remaining the esoteric domain of epigraphers. This has also been a trend in Slavic philology, but Slavic linguists (in particular) have been concerned precisely with these details, which means that they have had a greater interest in retaining variation when transcribing manuscript materials. A consequence of this focus on the significance of variation has been difficulty in using computer text corpora for retrieval and collation purposes, since searching for an item that may be spelled in several different ways is inherently more complicated than searching for a

single spelling. But full normalization of the textual material during input would lead to a loss of information that, while harmless for some disciplines, would render certain types of research (primarily linguistic and orthographic, but also some aspects of textual studies) impossible.

A particular problem concerns the use of superscription in medieval Slavic manuscripts, affecting letters, accent marks, musical notes, and other nonalphabetic characters, since most software and encoding systems are ill-equipped to handle not only the variation between superscribed and non-superscribed spellings, but also the nonlinear presentational aspects of superscription. One related problem in text encoding has been a traditional failure to distinguish informational units (characters) from presentational ones (glyphs), a distinction that is crucial for efficient text encoding and processing, especially when dealing with orthographic complexity of the sort found in medieval Slavic manuscripts. The crux of this problem is that minor variation is vital for some purposes, while it is an obstacle for others, and many of the reports in this volume address precisely the different—and sometimes contradictory—requirements of different types of research.

II. Over seventy persons from ten countries gathered in Blagoevgrad in the last week of July, 1995, in an effort to examine these problems and propose solutions to them. The present volume is intended to make the results of that conference accessible to a wider audience.

In the "State of the Art" session, David J. Birnbaum's (Pittsburgh) "How Slavic Philologists Should Use Computers" emphasized the importance of multiple use, structure, portability, preservation, and standardization in the development of document-encoding systems for Slavic philological research. Winfried Bader's (Tübingen), "Bible and Computer: A Brief History of the Last Ten Years" contrasted one of the oldest uses of computer-based text studies (biblical scholarship) with one of the youngest (Slavistic scholarship). Elena Paskaleva and Milena Dobreva's (Sofia) "New Tools for Old Language: Computer Processing of Bulgarian Texts" surveyed both the problems involved in the use of computers to process early Slavic texts and some of the attempts that have been made to resolve these problems, particularly in Bulgaria.

In the "Encoding Problems" session, David J. Birnbaum's (Pittsburgh) "Informational and Presentational Units in Early Cyrillic Writing" discussed theoretical and methodological prerequisites for developing character- and glyph-encoding systems that are flexible and powerful enough to satisfy the varying requirements of different Slavic researchers, while simultaneously providing a standardized framework capable of generating portable documents. Karsten Grünberg's (Heidelberg) "Transcription Rules for Old Church Slavonic Writing" examined transcription rules for representing the information in early Slavic manuscripts, and demonstrated some of the results achieved through the use of TUSTEP to process transcribed material in a practical research environment. Sebastian Kempgen's (Bamberg) "Implementing a Medieval Script on a Personal Computer" surveyed both general theoretical considerations and concrete implementation problems of font development in current environments.

In the "Text Processing: Problems, Methods, and Tools" session, Mary MacRobert (Oxford) reported on her application of the Text Encoding Initiative (TEI) guidelines to encoding a fourteenth-century Serbian Church Slavonic Psalter, drawing particular attention to such problems as concurrent hierarchies and overlapping elements. Aleksandr Moldovan's (Moscow) "О создании фонда письменных источников русского языка XI-XII вв. в Институте русского языка РАН" stressed the importance of coordination to avoid duplication of effort, and reported on theoretical and practical issues affecting large-scale encoding projects currently under way at the Russian Language Institute of the Russian Academy of Sciences. Irina Azarova's (St. Petersburg) "Проблемы и методика издания славянского перевода библии" reported on corpus-based research on fifteen hundred early Slavic biblical manuscripts currently being conducted by the Russian Bible Society.

In the "Manuscript Collation" session, Michael Bakker's (Amsterdam) "Computer Collation of Manuscript Transcriptions" discussed theoretical and practical issues involved in using *Collate*, a commercial Macintosh-based program developed in England for collating manuscript witness, in a large-scale Slavistic project. Tanja Ivanova, Nina Šojleva, and Irina Bolčeva's (Sofia) "Advanced Searching" demonstrated the use of *Advanced Searching*, a shareware program developed in Sofia to

provide *Collate*-like capabilities in a Microsoft Windows environment, in their research on the textual transmission of Černorizec Xrabr's "O pismenex."

In the "Fonts and Image Processing" session, Mimoza Majstoroska's (Skopje) "A Church Slavonic Alphabet for Reprinting Old Manuscripts Using a Microcomputer" discussed the graphic analysis of early Cyrillic writing and demonstrated the application of original optical character recognition software based on this analysis to medieval manuscript materials. Ljupčo Mitreski's (Skopje) "Fonts, Character Sets, and Church Slavonic Graphemes" discussed keyboard layout and other user-interface issues involved in editing early Cyrillic texts on a computer. Ivelin Stojanov (Sofia) also discussed the development of optical character recognition systems for early Slavic manuscript materials. William R. Veder's "Transcription and Edition" emphasized the importance of developing portable file formats and encoding methods to support collaborative work in Slavic philology.

In the "Problems in Running a Concrete Project: Computer-Supported Processing of Medieval Slavic Manuscripts" session, Anisava Miltenova's (Sofia) "Computer-Aided Analysis of the Macrostructure and Typology of Medieval Slavic Miscellanies" discussed both philological and informatic aspects of studying early Slavic manuscripts on a computer, with particular attention to the structural analysis of mixed-content miscellanies. Andrej Bojadžiev's (Sofia) "Paleographic and Orthographic Features in the Description of Slavic Manuscripts" addressed the identification and encoding of parameters for describing codicological, paleographic, orthographic, and linguistic features of early Cyrillic manuscript materials. Both papers in this session were presented within the framework of the Text Encoding Initiative (TEI) application of Standard Generalized Markup Language (SGML).

In the "Applications of Data Base Management Systems" session, Stana Jankoska, Dragan Mixajlov, and Ljupčo Josifoski's (Skopje) "A Database for Ancient Slavic Manuscripts in Macedonia" outlined the parameters used for describing early Slavic manuscripts in conformity with UNIMARC and other international cataloguing standards. Zdenka Ribarova (Skopje) and Kirill Ribarov's (Prague) "Computer Processing of Old Church Slavonic Manuscripts: Results and Prospects"

described the powerful STINO system, which is used in the Prague Old Church Slavonic dictionary project for database management and lexical analysis. Reinhard O. Lampe's (Heidelberg) "Problems Involved in Setting Up a Unified Comprehensive Catalogue of Slavic Texts" described the management of multilingual data from a computational perspective. Jurij Labyncev's (Moscow) "Международный исследовательский проект 'Slavia Orthodoxa et Slavia Romana': Взаимодействие славянских миров: Духовная культура Подляшья" outlined the plans for a broad-ranging encyclopedic study of materials in all languages that pertain to both Orthodox and Roman culture. This report on the general intellectual organization of the project set the background for Larisa Ščavinskaja's technical report on the use of relational and source-oriented databases (such as *Paradox* and *KLEIO*) and statistical analysis software in its implementation: "Основные задачи и принципы построения банка данных Международной славяноведческой программы 'История книжной культуры Подляшья.'"

In the "Case Studies: Types of Manuscripts" session, Monia Camuglia's (Pisa) "The Psalter: From the Oral Tradition to the Informatic One" described the application of the *Textual Data Base (DBT)* program to an analysis of textual variation of fifteen psalms in seven psalters. This report addressed issues involved in configuring *DBT* to work with early Cyrillic materials, and also the use of the *DBT* query system to analyze and navigate the text. Cynthia Vakareliyska's (Eugene) "Medieval Slavic Menologies On Line" outlined information retrieval desiderata for a menological database. Emilija Gergova's (Sofia) "Computational Analysis of Hymnographic Compendia" discussed both philological and computational problems involved in studying the multilayered structure of hymnographic sources, with particular attention to the October Menaion. Marija Šniter's (Plovdiv) "A System for Encoding Euchological Texts within the Parameters of the *IST* Computer Program" examined criteria for assigning cataloguing numbers to euchological texts, based on her analysis of some eighty-five mostly South Slavic records from the eleventh through eighteenth centuries. Francis R. Salter's (Bracebridge) "Some Suggestions Arising from a Personal Computer Analysis of the *Svetostefanska Xrisovulja*" reported on a project undertaken during the early stages of development of

computer support for Church Slavonic writing, and addressed issues of programming and data structures required to encode and manipulate the text both at that time and from a modern perspective.

In the "Multimedia" session, Dragomir Petrov and Elena Koceva (Sofia) reported on the development of the UNESCO-supported *St. Sofia* CD-ROM project (in collaboration with Velina Bratanova), which provides text, images (including medieval manuscripts), and audio material concerning all aspects of the development of early Bulgarian culture.

The final conference session was devoted to *KLEIO* and included presentations by both historians and philologists. Viktorija Tjažel'nikova's (Moscow) "A Russian Version of the Source-Oriented Software *KLEIO:* New Opportunities for the Treatment of Cyrillic Sources" outlined the features of *KLEIO* that support the encoding of structure and semantics together with physical data. Vladimir Tixonov's (Moscow) "Interactive Full-Text Processing in *KLEIO:* Scanning Criteria of Basic Terms (On Appeals by Disenfranchised People)" reported on the practical application of *KLEIO* to the scanning and content-based analysis of appeals by people whose civil rights had been revoked in early Soviet Russia (lišency). Igor Jušin's "A Model of Historical Sources in *KLEIO* as a Basis for Integrated Social Classification" outlined several traditional problems involved in the computational analysis of systems of social classification, and discussed how a source-oriented system such as *KLEIO* addresses them. Nadežda Romankova's (Smolensk) "Проблемы атрибуций анонимных старославянских сочинений и компьютер" discussed the use of statistical analysis to resolve the authorship of works possibly attributable to Kliment Oxridski. Adelina Anguševa's (Sofia) "A Computer Investigation of Medieval Prognostic Books by Means of *KLEIO*" discussed the use of *KLEIO* to compare kalendologia in Byzantine, Latin, and Slavic written culture. Margaret Dimitrova's (Sofia) "Greek and Latin Loanwords in the New York Missal and the *KLEIO* Computer Program" demonstrated how *KLEIO* can be used to generate indices of loan words and lists of their phonetic and morphological peculiarities, which can then be subjected to further study.

III. Some of the most interesting and promising parts of the conference are not reflected in this volume because they took the form of demonstration workshops, rather than formal reports. Winfried Bader (Tübingen) demonstrated the capabilities of *TUSTEP* for philological research, and William Veder and Michael Bakker (Amsterdam) demonstrated *Collate* and other Macintosh-based tools. Hands-on demonstrations were also held for *Advanced Searching* and two CD-ROM products, the *"St. Sofia" Project* and the Bulgarian National Library's *"Unique Balkan Manuscripts"* album of Slavic and other manuscript folios. Winfried Bader (Tübingen), David J. Birnbaum (Pittsburgh), Nicholas Finke (Cincinnati), Harry Gaylord (Groningen), and Anisava Miltenova (Sofia) conducted a half-day Text Encoding Initiative (TEI) workshop, followed by a separate hands-on session that provided the participants with an opportunity to use SGML editing software.

IV. A general consensus emerged among many participants at the conference that SGML, particularly in its TEI implementation, provides a framework that is potentially capable of satisfying the varying needs of different researchers, while at the same time affording a common, standardized model for document architecture and character and glyph encoding. Several conference participants engaged in significant text-encoding projects agreed to combine their efforts to adapt the TEI guidelines to the specific needs of Slavic philological research. Other concrete results of the conference were the immediate establishment of an anonymous FTP archive for early Cyrillic texts and related electronic files (ftp.pitt.edu in dept/slavic, maintained by David J. Birnbaum [Pittsburgh]) and plans for publishing a World Wide Web page (maintained by Ralph Cleminson [Portsmouth]). Colleagues from several countries volunteered to contribute electronic texts to the ftp archive, and we hope that this archive and World Wide Web page will provide long-term, stable, centralized locations for the collection and dissemination of relevant information and materials.

V. Publication costs required that the present volume be printed from camera-ready copy supplied by the individual authors, without further editing, which has led to considerable variation in terminology, style, systems of transliteration, and other editorial details. The editors consider this inconsistency a valuable illustration of the importance of developing standardized methods and systems

for document encoding and processing. Certain papers presented at the conference are not included in the present volume because camera-ready copy was not received in time; these papers are represented here instead by reprints of abstracts prepared by the authors as part of the original conference program.

VI. The publication of this volume, and the costs of the conference itself, were supported by grants from the International Research and Exchanges Board, with funds provided by the U.S. Department of State (Title VIII) and the John D. and Catherine T. MacArthur Foundation, by the Joint Committee on Eastern Europe of the American Council of Learned Societies, by the Open Society Fund, and by the Soros Center for the Arts (Sofia). None of the these organizations is responsible for the views expressed. The Organizing Committee (David J. Birnbaum [Pittsburgh], Andrej Bojadžiev [Sofia], Milena Dobreva [Sofia], and Anisava Miltenova [Sofia]) would also like to acknowledge the generous support of the Bulgarian Academy of Sciences, the University of Pittsburgh, the University of Sofia, and the American University of Blagoevgrad. *Author/Editor* Software for the SGML hands-on session was contributed by SoftQuad, Inc. (info@sq.com).

David J. Birnbaum
Department of Slavic Languages and Literatures
1421 Cathedral of Learning, University of Pittsburgh
Pittsburgh, PA 15213 USA
djbpitt+@pitt.edu
http://www.pitt.edu/~djbpitt

How Slavic Philologists Should Use Computers

David J. Birnbaum
Department of Slavic Languages and Literatures
1421 Cathedral of Learning, University of Pittsburgh
Pittsburgh, PA 15213 USA
djbpitt+@pitt.edu

KEYWORDS: multiple use, structure, portability, preservation, standardization

ABSTRACT: The present report discusses electronic text processing in Slavic philology from the perspectives of MULTIPLE USE, STRUCTURE, PORTABILITY, and PRESERVATION. It is argued that texts created with attention to these four issues are a much greater contribution to science than texts created in a proprietary fashion by a small group of users working in isolation. In fact, texts created without attention to these issues are guaranteed restricted use, a limited audience, and a short lifespan. Cooperative STANDARDIZATION is a valuable tool for achieving the goals listed above, and in thus making research more accessible to colleagues both today and in the future.

Slavic philologists use computers for many different purposes. Surely the most widespread computer application for Slavic philologists is word processing, which provides easier revision control and better font support than is available on typewriters. But while revision control and font support are unquestionably valuable resources, they represent only a small portion of the research advantages available through the use of computers. One premise underlying modern work in electronic text technology is that 1) *a computer is not just a better typewriter.* Four other premises, which are elaborated below, are that 2) *data should be separated from processing,* 3) *text for philologists is not just a sequence of characters,* 4) *text has structure that can be exploited by researchers,* and 5) *the value of scholarship is related to its accessibility.*[1]

The present report outlines some guidelines for improving the status and utility of electronic texts in Slavic philology. The somewhat provocative title, "How Slavic Philologists Should Use Computers," does not imply an attempt to persuade others to use a particular hardware or software product because it is allegedly better than what they are currently using, a sort of missionary work

[1]Some of these principles are adapted from Text Encoding Initiative (TEI) workshop materials originally prepared by Lou Burnard and Michael Sperberg-McQueen.

that is always misdirected. Rather, the goal of this report is to outline how to produce portable electronic text files that can be read on any hardware or software platform. Electronic texts may reasonably require certain *types* of hardware or software, but an electronic text that requires all users to purchase the same specific piece of hardware or software unnecessarily restricts itself to a limited audience. And since Slavic philologists are already a limited audience, it is all the more important to avoid any further restriction.

The present report discusses four goals that should not be controversial, and that should govern the way Slavic philologists use electronic texts. These are MULTIPLE USE, STRUCTURE, PORTABILITY, and PRESERVATION. By MULTIPLE USE, I mean that the format of an electronic document should not restrict the ways in which it can be used. This was the point of my earlier premise that *data should be separated from processing.* By STRUCTURE, I mean that we should not be restricted to operations like full-text searching, which treat a text as if it were an undifferentiated sequence of characters. Full-text searching is a valuable tool, but it is no less important to be able to search a full text transcription for information that occurs only within certain structural divisions of the work. This was the point of my premises that *text for philologists is not just a sequence of characters,* and that *text has structure(s) that can be exploited by researchers.* By PORTABILITY, I mean that the format of an electronic document should not restrict the platform on which it can be processed today, so that Slavists should be able to access one another's electronic texts even though we work on different platforms. By PRESERVATION, I mean that the format of a document should not restrict the platform on which it will be able to be processed in the future, so that today's texts will still be accessible even when all of today's platforms are long forgotten. When examined closely, it turns out that these four issues are interrelated, and that all are connected to the general concept of STANDARDIZATION. By STANDARDIZATION, I mean that the format of our electronic texts should belong to us all, as a scholarly community. This format should be stable, and nobody should own or control it in a way that would limit anyone else's ability to produce or access conforming documents. A standard for a scholarly community absolutely

cannot be a commercial tool that compels users to buy a particular manufacturer's products in order to obtain access to data, or that can be changed on the whim of a manufacturer.

MULTIPLE USE. If Slavic philologists use computers only as typewriters, and electronic texts only as an intermediate step toward paper texts, they greatly limit the utility of their research. To be sure, computer typesetting can increase productivity considerably, and Slavists who prepare printed editions on computer can produce more editions more quickly than those who do not. But the end result is still a paper edition; computers can increase the quantity, and perhaps the accuracy, of this type of work, but the nature of the product remains the same. Printed editions will continue to be a very important tool of our trade, and I do not by any means intend to suggest that they are obsolete, or that they are likely to be supplanted entirely by electronic texts. But printed editions are no longer the only tool in our philological workshop. In addition to improving our ability to perform the traditional task of producing paper editions, electronic texts can also support types of research that would not be practical, or perhaps even possible, with paper texts. At the simplest level, for example, we can use a single electronic file to prepare different types of printed output, such as texts for both diplomatic and critical editions, so that different readers with different interests can obtain different views of the identical physical data file. Rather than decide whether to normalize a text during transcription or not, it may prove more useful simultaneously to normalize and not to normalize, and then to generate different views on demand, depending on specific research tasks.

Electronic texts also provide new types of access to information. A paper edition presents a text in a single, specific order, usually corresponding to the order of the text in a manuscript witness. If we want to know what appears on folio 6 verso, we can find it almost instantaneously, because the arrangement of the printed edition corresponds to the order of folios in the original manuscript source. But what we cannot do easily with a paper edition, or even with an original source, is find all examples of a particular word wherever it occurs in the manuscript. We can obtain this information through linear searching, i.e., by starting at the beginning and reading until the end, but this is

time-consuming and impractical. The traditional scholarly shortcuts for this type of research are indices and concordances, but an index or concordance traditionally has to be prepared and published separately from the transcription of the manuscript. The task is not impossible, and people compiled indices and concordances to voluminous manuscripts before there were computers, but it represents a tremendous commitment of time and effort.

With an electronic text, on the other hand, the *search* feature of most word processors can find any form in the text in an instant, in many cases making an index or concordance unnecessary. And if a traditional index or concordance is required, a computer program that requires half a day to write and that occupies no more than a page or two of code can generate a sorted word list with or without context and with or without folio and line references, and can do so *for any manuscript.* In other words, a properly prepared electronic transcription does not provide merely a transcription of the text; it provides a searchable text, an index, and a concordance at the same time. These are not perfect indices or perfect concordances, of course, but their value is obvious to anyone who has used them.

The preceding uses of electronic text, as valuable as they are, are based primarily on the notion of text as a sequence of characters, a definition I criticized earlier. The searching one does in a word processor is unstructured; you enter a string of characters, and the search engine ploughs through the text from top to bottom, reporting matches as they are encountered. More sophisticated electronic text encoding can support STRUCTURED searching, such as the ability to match strings only when they appear in rubrics, or only when they appear in specific years in a chronicle manuscript, or only when they appear in specific texts in a miscellany, or only when they appear in direct speech, or only when they are not biblical citations, etc. A sophisticated electronic text can make it possible to extract a table of incipita from a full-text transcription automatically, a very useful task that is much more time-consuming with unstructured electronic text, not to mention with paper editions.

Electronic texts can also support hypertextual links, which allow the reader to move in any direction both within a single electronic text and among related electronic texts. Transcribed primary sources can be used in dynamic critical editions, with variable copy and control texts. Reference materials of all sorts, from dictionaries to bibliographies, can be accessed from within an electronic file. Hypertext links can associate parts of a transcription with parts of a scanned image, so that instead of just having facsimile and transcription on facing pages, as with the *Codex Suprasliensis* edition, links can be created between exact locations within those pages.

Any electronic text supports a variety of uses, and full-text searching comes for free in just about any word processor. But more sophisticated tasks, such as structured searching or hypertextual linking, require careful planning before and during transcription. I think that the main reason most Slavists do not prepare this type of rich, multiple-use electronic document is that they do not know how to do so. The trick is to encode electronic texts so that they include all the information needed for research, but to do so in a way that prevents information needed for one type of access from interfering with information needed for another type of access. An intelligent electronic edition requires not just philological expertise, but also informatic expertise, and most Slavists are not trained in informatics. I suggest that multiple-use documents are a goal worth pursuing, and that Slavic philologists should learn how to publish useful electronic texts, either on their own or through cooperation with informatics specialists. Otherwise, we fail to take advantage of many of the strengths of electronic text, and we needless restrict ourselves to just one or two blades of a very comprehensive philological Swiss army knife.

Portability. The body of primary materials used by Slavic philologists is far too great for any individual to encode alone, yet most Slavists use only electronic texts that they themselves or their closest collaborators have prepared. There are journals and publishing companies that are devoted to making paper publications in Slavic philology available to a worldwide community of scholars, but if Slavists share electronic texts at all, they do so through *samizdat*, and this community is in any case

very limited. Just as generations of Slavists have used paper publications by predecessors and colleagues to enhance their own research, modern Slavists should be able to use one another's electronic editions. And while many Slavists prepare electronic texts, but how many of us have ever published such a text? It is not because we think the electronic files are valueless, or we would not have bothered to prepare them in the first place. And it is not because we consider the transcription of primary sources unworthy of publication, since we publish such transcriptions on paper.

Perhaps in some cases Slavists may not realize that electronic texts can facilitate different types of research not only for their original compilers, but also for colleagues. But I think the main reason most Slavists do not publish electronic texts is that in most cases nobody else could read them. There are several variants of this problem:

1) Slavists create electronic texts using their favorite word processors and fonts, and anyone who uses different hardware or a different operating system or a different word processor or a different font is out of luck.
2) Slavists create plain ASCII files, which everyone can read, but with transcriptions that are so heavily normalized that whatever their textological value, they are frequently useless for linguistic or orthographic analysis. Such texts either do not contain any structural markup.
3) Slavists create rich files, but with structural markup that nobody else can use because it is in a propriety format. In order to distribute a more accessible ASCII file, the editor must then strip out valuable information entirely. Or the markup is retained, but it is mixed into the content with little or no formal organization.

That is, currently most Slavists do not prepare multiple-use documents. It is as if our community has forgotten, when it comes to electronic texts, that the value of scholarship depends not only on its quality, but also on its accessibility. Scholars do not normally publish printed books and articles in languages that none of their colleagues can read. If electronic texts are important enough to produce, they should be made accessible through the use of a format that is not constrained by hardware or

operating systems or word processors or fonts. The problem of document portability is comparable to the problem of multiple use, and a document that can be accessed only on a single system is clearly less useful than one that can be accessed on multiple systems.

Most humanities scholars who use computers have been told by colleagues at one time or another that if only we would use different computers or different operating systems or different application software or different fonts, all our problems would be solved. But the solution is not to legislate the hardware platform or operating system that others should use, since it is clear that there never will be complete consensus in this area. And it is not necessary to standardize hardware or operating systems or application software or even fonts in order to share electronic texts. All that is necessary is to standardize text structure. And it should be painfully clear to anyone who has tried to use the import and export filters that are built into standard commercial word processors that these filters are helpless at dealing with diplomatic transcriptions of early Cyrillic manuscripts. If electronic texts are to fulfill their potential as accessible contributions to scholarship, the solution is to encode them in a format that is hardware-independent, operating-system-independent, application-independent, and font-independent.

One framework that supports this type of document is Standard Generalized Markup Language, or SGML. SGML is an international standard endorsed by the national standards bodies of many countries, and it was developed precisely to cope with problems of document reuse and portability. But SGML is not the only possible solution, and other frameworks, such as LaTeX, are also available, although with certain limitations. What is important is not the specific model, but the fact that the model is stable, platform-independent, non-proprietary, and noncommercial. What an academic community must avoid is commercial, proprietary systems, such as those that require users to buy Microsoft Word (for Windows or Mac) or Word Perfect or ChiWriter or T^3. The biggest problem with these applications is not internal limitations, but the fact that document conversion is not guaranteed to be bilateral, and while it always possible to convert from a platform-independent architecture

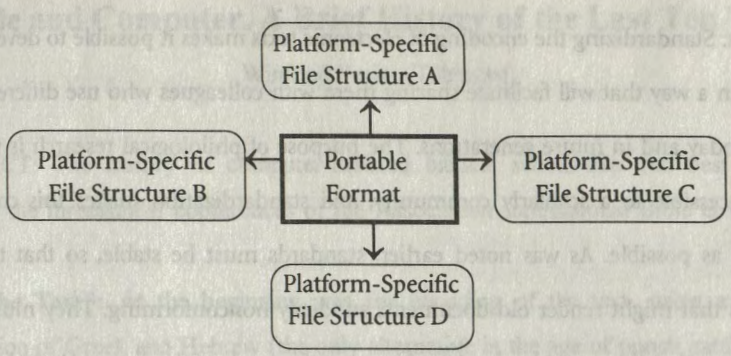
like SGML to another format, conversion from platform-dependent platforms is often impossible. If electronic documents are created in a system designed for portability and platform independence, they remain accessible to their original authors, while at the same time also becoming accessible to others.

The operations ultimately performed on electronic text files may include the preparation of critical editions, of diplomatic editions, or of concordances and indices, the performance of structured searching or statistical analysis, the implementation of intradocument linking (internal hypertext) or interdocument linking (external hypertext), or the preparation of documents for access on the World Wide Web. These tasks may at some stage involve the use of platform specific hardware or software resources, but this does not mean that it is necessary to build platform dependencies into the text files. Whatever the ultimate goals, if Slavists want to work collaboratively at encoding a truly substantial corpus of early Slavic manuscript materials in electronic form, it is imperative to create files with portability in mind.

In the following diagram, note that the arrows between the central, portable document format and the various platform-specific document structures point only in one direction. Files created in a way that depends on proprietary features of Platforms A, B, C, or D are essentially restricted to use on those platforms. Files created in the portable document format in the center, on the other hand, can be used on any platform.[2]

---

[2]This does not mean that it is impossible to create portable documents on local platforms. Rather, it means that it is necessary to take care to avoid proprietary features of those platforms when preparing documents for interchange.



As the figure illustrates, the issue of portability is similar to the issue of multiple-use documents. Just as a document that is designed to function only in a critical edition may be useless for linguistic analysis, a document that is depends on proprietary features of a particular platform may be unusable on other platforms. But a document prepared in the central, portable format can be transformed, through the one-directional arrows, into any desired local format, and the transformation can be designed to support any desired local task.

Preservation. Information that becomes inaccessible loses its value, and electronic texts can become inaccessible very quickly, due to rapidly changing standards. I prepared electronic files ten years ago that I can no longer print, because I no longer use the hardware or software on which these files depend. It is almost a law of nature that hardware will become obsolete and software will fall out of use, and if Slavists prepare electronic materials that depend on specific hardware or software, we shorten their utility and thus limit the value of our contributions to scholarship. This is why portable formats would be needed even if all Slavists today were to agree to use the same systems; it is not possible to predict what systems will be available ten years from now, or whether the systems of the future will be able to read today's files. The solution to this problem is obvious: electronic documents should be prepared with an eye toward longevity. And longevity is essentially diachronic portability, which means that if files are not dependent on single platforms today, they are less likely to be dependent on today's platforms tomorrow.

STANDARDS. Standardizing the encoding of electronic texts makes it possible to develop them for multiple uses in a way that will facilitate sharing them with colleagues who use different computer systems both today and in future generations. The purpose of philological research is to make new information accessible to a scholarly community, and standardization makes this community as comprehensive as possible. As was noted earlier, standards must be stable, so that they will not change in ways that might render old documents suddenly nonconforming. They must be coordinated, in the sense that they must be developed to serve the needs of the entire scholarly community. They must be non-proprietary, so that their use can never be restricted to single platforms. And they must be noncommercial, so that they will be available freely to anyone who needs them. Proprietary and commercial products can be valuable tools for manipulating and rendering files on local platforms, but they cannot be prerequisites for multiple-use, portable electronic text files.

CONCLUSION. A clever electronic text project loses much of its scholarly value if the only person who can use it is its author. Slavic philologists should use computers to make electronic texts serve as many uses, and as many users, as possible.

Bible and Computer. A Brief History of the Last Ten Years

Winfried Bader (Tübingen)

ABSTRACT: The history of computer-assisted biblical scholarship can best be told by surveying the threeannual conferences of the Association International Bible et Informatique (AIBI).

1985 "The Text": At the beginning was the encoding of the text, using a 7-bit-ASCII transcription of Greek and Hebrew (the only alternative in the age of punch cards) that today guarantees system-independence. The first projects undertaken were to produce word indices and concordances, for which the text was tagged with morphological information that could be used in lemmatization. These tags were added with the same 7-bit-ASCII encoding of the base text, using the lines of a file like records of a database.

1988 "Tools, Utilities": The utilization of a classical printed concordance by a scholar was the model for producing tools and utilities for access to the electronic text and its morphological tagging. Many of these tools were later distributed as commercial products.

1991 "Hermeneutics": New tools provide new goals. The possibility of conducting automated grammatical analysis raised a linguistic discussion about adequate computational grammar models. Definitions of word, phrase, clause, and sentence were discussed, bottom-up and top-down models appeared, and in several places programs for linguistic analysis were developed and implemented.

1994 "Desk and Discipline": The question now is: how does the use of computers influence our classical exegetical work?

Initial development in computer-assisted printed editions of biblical texts was slow and primitive, and the computer was first used like an old typesetting machine with WYSIWYG capabilities. Changes in perspective can be seen in by two computer-assisted editions currently being prepared: a new edition of BHS (Codex Leningradensis) and an edition of Qumran texts. Data entry is performed by scholars all over the world, and the texts will then be collected and processed automatically to generate files for the final printed edition. Using SGML-like encoding guarantees compatibility, interchange, and multi-use electronical editions both today and in the future.

New Tools for Old Language:

Computer Processing of Bulgarian Texts

Elena Paskaleva
Linguistic Modelling Laboratory
Centre for Informatics and Computer Technology
bl. 25a Acad. G. Bonchev St.
1113 Sofia, Bulgaria
e-mail: hellen@bgearn.bitnet

Milena Dobreva
Institute of Mathematics and Computer Science
bl. 8 Acad. G. Bonchev St.
1113 Sofia, Bulgaria
e-mail: dobreva@bgearn.bitnet

KEYWORDS: modern Bulgarian, old Bulgarian, encoding, NLP, TACT, SUPERLINGUA

ABSTRACT: The current state of computer encoding and processing of medieval Slavic texts compared to the modern Bulgarian is presented. The difficulties in using available software products (text editors; shareware products for statistical linguistic research on the example of TACT) are described. The basic conclusion is that the future work should be put in the frameworks of specialized linguistic environments, and one Bulgarian example - the SUPERLINGUA system is given.

ACKNOWLEDGEMENTS: The work on this paper is partially supported by grants on the projects I-426/94 and MU-IS 6/95 of the National Fund of Scientific Research in Bulgaria.

1. INTRODUCTION

For more than 30 years now the specialists in computer and language sciences have been working hard to meet jointly one of the greatest challenges of information science in the end of the twentieth century. The challenge is to combine two types of information technologies - the most ancient one, *human language*, and the most modern one in the field of linguistic engineering and linguistic resources - *Natural Language Processing.*

This ambitious task runs into difficulties that belong to several layers of knowledge processing depending on its goals and that affect the two basic processes: *text*

encoding and *text processing*. Concentrated on the text processing tasks, the researchers in the Humanities usually do not expect that they will have to overcome first the problems of the encoding, which are extremely hard in the case of Medieval texts.

The troubles caused by these complex tasks are of two types: *code troubles* and *knowledge troubles*.

The first type of troubles are connected with the ability of the encoding system of the processing tool, i.e. the computer, to represent the primary units of linguistic knowledge in the written languages: letters, and some their combinations, i.e. words.

The second type of troubles are connected with the processing of knowledge whose description is not (and can hardly be entirely) formalized so as to be processed by a computer. We have in mind the formal description of all levels of natural language.

We will present the current state of encoding and processing Bulgarian texts in the following sections.

2. TEXT ENCODING

The computer, invented in the USA, was naturally adjusted to process the English language. That is why there were no code troubles in the processing of modern English texts which were the first encoded and processed by computer.

The European languages subjected next to computer processing, in case they use the Latin alphabet, overcome the code troubles in a relatively easy way, with some stipulations on the distribution of the extended ASCII space for several additional symbols.

The situation of the languages using non-Latin alphabet is entirely different. The elements of this alphabet, the characters, can be located in the ASCII extension either

in full volume (e.g.Armenian) or partially (e.g. Greek), or they can be scattered in the primary ASCII and the extension (e.g. some variants of the Cyrillic code page).

The three approaches to solve the problem of encoding texts written in non-Latin alphabet, transliteration, creating variants of the ASCII table, and coining new encoding standards, are discussed below.

2.1. Transliteration

One of the first applications of computer processing of old languages - *Thesaurus Linguae Graecae* [HCY 91] - solves the input problem with the help of the simple transliteration of the Greek alphabet. Transliteration is used also in the project *Corpus dei Manoscritti Copti Letterari* [Orlandi 90]. It was applied when the first attempts to prepare concordances of Pushkin's poems were performed. Besides, it presupposes a strict correspondence between a non-Latin and a Latin letter or a letter combination (the opposite transition is not always possible) for all languages using the respective alphabet. In the case of a language that is alone in using the given non-Latin alphabet, like Greek, the unification of the transliterational standard is achievable to a greater extend. As for languages, like Slavonic, which share the Cyrillic alphabet, the standards of transliteration can differ even within the boundaries of one and the same language.

2.2. The conquest of the ASCII space

This approach is the only one used in the present moment in Bulgaria. The use of the extension of the ASCII table to encode Cyrillic (and other alphabets) appeared due to the natural desire of the people to work with their original alphabet, not with the transliterated version. The basic problem of this approach is the chaos of different code standards even in the same country (in Bulgaria, there are two standards - MIK and MSX) [Dobreva 93].

With Slavic languages using Cyrillic the number of standards exceeds the number of languages. Every computer Slavist has reluctantly come across this problem, assumingly using different drivers made in his country but he is helpless in his efforts to read Cyrillic texts sent by colleagues from other countries.

2.3. The Standards of the Future

The huge range of non-Latin alphabets lead to the idea to create a standard which will provide free space for all of them. Thus, the UNICODE standard was set up [UNICODE 92]. The 'piece' of Cyrillic symbols included into it does not look completely satisfactorily, especially in the presentation of the Medieval Slavic alphabet.

2.4. Medieval Slavic Texts and the Computer

All above listed inconveniences and the chaos in the standards hold true for modern texts. In modern writing system the elements of the alphabet have long ago been established and unified. In the computer processing of medieval texts there are also some particular hardships that appear, apart from those of language-specific character. They result both from the specific character of the object itself and the system which it forms.

2.4.1. Identification of the basic elements of writing - the characters

The former range of problems which happens to be extralinguistic, are related to the very physical registration of the original object - the character.

The difficulties coming from its identification are due to the type of the information carrier, i.e., the real medieval manuscripts, not their contemporary typographic editions. In addition to this, the modern concept of the 'standard' alphabet can not be applied to the medieval Slavic texts - the variation in the sets of letters and the functions of some of the graphemes are well-known to the researchers in the field.

2.4.2. Description of different levels of the language system

For the purposes of the researchers, different levels of detailization on the layout and writing peculiarities are needed. The ligatures, abbreviations, super-, sub- and in-script letters are very specific features of the Medieval texts, which can not be satisfactorily presented by the traditional text editors designed on the basis of knowledge about modern languages.

At a lexical and grammatical level the diversity becomes even wider but it is related rather to the processing than to the encoding.

The struggle to overcome system hardships is being fought not only by means of software and hardware but mainly by the organization of linguistic knowledge at character and orthographic level - comparison, unification and transformation of already registered elements, although specific information is being lost.

The means of solving the code problems of Medieval Slavonic texts can be arranged hierarchically depending on the rate of both system knowledge and human knowledge participate the recognition and determination of the code list. Thus the grapheme inventory of Medieval Slavonic manuscripts results as an outcome of procedures varying from the manual transliteration to using advances intelligent optical reading devices.

On the middle of the road we can locate procedures of composition and processing of old cyrillic characters in a graphic mode either by individual or standard software. The greater part of the experiments for processing Medieval Slavonic texts in Bulgaria is related to this procedure.

PCs were introduced in Bulgaria for mass usage comparatively late but still it was before the times of WINDOWS. That is why we can felicitously remember the

processing of a Medieval Bulgarian text by means of a special graphic editor for 8-byte computers within the frameworks of a student's diploma paper.

The standard text editor in graphic mode , like Chi-writer also allowed for Medieval Bulgarian texts to be entered. In this process the unification and standardization of manuscript variants is entirely borne by the authors of the investigation (the lack of common standards of using Chi-writer makes the exchange of texts among different people and/or organisations more difficult).

The most popular editors used in Bulgaria, Chi-Writer, T^3 and WINDOWS applications, all produce files in their own formats which makes the exchange of texts difficult.

3. TEXT PROCESSING

The scope of means leading to the solution of the problems not only of encoding but also of processing texts overlaps in general with the means of processing Modern Bulgarian language. Bulgarian computational linguistics which has its origins as far as 30 years ago has dealt to different extent with all levels of automated language processing (see Fig. 1).

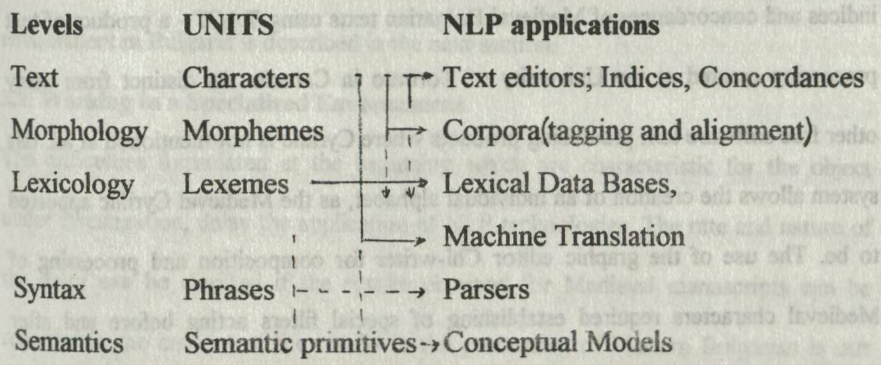


Fig. 1. Linguistic Levels, Basic Units and NLP Applications

The processing of Medieval Slavonic texts is disposed both in the highest part of this hierarchy and even stands over its top. The chaos in the field of coding of Medieval texts makes also considerably difficult their processing. Due to this reason, substantial part of the scientists' endeavours goes on seeking and testing a suitable software and hardware means for coding and entering texts. The hardware is represented in a broad variety of IBM-PC, Macintosh and even ATARI. It is the same with the software of text editors T^3, Chi-writer and Windows applications.

The deepest level of processing Medieval texts is the graphemic one and it comprises the word forms. The most advanced applications include composing indices, concordances and obtaining statistical information at the grapheme level. Most of the Bulgarian specialists investigating manuscripts devote much time to creating indices and concordances manually, no matter if they work in the computer or the pre-computer times. For example, some routine tasks with the processing of Latin texts, for which the computer processing is taken for granted, don't use the computer even if the materials are on a computer readable carrier.

3.1. Using Shareware and Commercial Software Tools

The co-author of this paper M. Dobreva conducted an experiment on creation of indices and concordances of Medieval Bulgarian texts using TACT - a product of text processing created at the University of Toronto in Canada. As distinct from many other free software text processing products where Cyrillic is not mentioned at all, this system allows the creation of an individual alphabet, as the Medieval Cyrillic appeared to be. The use of the graphic editor Chi-writer for composition and processing of Medieval characters required establishing of special filters acting before and after TACT operations. They realize the transition between the original text (Chi-writer) and

the internal TACT presentation for which characters of two odes are not permitted. For the purposes of obtaining the initial and final results on a laser printer, filters for the transition Chi-writer - WINDOWS and TACT - WINDOWS have been constructed.

This experiment illustrates a typical situation in the area of proceeing medieval Slavic texts: to be able to perform it for products which were not intended to be used for Medieval Cyrillic, one should create a set of special which change the encoding the above-mentioned text editors to encoding which can be processed by the program. TACT, designed with the idea that it could be used for texts written in different alphabets, does not allow encoding of one letter by two characters.

The situation described above briefly is very typical for any attempts to use commercial or shareware products for processing medieval Slavic texts. The necessity to write special programs for converting texts from one presentation to another is not welcomed by the linguists, who have to overcome difficulties like finding the encoding system used in a specific product and writing program code. Thus, the more appropriate approach is to construct friendly environments which naturally allow to work with different alphabets. The first large-scale project on creating such an environment in Bulgaria is described in the next section.

3.2. Working in a Specialized Environment

The difficulties formulated at the beginning which are characteristic for the object under investigation, delay the application of NLP technologies. The rate and nature of the delay can be clarified if the results obtained for Medieval manuscripts can be regarded in the context of the results in the processing of Modern Bulgarian in our country. One of the leading research groups in the area of NLP, the Linguistic

Modelling Laboratory, has covered the links from that scheme (see Fig. 1) for the modern Bulgarian language.

The general task is the creation of lingware for Bulgarian. The basic principles are:

- aiming at exhaustive models of linguistic knowledge from the text level via morphology and lexical data to syntax and semantics;
- attacking the deep language levels through gradual accumulation of linguistic knowledge from real large text corpora by means of an intelligent user-oriented interface;
- attacking the surface language levels via complete computer models.

Following these principles from the products constructed in the LML those belonging to more specific levels could be language independent, except for the case of constructing language-setting products.

That is how the system SUPERLINGUA [Paskaleva 94] is being elaborated and it is designed with friendly interface for the linguist. Within the framework of these achievements the Medieval Slavonic material can be processed after its suitable coding. If not on world scale, at least on a national scale the Bulgarian specialists would have to unify the standards used if it should be at the expense of a part of the precious extralinguistic information, i.e. historical, art, psychological and cultural aspects.

4. Conclusion

When a linguist who is oriented in the general problems of NLP becomes even vaguely aware of the above-mentioned range of problems, he is undoubtedly struck by the methodological mismatch among the character of the object, the research goals and the instruments of their achievement. If we have to find the reason of this mismatch, it is mainly in the complex and random nature of Medieval manuscripts and the

profoundness of ideas which an investigator can discover in them. The advanced computer technologies are not able to provide an instrument fine enough for this true wealth but we have to use effectively also the rough ones.

References

[Dobreva 93] M. Dobreva, *The Cyrillic Alphabet - one of the other Branches*, Invited lecture at the DIDOT seminar, Bilbao, Spain, October 1993.

[HCY 91] *The Humanities Computing Yearbook 1989/90*, ed. by Ian Lancashire, Clarendon Press, 1991.

[Orlandi 90] T. Orlandi, *The Corpus dei Manoscritti Copti Letterari*, In: Computers and the Humanities, Vol. 24, 1990.

[Paskaleva 95] E. Paskaleva, *Tagging Highly Inflective Languages (non-English, non-Latin alphabet, no morpho automation)*, Presentation at TELRI (Trans European Linguistic Resources Infrastructure) seminar, Tifany, Hungary, September 1995.

[UNICODE 91] *The UNICODE Standard*, Addison-Wesley, 1991.

APPENDIX

Bulgarian Publication on Computer Processing of Medieval SLavic Texts

М. Добрева, А. Бояджиев, *Проблеми, свързани с компютърното обработване на текстове, написани на православните славянски езици*, В: сб. Методология на математическото моделиране, 1993, стр. 180-185.

И. Христова, *Речник на словата на Климент Охридски*, С., 1994, see the Introduction.

M. Dobreva, *Applications of Computer Tools in Studying Medieval Slavonic Manuscripts*, S., 1995, 66 pp.

Informational and Presentational Units in Early Cyrillic Writing

David J. Birnbaum
Department of Slavic Languages
1421 Cathedral of Learning, University of Pittsburgh
Pittsburgh, PA 15213 USA
djbpitt+@pitt.edu

KEYWORDS: character, glyph, entity, character set, font, SGML, standardization, TEI, WSD

ABSTRACT: Slavists should encode whatever units of writing may be needed for their particular research needs. And they may do so in different ways, using different bit combinations, without sacrificing portability, as long as the encoding systems fulfill certain requirements. Such systems must be unambiguous and internally consistent, and they must be documented in a formal, standard way, with reference to standardized inventories. The Text Encoding Initiative (TEI) Writing System Declaration (WSD) provides the foundation for such a system.

Introduction

One of the most critical issues confronting Slavic philologists who work with electronic texts is how to encode the written units, or letter forms, found in these texts. While this looks like a single issue, it can be decomposed into two very different questions, which can be answered independently of each other:

1) How should the units in question be encoded?

2) Which units should be encoded?

These two questions are frequently confused, and this confusion is a cause for concern, since it may give the appearance of disagreement where none exists, and thus interfere with collaborative work. Despite the controversial nature of both questions, I think the definitive answers proposed here will not be controversial, because what is definitive about them is precisely their flexibility and accommodation. These answers are informed by many of the considerations mentioned in my "How Slavic Philologists Should Use Computers" (in this volume): *structure, portability, preservation,* and *standardization.*

Some Premises

Let me begin with six premises:

1) *Ontologically, an encoded electronic text is not information, but a representation of information.* Text encoding, like any written representation of information, may be considered a semiotic process, in which a series of signs is deployed to represent a series of relationships between symbol and meaning. The same physical coded character, which is to say the same series of bits in the machine, may mean one thing in one system of representation and something different in another. These representations are conventional, and a specific bit combination in the computer does not have an intrinsic meaning. It has a specific meaning in electronic texts only because users agree that it will have that meaning.

In practical terms, this means that the adequacy of a system for encoding characters normally does not depend on the specific bit combination used to represent each written unit in an original source. In many cases, various proposed systems for transcribing early Cyrillic writing represent the same basic inventory of original textual units, and differ essentially not in what they represent, but in how they represent it physically. If we concentrate for a moment just on these areas of overlap among encoding inventories, from an informatic perspective, these differences in the choice of sign that we associate with a specific original letter form become insignificant, because any such system that is internally coherent and unambiguous can be converted automatically to any another. Significant incompatibility arises only where the inventories of units that can be distinguished do not overlap, so that one transcription system is able to represent distinctions that cannot be represented in another.

2) *All encoding is interpretive.* Physically, the letter forms that we encounter in our manuscript materials can differ infinitely. Even in the most careful handwriting, if the same letter is written several times, minuscule differences will intrude. This is inevitable because handwriting is analogue, with an uncountable number of physical parameters that may vary in subtle ways, while an electronic text is digital, with a finite number of distinctive characters, or bit combinations.

In the same way that coded electronic texts are representations of information, original manuscripts are also representations of information. When we read a manuscript, each letter form represents certain information content, which we deduce through an interpretive mechanism, silently discarding variation that we consider irrelevant. We may choose, for example, to ignore the paleographic variation that we do not consider part of the information we wish to study, but the reason we do this is not because it is not information—for it may be very important for paleographers. Rather, we ignore it because it may not be relevant information for our purposes, which may involve the study of texts, rather than of paleography. From a character representation perspective, what I mean when I say that all encoding is interpretive is that we decode the physical signs in primary sources as representing certain information that we consider important for our analytical purposes, while a different scholar, with different analytical purposes, may choose to concentrate on a different interpretation of the same primary material. To encode a source is thus to decide what to encode from that source and what to omit, and this decision is an act of interpretation. It is not practical, and perhaps not even possible, to encode all of the information that is available in an original source. And it is certainly not desirable to do so; there is no point to encoding information that would not be an object of scholarly inquiry. And even if we could encode every bit of information we could find in an original source, just in case it might some day be an object of scholarly inquiry for someone else, it is likely that we would still fail to exhaust the information content of our original source, because future generations may well discover things we have overlooked. I will give a real example of such an occurrence presently.

3) *All transcription involves normalization.* Because all encoding is interpretive, all encoding necessarily involves normalization. By normalization, I mean that we neutralize differences in signs that we do not consider worth encoding. Normalization is destructive, because unless the neutralized distinctions can be recovered algorithmically (which is almost never the case with our materials), they cannot be recovered at all. "Destructive" is a powerful word, but I do not by any means intend

to suggest that normalization is bad. Rather, it is inevitable, it is useful, and the important thing is to make decisions about normalization consciously and for cogent reasons.

4) *Any attempt to legislate an inventory of encoding units will fail synchronically.* Slavic scholars differ in the distinctions that they will wish to retain while transcribing sources. For example, if one works with manuscripts in which narrow and broad omicron are not distributed in any significant way, one may view the difference as only of paleographic interest, and one may accordingly choose to encode these two letter forms identically. But as Andrej Anatol'evič Zaliznjak discovered not long ago, in some manuscripts this distribution reflects accentual and phonological features, so that the position of stress and the height of mid vowels can be inferred from the distribution of these letters even in manuscripts that contain no supralinear accent marks. In this type of manuscript, the distribution of narrow and broad omicron is not just paleographic, but also orthographic. This distribution does not obtain in all manuscripts that distinguish narrow and broad omicron, but it is not rare, and where it is meaningful, it is a very valuable object of study for historical accentology.

If we legislate that manuscript transcriptions must encode narrow and broad omicron differently because in certain manuscripts the difference is important for historical accentologists, we impose an unreasonable and impractical burden on those for whose research this difference is irrelevant. And the latter will refuse to implement the distinction in their transcriptions, and rightly so. But if, on the other hand, we legislate that manuscript transcriptions must encode narrow and broad omicron identically because the difference is rarely of lexical or textual significance, we disenfranchise historical accentologists, who will respond by defying the prohibition and encoding the information they need for their research. And they will be right to do so.

5) *Any attempt to legislate an inventory will fail diachronically.* Until the aforementioned studies by Zaliznjak, generations of Slavists had been unaware of the information that could lie hidden in the distribution of narrow and broad omicron. To draw a somewhat anachronistic conclusion, they were unaware that they might be able to research this particular question by encoding the distinction in

electronic texts and subjecting them to distributional and statistical analysis. Now, after reading Zaliznjak's comprehensive studies of this issue, as well as studies by his students, such as Elena Stadnikova-Tugaj, their discovery is so obvious that we are amazed that we never noticed it ourselves.

We should not be so arrogant as to assume that distributions of other letter forms that we consider cosmetic will not be shown to be linguistically significant in the future. Attempts to legislate a single inventory of units to be used in all encodings of all sources will fail diachronically because we cannot be certain that there are not informational distinctions in our materials that we have not yet discovered. In some cases, the distribution of what we think are insignificant variants may even be shown to be significant precisely through computer-assisted statistical analysis.

6) *The scholar should determine what to encode.* One consequence of the preceding observations is that different scholars will want—and need—to encode the same materials in different ways. They will do so because although they look at the same archeographic object, they ask different questions about it. And they will ask their questions whether we approve of them or not, and whether these questions then turn out to be wise or foolish will ultimately be determined according to the ultimate significance of the answers. Not only do we not now know all the answers, but we do not even know all the questions. If we are to enable others to make discoveries that have eluded us, we need a system for transcribing early Slavic primary source materials that will let people test new hypotheses. And if this is to be possible, we need to let individual scholars determine how to interpret their material and what to normalize.

Standardization

In my "How Slavic Philologists Should Use Computers" (in this volume), I stressed the importance of standardization, and a question naturally arises about how we can simultaneously standardize transcription and yet allow each scholar the flexibility to determine what to encode. I am going to answer this question by returning to the two parts of my original definition of the problem: how should the units of early Cyrillic writing be encoded and which units should be encoded?

The units of early Cyrillic writing should be encoded the way each user wishes to encode them, thereby providing the flexibility that is required if a common system is to be adequate for all scholars. But this encoding should be done within a framework that ensures portability and preservation, thereby providing the prerequisites for standardization. For example, because the treatment of eight-bit characters in different processing environments varies, portable documents should eschew eight-bit characters, as well as certain seven-bit characters whose values are also unstable across platforms. And any system of representation must be unambiguous, so that if we want to be able to distinguish, for example, a kamora from a palatalization hook during processing, we should not encode them identically, even if we do not require them to look different physically on our screens or printers.

The good news is that the specific bit combinations used to represent the written units of the original source do not have to be standardized in order for users to share materials. What does have to be standardized is a way to document different transcription systems, and this must be done in reference to a common term of comparison. Such points of comparison exist, and they are standardized, which is to say that they are registered international ISO standards.

One such standard is ISO/IEC 10646-1: 1993, a subset of which is better known as Unicode. ISO/IEC 10646-1: 1993 is a character set standard, which means that it is intended to represent units of information, or the units that may distinguish sound or meaning. Another such standard is ISO/IEC 9541-1: 1991, which is a glyph standard, which means that it is intended to represent presentational units, or units that are graphically different. The vernacular reference to "characters" in a font is incorrect from the perspective of information scientists engaged in standardization, since it glosses over the difference between characters and glyphs. There is no real need to dwell on this distinction, but it is worth noting that while there is often a one-to-one relationship between characters and glyphs, this need not always be the case. An accented letter, for example, may be stored in the information stream as a single character, but may be printed by fetching two glyphs from separate cells in

a font, and rendering one above the other. Or the accented letter may be stored as two units in the information stream, but fetched from a single cell in the font. And the keyboard has nothing directly to do with characters or glyphs: one keystroke may fetch multiple characters or glyphs, or multiple keystrokes may be needed to access a single character or glyph.

What is important for our purposes is that if we adopt a standardized way to document the relationship between our preferred individual systems of transcription and standardized character or glyph inventories, such as the ones I just mentioned, we make it possible to design engines to convert among encodings. If I receive a file in an arbitrary unfamiliar encoding, as long as it is accompanied by a standardized, formally structured file documenting the standardized character and glyph values of the encoding units, I can remap these files to my transcription system automatically.

A model for this type of system exists within the framework of the Text Encoding Initiative, or TEI, in the form of what is called a Writing System Declaration, or WSD. The WSD is a file formatted according to Standard Generalized Markup Language, or SGML. It is not the document transcription, but an auxiliary SGML file that records the transcription methods used in an associated primary document file.

Despite the obvious convenience of a standardized transcription system, standardization at this level is not a requirement from an informatic perspective. This is fortunate, because it seems unlikely that Slavists could reach agreement on this issue, not only because we are all familiar with our own systems and reluctant to abandon them, but also because we may have voluminous legacy documents that we do not wish to recode. As an alternative to changing transcription systems or recoding documents, one can prepare a Writing System Declaration for each transcription method, which can be distributed with the documents that use it. This solution does not require that anyone change transcription systems, it does not privilege any particular transcription system over any other, and it is capable of accommodating new transcription schemes as they emerge. And it is part of the multi-

national and multidisciplinary Text Encoding Initiative framework, which is to say that Slavists will not be the only ones employing it.

Which Units of Early Cyrillic Writing Should be Encoded

Just as we cannot hope effectively to legislate a single concrete transcription system that will prove adequate for all Slavic scholars, I do not believe that we can legislate an inventory of letter forms that all Slavists should encode either now or in the future. The information we encode will depend on what we consider valuable for our research. And this, in turn, will depend not just on the materials we study, but also on the questions we ask about these materials.

A model such as the Writing System Declaration does not require that we all use the same inventory of units. What is required is that we all use the same point of comparison for our inventories, such as registered character or glyph inventories. Just because a great sea of distinctions is available in these inventories does not mean that each of us is compelled to use them all. If you need to distinguish broad and narrow omicron in your transcription, your Writing System Declaration can point to the relevant items in the standardized inventories, and your colleagues will know what you are representing. If you conflate them, you will point to only one of these items, thereby documenting not only the meaning of each bit combination in your text, but also the fact you have not encoded distinctions in the width of omicron. The conflation cannot be undone, of course, but the problem is not to find a way to convey more information than the electronic text contains, but, rather, to find a way to transport the information that *is* there across platforms.

It is important to keep in mind that a very rich system that is full of units we may consider irrelevant for our own work is not a flawed system. Just as we do not feel compelled to eat everything on the menu in a restaurant, we are not compelled to fill our trays with encoding units that are not to our taste. *Flexible encoding is not about rules, but about opportunities.*

There is still room to discuss the inventory of units needed to encode a particular Cyrillic source for a particular analytical purpose. But we should recognize that this issue is a philological question, and not an informatic one, and it amounts to a restatement of the question of what information is important in an original source. Whatever our views on this philological issue, the informatic point is that even if do not agree on the units we need to encode, or on the bit combinations we wish to use to encode them, we can still agree to use a system that will make our texts interchangeable.

Transcription Rules for Old Church Slavonic Writing

Karsten Grünberg

Slavisches Institut der Universität Heidelberg,

Germany

ABSTRACT: One of the most important tasks for Paleoslavonics in the years to come will be to set up databases for church slavonic linguistic monuments and for fast access to the information they contain. A database for the study of church slavonic manuscripts should cover three essential areas: The first area should provide, in an adequate structure, data fields for the cataloguing of manuscripts, to contain the essential data from manuscript catalogues, e.g. signature, repository, age, scripture, water marks, contents, bibliographic information, etc. The second area should contain facsimiles in the form of computerized picture files, linked to the relevant entries in the catalogue database. Due to today's scanning technologies, it is possible to produce full colour facsimiles of manuscripts in a satisfying quality. These facsimiles can be published on CD-ROM, even with full colour facsimile of a manuscript still requiring enormous amount of disk capacity. The third area should be sets of text files, linked to both, the relevant catalogue database entries and the relevant manuscript facsimile picture files. These text files should provide the monument's text, encoded according to an unified transliteration standard, for further computerized processing. As long as there are no computer programs available which, by means of optical character recognition (OCR) algorithms, automatically produce sufficiently accurate text files from pictures, these texts have to be entered manually into the computers. Even today's sophisticated OCR software packages are unable to correctly "read" manuscripts - due to a low hit rate, they produce too many errors and thus require a lot of time consuming additional editing of the resulting text. As a result, it is still faster and more economic to directly type in the text by hand.

The present report offers for considerationand discussion a suggested set of general rules for the transcription and transliteration of Church Slavonic manuscript material.
We suggest the following general rules to be applied for text files:

1. Superscripts to be enclosed in slashes.
2. Tildes to be indicated by a preceding left and right brackets.
3. A word break at the end of a line to be indicated by a vertical bar.
4. Initials in a header to be marked by a preceding ampersand, at the beginning of a verse by a preceding at-sign.
5. The entries are to be based upon the page and line numbering of the original.
6. Additions to be enclosed in double quotes.
7. Operator's comments to be enclosed in braced brackets.
8. Ambiguous occurrences to be marked by a bracketed question mark.

Text files, produced observing those rules, offer the advantage that information, not required for a certain job, such as comments during compilation of an index, or orthographic details, and original chapter and verse numbering during reconstruction, can be suppressed for the current job and reactivated for other jobs. Therefore it is extremely important to use unique, unambiguous encoding within the text files, and to establish and observe generally accepted rules for transliteration. Without such transliteration rules, a flexible data exchange will be impossible. A project for such rules was submitted already in 1981, which now should be extended by codes for the accents.

Transliterationsrichtlinien für kirchenslavische Sprachdenkmäler

1. Zur Edition kirchenslavischer Sprachdenkmäler

Editionsarbeit in den Geisteswissenschaften umfaßt sehr verschiedene Bereiche. Editionsarbeit kann in der Herausgabe von Zeitungen, Zeitschriften oder Kongreßberichten, in der Publikation von Bibliographien oder Monographien wie auch in der Edition von älteren oder neueren Quellen geleistet werden. Hier wird näher auf die Edition älterer slavischer Texte eingegangen. In der Slavistik ist es bis heute üblich, slavische Texte nach dem Text einer bestimmten Handschrift abzudrucken, also diplomatische Ausgaben herauszugeben. Die philologische Arbeit beschränkt sich dabei auf das Lesen und Erkennen des Wortlautes einer Handschrift und dessen Abschreiben. Die Wiederherstellung (Rekonstruktion) des Urtextes, wie er einmal ausgesehen haben muß, als er das erste Mal als Text in der slavischen Literatur erschienen ist (sei es als Übersetzung aus einer anderen Sprache, sei es als Neuschöpfung der slavischen Literatur), wird vernachlässigt und der Nachwelt überlassen. Es bestehen bis heute keine allgemein gültigen Editionsprinzipien in der Slavistik, obwohl hoffnungsvolle Vorschläge in dieser Hinsicht unterbreitet worden sind.[1] Diplomatische Ausgaben helfen bei der anschließenden Rekonstruktion häufig sehr wenig oder sind so benutzerunfreundlich gestaltet, daß es sinnvoller ist, die Originalhandschriften zur Rekonstruktion heranzuziehen. Panzer führt als Beispiele die Ausgaben von Bodjanskij 1865[2] und Kuev 1967[3] an, in denen der Wortlaut ein und desselben Textes von mehreren Handschriften hintereinander abgedruckt wurde. Das weitere Bearbeiten des Textes einer solchen Ausgabe, nämlich seine Rekonstruktion, kostet einige Mühen; denn die Hälfte seiner Arbeitszeit muß der Philologe mit dem Blättern in diesen Ausgaben verbringen! Wenn schon ein buchstabengetreuer Abdruck erfolgt, dann sollte er doch wenigstens nicht Text für Text hintereinander, sondern Wort für Wort untereinander (zeilensynoptisch) erfolgen. Der Textvergleich würde sich um einiges vereinfachen.

Die Angewohnheit, Faksimile-Ausgaben herauszugeben ist, nicht von vornherein abzulehnen; sie ermöglichen immerhin einer breiten Öffentlichkeit den Zugang zu seltenen Handschriften oder frühen wissenschaftlichen Bearbei-

---

[1] vgl. Panzer 1987 und 1991.

[2] Bodjanskij hat hintereinander 16 Handschriften der *Vita Constantini* und 8 Handschriften der *Vita Methodii* gedruckt.

[3] Kuev hat insgesamt 73 Versionen des ЧЄРНОРИЗЄЦЪ ХРАБРЪ "О ПИСЬМЄНЄХЪ" hintereinander abgedruckt.

tungen, was in der Slavistik aus politischen Gründen immer ein Problem gewesen war und teilweise noch ist. Wenn jedoch photomechanisch reproduziert und auf philologische Arbeit verzichtet wird, so sollten diese Faksimile-Ausgaben in einer guten, das heißt leserlichen Qualität veröffentlicht werden, was keineswegs immer der Fall ist.[4]

Die wesentlichen Aufgaben einer Edition bestehen darin, erstens die Textgeschichte eines Textes so weit wie möglich zu dokumentieren und deren Nachvollzug zu ermöglichen, zweitens, die einzelnen Textvarianten der Wissenschaft zur speziellen Untersuchung jeglicher Art zur Verfügung zu stellen. Jegliche Information, die eine Handschrift enthält, muß möglichst im Rahmen einer Edition erhalten bleiben. Drittens sollte bei einer Edition ein Lesetext hergestellt werden, der einer breiten Schicht von Interessenten (Historikern, Theologen,Philologen, Studenten usw.) den Einstieg in die Problematik dieses Textes ermöglicht. Sicherlich ist es für einen Historiker oder Theologen uninteressant, ob die einzelnen Textvarianten für den Laut /i/ den Buchstaben и, ї, і oder ѵ (мирѧ, мїрѧ, мірѧ oder мѵрѧ), für den Laut /o/ den Buchstaben о oder ѡ schreiben (отъ oder ѡтъ). Für diese Gruppe von Interessenten sollte ein Text mit Vereinheitlichung von graphischer Variation und Einführung moderner Interpunktion zur Verfügung stehen. Für einen Sprachwissenschaftler mit Schwerpunkt Graphematik sind jedoch Fragen der graphischen Variation unentbehrlich und sollten ebenfalls berücksichtigt werden. Es wird daran deutlich, daß eine herkömmliche *statische* Edition in Form eines Buches, das einmal gedruckt unveränderlich bleibt (bis zu einer eventuellen zweiten Auflage), diesen Anforderungen nicht gerecht wird. Es wird vielmehr eine *dynamische* Edition benötigt, die durch einen maschinenlesbaren Text und automatisch anwendbare Normalisierungsregeln gewährleistet ist. So sollte bei den folgenden acht Hauptarbeitsgängen einer kritischen Edition darauf geachtet werden, daß jegliche Information, die der Editor bei der Bearbeitung der Texte von der Handschrift bis zur Publikation erhält, dem Benutzer zur Verfügung gestellt wird:

1. Sammlung und Zusammenstellung der Textzeugen
2. Textzeugen-Vergleich (Kollationieren)
3. Auswertung des Kollationsergebnisses (Genealogie, Stemmatisierung)
4. Zusammenstellung der editionsrelevanten Textzeugen
5. Herstellung des Editionstextes (Rekonstruktion)
6. Zusammenstellung der Apparate

---

[4] vgl. die nur äußerst schwer zu benutzende Ausgabe des Neuen Testaments aus dem Čudov-Kloster, herausgegeben von Lehfeldt 1989.

7. Registererschließung
8. Publikation von Editionstext, Apparaten und Registern

Da eines der Ergebnisse einer kritischen Edition sein sollte, den bearbeiteten Text nicht noch einmal veröffentlichen zu müssen, ist es notwendig, mehr als bei anderen Editionen (z. B. Zeitungen oder Monographien) darauf zu achten, daß sowohl im Text als auch im Apparat keine Fehler erscheinen. Um diesem Ideal so nahe wie möglich zu kommen, sollte der Computer als Hilfsmittel eingesetzt werden, der zwar keine philologische Arbeit ersetzen, aber Routinearbeiten, bei denen am häufigsten durch mangelnde Konzentration Fehler entstehen, fehlerfrei übernehmen kann. Durch die Möglichkeiten moderner Computertechnik erübrigen sich auch diplomatische und Faksimile-Ausgaben in Bücherform, ohne auf deren Inhalte verzichten zu müssen.

2. Der Computer als Hilfsmittel für Editionen

Der Computer hat auch in die Geisteswissenschaften Einzug gehalten und ist in seinen verschiedenen Einsatzgebieten nicht mehr zu entbehren. Das größte Einsatzgebiet ist sicherlich die reine Textverarbeitung zur Erstellung von Manuskripten, Referaten, Typoskripten, Korrespondenz usw. Der Computer setzt sich aber auch in anderen Anwendungsbereichen immer mehr durch und hat auch zu neuen geisteswissenschaftlichen Disziplinen geführt. Die Computerlingustik, die sich mit Linguistik und Datenverarbeitung befaßt, hat sich inzwischen an vielen Universitäten etabliert und bietet in Zusammenarbeit mit den anderen geisteswissenschaftlichen Disziplinen ein enormes Potential an Entwicklungsmöglichkeiten. Leider wird diese Möglichkeit der interdisziplinären Zusammenarbeit nicht immer in vollem Maße genutzt, aber genau darauf ist der Computerlinguist angewiesen. Er muß die technischen Voraussetzungen schaffen, um dem Linguisten mechanische Arbeiten, die bekanntlich viel Zeit kosten, abzunehmen. Der Linguist sollte immer nur Anwender und nicht Produzent von Computerprogrammen jeglicher Art sein, um nicht unnötige Zeit zu verlieren und damit seine eigentliche fachliche Qualifikation herabzusetzen. Eine erfolgreiche Kommunikation zwischen Computerlinguisten und Linguisten setzt voraus, daß der Linguist über solide Grundkenntnisse in der Computertechnik und der Computerlinguist in der Linguistik verfügt. Ein guter Informatiker ist noch lange kein guter Computerlinguist!

Inzwischen sind anfängliche Skepsis, aber auch erste Euphorie einer nüchternen Einschätzung der Möglichkeiten wie der Grenzen des Computereinsatzes gewichen. Da mit Hilfe des Computers große Datenmengen gespeichert, verwal

tet und bearbeitet werden können, sind bevorzugte Einsatzgebiete die Edition und Erschließung von Texten mit komplexer Überlieferung sowie sprachwissenschaftliche Fragestellungen. Im Bereich der Slavistik hat zuletzt Hansack eine gelungene Zusammenstellung dessen gegeben, was mit Computereinsatz möglich ist.[5] Bakker und van der Tak haben das in Oxford unter der Leitung von Robinson entwickelte Computerprogramm *Collate* zur automatischen Collationierung von Texten vorgestellt[6]. Bei der Wahl von Computerprogrammen sollte nicht nur auf die Leistungsfähigkeit des Programms, sondern auch auf die Bereitschaft der Hersteller zur Kooperation geachtet werden. Das beste Computerprogramm ist nutzlos, wenn bei Problemen nicht schnelle Beratung und effektive Abhilfe erfolgt. Die Linguisten sollten sich davor hüten, einen Wettbewerb der Programme zu veranstalten, den die Produzenten unter sich auf den entsprechenden Kongressen durchführen können. Vielmehr müssen die Linguisten, die nicht mehr als bloße Anwender von Computertechnik sind, die Produzenten dazu drängen, auf die Kompatibilität von Programmen untereinander, d. h. auf die Integration schnell zu handhabender Export- und Importfunktionen für Text- und Bilddateien, und auf den flexiblen Datenaustausch zwischen den verschiedenen Betriebssystemen zu achten. Es kann nicht im Sinn der Wissenschaft liegen, den Wettbewerb der Computergiganten IBM und Apple (Macintosh) auf dem Gebiet der slavischen Sprachwissenschaft weiterzuführen.

Eine Datenbank zur Erschließung kirchenslavischer Handschriften sollte drei wesentliche Bereiche umfassen. Im ersten Bereich sollten Datenbankfelder zur Katalogisierung von Handschriften kreiert werden, auf deren Struktur hier nicht näher eingegangen werden soll, da auf diesem Gebiet schon andere Vorarbeiten geleistet worden sind. Diese Datenbankfelder sollten die wesentlichen Informationen der gängigen Handschriftenkataloge enthalten, wie z. B. Signatur, Aufenthaltsort, Alter, Schrift, Wasserzeichen, Inhalt, wesentliche Literaturhinweise usw. Der zweite Bereich sollte Faksimiledarstellungen in Form von Bilddateien enthalten, die an den Katalogisierungseintrag der Datenbank geknüpft sind. Die Scanner-Technik ermöglicht inzwischen zufriedenstellende, auch farbliche, Faksimiledarstellungen von Handschriften, die auf einer CD–ROM der Öffentlichkeit zugänglich gemacht werden können, auch wenn farbige Darstellungen einer Handschrift noch eine erhebliche Speicherkapazität verbrauchen.

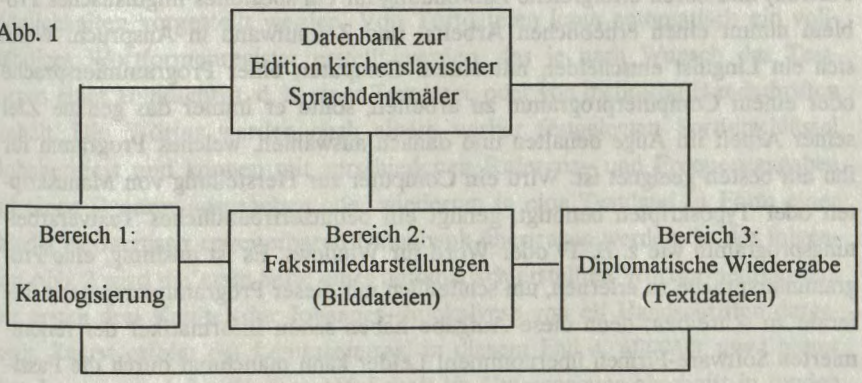
---

[5] Hansack, E.: Paläoslavistik und Computer: die Erschließung des kirchenslavischen Erbes, in: Anzeiger für Slavische Philologie, Bd. XXII/1 (1993), S. 87–96.

[6] Bakker, M. – van der Tak, J. G.: Collating Greek and Slavic Apostolos Manuscripts, in: Palaeobulgarica XVIII, 2 (1994), S. 32–49.

einer Handschrift noch eine erhebliche Speicherkapazität verbrauchen. Der dritte Bereich besteht aus Textdateien, die wiederum mit der Katalogisierungsdatenbank und der Faksimiledarstellung in Bilddateien verknüpft sein sollte. In diesen Textdateien wird der durch einheitliche Transskriptionsrichtlinien codierte Text für die maschinelle Weiterverarbeitung zur Verfügung gestellt. Solange die Scanner–Technik in ihrer Texterkennung (OCR–Phase) noch nicht in der Lage ist, zufriedenstellende Textdateien automatisch aus Bilddateien zu erstellen, müssen diese Texte per Hand in den Computer eingegeben werden. Insbesondere bei Handschriften ist die OCR–Phase bisher ungenügend entwickelt, da zu viele Fehler entstehen und eine sekundäre Korrektur so mühsam ist, daß bis jetzt ein sofortiges Eingeben der Daten per Hand effektiver sein dürfte. Wenn man bedenkt, daß eine Handschrift auf jeden Fall sorgfältig vom Anfang bis zum Schluß gelesen werden muß, dann können auch gleichzeitig bei der Lektüre die Finger auf einer bequemen Tastatur bewegt werden. Nach einer gewissen Übung können die Transskriptionsrichtlinien so verinnerlicht werden, daß eine zufriedenstellende Geschwindigkeit bei der Erstellung von Textdateien auch per Hand erfolgt. Die Verbindung zwischen Faksimiledarstellung und Textdatei ermöglicht eine fortlaufende Kontrolle der eingegebenen Daten, die auf einem guten Bildschirm auch parallel dargestellt werden können. So ist die ständige Transparenz von der Handschrift zum Text gewährleistet. Schematisch läßt sich der Aufbau einer Datenbank zur Erfassung kirchenslavischer Handschriften folgendermaßen darstellen:

Abb. 1



Im folgenden Kapitel soll ein Computerprogramm vorgestellt werden, das in dem Bereich "Diplomatische Wiedergabe (Textdateien)" während der Edition kirchenslavischer Texte sehr nützliche Hilfestellungen bieten kann.

БАН ЦЕНТРАЛНА БИБЛИОТЕКА

3. TUSTEP: Tübinger System von Textverarbeitungs-Programmen

Das "Tübinger System von Textverarbeitungsprogrammen" TUSTEP wurde am Rechenzentrum der Universität Tübingen in der Abteilung "Literarische und Dokumentarische Datenverarbeitung" unter der Leitung von Prof. Dr. W. Ott entwickelt. Es handelt sich bei dieser Software nicht um ein einzelnes Textverarbeitungsprogramm, wie dies bei den herkömmlichen Programmen wie Word für Windows, Signum, Nota Bene, T³ oder um eine Kombination von Textverarbeitungs- und DTP-Programm (Desktop-Publishing, z. B. Ventura, Pagemaker, Corel Draw), sondern um ein System von verschiedenen Programmbausteinen, die je nach linguistischem Problem miteinander verknüpft und durch Parameter modifiziert werden können. TUSTEP ist kein Programm, das für ein spezielles linguistisches Problem, z. B. Erstellung von Registern, Indices, Editionen, Bibliographien, hergestellt wurde. Es bietet jedoch die Möglichkeit, durch Erlernen der systeminternen Programmiersprache, sich selbst die notwendigen Programme für eine spezielle Problemstellung zu schreiben. Das hat den Vorteil, daß das Programm TUSTEP sehr vielseitig einsetzbar ist und den Anforderungen der in den Geisteswissenschaften doch sehr unterschiedlichen Vorgehensweisen gerecht wird.

Das Ziel von TUSTEP besteht darin, Probleme der wissenschaftlichen Textdaten-Verarbeitung mit knapperen und problemnäheren Anweisungen lösen zu können, als dies auf herkömmliche Weise, z. B. mit Programmiersprachen, möglich wäre. Das Erlernen einer Programmiersprache (z.B. C, Pascal oder Fortran) und deren erfolgreiche Anwendung für ein spezielles linguistisches Problem nimmt einen erheblichen Arbeits- und Zeitaufwand in Anspruch. Wenn sich ein Linguist entscheidet, mit einem Computer, einer Programmiersprache oder einem Computerprogramm zu arbeiten, sollte er immer das genaue Ziel seiner Arbeit im Auge behalten und danach auswählen, welches Programm für ihn am besten geeignet ist. Wird ein Computer zur Herstellung von Manuskripten oder Typoskripten benötigt, genügt ein benutzerfreundliches Textverarbeitungsprogramm wie z. B. T³ oder Word für Windows. Es ist unsinnig, eine Programmiersprache zu erlernen, um schließlich mit dieser Programmiersprache Referate zu schreiben; denn diese Aufgabe haben schon Informatiker der renommierten Software-Firmen übernommen! Leider kann manchmal durch die Faszination des Computers der eigentliche Zweck der Computer-Anwendung vergessen werden.

In TUSTEP werden für die wichtigsten Grundoperationen der Textdaten-Verarbeitung Programme bereitgestellt, deren Leistung der Benutzer über Parameter spezifizieren und die er in vielfältiger Weise für die Lösung verschie-

denster Aufgabenstellungen kombinieren kann. Der Begriff Textdaten-Verarbeitung soll das typische Einsatzgebiet von dem, was heute unter Textverarbeitung verstanden wird, abgrenzen. Auch mit TUSTEP können die notwendigen Funktionen zur Dokumenten-Erstellung wie *Eingabe, Korrektur, Formatieren, Drucken von Texten* ausgeführt werden, da diese in allen Wissenschaftsbereichen zum Zweck der Dokumentation und Publikationsvorbereitung benötigt werden. TUSTEP wurde aber als Hilfsmittel für diejenigen Wissenschaften entwickelt, in denen vor allem *Texte* Objekte der Forschung sind: Literatur-, Sprachwissenschaften, historische Wissenschaften und Bibliothekswesen. Der erste Schritt ist die Eingabe des zu bearbeitenden Textes in eine Textdatei nach vorher festgelegten Transliterationsrichtlinien, für die in Kapitel 4 ein konkreter Vorschlag gemacht werden soll. Nach der Eingabe des Textes steht die Textdatei für verschiedene Rechenoperationen zur Verfügung, z. B. zur automatischen Collationierung von Texten (Abb 2)[7], wobei die Texte zeilensynoptisch untereinander dargestellt werden können. Es wird von einem Basis–Text ausgegangen, in diesem Fall der Apokalypse–Text der Handschrift *N1*. Identische Lesarten werden mit einem Gleichheitszeichen gekennzeichnet (=), abweichende Lesarten werden notiert. Treten Abweichungen vom Basis–Text in mehreren Handschriften auf, so wird dies vermerkt (z. B. = [Rum]). Diese Textvergleiche können mit verschiedenen Basis–Texten durchgeführt, die Anordnung und die Auswahl der Texte dabei beliebig variiert werden.

Auf Textdateien können aber nicht nur Vergleichprogramme sondern auch Registerprogramme verschiedener Art zugreifen. Hier sollen kurz zwei Möglichkeiten vorgestellt werden. Von Textdateien kann automatisch ein vollständiges Wortformenregister erstellt werden, das je nach Wunsch das Textcorpus einer Handschrift, d. h. einer Textdatei, oder von mehreren Handschriften enthält. Die Wörter werden nach einem vorher festgelegten Sortierschlüssel alphabetisiert und können mit verschiedenen Referenz– und Frequenzangaben auf einen Drucker ausgegeben oder wiederum in eine Textdatei in Form einer ständig automatisch erweiterbaren Datenbank übertragen werden. In der folgenden Abb. 3 wird die erste Seite eines automatisch erstellten Wortformenregisters der ersten drei Kapitel der Johannes–Apokalypse von elf Handschriften dargestellt. Zuerst erfolgt der Lemmaeintrag, in diesem Fall ⍺, ⍺ΓΓЄΛΟΥ usw.; hinter dem Lemma wird die absolute Häufigkeit im Untersuchungsabschnitt in runden Klammern angegeben, in diesem Fall kommt die Konjunktion ⍺ 26 mal und das

---

[7] In Abb. 2 wird der Text *N1* mit neun anderen Texten (*T122, Rum, T119, T121, D316, U82, U310, T710, C*) verglichen. Es handelt sich hierbei um den Text aus der Johannes-Apokalypse, Kap. I, V. 1 (Auszug).

Abb. 2: Textvergleich (Basis *N1*): Kap. I, 1 der Johannes–Apokalypse

1.4	ДПДКОЛИѰИИ	И(ИСѲ)СД	Х(РИ)С(Т)Д	ЮЖЕ	ДДСТЬ
[T122]	ДПОКОЛИѰИСЪ	И(ИСѲ)СЪ	Х(РИСТ)Ъ	=====	=====
[Rum]	ДПОКДЛИПСИ	И(И)С(Ѳ)СЪ		=====	=====
[T119]	=[Rum]	=[T122]	Х(РИСТ)Д	=====	=====
[T121]	=[Rum]	=[T122]	=[T119]	=====	=====
[D316]	=[Rum]	И(ИСѲ)С	=[T119]	=====	ДДСТЬ
[U82]	=[Rum]	=[D316]	=[T119]	=====	=====
[U310]	=[Rum]	=[D316]	=[T119]	=====	ДДЛ
[T710]	ДПОКДЛИѰИ	И(ИС)Ѵ(С)	=[T119]	ѠЖЕ	ДДДЕ
[C]	=[T710]	=[D316]	=[T119]	=[T710]	=[U310]

1.5	ѭмѲ	Б(ОГ)Ъ	ПОКДЗДТИ	РДБОМЪ
[T122]	ЄмѲ	Б(О)ГЪ	========	======
[Rum]	=====	=[T122]	========	======
[T119]	=[T122]	======	========	======
[T121]	=[T122]	======	========	РДБОМ
[D316]	=[T122]	=[T122]	========	=[T121]
[U82]	=[T122]		========	РДБОМЬ
[U310]	=[T122]	======	========	======
[T710]	=[T122]	=[T122]	========	=[T121]
[C]	=====	======	========	=[T121]

1.6	СВОИМЪ	ИМЪ	ЖЕ	ПОДОБДШЕ	БЪТИ
[T122]	======	ИМ	==	========	======
[Rum]	======	===	==	========	======
[T119]	======	===	==	========	======
[T121]	СВОИМЬ	ИМЬ	==	========	======
[D316]	СВОИМ	===	==	ПОДОБДЄТЪ	======
[U82]	=[D316]	=[T122]	==	ПОДОБДЄТЬ	======
[U310]	======	===	==	ПОДОБД	======
[T710]	======	=[T122]	==	ПОДОБДЄТ	======
[C]	СВОИ	=[T122]	==	=[U310]	======

1.7	ВЪ	СКОРЪ		СКДЗД.
[T122]	==	СКОРЄ И		======
[Rum]	==	ЗБОРЪ	=[T122]	======
[T119]	==	=====	=[T122]	======
[T121]	В	=====	=[T122]	======
[D316]	==	=====	=[T122]	======
[U82]	==	=====	=[T122]	======
[U310]	=[T121]	=====	=[T122]	НДЗНДМЕНД.
[T710]	ВО	=====	=[T122]	=[U310]
[C]	=[T121]	=====	=[T122]	=[U310]

N1: Sammlung Nikolskij (F. 32), No. 1, 13. Jh., BAN
T122: Sammlung Sergius-Dreifaltigkeits-Kloster (F. 304), No. 122, 15. Jh., RNB
Rum: Sammlung Rumjancev (F. 256), No. 8, 14. Jh., RNB
T119: Sammlung Sergius-Dreifaltigkeits-Kloster (F. 304), No. 119, 14. Jh., RNB
T121: Sammlung Sergius-Dreifaltigkeits-Kloster (F. 304), No. 121, 14. Jh., RNB
D316: Sammlung Archangelsk (F. 7), No. 316, 16. Jh., BAN
U82: Sammlung Uvarov, No. 82, 16. Jh., GIM
U310: Sammlung Undolskij (F. 310), No. 1, 15. Jh., RNB
T710: Sammlung Sergius-Dreifaltigkeits-Kloster (F. 304), No. 710, 15. Jh., RNB
C: aus dem Neuen Testament des Čudov-Klosters, 14. Jh., in: Lehfeldt, W.: Neues Testament des Čudov-Klosters. Köln 1989.

Abb. 3: Wortformenregister

```
Δ (26)  0.24%
    I,15<T120> I,15<T121> III,15<T121> III,16<T121> I,20<U310>
    II,2<U310> II,26<U310> III,9<U310> I,20<T710>
    II,2<T710> II,17<T710> II,27<T710> III,9<T710>
    I,20<Tl> II,2<Tl> II,17<Tl> III,9<Tl>
    III,9<U82>
    I,10<SSMD> I,20<SSMD>
    I,20<C> II,2<C> II,8<C> II,17<C> II,18<C> III,9<C>
ΔΓΓЄΛϑ (6)  0.04%  II,11<U82> II,12<U82> II,18<U82> III,6<U82>
    III,7<U82> III,14<U82>
ΔΓΓ()ΛИ (6)  0.04%  I,20<T120> I,20<U310> I,20<U82> I,20<SOL>
    I,20<D316> I,20<C>
ΔΓΓ()Λ()ЖЬ (1)  0.01%  I,1<C>
ΔΓΓ()ΛОЖ (2)  0.03%  I,1<T120>
    I,1<D316>
ΔΓΓ()ΛОЖЪ (1)  0.01%  I,1<SOL>
ΔΓΓ()ΛОЖЬ (2)  0.01%  I,1<T710> I,1<U82>
ΔΓΓΛϑ (1)  0.01%  II,1<T121>
ΔΓΓ()Λϑ (59)  0.38%  I,20<T120> II,1<U310> II,8<U310> II,12<U310>
    II,18<U310> III,1<U310> III,7<U310> III,14<U310> II,1<T710>
    II,12<T710> II,18<T710> III,1<T710> III,7<T710> III,14<T710>
    I,20<U82> II,7<U82> II,8<U82> III,1<U82> I,20<SOL> II,1<SOL>
    II,7<SOL> II,8<SOL> II,11<SOL> II,12<SOL> II,18<SOL> II,29<SOL>
    III,6<SOL> III,7<SOL> III,13<SOL> III,14<SOL> I,20<D316> II,1
    II,7<D316> II,8<D316> II,11<D316> II,12<D316> II,17<D316>
    II,18<D316> II,29<D316> III,1<D316> III,6<D316> III,7<D316>
    III,13<D316> III,14<D316> I,20<SSMD> II,1<SSMD> II,7<SSMD>
    II,8<SSMD> II,12<SSMD> II,17<SSMD> II,18<SSMD> II,29<SSMD>
    III,6<SSMD> III,7<SSMD> III,13<SSMD> III,14<SSMD>
ΔΓΓ()ΛЪ (2)  0.01%  I,20<T710> II,1<U82>
ΔΓΓ()ΛЫ (6)  0.04%  III,5<U310> III,5<T710> III,5<U82> III,5<SOL>
    III,5<D316> III,5<SSMD>
ΔΓΓ()ΛЬ (1)  0.01%  I,20<SSMD>
ΔΓΓ()ΛЮМ (2)  0.02%  I,1<U310> I,1<SSMD>
ΔΔЄΛΦИИСКЫЮ (1)  0.01%  III,7<T6>
ΔΔОВЫ (4)  0.03%  I,18<U310> I,18<T710> I,18<Tl> I,18<C>
ΔΔОВЬ (1)  0.01%  III,7<SSMD>
ΔΔϑ (7)  0.04%  I,18<T6> I,18<T120> I,18<T121> I,18<U82> I,18<SOL>
    I,18<D316> I,18<SSMD>
ΔЖЄ (1)  0.01%  III,4<T710>
ΔЖЬ (1)  0.01%  II,10<SSMD>
ΔЗ (9)  0.06%  I,9<T120> I,17<T120> III,9<D316> I,9<C> II,6<C>
    II,27<C> III,9<C> III,19<C> III,21<C>
ΔЗЪ (79)  0.51%  I,8<T6> I,9<T6> I,17<T6> II,6<T6> II,23<T6>
    II,27<T6> III,9<T6> III,10<T6> III,19<T6> III,21<T6>
    I,8<T120> II,6<T120> II,23<T120> II,27<T120> III,9<T120>
    III,10<T120> III,19<T120> III,21<T120> I,8<T121> I,9<T121>
    I,17<T121> II,6<T121> II,23<T121> II,27<T121> III,9<T121>
    III,10<T121> III,19<T121> III,21<T121> I,9<U310> II,6<U310>
    II,27<U310> I,8<T710> I,9<T710> I,17<T710> II,6<T710>
    II,23<T710> II,27<T710> III,10<T710> III,19<T710>
    III,21<T710> I,8<Tl> I,9<Tl> I,17<Tl> II,6<Tl> II,23<Tl>
    II,27<Tl> III,9<Tl> III,19<Tl> I,8<U82> I,9<U82> I,17<U82>
    II,6<U82> II,23<U82> II,27<U82> III,9<U82> III,10<U82>
    III,19<U82> I,8<SOL> I,9<SOL> II,6<SOL> II,23<SOL> II,27<SOL>
```

Substantiv ΑΓΓΕΛΟΥ 6 mal vor. Hinter der absoluten Häufigkeit wird die relative Häufigkeit in Prozent zum Gesamtwortformenbestand des Untersuchungsabschnitts angegeben. Hinter der relativen Häufigkeit erfolgt die Belegstellenangabe, wobei hier der Kapitel– und Verszähler als Referenzangabe gewählt wurde, es kann jedoch auch auf die originale Seiten– und Zeilenzählung der Handschrift Bezug genommen werden. Hinter der Referenzangabe steht die Sigle der Handschrift in spitzen Klammern, auf die sich die Referenzangabe bezieht. Solche automatisch erstellten Wortformenregister sollen keinen Index, kein Glossar oder sogar ein Wörterbuch ersetzen. Sie bieten jedoch eine umfassende Arbeitsgrundlage, auf die der Lexikograph aufbauen kann.[8]

Eine weitere große Hilfe für die Bearbeitung von Texten ist der sogenannte KWIC–Index, der ein Wortformenregister mit dem dazugehörigen Kontext erstellt. Auch hier sind verschiedene Arten von Referenz– und Frequenzangaben denkbar. In dem hier vorgelegten Beispiel (Abb. 4) handelt es sich um die erste Seite eines KWIC–Index der Textfamilie *b* (*C, TI, U310, T710*) im Untersuchungsabschnitt Kap. I–III der Johannes–Apokalypse. Hinter dem Lemmaeintrag erfolgt die absolute und relative Häufigkeitsangabe in Klammern. In den nächsten Zeilen werden die Belege in ihrem Kontext angeführt. Die Referenzangabe erfolgt hier nach dem Seiten– und Zeilenzähler des Originals, kann sich aber auch auf den Kapitel– und Verszähler beziehen. Hinter der Referenzangabe erfolgt in eckigen Klammern die Sigle der Handschrift. Das Lemma, auf das sich der Kontext bezieht, wird in der Mitte der Zeile zentriert. Der Kontext ist hier so gewählt, daß die Umgebung der jeweilige Zeile, in der das Lemma steht, mitausgedruckt wird. Es können jedoch auch die ersten zwei Wörter vor und hinter dem Lemma, oder der gesamte Vers gewählt werden. An dieser Stelle soll noch einmal darauf hingewiesen werden, daß die verschiedenen Programme zur Erstellung von zeilensynoptischen Vergleichen, Wortformenregister und KWIC–Index auf die oben beschriebenen Textdateien automatisch zugreifen, ohne sie zu verändern. Wenn Textdateien unter einheitlichen Kriterien (Transliterationsrichtlinien) erstellt werden, können diese Programme systematisch erweiterbare Datenbanken erstellen, die nicht nur für Editionen sondern insbesondere auch für die Herstellung von Wörterbüchern von großer Bedeutung sein können.

---

[8] Diese automatische Registererstellung hat sich bei der Edition des litauischen Kleinen Catechismus D. M. Lutheri durch Panzer 1993 bewährt.

Abb. 4: KWIC–Index

```
ѧ (5=0.31%)
1.4 [Tl]                    глаѳ() ѧ апокалипси и()с()с х()с()ѧ юже
2.37 [C]                           ѧ свѣтилник :з: седмь ц()рквъ
3.23 [Tl]  и седми ц()рквь сѵть ѧ свѣтилникъ
2.2 [U310]                   сѵть ѧ свѣтилникъ :з: седмь ц()рквъ
2.2 [T710]                   сѵть ѧ свѣтилникъ :з: седмь ц()рквъ
агг()ли (3=0.19%)
2.1 [U310]  ѧ сѵть седмь звѣздъ агг()ли седми ц()рьковъ
2.1 [T710]  ѧ сѵть седмь звѣздъ агг()ли седми ц()рьковъ
2.36 [C]                   звѣздъ агг()ли седми ц()рквъ сѵт
агг()лом (2=0.13%)
1.3 [U310]                         агг()лом своимъ рабѵ своемѵ иоан
1.3 [T710]                         агг()лом своимъ рабѵ своемѵ иоан
агг()л()мь (1=0.06%)
1.10 [C]                           агг()л()мь своим рабѵ своемѵ
адовы (4=0.25%)
2.30 [C]       ключа см()ртныѫ и адовы
3.17 [Tl]      ключа см()ртныѫ и адовы
1.37 [U310]    ключа см()ртныѫ и адовы напиши ѵбо ѫже видѣ
1.37 [T710]    ключа см()ртныѫ и адовы напиши ѵбо ѫже видѣ
аз (1=0.06%)
1.33 [C]                           аз иоанъ брат ваш и обещник
азъ (7=0.44%)
2.26 [C]      мнѣ гл()а не боиса азъ
3.13 [Tl]     мнѣ гл()а не боиса азъ
2.5 [Tl]                           азъ есмь алфа и о гл()ть г()с()д
1.31 [C]                           азъ есмь алфа и о гл()тъ г()с()ь
1.17 [U310]              вседръжитель азъ иоаннъ брат
1.17 [T710]              вседръжитель азъ иоаннъ брат
2.7 [Tl]                           азъ иванъ братъ вашь и обещникъ
азь (4=0.25%)
1.16 [U310]             еи аминь. азь есмь алфа и о гл()ть г()ь
1.16 [T710]             еи аминь. азь есмь алфа и о гл()ть г()ь
1.35 [U310]  мнѣ гл()а не боиса азь есмъ пръвыи и послѣдниі
1.35 [T710]  мнѣ гл()а не боиса азь есмъ пръвыи и послѣдниі
алфа (4=0.25%)
1.16 [U310]  еи аминь. азь есмь алфа и о гл()ть г()ь б()ъ иже
1.16 [T710]  еи аминь. азь есмь алфа и о гл()ть г()ь б()ъ иже
2.5 [Tl]                  азъ есмь алфа и о гл()ть г()с()дь б()ъ
1.31 [C]                  азъ есмь алфа и о гл()тъ г()с()ь б()ъ
амин (7=0.44%)
1.30 [C]  вса племена земнаѫ еѫ амин.
2.29 [C]       есмь в вѣки вѣком амин.
2.1 [Tl]  держава в вѣки вѣкомь амин.
2.4 [Tl]  вса племена земнаѫ еи амин.
1.36 [U310]     есмь в вѣк вѣком амин.
1.36 [T710]     есмь в вѣк вѣком амин.
1.27 [C]                   вѣком амин. се градет съ облакы
аминь (5=0.31%)
3.16 [Tl]             вѣки вѣкомъ аминь.
```

4. Transliterationsrichtlinien

Für die Erstellung von Textdateien sind einheitliche, allgemeingültige Transliterationsrichtlinien notwendig, mit denen alle notwendigen Informationen einer Handschrift eindeutig gekennzeichnet werden, um später zum Vergleich automatisch abgerufen und verglichen werden zu können. Hier stellt sich die Frage, welche Informationen für eine spätere Bearbeitung von Bedeutung sind und welche Informationen vernachlässigt werden können. Zwei Beispiele mögen die Problematik veranschaulichen:

a) In altkirchenslavischen Handschriften werden die Wörter am Zeilenende ohne erkennbare Regeln getrennt: (Ende Zeile 1) СВО (Anfang Zeile 2) ИМЪ für СВОИМЪ. Ein Eintrag in eine Textdatei ist auf zwei Arten denkbar. Die Trennung wird vernachlässigt, und das Wort wird zusammengeschrieben als СВОИМЪ notiert. Dabei gehen aber alle Informationen über die Trennung in der Textdatei verloren. Die zweite Möglichkeit besteht in der Notation mit einer eindeutig als Trennungszeichen definierten Markierung, z. B. СВО | ИМЪ. In diesem Fall bleibt die Information der Worttrennung erhalten und kann nicht nur innerhalb einer Handschrift sondern auch im Vergleich mit anderen Handschriften untersucht werden.

b) In altkirchenslavischen Handschriften begegnen sehr häufig Superskriptionen, die häufig am Wortende, aber auch in der Mitte eines Wortes anzutreffen sind. Für die Rekonstruktion eines Textes ist die Frage der Superskription nicht relevant und so kann die Superskription aufgelöst werden. Damit gehen aber jegliche Informationen darüber verloren, wann und wie oft die Superskription in einer Handschrift oder im Vergleich mit anderen Handschriften verwendet wird. Die Superskription kann z. B. mit zwei Schrägtrichen eingegrenzt werden (ѡ/т/ für оТЪ oder хр/с/тс für ХРИСТОСЪ).

Die folgende Aufstellung enthält allgemeine Richtlinien zur Erstellung von Textdateien:

1) Superskriptionen werden durch Schrägstriche gekennzeichnet:
   ѢЫ mit Superskription х wird als ѢЫ/х/ notiert.
2) Tilden werden durch zwei vorausgehende runde Klammern wiedergegeben: Г$\tilde{\text{Л}}$ЄТЪ wird als Г()ЛЄТЪ notiert.
3) Die Worttrennung am Zeilenende wird durch einen senkrechten Strich | gekennzeichnet.

4) Initialen werden bei einer Überschrift mit vorangehendem &, bei Versanfang mit vorangehendem @ gekennzeichnet.
5) Die Einträge erfolgen nach der Seiten- und Zeilenzählung des Originals. Die Kapitel- und Verszählung des Originals wird mit zwei senkrechten Strichen begrenzt: |Г| bedeutet dritter Vers. Die standardisierte Kapitel- und Verszählung wird in eckige Klammern eingegeben: [II, 2]
6) Hinzufügungen werden durch Anführungszeichen eingegrenzt:
Ms. *U81;* [I, 9] ... в тєрпѣньи "і()с х()с()въ бъіхом во ѡтоцѣ" наричаємѣмь патмѣ ...
7) Kommentare des Textbearbeiters werden in geschweiften Klammern notiert: свѣ{verbessert aus и}дѣтєль
8) Nicht eindeutige Lesarten werden mit einem in runden Klammern stehenden Fragezeichen gekennzeichnet: (?)

Der Vorteil von solchen Textdateien besteht darin, daß Informationen, die für den gegenwärtigen Arbeitsschritt nicht benötigt werden, wie z. B. den Kommentar des Textbearbeiters bei der Registererstellung, oder orthographische Distribution von aksl. и, і, ї und о, ѡ sowie originale Vers- und Kapitelzählung bei der Rekonstruktion, unterdrückt und bei Bedarf wieder abgerufen werden können. Es ist von außerordentlicher Wichtigkeit, daß erstens die Markierungen in den Textdateien eindeutig sind und daß allgemeingültige Transliterationsrichtlinien geschaffen werden, nach denen Textdateien erstellt werden sollen. Ohne solche Transliterationsrichtlinien ist ein flexibler Datenaustausch nicht realisierbar. Schon im Jahr 1981 wurde ein Vorschlag für solche Richtlinien unterbreitet, der im Anhang 1 mit Ergänzungen für die Akzente gegeben wird[9].

Eine konkrete Anwendung in einer Textdatei wird im folgenden Beispiel gegeben. Der Anfang der Johannes–Apokalypse aus der Gennadius–Bibel (Abb. 5) wird in einer Textdatei (Abb. 6) im Anhang dargestellt. Selbstverständlich können diese Textdateien auch automatisch in altkyrillische Buchstaben übertragen werden. Für den Datenaustausch mit anderen Programmen anderer Anwender, ist es jedoch sinnvoll, Textdateien im ASCII-Format zu erstellen. In solche Text–Dateien können auch Akzente integriert werden, die je nach Bedarf des Anwenders wieder eliminiert werden können, da die Codierung der Akzente zu sehr umfangreichen und unübersichtlichen Textdateien führen können, für einen Akzentologen sollten sie dennoch zur Verfügung stehen. Als Beispiel wird

[9] Hilandar Monastery and other repositories of Medieval Slavic Manuscripts (Appendix 2: Proposed Transliterations for Computerization of Medieval and Modern Cyrillic), in: Полата Книгописная Bd. 5 (1981), S. 89.

der Beginn der Johannes–Apokalypse aus der Ostroger–Bibel (Abb. 7) in einer Textdatei (Abb. 8) im Anhang mit Akzenten dargestellt.

Das oben beschriebene System wird als Hilfsmittel zur Edition der altkirchenslavischen Johannes–Apokalypse erfolgreich eingesetzt und die Text–Dateien stehen jederzeit über E–Mail anderen Benutzern zur Verfügung.

Literaturhinweise:

Bakker, M. – van der Tak, J. G.: Collating Greek and Slavic Apostolos Manuscripts, in: Palaeobulgarica XVIII, 2 (1994), S. 32–49.

Bodjanskij: Житие блаженаго отьца нашего и учителя, Мефодия, архиепископа моравьска, in: Чтения в императорском обществе истории и древности российских при Московском университете. Кн. 1 (1865), S. 1–97.

Hansack, E.: Paläoslavistik und Computer: die Erschließung des kichenslavischen Erbes, in: Anzeiger für Slavische Philologie, Bd. XXII, 1 (1993), S. 87–96.

Hilandar Monastery and other repositories of Medieval Slavic Manuscripts, in: Полата Книгописная, Bd. 5, S. 43–92.

Kuev, K.: Черноризец Храбър. Sofia 1967.

Lehfeldt, W.: Neues Testament des Čudov–Klosters (=Bausteine zur Geschichte der Literatur bei den Slaven). Köln 1989.

Ott, W.: Computerunterstützte Edition, in: Edition 3 (1989), S. 157–176.

Panzer, B.: Zum Text der Method–Vita, in: Die Slawischen Sprachen. Bd. 13 (1987), S. 61–73. Und in: Symposion Methodianum. Beiträge der Internationalen Tagung in Regensburg (17. bis 24. April 1985) zum Gedenken an den 1100. Todestag des hl. Method. Hrsg. v. K. Trost, E. Völkl, E. Wedel. Neuried 1988, S. 647–55.

Ders.: Zur Edition kirchenslavischer Texte, in: Die Slawischen Sprachen. Bd. 28 (1991), S. 83–102.

Ders.: Der Kleine Catechismus D. M. Lutheri – Mážas Katgismas D. Mertino Lutteraus, Besorgt von D. Henrich Lysius Tillsitt 1719. Deutscher und Litauischer Text nach der Handschrift Ms. Slav. Quart. 4 (Berlin/Krakau). (= Heidelberger Publikationen zur Slavistik, Linguistische Reihe, Bd. 5). Frankfurt/M. 1993.

Anhang: **Vorschlag für Transliterationsrichtlinien**

a	Ⱑ	jh	Ⱌ
b	Ⱏ	sh	Ⱎ
v	Ⱒ	wt	Ⱏ
g	Ⱓ	7	Ⱐ
d	Ⱔ	y	Ⱑ
d2	Ⱘ	6	Ⱒ
e	Ⱙ	e2	Ⱓ
je	Ⱚ	ju	Ⱔ
zh	Ⱛ	ja	Ⱕ
z	Ⱜ	eh	Ⱖ
z2	Ⱝ	8	Ⱗ
1	Ⱞ	j8	Ⱘ
1h	Ⱟ	q	Ⱙ
i	Ⱑ	jq	Ⱚ
j	Ⱑ	8h	Ⱛ
k	Ⱑ	k3	Ⱜ
l	Ⱑ	p2	Ⱝ
lj	Ⱑ	&	Kapitelüberschrift
lh	Ⱑ	@	Versanfang
m	Ⱑ	"..."	Hinzufügungen
n	Ⱑ	{...}	Kommentare
nj	Ⱑ	[I,1]	Kapitel-, Verszähler
nh	Ⱑ	|	Zeilenende
o	Ⱑ	()	Tilden
oh	Ⱑ	/.../	Superscriptionen
p	Ⱑ	:x:	Zahlzeichen zwischen Doppelpunkte gesetzt.
r	Ⱑ		Steht über dem Zahlzeichen eine Tilde, so wird
s	Ⱑ		sie mit zwei vorausgehenden Klammern
t	Ⱑ		gekennzeichnet, z. B. $\overrightarrow{\text{I}}$ als :()1:
c2	Ⱑ	%1	◌́ (оксия)
u	Ⱑ	%2	◌̀ (вария)
ou	Ⱑ	%3	◌̾ (оковавы, смычец)
uh	Ⱑ	%4	◌̑ (кендима)
f	Ⱑ	%5	◌͡ (камора, периспомень)
th	Ⱑ	%6	’ (апостроф)
x	Ⱑ	%7	◌̔ (дасия)
c	Ⱑ	%8	◌̓ (звательцо)
ch	Ⱑ		

Abb. 5:

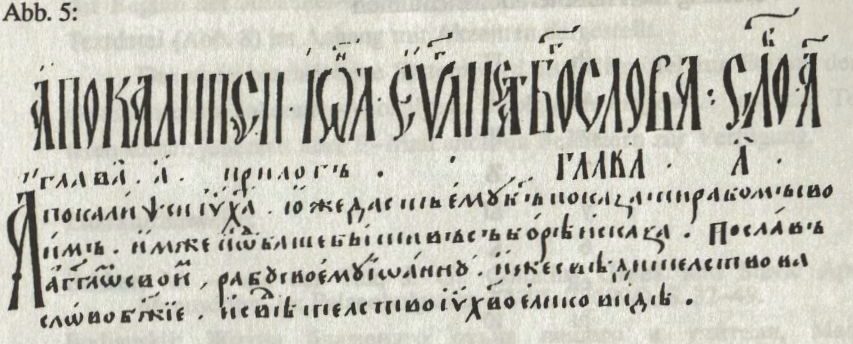


Abb. 6:

```
1.1     &apokalipsi ^. 1oh/()n/a ^. eu/()g/lista b/()g/oslova^.
        sl/()v/o :()a:

1.2     glava. :()a:. prilog7. glava. :()a:.

1.3     [I,1]@apokalip2si  iu() x()a juzhe dast6 emuh b()7 pokazati
        rabom7 svo|im7.

1.4     imzhe poh/d/bashe byti v7 s7boreh i skaza. @poslav7

1.5     ag()glohm svoim, rabuh svoemuh 1ohannuh. [I,2]izhe
        svehditelstvova

1.6     slohvo b()h1he i sveh/d/telstvo 1uxvo eliko videh.
```

Abb. 7:

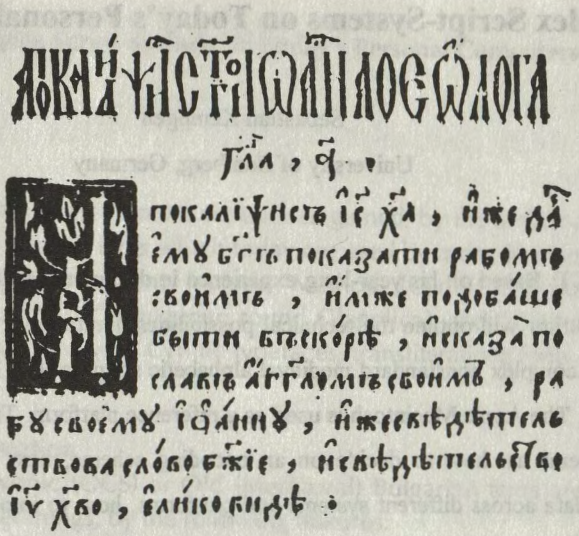


Abb. 8:

```
1.1     &apokal%1ip2i/()s/ s()tgo ioha/()n/na the%5%1ohloga
1.2     gl/()a/v, :()a:·
1.3     [I,1]@apokal1hp2is7 %51s() x()a, %5ihe da/()s/
1.4     %5emu b()g7 pokazati rabom7
1.5     svo%5m7, %5%1imhe podob%1awe
1.6     byti v7 sk%1oreh, %5i skaz%2a po|slav7
1.7     %5%1agg()lom7 svo%5im7, ra|buh
1.8     svo%5emuh %51%5o%1annuh, [I,2]%5izhe svehdehtel6|stvova
1.9     sl%1ovo b()zh1he, %5i svehdehtel6stvo
1.10    %51()u x()vo, %5el%1iko videh·
```

Comlex Script-Systems on Today's Personal Computer

Sebastian Kempgen

University of Bamberg, Germany

ABSTRACT: Based on his year-long experience in designing an Old Church Slavonic font, the author will outline the technical possibilities, limitations, and perspectives of realizing a complex nonstandard medieval alphabetic script system on today's personal computers. The Apple Macintosh is used as a reference platform. The report will cover such problems as: how to decide on an encoding scheme, how to enable users to exchange data across different systems and platforms, how to implement superscripts and charactervariants, design issues concerning the appearance of characters themselves, etc.

Complex Script-Systems on Today's Personal Computers

0. Introduction

The present paper is based on the experience gained by the author over the last ten years from designing fonts which today are used by many Slavists, especially in Western Europe. These fonts include "Method" (Cyrillic Old Bulgarian/Old Church Slavonic), "Kirill" (Bulgarian round Glagolica), and "Glagol" (Croatian square Glagolica) along with Cyrillic typefaces, transliterations fonts etc.

1. OCS – The Situation

Old Church Slavonic (OCS) or Old (Mediæval) Bulgarian texts are characterized, among other things, by the following features:

- No normative alphabet or orthography;
- No clear-cut distinction between grapheme and allographs;
- Many diacritics above lowercase and uppercase letters;
- Complete alphabet used as superscripts;
- Many ligatures.

Examples illustrating these points are well-known to every specialist. In fact, each scribe or school may be considered to use his/its own system applying more or less strict writing rules to the text in question. The farther we abstract from individual manuscripts, the closer we approach the ideal OCS alphabet which text-books and grammars present, although hardly any single text can be said to implement it one hundred percent.

Consequently, a font-designer has to make a decision about what he is aiming for: is it an individual text which he wants to electronically represent or is it the normalized abstract orthography and appearance of an OCS grammar? If a grapheme has several letterforms, are both to be included in the font or not? Which one has to be chosen as the basic one? Is it the older form, for example, or is it the more common variant? How should diacritics and superscripts be realized?

2. Font Technology

The basic distinction here is between bitmapped fonts and outline fonts. Ten years ago, bitmap fonts were the standard, when dot-matrix printers were the most common printers and laser-printers did not yet exist. In bitmap fonts, each character is represented at a fixed point-size by a collection of pixels. Today, of course, outline (or vector) fonts are the state of the art. The difference between

bitmap fonts and outline fonts is comparable to using a painting program (which turns single dots on or off) or a drawing program (where one draws objects which can be re-sized at will). In outline fonts, the designer draws the outline of a character, setting corner and curve points and connecting them with lines. While this sounds simple for a single character, it becomes a very demanding task for a complete alphabet. The premier font-creating package is called Fontographer; other programs (such as FontStudio) have practically disappeared or occupy a niche only (such as URW's fine Ikarus program). Using such a font-editor, the designer generates the end-user product, for example a PostScript Type 1 font or a TrueType font. When these fonts are being used, the outlines contained in the font files are scaled to the appropriate size (say 12 points) and filled with pixels, depending on the resolution of the output device. Laser-printers, for example, have at least four times the resolution of standard computer monitors. Consequently, characters are printed with much finer detail than represented on-screen. Bitmaps may or may not be required with fonts, depending on the platform and the font format. Commercial fonts usually come with a range of hand-edited or polished bitmaps for perfect on-screen representation, especially in small point sizes.

Font files as used by Fontographer usually contain 2^8 = 256 slots of which 225 positions may normally be used. Although font files with 512 or 1024 characters are possible, creating fonts with that many characters is impossible if we deal with single-byte languages. Thus, in all practice, we are limited to standard 8-bit fonts. If the number of characters required exceedes this limit, as is easily the case with OCS, several fonts habe to be created with each one being dedicated to a specific task. For example, one font can contain the basic alphabet and diacritics, another one superscripts and so on.

3. OCS Fonts on Today's PCs

3.1. Encoding Strategies

Font development is always done under specific historic conditions. The basic and simplified decision algorithm when creating a font for a specific language or script can be summarized as follows: Is there a standard? No – Create new encoding. Yes – Is the standard practical? Yes – Follow standard. No – Create new encoding.

For OCS, the answers will be "no" or "yes"/"no", in both cases resulting in the advice to develop a new encoding. It is reasonable to base an OCS encoding on the encoding chosen for Scientific Cyrillic, and this in turn is based on ASCII or its variants. After an encoding for OCS has been decided on, Glagolitic can also be encoded: ASCII → Scientific Cyrillic → OCS → Glagolitic. The author himself, for example, created his own encoding scheme for scientific Cyrillic in 1985; from this in 1988 the encoding for Method, the OCS font, has been derived. And on the encoding used by Method, the corresponding Glagolitic

fonts, created in 1992, are based. This ensures – as far as possible – compatibility between the different scripts.

When choosing an encoding for new characters or scripts, several principles can be followed when starting from and modifying the standard Latin 8-bit-encoding:

• Vowel on vowel, consonant on consonant

The reason for this is two-fold: it is a mnemotechnic aid to remember where the characters have been placed on the keyboard, and it ensures maximum compatibility with existing hyphenation routines.

• Character pairs on character pairs

This principle is also obvious, but it is not always followed when it could be followed; and sometimes it is impossible or impractical to follow. Again, there are two main benefits: keystroke pairs remain pairs (lowercase vs. uppercase), and functions such as "small caps", which most wordprocessors offer, will continue to work.

• Basic ABC into lower ASCII

The reason for this is to make access to the most frequently used characters as convenient as possible, because the lower ASCII character set can be completely accessed on nearly every keyboard. Variants and accented characters can be put into higher ASCII. That less often used characters are somewhat harder to type than the basic set is a principle which underlies many keyboard drivers and thus is nothing specific. (When designing keyboard drivers, this principle becomes "Put basic ABC on normal layers, additional characters on option-keys.)

• Phonetic or graphic similarity

When Cyrillic is mapped to a keyboard layout, phonetic similarity may be used. Thus, for example, Cyrillic A will be on A, B on B, D on D and so on. This will, of course, depend on the phonetic values usually associated with characters. On the German keyboard, for example, you have characters *ä, ö,* and *ü.* Putting Cyrillic *я, ё,* and *ю* on these keys is an obvious solution for German users.

• Design for local keyboard driver

This principal is especially popular with Macintosh users, because the Macintosh has always had very convenient keyboard drivers. On the Macintosh, the Option (or Alt in PC terms) key by itself is not used to issue commands but is used to enter foreign characters, mathematical symbols etc. in a very straightforward manner. For example, on most keyboard drivers, option-a will produce *å,* option-c produces *ç,* option-p produces π, option-z produces Ω (which, of course, is the last character of the Greek alphabet) and so on. Macintosh keyboards thus have four layers: lowercase, shift, option, shift-option. As a keyboard has around 50 keys, this allows access to 200 characters. Add the five accent keys which work as dead keys to enter accented characters, and the

complete 8-bit character set can be can accessed very easily. Typing ALT-0 + decimal value has never been necessary on the Macintosh.

This feature of Macintosh keyboard drivers makes it clear why designing fonts for local keyboard drivers is so popular on this platform. Option keys produce alternate or related characters, and if a foreign language has more characters than the basic English alphabet, it is natural for users to access less often used characters by using the option key.

There is a drawback to this solution, however: what may seem logical on one keyboard, may not be true for another local keyboard driver. While many option-keys are identical on all international keyboads, most shift-option keys are not. When designing for the US keyboard, a certain combination may not make much sense on the French keyboard, for example. But as keyboard drivers can be easily created, copied, modified and installed by end-users on the Macintosh, a simple solution is to include the keyboard driver for which a font was designed with the font itself.

3.2. Character Design

Character designs have a "look and feel", and this is also true for OCS. Different countries have different traditions with respect to the printing types they use, so what is "normal" for a user in one country will not necessarily be of the same quality to another user. Font designers usually call OCS letters a "black letter design" because of the thick black stems which mediæval scripts exhibit giving OCS fonts a certain "weight".

When designing a general-purpose OCS font, it is reasonable to start from the printing types as used in standard grammars, text-books etc. because the look of these characters is what most normal users will expect from an electronic rendition.

However, specialists might also want to reproduce the look and feel of a specific text they are working on. Consequently, fonts can "look" Russian, Serbian, or Bulgarian, and/or 12th, 15th, 17th century, for example. Taken to the extreme, one font could be created for every single text.

After the basic decisions have been made, the process of recreating characters electronically sounds simple: select the letter form for each characters which seems most typical, use a scanner to bring a picture of the character into the computer, then use a font-editor to trace the outer edges of the character. While this can be done quickly for a single character, expertise, skill and a trained eye are required for complete character sets of commercial quality. High-quality fonts must, for example, exhibit even character size, spacing, constant "colour" etc. when printed. The amount of work that goes into designing a new font from scratch has been compared to writing a book. Also, hardware requirements should not be forgotten: font design requires the use of the largest possible monitor (anything else but a 21" flat screen is too small), a fast CPU with a large memory (RAM), and a fast PostScript-printer.

3.3. Diacritics

In standard encoding schemes for Western languages and now for Eastern Europe, all foreign characters are preaccented, that is, are available as such. Coupled with corresponding keyboard drivers, this, of course, is very convenient to users, and it is also the only way to achieve a typographically precise placement of the diacritic in every instance. It is important to remember that by pressing first ` and then i on the keyboard, a user does not tell the computer to put the grave accent on the i, rather, he tells it to call up the preaccented ì (notice the absence of a dot over the i!).

In contrast to standard modern languages, OCS has a very large number of different diacritis which can be combined with most base characters. To construct every conceivable combination of base character plus diacritic as such would result in a very large character set which could hardly be handled any more. Literally thousands of combinations have to be taken into account, because diacritics may sit centered over a character, centered between characters, or at the right or left side of a character. Consequently, font design will treat this problem by using so-called "flying accents". Flying accents are also called nonspacing accents, because they have a zero width, overlap to the left and can be combined with every base character. Using such a technique reduces a character set in a most dramatic way: every diacritic is needed only twice, once for lowercase and once for uppercase characters. However, this reduction comes at the expense of typographical precision. As base characters differ in width, flying accents will not fit equally well over each of them. Usually, they are constructed in such a way as to fit perfectly over characters of normal width. To achieve a better fit over narrow or broad characters as well, kerning can be built into the font. Kerning is a technique that moves two characters closer together or farther apart; one of the characters, can, of course, be also a diacritic. Kerning works within a single font, not between different fonts. Thus, it is impossible to simply construct a font which has only diacritics and nothing else and still kerns with base characters selected from other fonts. For Latin characters, the kerning tradition varies in different countries. Commercial fonts have from one hundred to one thousand kerning pairs ("Method", the author's OCS font, has nearly three thousand pairs, that is, it has been extensively kerned.)

While the combination of flying accents and kerning is certainly the right solution for OCS, application programs have to be considered. Unfortunately, even major word-processor do not care about the kerning available in fonts and output text without any kerning. Page-layout programs, though, always support kerning. On Windows machines, another problem has surfaced: because flying accents are not used in standard fonts at all, programs and printer drivers may not expect zero-width characters and may not (yet) handle them correctly. Thus, it is possible that a zero-width character is displayed on-screen as if it had a positive width, resulting in blank space to the right, or flying accents may be visible on-screen, but do not print or are cropped (the combination of Word for

Windows 6.0 and the PostScript printer driver for Apple LaserWriters seems to work without problems in this respect).

3.4. Superscripts

For superscripts, in principle the same is true as has been said about diacritics. Raising and lowering of individual characters has always been a feature of word-processors. It is obvious that this approach is too limited to handle a complete superscript alphabet as required for OCS. Consequently, superscript characters have to be implemented as a special font. Again, the same solution can be applied: superscripts can be constructed as flying accents, i.e. zero-width characters which operlap the character written before them. The simple solution that the author of this paper chose is to create a separate font and replace all uppercase characters with the corresponding superscript letter. This makes it very easy for users to enter superscripts, because they do not have to learn any new key combinations. Also, superscripts can be searched for simply as capital letters from a specific font.

4. The Future

4.1. Part I: Encoding

Clearly, the future of encoding is Unicode. Unicode is an encoding scheme with enough room for all the world's characters (most of them, anyway). The Unicode standard is being developed by The Unicode Consortium, founded and supported by several hardware and software manufacturers. For an excellent article about Unicode and its backgound see SHELDON (1991).

The Unicode standard has been published in two thick volumes, the first one being devoted to alphabetic characters, while the second one is devoted exclusively to Asian characters. How does Unicode treat OCS? Basically, Cyrillic OCS is thought of as modern Cyrillic plus x historic characters. It becomes immediately apparent that there are strange elements in this encoding which should never have been encoded as such (for example, izhica with double grave accent). On the other hand, many special characters are missing from this encoding, as any specialist in the field will point out. Glagolitic has not been standardized at all; it is mentioned shortly in the manual as a candidate for future inclusion though.

The current Unicode encoding, then, is not yet the standard expert Slavists are looking for. The good thing about Unicode is, however, that anybody can submit a proposal for the inclusion of new characters to the Unicode Consortium (the rules and guidelines for this are outlined in vol. 1 of the manual). So there is a way out of current limitations, if these suggestions are accepted.

Right now, we are seeing the first steps in implementing Unicode on PCs. Windows 95 is said to support Unicode encoding in (only?) five of its APIs (Application Programming Interface), while Apple says its Unicode support will be

"extensive" in "Copland" (System 8), planned for release in 1996, with "true Unicode support in major system components".

Before embracing Unicode as the solution to all encoding problems, however, one should be aware of the consequences Unicode will have. For *complete* Unicode support, all operating systems will have to be rewritten, all current software will have to be replaced with new versions, and, last but not least, Unicode fonts will have to be produced. Also, very powerful PCs will be required to run these Unicode-based operating systems. Hardly any of todays's PCs will be powerful enough or even be in use when full Unicode support finally becomes a reality. It seems realistic to the author to predict that widespread use of Unicode is at least five years away. Clearly, then, Unicode is no solution for today.

4.2. Part II: Technology

Future technology that is emerging now includes QuickDraw GX. QuickDraw GX is a Mac OS extension which takes care of font management, printing etc. It has been introduced as an optional feature with System 7.5 and will be an integral part of the next major Mac OS version which has been announced for 1996.

QuickDraw GX introduces a new font format, TrueType GX fonts. While older fonts (Type 1 and TrueType) continue to work (Type 1 fonts have to be converted first, though), only GX fonts can support new features such as automatic ligatures, automatic context-based allographs, automatic small caps – all in one font. The GX font format has been developed by Apple primarily for better handling of Arabic and similar scripts.

It seems that the new features of GX fonts could also be put to good use in the case of Fraktur fonts, mediæval fonts such as OCS etc. Although the technology of QuickDraw GX fonts seems promising, it has also been called a case of "overengineering". Acceptance of QuickDraw GX by end-users has been slow because of the memory requirements, some bugs in the initial release, and because powerful Macs are needed to perform the necessary calculations. Also, programs have to be rewritten to make use of special GX features, and few programs support GX at present. MS-Word for example, the de-facto standard in the scientific community, does not support special GX features. Adobe, inventors of PostScript Type 1 fonts and publishers of PageMaker, will also not support GX features which they say can be duplicated by using expert Type 1 fonts. The reason is simple: GX fonts are TrueType fonts, and Adobe does not supply TrueType fonts.

Another limitation is that so far QuickDraw GX is tied to the Macintosh and not a cross-platform technology, like other system software components from Apple (QuickTime and QuickDraw 3D are cross-platform, for example). Thus, data transfer between platforms is further complicated by using Quick-

Draw GX fonts. Also, Fontographer (v. 4.1), the premier font editor on the Mac, does not at present support the generation of TrueType GX fonts.

In spite of all this, Macintosh users still have to reckon with QuickDraw GX, because, as already pointed out, it will be an integral part of the next major Mac OS version.

5. The User

As the present conference proves, many Slavists care about standards. But what about "normal" scientific users who are writing books and papers on their PC and print final output on their laser-printer? What do these users want? From his own experience, the author knows that many users of foreign language fonts do not care about encodings, standards or data exchange with others. They simply want a solution which is convenient, that is, easy for them to use, one which fits in with their localized software (including keyboard driver). This is especially true for Europe with its many different countries and keyboards. Not all users are ready to switch keyboard drivers just to write a word or two in Cyrillic.

In Europe, the urge to move to encoding standards has been much greater within companies whose daily business relies on being able to exchange data between machines and platforms painlessly.

Every current software solution has an installed user base and following. To move these users to a new standard can be a very difficult and time-consuming task, and will always be only partially successful. A radical solution would, of course, be to offer only standard-encoded solutions after such a standard has been arrived at, sacrificing backward-compatibility.

6. Data Exchange

Exchange of data (text files, databases etc.) has previously presented problems for users. As far as Western languages are concerned, these problems are now a thing of the past. Apple bundles a System extension called "PC Exchange" with Macintosh System software; this allows PC disks to be mounted on the Mac desktop, PC files to be dragged to and from the Mac hard-disk etc. so that there is no difficulty in loading PC files into the Mac. Similar utilities now exist for Windows PCs giving those users access to Mac disks and files. Allowing PC disks and files to be manipulated as if they were Mac disks and files does not change anything in the content of the files, though, that is, so far no file conversion takes place. Fortunately, most cross-platform Mac programs directly read in texts made with their counterpart on Windows PCs and vice versa, re-encoding all characters as required on the fly. Also, there are data transfer utilities with filters for hundreds of application combinations.

If, however, Slavic or Baltic languages, which have a different encoding from Western languages, have been used in texts, these filters and conversion

procedures render complete nonsense because they treat all characters as if they were from the standard Western encoding scheme. This is also true of OCS texts written using today's OCS fonts. Consequently, other means of moving OCS text intact from a PC into the Mac and back again have to be found.

Fortunately, a solution is available on the Macintosh. Apple has shipped a program called Apple File Exchange (AFE) with System software 7.1. This is an application program which serves a similar purpose as the control panel PC Exchange which ships with System 7.5 and supersedes AFE. AFE reads PC disks and allows file transfer from a PC disk to the Mac hard-disk or vice versa. This in itself is, as we have seen, not enough. There is, however, a little-known add-on to AFE, called Text-to-Text Converter. This gem converts character codes on the fly while copying a text file, and, best of all, these translation tables are user-definable. They simply consist of two lists of corresponding ASCII or hex values. If the program encounters a character having the first value, it replaces it with a character having the second value. Furthermore, not only one-to-one replacements are possible, but also one-to-many replacements. This feature can be used, for example, to flag problems in the conversion process which could later be worked on by using a second conversion table or by a search-and-replace operation from within a word processor.

The author of this paper has already created translation modules for Cyrillic between Windows and Macs, and for Latin Eastern Europe (EE encoding to CE encoding). A conversion module for corresponding OCS fonts would be possible, if users required it.

There are, however, two limitations connected with this approach. The first one is that it works on plain ASCII text only. As this includes RTF files (Rich Text Format) which preserve character and document formatting, this really is no problem. The RTF format is supported by most major word processors, making this an ideal means of exchanging texts between platforms and programs. The second limitation lies in the fact that these converters do not work context-sensitively; thus, it is impossible to translate only selected words in a mixed-language text. Again, this does not present a problem if the source text is OCS only.

7. Conclusion

If we ask ourselves what can be achieved today in terms of standardization, the author would reach the following conclusion:

Possible and practical today are platform-dependent standards, that is one standard for Windows machines and one standard for Macintoshes. Utilities such as AFE (Apple File Exchange) with its Text-to-Text converter can be used for data exchange between platforms. The author of this paper is convinced that these platform-dependent standards will be needed at least for several years to come before a common Unicode-based standard can gradually replace them.

There can be no doubt, however, about the fact that not all current or future users care about standards or data exchange. They want the most convenient solution for their individual working habits, their local keyboard and software, and they want it now.

8. References

The Unicode Consortium:
1991-92 The Unicode Standard. Worldwide Character Encoding. Version 1.0. Vols. 1-2. Addison-Wesley: Reading, MA.

Apple Computer:
1995 Inside Macintosh: QuickDraw GX Printing. Cupertino.

Apple Computer
1995 QuickDraw GX Font Formats and The TrueType Font Format Specification. Cupertino.

Sheldon, K.M.:
1991 ASCII goes global. *Byte* 7, 108–116.

A TEI View of MS Pec 68

C. M. MacRobert

Lady Margaret Hall, Oxford

Keywords: Church Slavonic, CONCUR, Cyrillic, Multiple Hierarchies, Psalter, Rubrics, Superscripts, TEI, Variant Readings.

Abstract: This paper discusses problems of transliteration (diacritics and rubrics), coordination and description of original and corrected variant readings and representation of multiple textual hierarchies in a TEI mark-up of a Church Slavonic Psalter.

The purpose of this paper is to draw attention to three types of problem which arise when the mark-up conventions of the Text Encoding Initiative[1] are applied to a medieval Cyrillic manuscript, the 14th-century Serbian Church Slavonic psalter MS 68 from the Library of the Patriarchate at Pec. The immediate aim of my work on this manuscript is to produce camera-ready copy for a critical edition with various indices; a full electronic edition is a more distant and doubtful goal, whose realization depends in part on responses to the problems raised here.

My first type of problem will be familiar to anyone who has tried to produce a computer-readable version of a medieval text written in Cyrillic: how is one to represent the writing system in a way which is unambiguous and not liable to corruption in electronic processing and interchange? A transliteration of medieval Cyrillic for computing purposes was proposed in 1981,[2] and this is adequate to represent the letters used in MS Pec 68; in

[1] Edd. C. M. Sperberg-McQueen and Lou Burnard, *Guidelines for Electronic Text Encoding and Interchange*, I-II, Chicago - Oxford, 1994.

[2] '11-13 April 1981: The Hilandar Conference. Recommendations made by the Working Groups. Appendix 2', *Polata k"nigopis'naja*, 5, 1981, p. 89.

order to mark the diacritic marks, breathings and accents, which are associated with most syllables in the text, I have for the present adopted the conventions of the *Thesaurus Linguae Graecae*. The main advantages of this transcription are that it is concise and easy to use, and that it is relatively familiar or transparent to colleagues. The entity markers which TEI uses to indicate non-standard letters and other graphical features are more elaborate and would result in lengthy text files, but their over-riding merit is that they are explicit and safely transportable.

The devices which guide the eye through a medieval manuscript (and so are comparable to the use of capitals, paragraphing and spacing in modern printed works) can be represented, sometimes in rather *ad hoc* fashion, within the present TEI conventions, but may warrant the introduction of additional mark-up tags because they were such standard features in their time. For instance, rubricated headings and initials can be marked as highlighted text, but it might be preferable to have a specific rubric tag. Words, usually but not always *nomina sacra*, which regularly appear in shortened form under the *titlo* present a more delicate problem. It would be possible to indicate them with the current TEI abbreviation tag; but this is undesirable, not simply because it may not be possible to expand all of them with certainty, but rather because their shortened spelling with the *titlo* should probably be regarded as the correct one. For this reason I have used the asterisk to stand for the *titlo* and treated spellings such as 'b*6' and 'g*6' as standard; a TEI entity would once again be more explicit but more wordy. It would also be an advantage, instead of my present use of braces to mark superscript letters, to have a mark-up device[1] which could be deployed both when a consonant is written superscript under the *titlo*, e.g. 'c{s}*r6', and when a word is spelt regularly with a

[1] Lou Burnard has suggested to me that letters regularly written superscript could be marked as graphical entities, comparable to accented letters; this solution seems to reflect satisfactorily their status as a feature of orthography.

superscript unadorned by the *titlo*, e.g. 'lju{d}'; this would still leave one free to mark with the abbreviation tag the occasional clearcut instances of abbreviation, such as 'nashe{g}' at the end of a line.

The second type of problem arises from those peculiarities of MS Pec 68 which make it worth studying. This manuscript has been subjected to minute and comprehensive correction from an earlier to a later version of the psalter text[1]; but because the corrector's physical interventions were limited to what was strictly necessary, the original readings are often still visible, and where they cannot be read, knowledge of the early textual tradition will usually supply what has been erased. It is in fact possible to derive from the manuscript two related but distinct versions of the text. One straightforward way of representing them is to create two separate text files; the drawback is that, while each may be perspicuous in its own terms, the precise points at which each version coincides with or diverges from the other remain implicit. TEI offers the option of producing a composite text file, in which original and corrected readings are directly associated with each other within critical apparatus tags. Moreover, these tagged structures allow one to give details of the way (erasure, overwriting, superscript addition) in which changes have been effected and to adduce the support of other witnesses. Thus TEI affords considerable scope to the editor to include in the text file detailed information, textual and physical, about the manuscript; the problem then is to determine the optimum amount of information, to hit the appropriate balance between explicit, unambiguous detail and verbose complexity. For instance, where the original scribe wrote,'smetoshe', the corrector imposed the variant 'smuht1hshe' which was the norm in his version of the text; but because he approached the manuscript with respect and tried to make his alterations minimal

[1] C. M. MacRobert, 'Two for the Price of One: the Psalter MS Pec 68', *Oxford Slavonic Papers*, NS xxii, 1989, pp. 1-33.

and inconspicuous, he achieved this change by modifying two non-contiguous vocalic letters. One way of representing this is:

```
<app type='substantive'>
    <rdg hand='scribe' wit='S6 D B36' varSeq=1>
        <hi>s</hi>m<del rend='overwriting'>e</del>
        t<del rend='overwriting'>o</del>she</rdg>
    <rdg hand='corr' wit='B34 K M' varSeq=2>
        <hi>s</hi>m<add place='inline'>uh</add>
        t<add place='inline'>1h</add>she</rdg></app>
```

This mark-up is, I hope, reasonably explicit; but unless one has an unusually accurate and persevering editor and software capable of interpreting a file of this kind, I question whether it is more helpful than a less ambitious presentation in two complementary files:

Scribe's file:

```
<hi>s</hi>m<del>rend='overwriting'>e</del>
    t<del>rend='overwriting'>o</del>she
```

Corrector's file:

```
<hi>s</hi>muht1hshe
```

In this instance it is arguably not even necessary to mark the alterations in the corrector's file, since the resulting letter forms, here represented by 'uh' and '1h', were not employed by the original scribe.

My third type of problem is the most challenging to TEI, because it has to do with ways of marking overlapping multiple hierarchies within the text. An edition of a medieval source which aims to be useful for philological purposes has to give formal recognition to at

least two concurrent structures, that of the text and that of the manuscript. Anyone who has worked with Slavonic manuscripts knows that the division of the text into lines on the page can be crucial to the interpretation of spelling, because most scribes observed a rule that the written line should end with a vocalic letter. In a manuscript such as Pec 68 physical layout is still more important because it can signal the interventions of the corrector: his activity is betrayed, for instance, by the addition of letters in the left-hand margin before the line, or by the appearance of a letter overwritten in black (though sometimes accompanied by a surviving red diacritic) in the space reserved for rubricated letters at the beginning of the line. At the same time, however, it is essential to supply the canonical divisions of the Psalter into psalms, verses and versicles, in order to facilitate comparison with other versions of the text. In order to give an adequate account of the text in MS Pec 68 it is also necessary to mark the usual Eastern Orthodox division of the Psalter into sequences of psalms, known as *kathismata*, because sets of liturgical texts, hymns and prayers, are intercalated between groups of three (or in one instance two) kathismata.

TEI can cope with this degree of structural complexity, though it does require the encoder to give priority to one hierarchy over the other. Thus in my draft mark-up the canonical textual division of the Psalter has the status of overt hierarchy, while pages and lines are indicated by milestone tags, whose function is to mark a change from one successive portion of text to another without expressing any hierarchical relationship between them. It would be laborious but not difficult to convert a text file so constructed into a version in which the physical divisions of the manuscript were the predominant hierarchy, if this was needed, for instance, in order to reproduce the layout of the manuscript in printed form.

But this is not all. In addition to the physical organization of the manuscript and the textual hierarchy which serves the editor's needs, two other structures can be discerned which are in principle distinct from the editorial one, although they coincide with it to a considerable extent. The more interesting of the two is the scribe's conception of the text, as implied by his use of rubrication and spacing. The scribe naturally observed the Eastern Orthodox divisions of the Psalter which were standard in his time: that is, into kathismata, psalms, and versicles, but not into verses except where this was appropriate to liturgical practice, as in pss. 50 and 118 and the canticles appended to the Psalter. Moreover, his notion of the versicle did not invariably coincide with the canonical one and might even conflict with canonical verse divisions. For example, in ps. 47 one of his versicles spans canonical verse 6, versicle b, and verse 7, versicle a (as well as a page break). In TEI terms such a conflict between editorial and scribal hierarchical structures has either to be indicated implicitly, as it is in the manuscript, by the mark-up of physical devices such as rubrication, or has to be handled by means of lengthy and elaborate tagging, perhaps using the CONCUR feature which, the editors of the *Guidelines* warn, is not normally supported by SGML parsers. If the corrector's view of the text is also taken into account, the picture is likely to be further complicated, as he attempted to revise the division of versicles in the direction of the canonical pattern. So at the place in ps. 47 discussed above he superimposed a rubricated letter in mid-line at the beginning of verse 7. A mark-up which aimed to make explicit the discrepancies between editorial, scribal and corrected versicles while bypassing the points at which they coincide might look (in a simplified version) something like this:

<(ed.)seg type='verse' n='6'><!.....>

<(scribe)seg type='versicle'>

<(ed. and corr.)seg type='versicle' n='b'>

<lb n='17'><seg type='prosodic unit'><app type='substantive'>

<rdg hand='scribe'><hi>s</hi>metoshe</rdg>

<rdg hand='corr'><hi>s</hi>muht1hshe</rdg></app>

se</seg> <seg type='prosodic unit'><app type='orthographical'>

<rdg hand='scribe'>po{d}vizhashe</rdg>

<rdg hand='corr'>po{d}vizashe</rdg></app>

se</seg>

</(ed. and corr.)seg>

</(ed.)seg>

<(ed.)seg type='verse' n='7'>

<(ed. and corr.)seg type='versicle' n='a'>

<seg type='prosodic unit'><app type='substantive and structural'>

<rdg hand='scribe' wit='D B36' varSeq=1>

<del rend='erasure and overwriting'>i t</del>re-</rdg>

<rdg hand='corr' wit='S6 B34 K M' varSeq=2>

<gap reason='erasure' extent='1 letter'><add><hi>t</hi></add>

re-</rdg></app>

<pb n='25r'><lb n='1'> pet6</seg> <seg type='prosodic unit'>prije)t6 je)</seg> :

</(ed. and corr.)seg>

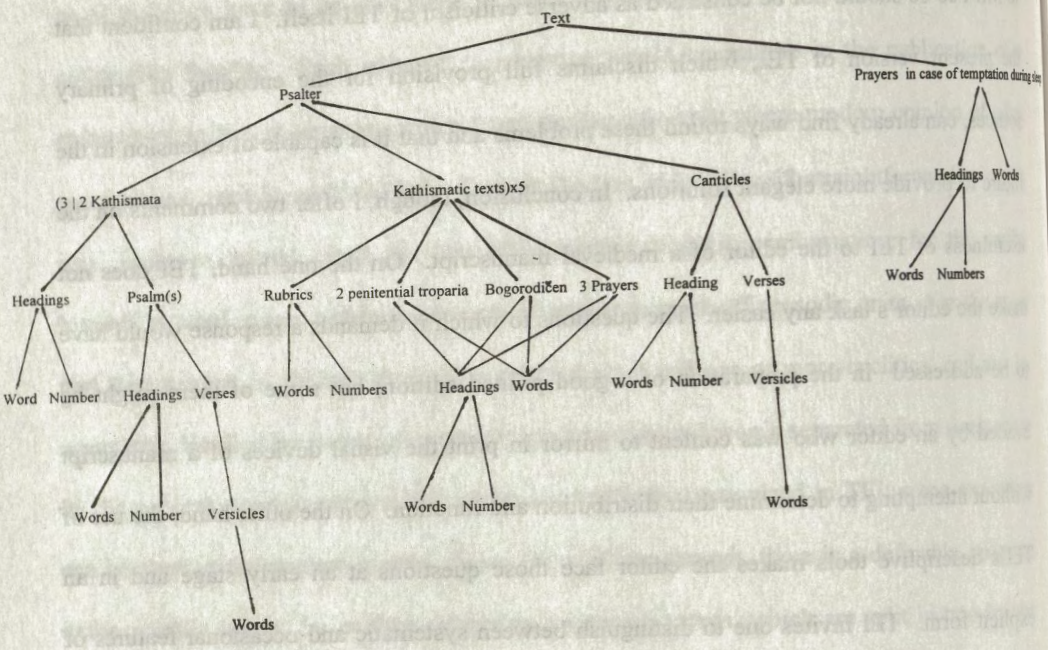
</(scribe)seg>

<!.....></(ed.)seg>

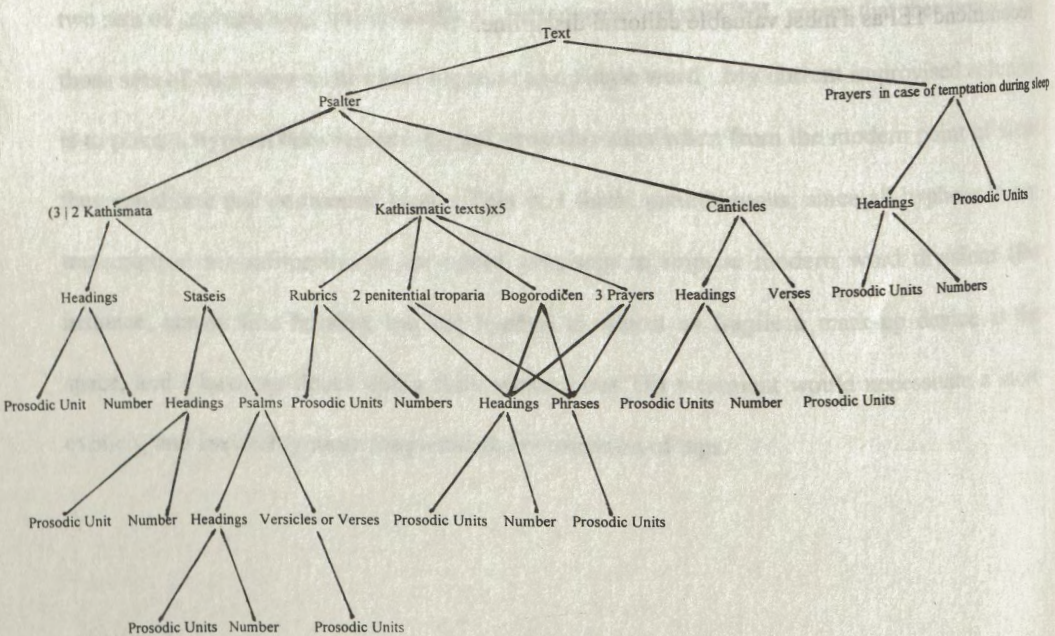
Just as canonical textual divisions are needed for purposes of reference, so modern word divisions have to be introduced to make indexing and the analysis of occurrence and collocation feasible. Such editorial intervention would be routine in the publication of a manuscript written in *scriptio continua*, and problematic only where modern opinion of what constitutes a word is itself divided. To treat the text of MS Pec 68 straightforwardly in this way, however, would entail the loss of interesting linguistic information, for the scribe himself inserted spaces within the line which appear to mark off prosodic units. For the most part they consist, in modern terms, of a stressed word and one or more enclitics, and can be adequately handled by means of segment tags placed round words separated from each other by the modern spacing conventions which are implicitly recognized in TEI; some examples can be seen in the passage quoted above. In addition, though, there is a definable range of items, usually treated by modern editors as compound words, which are split in two by the scribe, presumably because he read them with two stresses. These are difficult to mark in an economical way without introducing ambiguity, since the mere absence of a space between two sets of segment tags would hardly be sufficient to tell a SGML parser that the contents of those sets of tags were to be taken together as a single word. My current improvised solution is to place a hyphen between two tagged prosodic units when from the modern point of view they constitute one compound word. This is, I think, unambiguous, since all hyphens in my transcription are editorial and are added precisely to impose modern word divisions (for instance, across line breaks); but the hyphen is almost as fragile a mark-up device as the space, and I have no doubt that a fully worked out TEI treatment would necessitate a more explicit, and inevitably more longwinded, combination of tags.

The attention paid here to problems encountered in the application of TEI conventions to MS Pec 68 should not be construed as adverse criticism of TEI itself. I am confident that the present version of TEI, which disclaims full provision for the encoding of primary sources, can already find ways round these problems and that it is capable of extension in the future to provide more elegant solutions. In conclusion, though, I offer two comments on the usefulness of TEI to the editor of a medieval manuscript. On the one hand, TEI does not make the editor's task any easier. The questions to which it demands a response would have to be addressed in the preparation of a good printed edition; but some of them might be evaded by an editor who was content to mirror in print the visual devices of a manuscript without attempting to determine their distribution and function. On the other hand, the use of TEI's descriptive tools makes the editor face those questions at an early stage and in an explicit form. TEI invites one to distinguish between systematic and occasional features of the text, and to explain their presence as far as this is possible. I have already learnt a good deal about my text simply by attempting a TEI treatment of four pages. On this basis I would recommend TEI as a most valuable editorial discipline.

Editorial Structure of MS Pe668



Scribal Structure of MS Pe668



Processing Slavonic Bible Texts in the Russian Bible Society (Sankt-Petersburgh branch)

Irina Azarova (Sankt-Petersburgh)

In Sankt-Petersburgh branch of the Russian Bible Society the first stage of the ambitious Slavonic project draws to its close. The project aims at the critical edition of Slavonic texts according to John embracing the whole Slavonic Bible tradition. For this purpose some 1200 mss should be processed especially the ancient types of liturgical lectionaries which are supposed to be original widespread Biblical books on Slavonic soil which for sure include parts of the primary Cyrillo-Methodian translation.

On the present stage of the Project we seek for text-types (according to Colwell termninology) giving us the outline of the main Slavonic centres for the Bible translation and copying. The interlinks which indicate that various manuscripts from this group served as sources in a process of the manuscript compilation when more than one protograph was used. The text-type is associated with a number of important textual coincidences which can be measured statistically. We use Colwell's prophile metod for manuscript which was modified by Anatolii Alexeev. We are fully aware that our technique bases on several hypotheses which may be verified only later considering the output of grouping procedure. The results of some former applications of this method (Alexeev 1985, 1986; Gorina 1994) conform with tne principal textological conceptions and it may be thought to be a powerful reliable method of the textological analysis.

The input data base on the sample from the Gospel according to John (4 chapters: 13:31-18:1) read at the Holy Thursday. This passage may be unsufficient for the whole text criticism but it can prove differential for this part of the Bible. In our analysis we include mss dating from XI to XVI centuries.

The data are collected in various places of Russia as well as other countries where Slavonic Bible manuscripts are kept. On this stage the sample text of the manuscripts is compared with the corresponding text of the Ostromir Gospel to fix differences in the processed manuscript. Text spelling variations which coincide with those written in the special instruction are not kept in lists but language and orthographic peculiarities are fixed in mss descriptions. The use of computer on this stage is impossible on the full scale taking into acount the number and training of collators, but the basic stock for variation units was prepared on the computer by Natasha Gorina in her thesis. She processed the above mentioned sample from 50 mss having the utmost importance. As to the number of variation units she saw that the rate of their expanding have decreased during the process of analysis that implies that the level of the prior saturation was achived. The number of basic variation units is near 2500 (among them some 1000 individual units which are found only in one manuscript). In every variation unit all possible readings were fixed (see the sample of readings representation) from the basic stock of 50 mss. Then for other manuscripts special matrices were filled in which cells represent variation units, it is presupposed that the cell has a value equal to 1 (reading from the Ostromir Gospel), in another case the cell should be filled with the corresponding number of a reading presented in the basic stock. If a manuscript contains some alternative reading it is added at the special point of the description list. All additional readings are considered afterwards to

complete the whole stock of readings for variation units. Matrices are fed into commputer by means of a special procedure which connects the ms description with the sigla and fill the position for the variation units with appropriate values (lacunae in mss are specially marked).

At the end of this process the preparatory stage is terminated. The whole matrix comprises about 1200 lines for manuscript representation and 2500 (or more) columns for variation units, cells are filled with readings numbers. We understand that on the preparatory stage the data may only contain some mistakes. For this purpose we intend to make spot check to evaluate the number of mistakes in the representation.

On the next stage the computer processing of the original matrix will begin. For every pair of mss the ratio of nearness is computed which is equal to the number of coincidences between two mss related to the full number of variation units in the stock excluding individual readings (which can tell us nothing about the type of manuscript realation but could be very valuable in textological analysis) and lacunae in both mss. The ratio is expressed in per cent. After full computing the new square matrix is done, with the number of lines and columns to be equal to the number of processed mss. Each cell standing in (i, j) position represents the nearness ratio between i-th ms and the j-ms (the main diagonal should be empty). From this procedure it is clear that the readings are to be marked up (as individual, lacunal, etx.), for this indices are inserted in the list of readings (after the asterisk symbol) for variation units. It is important that the classification signs are clear-cut and unambiguous. We use such notation as "om" for omissions, "0" for lacunae in the texts, "a" for additions, "t" for transpositions. "u" for unica and some others concerning the linguistic level of interchange (derivational, lexical, syntactic) and some well-known processes of the text change (such structural

units as prepositions, particles and conjunctions, lexical changes, etc.). These linguistical marks can be used in various extent, for example, marks of individual reading and lacunae are obligatory, the other are optional.

Optional marks were inserted for the sake of seeking differential features in various readings. Natasha Gorina tried to differentiated readings on the basis of their contribution into the grouping distribution. She analyzed the division into groups basing on all readings and afterwards compare it with divisions obtained by various types of optional marks. The results were rather curious because those having maximal conformance with general picture were text transpositions and insertions and deletions of sintactic words, and relatively lesser conformance – in lexical changes which sjow mostly general editions. This results are not ultimate but they seem to us rather interesting because they prove that the least reasonable units can be most informative.

To make groups of mss we use several computer programs. One of them sorts the matrix replacing columns and lines in order to posit mss with the highest values of nearness more closer in the matrix dimension. The groups are lined around the main diagonal and may be picked by a schgolar. This analysis was successfully applied by Anatolii Alexeev (1985, 1986). The other procedure is the usual clustering analyses controlled by parameters which make it possible to receive more compact or more scattered groups. Natasha Gorina investigated this program on the data of 50 mss and chose the more appropriate values of parameters (they reflect the metric characteristics of the "space" of relations between mss). We think that we are to use several grouping procedures to make their results more reliable. This stage should be finished in near future and in September, 1995 in Sankt-Petersburgh the results will be discussed. We

would be glad to hear criticism (we'll try to present to all concerned colleagues the written material).

References

Alexeev 1985: Alexeev A.A. Projekt tekstologičeskogo issledovanija Kirillo-mefodievskogo perevoda Evangelija. Sovetskoje slavjanovedenije. M., Nr. 1, 1985.

Alexeev 1986: Alexeev A.A. Opyt tekstologičeskogo analiza slavijanskogo Evangelija. Palaeobulgarica X (1986), 3.

Colwell 1969: Colwell E.C. Study in Methodology in Textual Criticsm of the New Testament. Leiden: Brill

Gorina 1994: Gorina N.L. Metodika tekstologičeskogo issledowanija rukopisej (na materiale slavjansko-ruskix spiskov Evangelija XI-XV vv.). Avtoreferat dissertacii na soiskanie učenoj stepeni kandidata filol. nauk. Sankt-Peterburg. 1994.

О создании фонда письменных источников русского языка XI-XII вв. в Институте русского языка РАН

Александр Молдован

Институт русского языка РАН, Москва

РЕЗЮММЕ: Лингвистическое источниковедение в Институте русского языка РАН традиционно было ориентировано на введение в научный оборот письменных источников путем их научного издания с палеографическим аппаратом и указателем слов и форм. Условия академического института позволяли сосредоточить усилия коллективов ученных на издании крупных рукописных памятников, наиболее показательных в лингвистическом отношении и имеющих общекультурное значение (Синайский патерик, Изборник 1076 г., Успенский сборник XII—XIII вв., Выголексинский сборник, Мстиславово евангелие и др.), что способствовало определенному продвижению в исторических науках.

И все же в теоретическом отношении такая работа скорее лишь фиксировала существующий уровень знаний о рукописи, чем определяла перспективу ее изучения. Отсутствие организации планомерной работы с источниками и традиционное печатное представление результатов их изучения не позволяло сохранить единство исторической проблематики, взаимосвязанность ее элементов. Итогом этого стала разобщенность работ в области исторического языкознания, при которой дублируются наиболее трудоемкие и дорогостоящие операции, связанные с созданием вручную рабочих картотек по одним и тем же письменным памятникам (например, картотеки "Словаря древнерусского языка XI—XIV вв.", картотеки по исторической грамматике и историческому синтаксису и др.). Само нынешнее технологическое несовершенство работы по изданию рукописных текстов существенно замедляло ее и затрудняло публикацию.

С 1994 года в Институте начата работа по созданию компьютерного фонда письменных источников русского языка XI—XVII вв. В качестве основных задач определены:

— разработка методики компьютерной подготовки, систематизации и хранения текстовых данных на материале наиболее крупных неизданных текстов; изучение возможностей информационного использования текстов в их машинной форме;

– установление связей с другими научными центрами и отдельными специалистами; унификация технических приемов, определение порядка обмена готовыми материалами.

Предполагается, что работа по созданию компьютерного текстового фонда будет осуществляться по двум направлениям:

1. Введение в компьютер древнейших списков с соблюдением всех лингвистически релевантных особенностей графики и орфографии..

2. Создание банка текстов, сохраняющихся только в более поздних списках, на основе их текстологического исследования, с разночтениями основных редакций.

Источники по истории русского языка являются частью единой письменности православных народов. Поэтому работа по русским текстам и спискам должна быть интегрирована в общую международную программу переноса славянских рукописных текстов на машинные носители. В итоге должна быть создана единая информационная база для исторических и историко-филологических исследований.

TRANSCRIPTION AND EDITION

(abstract)

William R. Veder

University of Amsterdam

Transcription of Old Slavonic poses two problems which as yet have no practical solution:

(1) reduction of graphs to graphemes, and

(2) portability of machine readable manuscripts and texts. Both have a direct bearing on editions, which must be viewed as collective, not individual, enterprises. At this level, they are complicated by another problem:

(3) orthography of texts (as distinct from manuscript witnesses).

This presentation reviews several mutually exclusive solutions to these problems:

(1) tagging and commenting vs. reference to facsimile,

(2) use of programmable converters vs. standardization of text encoding,

(3) eclecticism vs. normalization.

Computer Collation of Manuscript Transcriptions

(abstract)

Michael Bakker

University of Amsterdam

This presentation demonstrates and discusses the automatic collation of computer transcriptions of manuscripts. Through data abstraction according to specified rules, significant variant readings are isolated from the huge amount of spelling variants, so that a text that is independent of the manuscripts can be constructed.

I concentrate on how the computerized manuscript transcriptions are created in practice and on the requirements they must meet. In addition, I discuss strategies for error-checking and economizing on time.

ADVANCED SEARCHING

Plamen Dimitrov Dimitrov - Sofia, Ljulin, bl.401, vh.G, ap.87, phone (359-2)254252
Irina Krasteva Bolcheva - Sofia 1680, kv. Krasno selo, bl. 223, vh.G, ap.116 phone (359-2)595592
Tanja Dontceva Ivanova - Sofia 1618, bul. "Tzar Boris III" No128, bl.11,vh.1, ap.1, phone (359-2)566153
Nina Hristova Shoileva - Sofia 1000, ul. "Khan Krum" 15 A, phone (359-2)876768

Keywords: textological methods; textual division; searching and collation; visualization of scanned manuscript; getting the misreadings; procedural algorithm

Abstract:*In the paper we want to represent the use of Advanced Searching version 2.0 for comparison of witnesses of* ***O pismenex*** *by Cernorizec Xrabr. Our program is a Microsoft Windows application that works with Collate compatible files. The main goal of this program is to search and collate parts of different texts on three levels - orthographical, morphological and lexical.*

We are a group of students from Sofia University "St. Kl. Ohridski" and from Technical University. We created the **Advanced Searching**, a program which can search and compare.

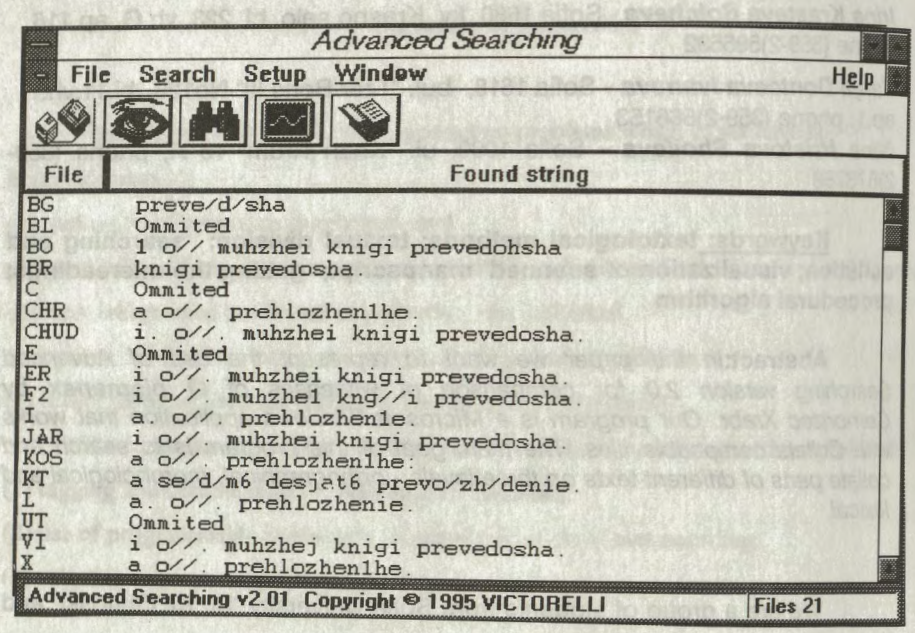
Why **Advanced Searching?** We realised that we needed a PC program, similar to **Collate®** which works only with Mackintosh. So we created the **Advanced Searching v.1.0** and it works with **Collate®** compatible files' format.

The version that we want to represent in this paper is a version **2.0** which has advanced possibilities for searching and visualization of the texts.

Generally, the **Advanced Searching** searches and compares preliminary defined parts of different texts. The coordinates of these parts are determinated by textological methods.

There are two levels of a textological text division used in **Advanced Searching**. The basic level (marked with "**U**"- Unit) denotes a paragraph from

the chosen text. The second level (marked with "**S**"- Syntagma) denotes a smaller part of the text which could be a whole sentence, a part of a sentence or just a word.



In our practice we use this program for the more precisely collation of witnesses of the tractate "**O pismenex**" by Cernorizec Xrabr.

By now the number of the working units is limited - there are 15 units only. The syntagmas are not limited in their numbers.

The searching of syntagmas and their visualization on the screen help us to find differences in the word order in the manuscripts and to discover differences in the level of lexemmas.

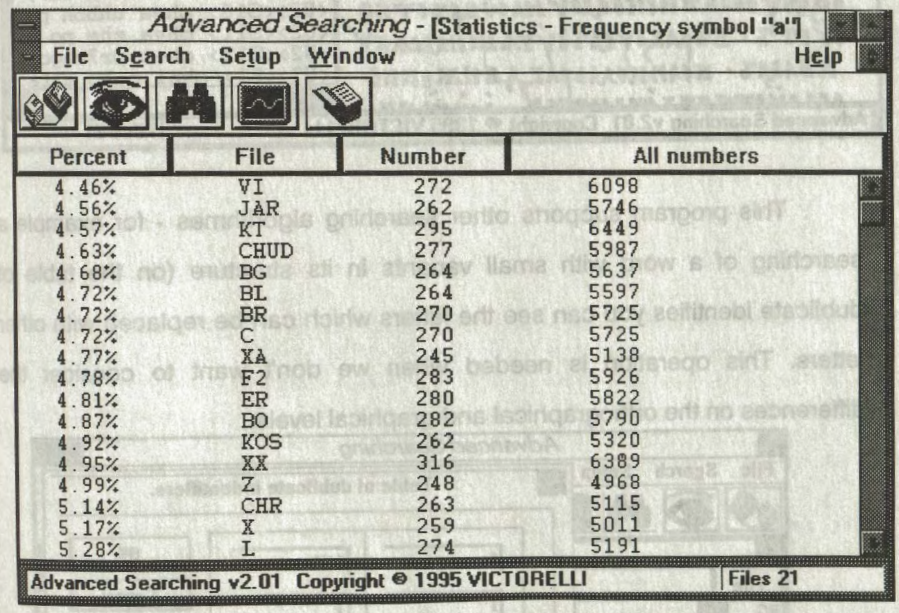
The main goal of this program is to collate parts of texts and to get the different readings (*misreadings*) on the orthographical or graphical levels.

The prosedure of the searching and collating is the following:

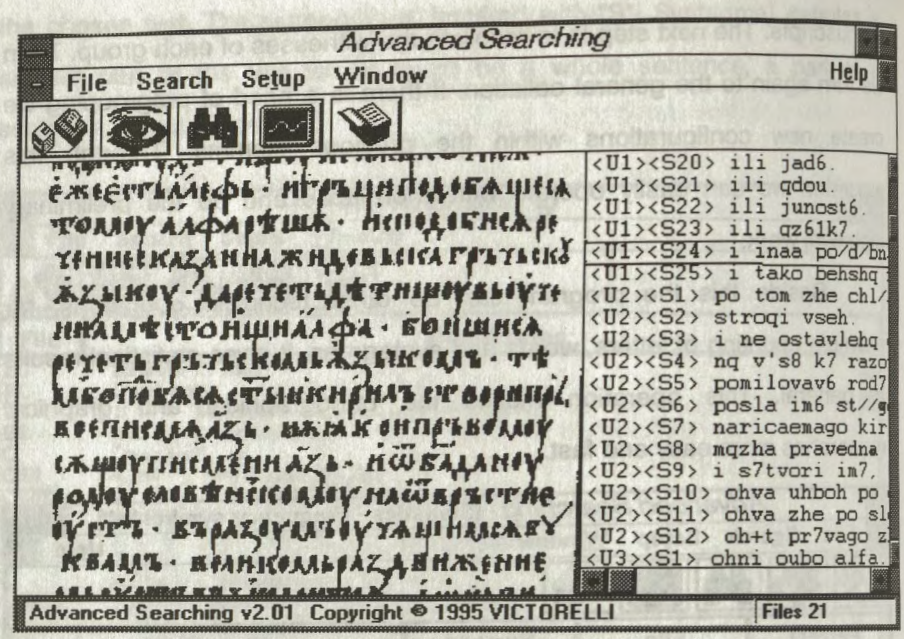
First we make a collation of the whole stemma. After getting the results we register the misreadings within the separate groups of the

manuscripts. The next step is to recollate the witnesses of each group. Then we turn again to the general collation. If there is a need of pregrouping we create new configurations within the previous stemma. Through this algorithm we can easily operate with short texts and do the preliminary technical work.

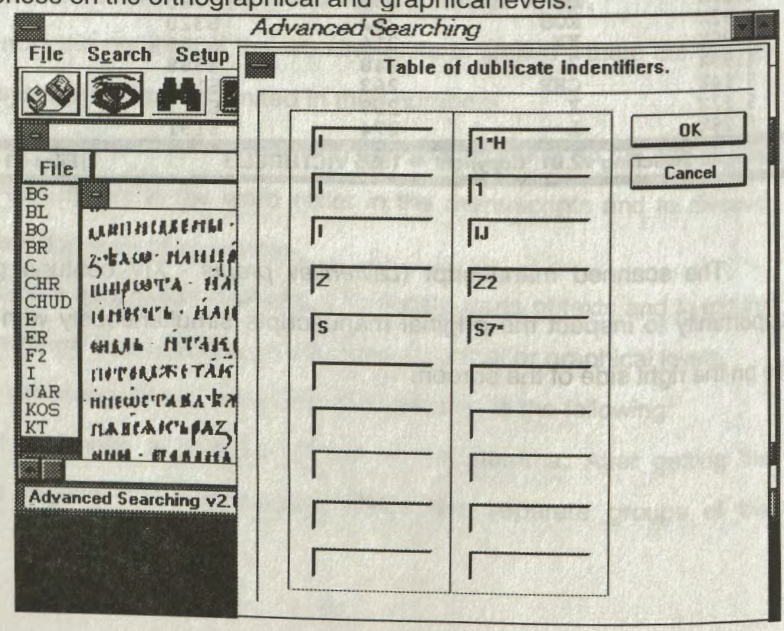
Beside this the program can do other technical operations, for example counting of letters, words and syntagmas. It gives statistical results in general. This operation makes the orthographical and graphical researches more easy and fast.



The scanned manuscript (*Lavrentiev prepis* - XIV century) gives an opportunity to inspect the original manuscripts simultaneously with the text file on the right side of the screen.



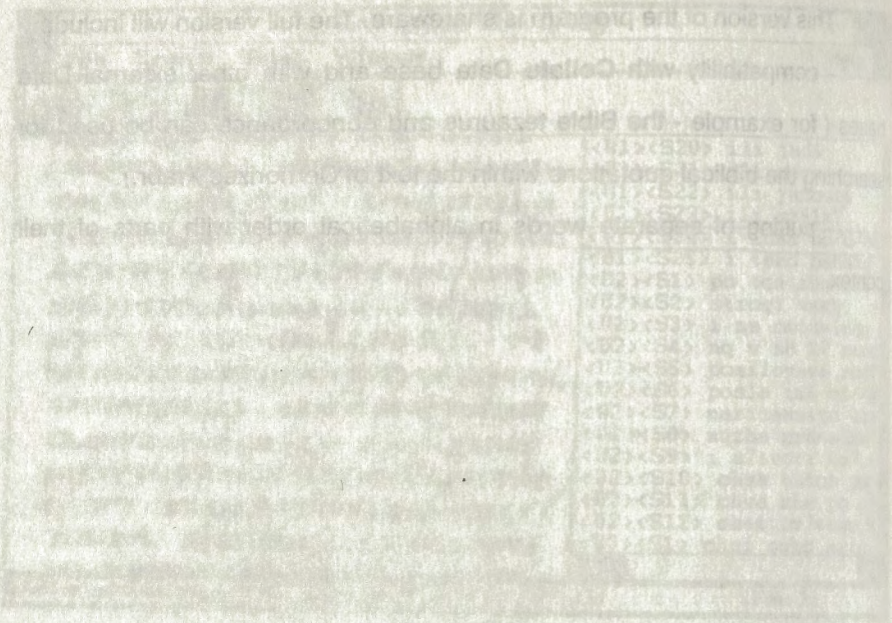
This program supports other searching algorithmes - for example a searching of a word with small variants in its structure (on the table of dublicate identifies you can see the letters which can be replaced with other letters. This operation is needed when we don't want to consider the differences on the orthographical and graphical levels.

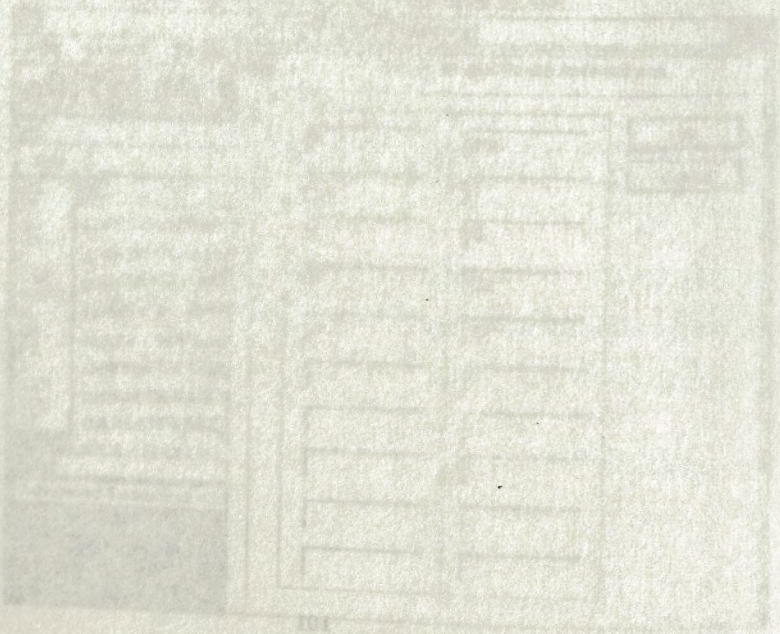


This version of the program is shareware. The full version will include:

- compatibility with **Collate** Data base and with other external Data bases (for example - the Bible tezaurus and concordance can be used for searching the biblical quotations within the text of Cernorizec Xrabr.).

- putting of separate words in alphabetical order with parts of their contex.





CHURCHSLOVENIAN ALPHABET FOR REPRINTING OLD MANUSCRIPTS USING MICROCOMPUTER

Mimoza Majstoroska,
ul Sava Kovacevic 38 II/4 91000 Skopje, Macedonia

Abstract

Advances in computer technology made possible the editing of Churchslovenian alphabet on medieval manuscripts. Geometrical analysis for each sign is made to find mathematical principles of the style. Those principles are used to form signs useful for applications that works under Microsoft Windows Operating System. A PC is suitable for above task.

According to the results of geometrical analysis an experiment has been devised for optical character recognition of churchslovenian letters. For that purpose a C-program has been developed (using C programming language). This program is in the early stage of development. Churchslovenian letters have their own feature characteristics and the program learns them. This first version produced several interesting concepts which should be developed further.

At this moment all efforts are being directed to obtain a cleaner picture of the manuscript. As the manuscripts are very old they are always "dirty" and their 'scan pictures' are dirty too. It's very important to clean them to get a good base for further processing.

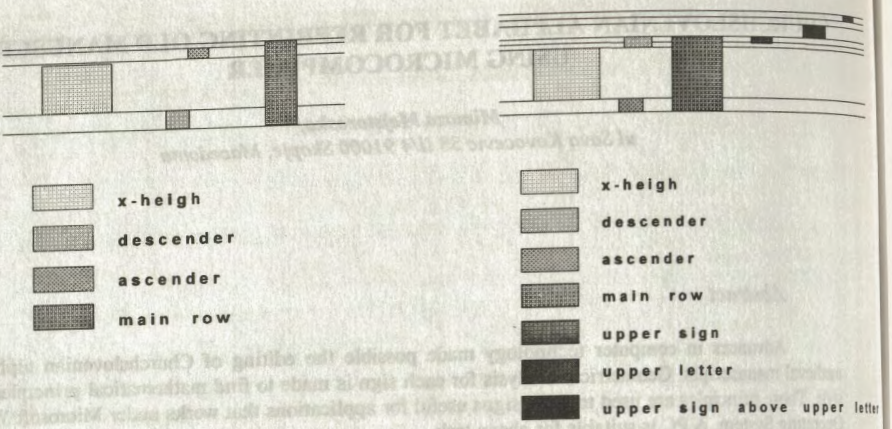
1. Introduction

The need for communication between people was present since the origin of the civilizations. At the beginning they used some signs to express their mind. The process of transformation from intellectual content into signs is called coding. The nature of the codes might be likened a grimace, a voice, a picture, a letter etc. The phenomenon of the letters produced a revolution in communication process. Further important moments in this revolution are: the occurrence of photography, printing machines, radio, television, computers and so on.

A letter or a sign is a product of an intelligent creative process, which brings an organized system of visual communication to the reader. Its development begins from pictorial forms through ideographical ones and is now accepted as a norm.

Modern computer systems use all those types of forms. The most used computer graphical environment is Microsoft Windows for personal computers. Its icons are a stylized form of ideographical or maybe pictorial forms. The Latin alphabet is used extensively at the level of computer work.

But even this alphabet is not able to satisfy some local and specialized needs. The many of the people in our region have problems with the use of Cyrillic letters in computer applications. A few applications have been carried out for use with Churchslovenian alphabet. This is a "dead", not an alphabet in active use and its structure as a composition is more complex than many others alphabets. For example English alphabet is double lined one, but churchslovenian is four lined alphabet. The following sketch shows what it means:



Sketch 1.2. A guide net for English and Churchslovenian alphabet

The facts mentioned above: "dead" alphabet and "complex" alphabet do not mean that this alphabet should be forgotten. The workers with old manuscripts ask for existing on a proper churchslovenian alphabet. The next chapter presents a sample for creation of a churchslovenian alphabet.

2. Churchslovenian alphabet creation

To create a set of signs some common characteristics of that set are needed in advance. The most important characteristics are: the ratio of the height and the weight of every sign, the line style and the internal geometry of the lines. There is always a modular ratio in the sign's proportion. Five rules in every art creation are: symmetry, modulation, harmony, rhythm and proportion. These are always present. Their empirical pointers help in the process of alphabet creation and later in recognition process. When this geometrical analysis for all signs is done it is possible to make a formula for guiding the process of drawing signs forms.

A procedure of creating letters begins with visual studying of different manuscripts written with this alphabet (to find the optimal form for the letters). Churchslovenian alphabet is handwritten, so it is necessary to analyze several different manuscripts, to find the mathematical idea in geometrical shape of each sign. It is also important to notice a composition of an old manuscript to invent its main elements and their order. An important fact is the material they were written with. The most used material was the feather, and because of that the stroke-line may vary from 1 to 4 units, where a unit is a dot. Another important fact is the shape of the letters. Some of them pass the base line below, some of them pass the x_height above. Some of them are thin and tall, others are too wide and so on.

To discover churchslovenian common characteristics different alphabets (from period of 12^{th}-13^{th} century) are processed and the letters from these different alphabets are compared to find the optimal shape. This shape is written in mathematical notation what means some formulas are defined and with them each sign is reconstructed. In the above

algorithms can be developed for the drawing and correcting of all letters. The following algorithm is an example for defining a letter from churchslovenian language:

```
moveto (x11, y11);
arc (x11, y11, stangle11, endangle11, r11);
arc (x12, y12, stangle12, endangle12, r12);
arc (x13, y13, stangle13, endangle13, r13);
arc (x14, y14, stangle14, endangle14, r14);
moveto (x21, y21);
lineto (x22, y22);
lineto (x23, y23);
lineto (x24, y24);
lineto (x21, y21);
```

The above is an algorithm for the drawing of the letter O. For this purpose it's necessary to estimate all the parameters. It is not possible to use those defined signs in standard programs. To make them useful in applications like WINDOWS further transformations with standard programs are necessary At the end it results in an alphabet as it is shown in APPENDIX *A*.

3. Manuscript processing for optical character recognition

Although churchslovenian manuscripts are not in great number and it seems there will not increase in the future. They are enough, they require attention from Slavic specialists and other professionals. As many other professions ask for computer support, churchslovenian researchers ask for this support too. Their needs and problems are specific, so an appropriate way should be found to resolve them. There are many difficulties in churchslovenian manuscripts processing. The following are the main ones:

1. Churchslovenian alphabet consist of a large number of signs (characters) at least double or more number than the English alphabet.
2. There are many upper letters, upper signs and shortcuts above the base row. They are small, but they produce a confusion in the computer's "mind".
3. There are many initials - large decorated letters and other decorations. These produce additional problems also.
4. Because they are very old, churchslovenian manuscripts are sometimes ruined, they are very often "dirty" and fuzzy. **In the next chapter a method to get clearer picture of manuscripts as a better base for their processing is presented.** It's not only important for the recognition process, but it's useful for the storage of the documents too.
5. Churchslovenian manuscripts are handwritten ones, they have a variety of stroke widths, orientations, sizes, positions and shapes.
6. Churchslovenian characters are more complex than today's Cyrillic's letters.

The above specifics make impossible to use some standard processing programs which are designed mostly for the English alphabet. They are not very adaptable even for the Cyrillic alphabet, further, for manuscript processing they are not useful at all.

So a C-program concept is made as a working base with old manuscripts. The main point in this program is to use modularity rules in tandem with geometrical analysis of churchslovenian alphabet.

Start point of this program is a scan picture of the manuscript. Here will be given the preprocessing of the picture. When a picture appears on the monitor it is written in a data base in a new way with code appropriate for further elaboration.

At first the picture should be subdivided. The computer can divide it by horizontal (row) scanning. Scanning the picture, when a row without "black" pixels appears it means it is the end of a division. This condition of "no black pixels" should not be strictly respected. There always some unexpected "black" pixels, but they are not in position to form a text row. They usually make a fuzziness in the picture. So the condition "no black pixel" is substitute with some x-value toleration. The size of this x-value is such that it cannot form a matrix row of written text. After studies for sizes and shapes of the alphabet's signs and experimental tests the x-value for horizontal rows is considered to be one module.

After the separation of a whole, it should be as the following:

- a whole is a base row
- a whole is an upper row
- a whole is upper sign
- a whole are incidentally connected rows (maybe because of the initial letter)

Studying the vertical size of the whole and using the modulation rules it is possible to find out what case is it.

First, second and third case are OK and there no problem for processing to go on. The last case needs an additional step to remove the initial letter. Because the initial letters are always in the beginning of the row using contour following method the first object in the row is bounded and then removed. Now the cursor goes back again at the start position of the whole.

After the separation of a whole, processing changes the direction from horizontal to vertical and now scans the whole by columns, to separate each sign in the row.

The concept for recognition is based on the results achieved by geometrical analysis. Relevant factors in this recognition are:

- matrix width,
- matrix height,
- procentual matrix black and white ratio
- appearance of black pixels in some characteristic part of the matrix.

Combining several of the above pointers, an unknown sign passing through several filters is recognized as one specific character.

This program is testing both the original manuscripts and the texts written by created alphabet. In both cases this algorithm will give satisfactory results in separation and extracting the signs, but in the process of recognition the results can vary considerably. Written texts are recognized very well, but original manuscripts can't be recognized satisfactory. There are many additional things should be done helpful to the main aim:

- making a vocabulary on churchslovenian words
- making a library of the most often using shortcuts
- develop symmetry method
- using some methods for picture processing to get better quality of scan pictures.

At times the idea for picture processing to recognize each churchslovenian sign sounds not very encouraging, but if the Chinese have their courage to recognize their

letters which number is at least over 3000 at first hand and over 7000 at second hand, than this idea is not so hopeless.

Also a new method based on the neural networks is under development with the same aim.

The most encouraging thing in this work is that the number of words in churchslovenian language is not so huge as in other active languages also, it is not so difficult to make a churchslovenian vocabulary. Many OCR programs use vocabulary for better recognition.

4. Cleaning the picture

One of the main things in work with old papers is problem of preprocessing the scan picture of old manuscripts. It's very important to store in documentation form a "clear" picture not for old papers only, but for other important subject which is not obtainable in it's original form.

At first, several different methods of scanning the picture (photography) should be tested.

- Scanning the picture by black and white method
- Scanning the picture in 16 scale gray
- Scanning the picture in 256 scale gray
- Scanning the picture in color

Comparing the result pictures mentioned above some differences are noted among them as it is shown in APPENDIX *B*.

The first method black and white is the fastest one, but not very precise. Sometimes it's useful during the working process, because it's fast and it gives the most evident review of the manuscript.

Second one is a little more precise than the first, but not enough for serious work. So it's not recommended for universal work. It can be used only for printed texts, or clear, not ruined manuscripts.

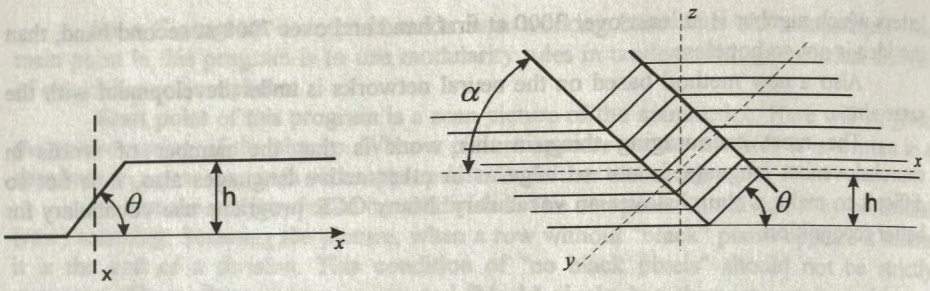
The third one is the most interesting one. It gives the best base and the most opportunities for further work. Some directions for picture processing will be shown here.

Color scanning generally is the most complex method, but at the end, by nature we want to get a gray scale picture, or a black and white one.

Gray scale scanning gives a large field for additional intervention in the picture to achieve an image most convenient for main processing.

One of simple methods is cutting some of the threshold value. In that way a value can be chosen lighter or darker from gray scale size until a satisfied result and picture is achieved.

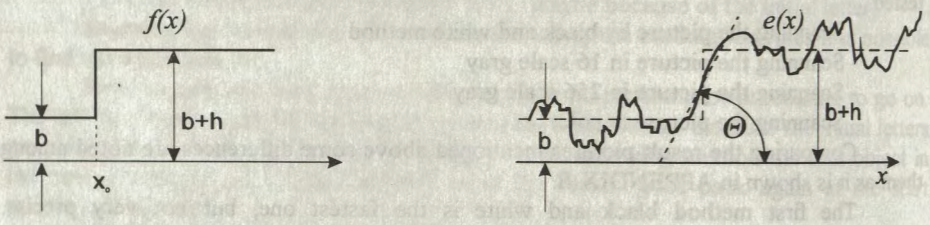
This method is based on the fact that an edge is local discontinuity in image luminance or amplitude level. One- and two-dimensional edges are represented as a ramp increase in image amplitude level from a low to a high level:



Sketch 3.4. One-dimensional and two-dimensional edge

In the one-dimensional case the edge is characterized by its height, the slope angle, and x coordinate of the slope midpoint. An edge exists if both the slope angle and height are larger than specified critical values. For the two-dimensional example the orientation with respect to the x-axis is also of importance.

An ideal edge is represented in the first part of the following sketch. But in real no edge has this shape. A real edge has the shape as the second part in the sketch.



Sketch 5.6. An ideal and a real edge

An ideal edge detector processing the image regions should produce an edge indication localized to a single pixel located at the midpoint of the slope. A threshold operation is performed to determine the pixel location of significant edges. A negative going edge exists if:

$$G(j,k) < T_L\ (j,k)$$

and a positive going edge exists if:

$$G(j,k) > T_U\ (j,k)$$

where T_L (j,k) and T_U (j,k) are lower and upper threshold values, respectively. These threshold values may be made spatially varying to compensate for gross spatial luminance changes. Threshold selection is one of the key issues in edge detection. A threshold level set too high will not permit detection of low-amplitude structural image elements. Conversely, a threshold level set too low will cause noise to be falsely detected as an image edge. An edge location map E(j,k) is often generated to indicate the positions of edges within an image. For example, ali positive edge locations could be indicated by

white pixel values against a black background. Alternatively, positive going edges could be noted by white pixels, negative going edges by black pixels, and nonedge regions by midgray pixels.

APPENDIX *C* shows some test result pictures for different threshold values.

5. Conclusion

The realization on this theme needs help and support from different professions:

- Researching part of the work, some analyses are under the field of geometry and art, as the churchslovenian alphabet is an art letter.
- Making this alphabet in printing font is in range of typography and graphic profession.
- Because this is a "dead" letter there are need from paleography and scientists for churchslovenian alphabet
- The aim this alphabet to become a computer font it is under the field of computer specialists.

Each of these professions has its own rules. Sometimes is difficult to coordinate all of them in the same time. Only a friendly dialogue in the spirit of tolerance will bring a positive result.

This seminar I hope will help in that direction.

6. Literature

[1] Dr Popić R., dipl.ing. Lolić B, prof Dr Afgan N., *Naučno - tehnički rečnik englesko - srpskohrvatski,* Privredni pregled - Beograd, 1983

[2] Höcker O., *Ukrasno pismo*, Školska knjiga - Zagreb, 1965

[3] Mesaroš F., *Tipografsko oblikovanje*, Viša grafička škola - Zagreb, 1981

[4] Жиљак В., *Сшолно издавашшво*, Дечје новине - Горњи Милановац, 1989

[5] Pijuković N., *Priručnik za ručne slovoslagače*, Grafički Školski centar - Beograd, 1973

[6] Đorđić P., *Istorija ćirilske paleografije*, BIGZ Beograd, 1975

[7] Mr. Vojčić V., *Uvod u C jezik*, Institut za Nuklearne Nauke - Boris Kidrič - Vinča, 1988

[8] Zurovac M., *Umetnost kao istina i laž bića*, Budućnost, Novi Sad, 1986

[9] Алтбауер М., *Синајски Псалшир*, Македонска Академија на науките и уметностите - Скопје, 1971

[10] Klages Ludwig, *O rukopisu*, štamp."Ognjen Prica" Zagreb, 1968

[11] Harvey M., *Lettering Design*, Bodley Head GB, 1975

[12] M. Cohen, *The art of writing*, print in FRGermany, Paris UNESCO, 1965

[13] Publisher's Type Foundry *User's Manual,* ZSoft Corporation, 1987

[14] *WFN Type Guide* Corel Systems Corporation, 1990

[15] *Corel Draw User's Guide,* Corel Systems Corporation, 1990

[16] Jakupović A., Milović R., *AUTO CAD,* Tehnička knjiga, Beograd, 1990

[17] Druet R., Herman G., *La civilisation de l' ecriture,* Librariric Artheme, Fayard dessain et Tolra, 1976

[18] Shi-Kuo C., *Principles of Pictorial Information System Design*, Prentice Hall Inc., New Jersey, 1989

[19] Ballard D., Brown C., *Computer Vision,* Prentice Hall Inc., New Jersey, 1982

[20] Harrington S., *Computer Graphics A Programming Approach*, McGraw-Hill Int. Book Comp., Singapore, 1983

[21] Schildt H., *Turbo C: The Complete Reference*, Mahomet, Illinois, 1987

[22] Kernigham B., Ritchie D., *The C Programming Language,* Prentice Hall Inc., New Jersey, 1978

[23] Spisanie *Computer,* IEEE Computr Society, Los Alamitos, july 1992

[24] Таранов, Н.Н., *Рукописни шрифт*, Виша школа, Лавов 1986

[25] Spisanie *Byte,* april, 1991

[26] Petrović Đ., *Kompozicija arhitektonskog oblikovanja,* Naučna knjiga, Beograd, 1972

[27] Tehnical Reference for Laser Jet III

[28] Bojković Z.S., *Digitalna obrada Slike*, Naučna knjiga - Beograd, 1989

[29] Mesaroš F., *Grafička enciklopedija,* Tehnička knjiga, Zagreb, 1971

[30] Десподова В., Славева Л., *Македонски средновековни ракойиси,* Институт за истражување на старословенската култура, Прилеп, 1988

[31] Угринова Р., *Учебник йо сйарословенски јазик,*

[32] Левочкин И.В., Инициjали XI-XVI веков, Книга, Москва, 1983

[33] *Word for Windows*, Sarajevo

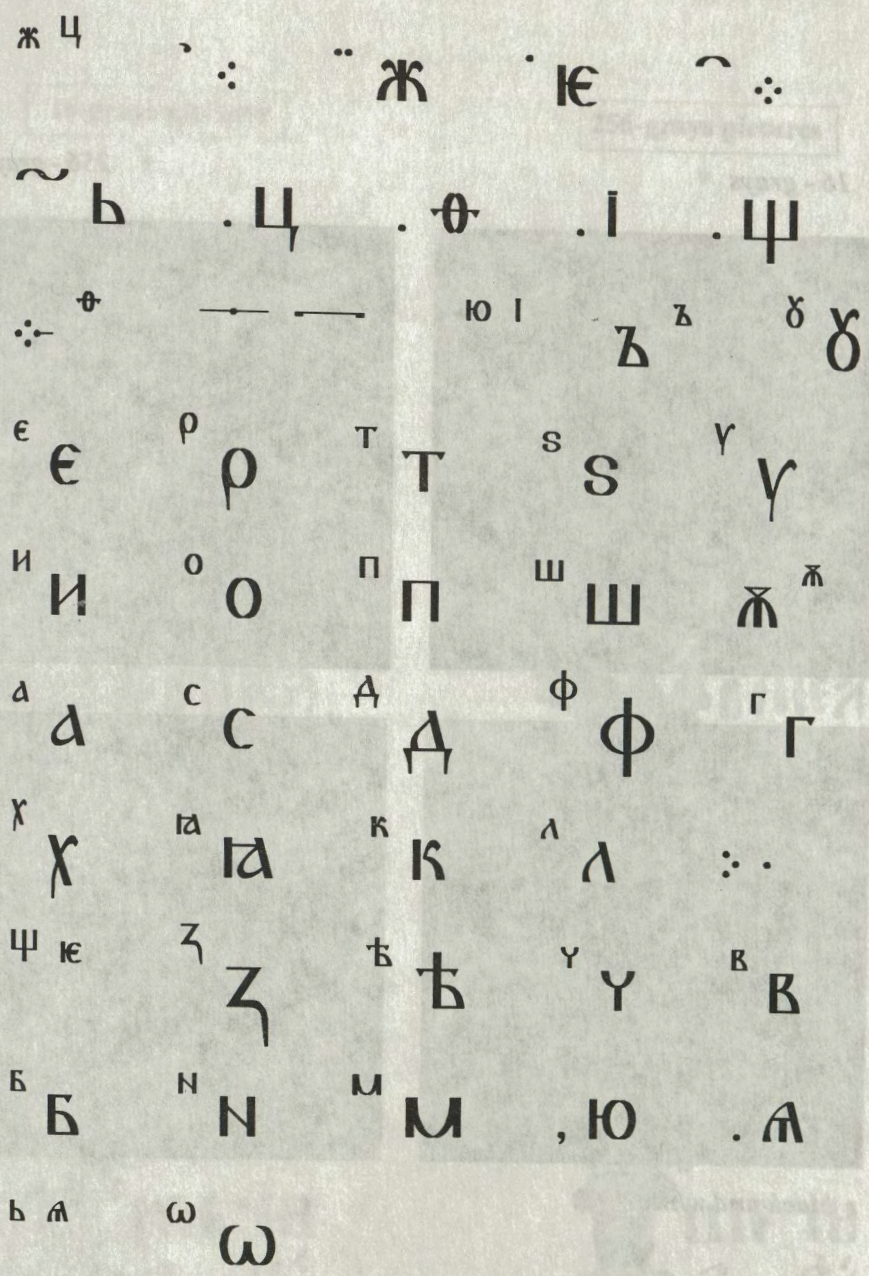
[34] Rumelhart D.E., McClelland J.L., *Parallel distribuited processing: Explorations in the microstructure of cognition,* Cambridge, Press/Bradford Books, 1986

[35] Rumelhart D.E., McClelland J.L., Explorations in *Parallel distribuited processing: A handbook of Models,* Cambridge, Press/Bradford Books, 1988

[36] Chen-Tsun Chuang, Lin Yu Tseng, *A Heuristic Algorithm for the Recognition of Printed Chinese Characters,* IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, VIL 25. NO4, APRIL 1995

[37] William K. Pratt *Digital Image Processing*, John Wiley & Sons, Inc. 1978

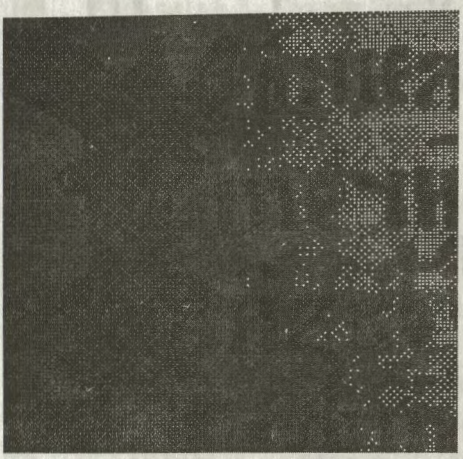
CHURCHSLOVENIAN CODE NUMBERS



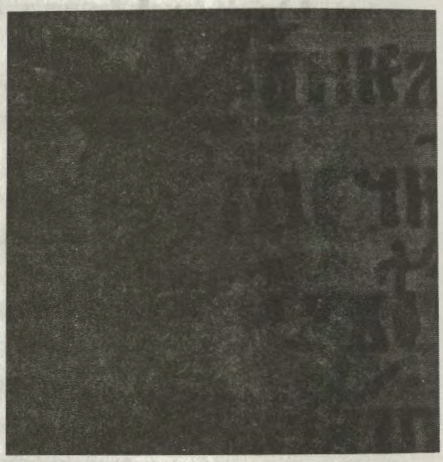
APPENDIX *A*

DIFFERENT SCAN METHODS

16 - grays



256 - grays





black and white



color

PROCESSING THE PICTURE WITH DIFFERENT THRESHOLD VALUES

16-grays pictures



threshold value = 120



threshold value = 100



threshold value = 90

256-grays pictures



threshold value = 100



threshold value = 90



threshold value = 80

Ljupčo Mitrevski
Institute of ancient Slav culture - Skopje, Macedonia

FUND, FONTS AND CHARACTER SET OF CHURCH SLAVIC GRAPHEMES

Working on the Church Slavic texts, the application of modern computer techniques is needed for more complete, more exact and more large research and presentation the medieval Slavic manuscripts of the most different aspects of exploration.

Certainly, the most interesting for us is the linguistics aspect of research and study of the Church Slavic texts. Because of specific needs in that researches, since 10 years was actual one new manner of research in our scope. In essence, at that time the degree of the development of computers and software, that we could use, enabled the processing of the texts in one higher degree. So, it is important to say that with this manner of application of computers, the science approach and methodology of work were not disturbed.

From one solid creation of the Church Slavic alphabet for use in the computers, we succeeded to create a small additions that enabled phonetics-phonology and morphology analyses of the text. In the same way, the making of phrases of the inputted text was computerized on the whole.

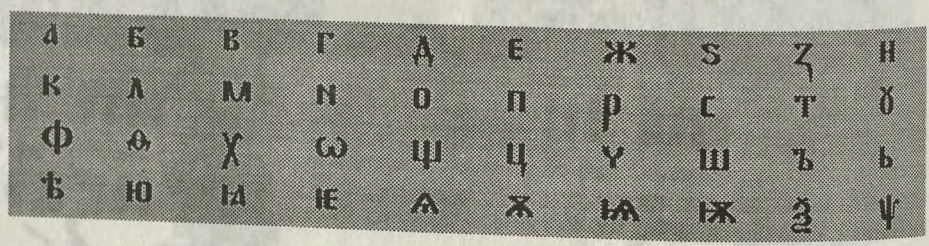
After creation Church Slavic alphabet (font), we are puting the Church Slavic text we need, in its readable form. Than, we are making the possible and needed linguistics analyzes in this text, we make a glossary at, and the end, the text is printed.

All that is said up is related by the more old versions of computer programs. The new, more modern versions of the same programs, change the manners of solving the problems of use the computers in processing churchslavic texts.

Because of, we will stay to the preliminary solutions related with our problems and the application of computers. More concretely, we will stay at the start of our needs - to create of necessary Church Slavonic fonts, their disposition on keyboard and the manner of its use in the following process of analysis the texts. All these phases are linked with nowadays the most spreaded programs in the world working at "Windows" environment.

NECESSARY FUND OF CHURCHSLOVENIAN GRAPHEMES AND SIGNS

It's known that we need more than 50 Church Slavic or graphemes.



The basic computer keyboard almost allows this fund of graphemes to take its place on the keys.

But, when this main fund accepts some variant forms of some graphemes (ѡ, ꙁ, ѳ, ѧ, ꙗ, ѕ, ї), some the most often used ligatures (ꙋ, ѹ, ҁ, ҁ, ҥ) several graphemes from Greek alphabet whose also very often take a part in churchslovenian text, so and a great number of punctuation signs (. · ⁘ ⁖ : ҃ ҄ ҅ ҆) and different types of titlas (◌̾ ◌̇ ◌̑ ◌̀ ◌́ ◌̃ ◌̇ ◌̑ ◌̀ ◌́ ◌̃) the keyboard becomes too small to keep all those signs.

This problem becomes more complex with the need for use of this whole mentioned fund as superlinear (over-letter) and supersuperlinear (over over-letter). For example: ро҆у҃нъ ҉с҃.

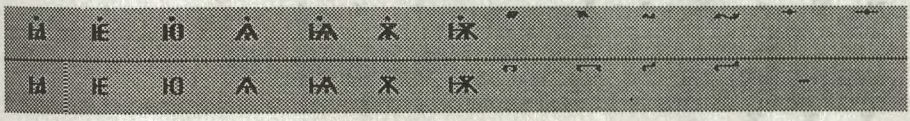
So, it becomes impossible to think for a complete solution based on one font, but we had to find another multidimensional solution for Church Slavic font.

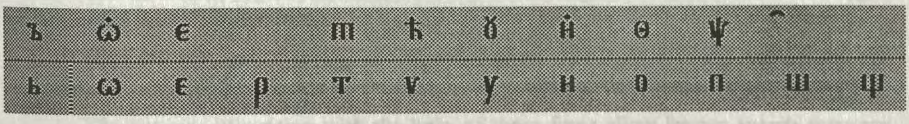
FONTS OF CHURCHSLOVENIAN ALPHABET

After several combination were made, our final solution we found in creation of two fonts with main aim to take place for all churchslovenian fund of graphemes and other signs. In the same time we take care for further use in text processing.

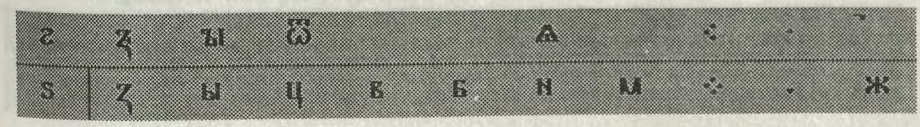
1. MAIN CHURCHSLOVENIAN FONT ***(STARSLO)***

This font consists all letters from churchslovenian alphabet. These letters take place in the base keyboard. Among them we put several alternative forms of some graphemes, graphemes from Greek alphabet, punctuation signs, so and upper signs (titlas, paerocks, etc.). After that, on the keyboard there are left about 10 keys unused. These keys can be used with other graphemes or their alternative forms, depend on the needs.









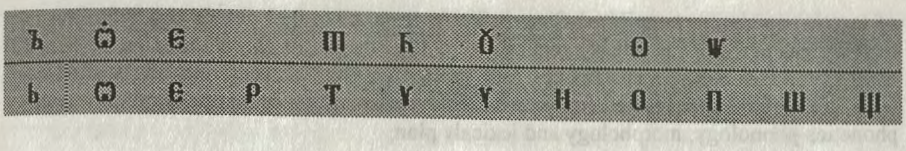
Alternative codes we used to put superlinear letters, signs and several the most used ligatures. Upper letters appear in both forms as separate letters: **the form of the normal letter and the form of the same letter with a title above it.**

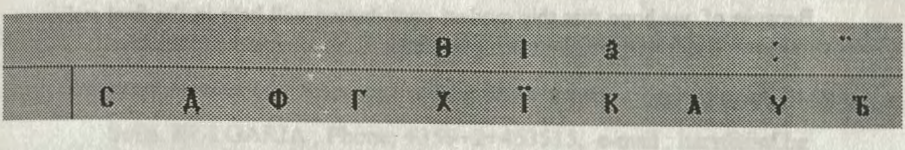
0131	0132	0133	0134	0135	0136	0137	0138	0139	0140	0141	0142	0143	0144	0145	0146
ѧ	ѧ̑	ѣ	ѣ̑	ѥ	ѥ̑	ѓ	ѓ̑	ѧ	ѧ̑	є	є̑	ж	ж̑	ҙ	ҙ̑
0147	0148	0149	0150	0151	0152	0153	0154	0155	0156	0157	0158	0159	0160	0161	0162
		ѕ	ѕ̑			н	н̑	і	і̑			к		ѧ	ѧ̑
0163	0164	0165	0166	0167	0168	0169	0170	0171	0172	0173	0174	0175	0176	0177	0178
м	м̑	н	н̑	о	о̑	ѡ	ѡ̑			п	п̑	р	р̑	с	с̑
0179	0180	0181	0182	0183	0184	0185	0186	0187	0188	0189	0190	0191	0192	0193	0194
т	т̑				ѹ	ѹ	ѹ	ф	ф̑	х	х̑	ц	ц̑	ч	ч̑
0195	0196	0197	0198	0199	0200	0201	0202	0203	0204	0205	0206	0207	0208	0209	0210
		ш	ш̑	щ	щ̑	ъ	ъ̑	ѣ	ѣ̑			ю	ю̑	ѧ	ѧ̑
0211	0212	0213	0214	0215	0216	0217	0218	0219	0220	0221	0222	0223	0224	0225	0226
ѩ	ѩ̑	ѫ	ѫ̑	ѭ	ѭ̑			ѣ	ѣ̑	ѣ	ѣ̑	ы	ы̑	ъ	ъ̑
0227	0228	0229	0230	0231	0232	0233	0234	0235	0236	0237	0238	0239	0240	0241	0242
ъ	ъ̑	ѳ	ѳ̑	э	э̑	.	,	⁘	⁖	⁝	–		ѧ	ѧ̑	
0243	0244	0245	0246	0247	0248	0249	0250	0251	0252	0253	0254	0255			
ҍ	ҍ	ҥ	э												

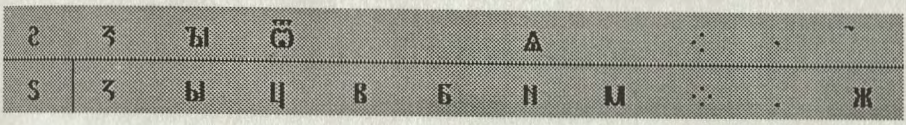
2. CHURCHSLOVENIAN FONT OF CAPITAL LETTERS (STARBIG)

When a preliminary solution for base Church Slavic alphabet (font) was made, very easy we made a font for capital Church Slavic letters (Starbig). The keyboard order in this font (Starbig) is the same with the keyboard order for base Church Slavic alphabet (Starslo). The difference between these two fonts is in the letter size and their style appropriate to the capital churchslovenian letters.









We know that the capital letters in the medieval Slavic manuscripts are written in that manner that with the small letters the common thing is the top settlement, in regard to, all letters (capital and small) hang to top settlement line. Because of, for us, as more logically was to put on the solution to write n its whole computer use:

Гѣ Въωно врѣмѧ

Font **Starbig** all another signs can use of the base Church Slavic font Starslo (punctuation, titles, etc.)

Font **Bigstar** in essence is the same font Starbig, and only difference among them is in the level. Font **Bigslo** has common base line with font Starslo:

Гѣ Въ ωно врѣмѧ

Diacritical marks and diacritical letters are at the higher level, because of the bigger altitude of the letters of this font.

We think that font Starbig, for difference of the font Bigstar, is more acceptable for us, because it is the real reflection of the manner of writing of the churchslavonic texts.

The process of writing with the help of these fonts allow bigger speed, accuracy and credibility. The letters of the base alphabet we can write directly on keyboard. That is more practically to write the diacritical signs and the diacritical letters (with the alteration), than to write this letters and signs with lifting for one level more. The diacritical signs and letters we write straight after the letter on which the diacritical signs and letters we need to write. That manner is conforming at the order of writing and reading the words.

Begining of that, our propositions for unification of Church Slavic letter in the computers, is reflectection of our needs and the computer possibilities. We think that in this manner, with the help of the two fonts, we will get better and more practical manner of graphics presentation of Church Slavic texts. So, the form of graphems and tehniques form

of text is too similar as original Church Slavic texts. The next use of text written at this manner, enables more practicaly realization of the analysis of the churchslavonic texts by computers. This solution enables not only credibl graphics presentation of churchslavonic texts, but with that are created conditions for the next working on the texts of more aspects: phonetics-phonology, morphology and lexicaly plan.

Because of, we hope that this modest contribution will help in the final realization of unification and codification of Shurch Slavic alphabet in computer processing.

OPTICAL CHARACTER RECOGNITION OF HISTORICAL DOCUMENTS

STOYANOV Ivelin Peev
Bulgarian Academy of Science, Institute of Information Technologies, bl.29A
Sofia, BULGARIA , Phone: +359+2+453515, e-mail : hary@iinf.bg

Keywords : character recognition, Fourier transform, neural networks, document reading

ABSTRACT

The proposed approach is a part of a global project for a historic dictionary of the Bulgarian language. The required efficiency of old historic documents researches leads to the necessity of an automation of document entering. The graphical image of a document is captured by a grey scale scanner. The symbols, which the system has to recognise are separated from the background by the means of an adaptive thresholding method, which neutralises the negative influence of some spots - from ink, etc. A heuristic algorithm segments the image to single symbols (characters). The Fourier transform is used for character's representation. The Backpropagation Neural Network (**BPNN**) is used as a classifier. It has as an input, the Fourier transformed shape of the unknown symbol. At first, the **BPNN** is learning to recognise some set of symbols. As the reading process is hold, the interactive learning and self-learning of the classifier are carried out. The post-processing includes recognised characters representation in the same view, as that in the graphic image. The achieved correct reading is about 90-95%, depending on the document quality.

INTRODUCTION

The present paper introduces You to an Optical Character Recognition (**OCR**) System that is being in process of developing, specially designed for the aims of the automatic historic documents reading. It is a part of a Global Project for a historic Dictionary of the Bulgarian Language, developed by specialists from the Institute of Information Technologies, the

Institute for Bulgarian Language and the Institute for Bulgarian Literature, all from the Bulgarian Academy of Science. The Project is opened for other participants as well.

As we know, the **OCR** is an old problem, which is solved for some particular cases, as typewritings and printings reading, where the graphic image of the characters - their shapes - are previously defined. There are some other **OCR** systems that could be learned to recognise some non-standard font, such as Omnipage, ProLector, Kurzweil-5200. All the above mentioned systems in the case of Historical Documents have some disadvantages, and therefore it is difficult they to be applied in such projects. The problems are as follows [DOBREVA]:

- the medieval documents consist of characters, which are constructed of many additional symbols, as accent, tilde, not defined in classical **OCR's** ;

- these diascript, superscript, subscript symbols form an alphabet with more letters than the usual ones;

- multilingual and multialphabetical environment;

- they are not opened systems and cannot be included in automatic systems.

The above problems, as well as the research interest for new methods application are the motivation for the current **OCR** developing.

An **OCR** System consists of the following subsystems: Image acquiring, Preprocessing, Classification, Postprocessing, which will be discussed in the rest of the chapters.

I. IMAGE ACQUIRING

The document for an automatic reading is scanned by a halftone scanner with a resolution of **200 dpi**. The results of the researches demonstrate, that such a scanned and visualised on the screen image is readable for the human eyes, therefore it has to be enough for an automatic reading. (Fig.1) The halftone scanning is chosen, because the faded documents,

as well as the other hard readable ones, require special preprocessing, which have the best results on a halftone image.

The other methods of Image acquiring are also available, as the system could work with the most widely used method of graphic representation - the **TIFF** standard. So, a TV-camera, with a framegrabber and a special software could capture the image of some document in order to be saved in the **TIFF**-format. In this way the **OCR** system could read it and deal with it. The using of the **TV**-camera makes possible the application of some special optical filters, which, for instance, reject the visible light and pass the ultraviolet one. Such ultraviolet rays could extract information, invisible for our eyes, and in the case of faded ink, we or the computer system could restore the document.

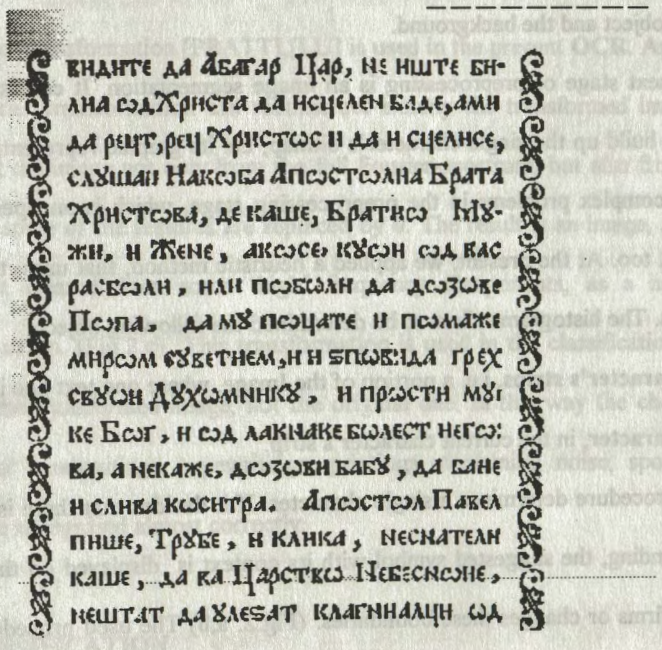


Fig.1 A Page from {Abagar} printed in 16-th century

II. PREPROCESSING

The next important stage is the preprocessing. There are some steps in the preprocessing.

The first one is the image improvement. It includes smoothing, spot clearing and so on.

The second one is the thresholding or separating which is a process of seeking a threshold - limit of grey-level that divides the character's points from the background's points. Usually the 'black' points with grey-level near to **0** belong to the characters and the 'white' points with grey-levels near to **255** belong to the background. So, to define this threshold we use an adaptive thresholding method, developed by [SHAPIRO]. It uses a derivative where a bigger two-dimensional derivative in some point suggests a bigger probability of a limit between the object and the background.

The next stage of preprocessing is an image segmentation. It consists of determining which points build up the single character's bitmap, i.e. its graphic representation. This is one of the most complex problems in the preprocessing stage, which is an open question in the present **OCR** too. At the present, we applied a heuristic method, that uses the histograms by X and Y axes. The histograms allow to be determined the following objects:

1. **character's strips**, i.e. a portion of the Image, where one text line is located,
2. **character**, in the current character's strip.

The procedure determines a single character. If it decides that there is a possibility for incorrect bounding, the suggested symbol with its context is displayed on the screen and the operator confirms or changes these boundaries. (Fig.2. a,b) The used procedure is one of the easier, but not the most efficient way for segmentation, especially for the case of ancient documents, where the characters are not full-connected (diascript, superscript, subscript symbols).

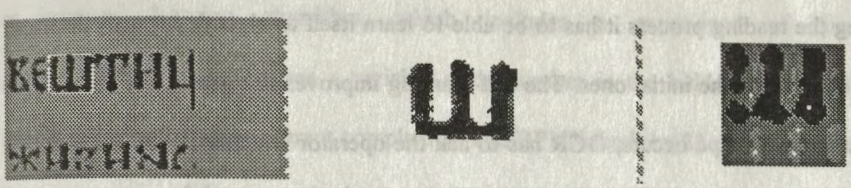


Fig. 2 The Preprocessings:

(a) Segmentation (b) Captured Character (c) Fourier Transform

Other possible preprocessing operation is: the extracting or computing of some symbol's characteristics, which could be used in the stage of classification. Such ones could be as follows: 2D means, 2D deviation, and other higher statistics' moments, some transformations and so on.

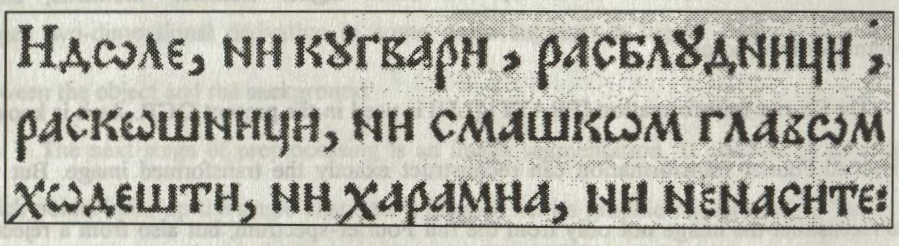
The Fourier transformation [PRATT],[LU] is used in the present **OCR**. As it is known, the inverted Fourier transformation can reconstruct exactly the transformed image. But we could reconstruct the image not only from the full Fourier-spectrum, but also from a rejected spectrum, where some of the features are replaced by **0**. The result is an image, similar to the original one, but cleared from some 'high' frequency components, as a noise, curves, convexities, and so on. (Fig.2.c) This transformation is used in the classification subsystem, which classifies the transformed image, not the original one. In this way the characters from different kinds of handwritings, typewritings, printings, containing noise, spots and other damages could be recognised almost correctly.

III. CLASSIFICATION

Many kinds of classificatories exist. The applied one has to be able to be learned and to learn itself. The initial learning has to train the classifier to recognise the symbols at first. After

that, during the reading process it has to be able to learn itself to classify correctly the shapes of characters near to the initial ones. The self-learning improves the process of classification. When an unknown shape occurs, **OCR** has to ask the operator whether to create a new class of symbols or to update the set of some existing one.

The Neural Networks (**NN**) [WASSERMAN], [ZEIDENBERG] are one of the best classifiers, which could do all of the above. They could be learned by a teacher or by themselves - it depends on the used **NN** structure and the learning algorithm. During the learning process **NN** creates associations between the presented input pattern and the Output, at each training cycle.



(a) Image with text

(Щ)ДШЛЄ (С) НИ КУГВ(С)РИ , РАСБЛ(Є)ДНИЦИ ,
РАСКШШНИЦИ , НИ СМАШКШМ ГЛАВШМ
(УС)ДЄШТИ (С) НИ (У)АРАМИА , НИ НЄНАСИТЄ

(b) Recognised text, displayed with font Izhica, the wrong recognised characters are in brackets.

Fig. 3 Recognition of 3 text lines from { Abagar }

The **NN** used at the present **OCR** is a 3-layer Perceptron with BackPropagation algorithm for training (**BPNN**). We chose it because it demonstrates a good capacity for classification, and it could be retrained at the recognition process. But the standard Backpropagation algorithm doesn't match the complexity of the problem. Therefore, we used

an improved learning algorithm, which includes a supervisor of the training process[STOYANOV]. It uses a variable node number in the hidden layer, which allows learning of problems with different complexity. The **BPNN** feeds by **21** of the features from the Fourier spectrum of the recognised symbol. The hidden layer has maximum of **50** neurones. The output layer has **60** neurones, corresponding to the number of recognised characters.

IV. POST-PROCESSING

At present the recognised characters are organised in a text file with 8 bits representation per a single character. We intend to represent the recognised characters in the followed 3 levels:

1. **Image.**

It contains the scanned document without any processing. It will be used as a base for the next levels. It also could be used for CD Rom creation and so on. The problem with **Image** representation will be the extremely large size of Data Base.

The main users of such a representation could be palaeographers.

2. **Document.**

This 2-nd level will reduce significantly the required space for representation. It contains the document in a following format:

{ [**Char**1] , [**Char**2], ... }, where

Char = { (**X**,**Y**) position of the Character in Original Image,

Char **Number**, Char **Type**,

Pointer to Data Base of Character's **Shapes**

}

As it could be seen, the original shapes of characters are stored and the document will look just as the original one, when it is displayed on the screen. Perhaps the orthographical researchers will be the users of such a representation. It could be used for printing too.

3. **Text.**

The last level of representation takes minimum of memory. It will be helpful for linguistic researches. The recognised image will be represent by a string , or a sequence of characters coded into 8 bits.

V. RESULTS AND CONCLUSIONS

The **OCR** is tested with an {Abagar} printed in 17-th century. A page of it is used for **OCR** training and other 2 pages are tested for recognition. For our pleasant surprise the classification achieved about of 95% of correct reading, nevertheless the system is yet at a stage of developing.

The future plans are as follows: The first main problem we have to get through is an automatic symbol segmentation. The other problems are efficient character representation after the recognition process, a proper data base for document representation, dictionary making and so on.

Our purpose is to create an **OCR**, that can read not only Bulgarian Symbols, but other ancient alphabets also: Latin, Greek and so on. Such an OCR System will help a lot of You to make Your investigations faster.

REFERENCES

[SHAPIRO] V.Shapiro, P.Veleva, V.Sgurev, "An Adaptive Method for Image Thresholding" 11-th IAPR , 1992, 696-699

[DOBREVA] M.Dobreva "Applications of Computer Tools in Studying Medieval Slavonic Manuscripts", Sofia 1995

[STOYANOV] I.Stoyanov at all "Off-Line Signature Verification by Beural Network Pressure Analysis" Int. Conf. on NNACIP'94 Mexico, 1994, Proc pp. 74-79

[WASSERMAN] Ph.Wasserman "Neural Computing - Theory and Practice", Auza Research, New York, 1989

[ZEIDENBERG] M.Zeidenberg "Neural Networks Models in Artificial Intelligence" USA Ellis Horwood, 1990

[LU] Yi Lu, Steven Schlosser, Michael Janeczko "Fourier descriptors and handwritten digit recognition" Machine Vision and Applications, 1993, Springer-Verlag, pp.25-34

[PRATT] Wuilliam Pratt "Digital Image Processing" , New York, 1987, Wiley-InterScience Publication

Computer Assisted Analysis of the Macrostructure and Typology of Medieval Slavic Miscellanies

Anisava Miltenova
Institute of Literature BAS, 52 Shipchenski prohod Str., bl. 17, Sofia 1000
ANMILTEN@BGEARN.ACAD.BG

Abstract: This paper addresses the presentation of problems of data entry according to typology of manuscript (menologia, liturgical manuscripts, mixed content miscellanies, etc.), with attention to selected features (options) for computer description, classification schemes, the universality of the template, terminology, key words, and related issues. Particular attention is devoted to mixed-content miscellanies, which are the source of particularly complex descriptive problems. The specific problems of the co-existence of the separate texts (cycles and segmentation), as well as aspects of their processing are discussed and examined, with suggestions for the SGML facility application).

The study of the structure and composition of Medieval Miscellanies has one of the most significant topical trends in Slavic studies over the last decades. The International Symposium *"Littérature moyenne grecque et littératures slaves. Corrélations dans les recueils manuscrits"* in Thessaloniki 1979[1] was an expression of growing interest in contemporary studies of medieval miscellanies.

Since that time we have seen new achievements in the field of typological analysis of composition, and in revealing the chronology of translations and original texts in miscellanies and their distribution. These achievements were based both on new source materials, and on revisions of existing sources. A number of editions of valuable MS monuments came out. What was established was the fact that in order to consider systematic approaches to the *Repertorium* of medieval miscellanies, we need a clear and internally non-contradictory classification of macro-genre structures. The traditional approach to the composition of MS

[1] "Cyrillomethodianum", V, Thessaloniki, 1981.

was overcome - an approach in which every article in a miscellany is described or published separately (with or without citation of the Byzantine sources). This approach was replaced with a greater attention to the role of the *"convoy"* (D. S. Lichachov[2]), i.e. preceding texts and subsequent texts. In order words concentration on the typological function of proximity of texts came to be applied by a number of scholars in their work on various types of medieval works. This led to an appreciation of the role of compilations for Slav authors, who adapted, interpreted and adjusted the texts of every component article in accord with the general trends in the MS as a whole (R. Picchio[3]).

In contemporary medieval studies there has been a trend to recognize a relationship between codices and their component text at every level: chronology of translations and original works, the history of concrete texts, genre structure, stylistic characteristics, and linguistic features in connection with their creation. In other words the object of study is an overall complex manifestation of the written word. This new perspective stems from the need to reveal the interrelationship between the composition of a miscellany and its separate texts, the process of including the texts in anthologies, and the history of earlier forms of anthologies as they pass through the redaction of various texts, cycles, or entire compositions. But the specific variations in most of the monuments of the Slavic MS heritage at all levels of MS texts, from orthography and stylistic elements through concrete and deliberate redactions, compels researchers to undertake complex and intricate analysis of large MS texts, which pose difficulty for linguistic analysis as well as for synchronic and diachronic comparison.

---

[2] *Лихачев, Д.* С. Текстология (на материале русской литературы X- XVII веков). 2-е изд. Л., 1983, 232-248.

[3] *Picchio, R.* Compilation and Composition: Two Levels of Authorship in the Orthodox Slavic Tradition. – Cyrillomethodianum, V, Thessaloniki, 1981, 1-4.

My current work aims at providing an answer to the practical and methodological issues I have just outlined the application of computer aided methods to the analysis and processing of medieval texts in Slavic miscellanies. This project arose several years ago in the course of work on the description of manuscripts within the Bulgarian cultural heritage, and in particular in experimental work in the use of electronic media, so as to allow distribution and access through modern information networks. Data bases of comparable magnitude, consisting of bibliographic descriptions on the one hand, and an impressive mass of full texts, on the other, already exist in the field of Classical philology, and are maintained and enriched to this day (e. g. *The Thesaurus Linguae Graecae* - planned and carried out in 1972 at the University of California, USA; *Isocrates* - created in 1985 at Brown University, USA; *Perseus* - created in 1988 at Harvard and Boston University, USA, etc.). No Slavic country has ever devised and begun any such project, due to the absence of software, that would support the processing of Old Slavic texts as well as the specific features of the MS tradition, which cannot easily be subjected to standardization. There have been a number of attempts been to catalogue Slavic MS with software used for library data bases (as was the case with CIBAL project, based on CDS ISIS), but such applications have a fixed score considering the number of fields and their lengths. Hence, an efficient form of MS description, reflecting as much as possible all relevant features, and allowing simultaneously the search for the optimal number of elements at all levels, from the codicological description of MS to texts contained in codices, calls for a principally new solution.

The outcome of the efforts of the team of authors (D. Birnbaum, A. Boyadzhiev and A. Miltenova, who compiled the model and the principles of the description; B. Van Dijk, M. Dobreva and H. Gaylord, software design) was a system based on the SGML standard

(adopted by ISO), within the framework of the TEI (Text Encoding Initiative). The team ascertain the fact that much of the data that are obligatory in the description of MSS are missing in the classes of texts described within the TEI framework. In this sense the design and inclusion of the new elements in the description is a contribution to the TEI format and the application of SGML to Slavic philological studies.

Computer processing of the content of Medieval miscellanies can be applied where the scholar is faced with a problem combining a variety of features, related to miscellanies as a whole in one place on the one hand, and to the other features, related to the history of a concrete text or a group of texts, on the other. Suitably encoded computer texts would allow the tracing all forms and stages of the existence of a given text as well as the development of structural models of various types of medieval miscellanies. In this project we selected the following elements in the description of texts: Computer processing of the content of Medieval miscellanies can be applied where the scholar is faced with a problem combining a variety of features, related to miscellanies as a whole in one place on the one hand, and to the other features, related to the history of a concrete text or a group of texts, on the other. Suitably encoded computer texts would allow the tracing all forms and stages of the existence of a given text as well as the development of structural models of various types of medieval miscellanies. In this project we selected the following elements in the objective description of texts:

I. ManuscriptContentDescription

TextType: original, compilation, translation

TextStyle: narrative, non-narrative

NumberTexts

ManuscriptCreation

ManuscriptDate

ManuscriptPlace

ManuscriptSource

Greek

Other

Translation

Num

Date

Protograph

Antigraph

Apograph

LitRedaction

Num

Date

II. ArticleContentDescription

TextType: original, compilation, translation

TextStyle: narrative, non-narrative

ArticleName

ArticleAuthor

ArticleData

TextLocation

LocNumber

Neighbor

Type: following, preceding

Position

LocFolio

LocLine

Source

Type: Greek, other

Translation

Num

Date

Protograph

Antigraph

Apograph

LitRedaction

Num

Date

ChurchCalendar

III. Text

Title

Incipit

Explicit

The problems that arise in the application of computer aided analysis of medieval miscellanies are two-fold:

A. Problems of the macrostructure of various types of miscellanies.

B. Problems of description and identification of text articles.

A. The following general principles can serve as a basis for the preliminary classification of miscellanies:

1. Classification according to functions, where we find general miscellanies or liturgical collections, and what are known as *cheti-sbornitci*, or miscellanies for non-liturgical purposes and readings. Underlying them is a practical purpose, in most cases serving the function of the liturgy, and their connection with practical needs – prologia, prayer books (euhologia), nomocanons, hymnographic miscellanies, etc. The structure of liturgical collections is

subordinated to the daily life of the church, e. g., the Sermons in the Miscellany of John Chrysostomos, linked with the Feast of Easter (the so-called Long Lent) and days dedicated to feasts of the saints within the church calendar. *Cheti* miscellanies consist chiefly of texts that are not subordinated to the daily life of the church. Naturally, the translation of entire miscellanies from one macrogenre to another are quite frequent, not always allowing a clear borderline between them.

The first type, liturgical miscellany, usually follow the calendar of Church feasts (according to the synaxarion)[4]. They are known according to their names, e. g. hagiographic or homiletic miscellanies, as well as panegyrics, menologia, menaea, etc. The appearance of hagiographic and panegyric miscellanies with a constant structure coicides with the earliest period of Old Bulgarian literature (e. g. Codex Clozianus, Codex Suprasliensis, or MS that were restored later, such as the Mihanovic homiliary, Jagic's Miscellany of John Chrysostomos, German's Miscellany etc.). These miscellanies contain sermons for the major feasts of the year, as well as lives of saints and texts linked with the triod cycle. Throughout this early period the composition of these miscellanies depends on the Stoudite Typikon. There is no consistency in the order of the separate works, nor with the church feasts; one and the same sermon can be linked to different dates and various feasts. From a stylistic and compositional point of view, Old Bulgarian hagiographic and homiletic works are less dependent on their component texts, are subject to redaction, and are frequently compilation of one and more sources. The leading element in their computer analysis could be the data

[4] *Иванова, К.* Агиографско-хомилетичните сборници с устойчив състав в южнославянските литератури. – Cyrillomethodianum, V, Thessaloniki, 1981, 11-21; Vorläufiger Katalog Kirhenslavischer Homilien des beweglichen Jahreszyklus. Aus Handschriften des 11. – 16. Jahrhunderts vorwiegend ostslavischer Provenienz. Zusammengestellt von *T. Certorickaja*, unter der Redaktion von *H. Miklas*. Westdeutscher Verlag, 1994, 38–57.

from the option CHURCH CALENDAR, attaching the rest of the options to it: ARTICLENAME, ARTICLEAUTHOR, SOURCE, TRANSLATION, PRORTOGRAPH, ANTIGRAPH, APOGRAPH, LITREDACTION, together with the real TITLE, INCIPIT and EXPLICIT of the text. The same principle is valid for prologia, menaea etc.

The contents of miscellanies with a variable composition (usually non-liturgical works) are united in cycles according to thematic or stylistic principle.[5] The structural models of variable miscellanies fall into two basic categories: the core and those around the core. The core works are in fact a relatively stable combination of articles, with close historical and literary features, which function as a compositional core in the MS. Despite this shared context the core of every MS differs stylistically or liturgically, so that every miscellany has its own style. From a somewhat different perspective, the core of a certain type of miscellany is essentially an abstraction of common structural features and redactions, without necessarity standing for any concrete MS, a supposed protograph, or hypothetical proto-miscellany. A potential structural model, which combines the cores of two and more structural models, the so-called contaminated type, is also possible. Miscellanies reflecting the core of typical works in an extended or reduced form also fall here, as long as their content contains sufficient textological proof of a common origin. The problem of succession in the arrangement of the separate articles is no less essential; however, it is a secondary criterion determining the composition type of a given miscellany.

In this sense the concept of a "surrounding background" has a more general, abstract meaning within the framework of a given structural model, based on a common core of

[5] *Милтенова, А.* Към методиката на изучаване на сборниците със смесено съдържание в старите южнославянски литератури. — In: Studia slavica mediaevalia et humanistica. Riccardo Picchio dicata. Roma, 1986, 517-526.

articles. The correlation between the core and the surrounding context is realized at several levels: conceptual-aesthetic, functional, linguistic and stylistic similarity, thematic and subject patterns, etc. For example the scribe who make the copy of the Miscellany of Panagjurishte, 16th cent. (No 433, National Library - Sofia) has added two texts to the original composition, consisting of the Svjatoslav Miscellany of 1076: the Sermon concerning the reading of books and Precepts and Advises from father to his son. The compiler of the Miscellany of mixed content No 149, 16th cent. (National Library - Vienna) has include in the beginning the Life of Alexis, The Holy Man, and at the end the Rules of the Nomocanon. They are the only difference between the two MS, where 36 works are common and fully coincide in their arrangement and redaction. Evidently the scribes, who were also compilers felt that the composition of miscellanies as something flexible and open, in spite of the stable core.

3. Classification of miscellanies according to content (a thematic principle). This classification is based on the characteristics of miscellanies according to the prevailing themes of the works, included in them: historical, legal, hymnographic, apocryphal, etc. Titles known as "monastic miscellanies" (*monasheski sbornik*), which are addressed to a certain public, fall here. These titles do not indicate content, because "monastic miscellany" might contain orations or The Ladder of Paradise by John Sinaiticus, or the Dioptra by Philip Monotropus, or patericon stories, etc. In other words, the name of the MS does not give information about its content, but only about the intended readership.

4. Classification of MS miscellanies according to genres: paterica works, menaea, chronicles, prologues, panegurics, palaia, damascenes, etc. This is the most comment principle in the traditional descriptions of MS.

Florilegia, as well as miscellanies apophthegmata, eratoapocriseis, etc. pose specific problems for computer aided processing of miscellanies. The terms denoting various types of miscellanies in the Slavic middle ages should be included in the list of KEYWORDS in the scheme shown.

The aim of the *searching programs* compiled on the basis of material from Medieval MS are follows:

- an index of the titles of works (scientific titles), allowing the retrieval of all remaining data for every work: MS, chronology, repository etc., together with original title, incipit and explicit;

- an index of the titles of articles and their authors, together with the remaining information on the concrete MS (repository, source, translation, literary redaction, protograph, antigraph, data on the history of the text, days on the church calendar, bibliography);

- an index of authors across the full library of manuscripts;

- links between the list of keywords and the concrete texts, etc.

What is the extend outcome of the computer analysis of the content of Medieval miscellanies? In the first place the data will become the basis of a *repertorium* of medieval Slavic texts (original works and translations), which would include exhaustive information about the separate texts, as well as information about the MS where these texts are contained. In the second place it would be possible to undertake the computer-assisted comparison of the composition and structure of a separate miscellany with the composition and structure of other MSS. In this way, the typological similarity and differences among miscellanies of one and the same type will be simplified, and it will be easier to study the history and distribution of texts

within separate cycles and entire compositions. Last, but not least, this is the first attempt through computer aided means to link together real TEXT from MSS (TITLE, INCIPIT, EXPLICIT), scientific description, and the data from study. This is a necessary prerequisite for the preparation of texts for publication or inclusion in a scientific monograph.

REFERENCES

Иванова, К. Агиографско-хомилетичните сборници с устойчив състав в южнославянските литератури. — Cyrillomethodianum, V, Thessaloniki, 1981, 11-21.

Иванова, К. Класификация, типология и каталогизация на някои типове сборници в балканската славянска ръкописна традиция. — В: Славянска палеография и дипломатика. 2. Лекции от втория семинар по славянска палеография и дипломатика. София, август 1983 (CIBAL. Balcanica III. Études et documents 4). С., 1985, 173-182.

Лихачев, Д. С. Поетика древнерусской литературы. М. 1979.

Лихачев, Д. С. Текстология (на материале русской литературы X- XVII веков). 2-е изд. Л., 1983.

Милтенова, А. Към литературната история и типология на сборниците със смесено съдържание. — Старобългарска литература, 7, 1980, 22-36.

Милтенова, А. Към характеристиката на българската литература през XV—XVII век. (Сборници със смесено съдържание). Канд. дисертация (машинопис). С., 1982.

Милтенова, А. Към методиката на изучаване на сборниците със смесено съдържание в старите южнославянски литератури. — In: Studia slavica mediaevalia et humanistica Riccardo Picchio dicata. M. Collucci, G. Dell Agata, H. Goldblatt curantibus. Vol. 2. Roma, 1986, 517-526.

Черторицкая, Т. В. Торжественник и Златоуст в русской письменности XIV—XVI вв. — В: Методические рекомендации по описанию славяно-русских рукописных книг. Вып. 3. Ч. 2. М., 1990, 329-381.

Vorläufiger Katalog Kirhenslavischer Homilien des beweglichen Jahreszyklus. Aus Handschriften des 11. – 16. Jahrhunderts vorwiegend ostslavischer Provenienz. Zusammengestellt von **T. Čertorickaja**, unter der Redaktion von **H. Miklas**. Westdeutscher Verlag, 1994, 38–57.

Picchio, R. Models and Patterns in the Literary Tradition of Medieval Orthodox Slavdom. – In: American Contributions to the VII International Congress of Slavists. Warsaw, Aug. 21-27, 1973. Vol. II. ed. V. Terras. (Slavistic Printings and Reprintings 296). The Hague/Paris, 1973, 439-467.

Picchio, R. Compilation and Composition: Two Levels of Authorship in the Orthodox Slavic Tradition. – Cyrillomethodianum, V, Thessaloniki, 1981, 1-4.

Paleographic and Orthographic Features in the Description of Slavic Manuscripts

Andrej Bojadžiev

University of Sofia

Keywords: Slavic codicology, paleography, linguistics, features in the descriptions, SGML, TEI.

Abstract: Establishing a set of well-organized rules in the field of Slavic codicology, paleography, orthography, and linguistics for computer descriptions of medieval manuscripts and texts has long been a complex problem. Manuscripts from different times and places may impose different encoding requirements, and Slavists are far from agreement about how to manage these differences. The paper addresses the presentation of obligatory and optional features on two levels: discussion of their place in the system and techniques for their encoding and presentation in electronic form. They are made mainly with the help of Document Type Definition files generated on the principals of Standard Generalized Mark-up Language and the Guidelines of the Text Encoding Initiative.

General remarks

By establishing a descriptive system designed to account for the range of codicological peculiarities of Slavic manuscripts we will be able to gain valuable extra evidence concerning as to their dating, provenance, and later individual history -- evidence that might not be forthcoming from purely textual, linguistic or art-historical evidence. We may acquire a broader knowledge of the many sectors of the society and economy concerned in one way or another with the production, patronage, use, exchange, trade, and stewardship of books. It is very important to investigate the books according to the relationship of all these elements together -- writing, graphics, orthography, text segmentation and text history, illumination, binding, collation of quires, etc.

All these aspects are increasingly acknowledged as important parts of our study of medieval Slavic culture and some of them have a traditional place in all descriptions of Slavic manuscripts.

The problem is not whether to include them in our descriptions, but what parts of them to include. Must we deal only with what is well-known, and just describe these facts briefly, or should we include some data from the fields like binding or sewing techniques. This problem is not just a matter of differentiation between check-list, short catalogue and analytic description. It is the question of what we are searching for: only a description or a thorough investigation.

There are several philological tasks in the description of Slavic manuscripts that have to be resolved before we sit in front of our computers. They are:

1. *The inventory of units.*

2. *The terminology of units.*

3. *The structure of units.*

The situation in Slavic paleography and codicology could be divided into two sets of problems:

1. We may know too little when we begin our study to be sure what kind of elements is relevant. After careful study it may turn out that some can be omitted. That is the situation in the Slavic codicology. We know very little about the sewing techniques, clasps, spine or treatment of the binding edges in the Slavic manuscripts. What we do know for sure is that some of these elements are significant for typology, dating or history of the manuscripts. So we must include these features as well in our description.

2. We know that some other elements are relevant and we know more about them, but it may not be easy to systematize them. This is the case of Slavic writing, orthographic systems, language description or systems of illumination.

The first case in the Slavic manuscripts, elements about which our knowledge is inadequate, is solved very often in two ways:

1. By including only limited information about the object. For example the description of binding may be made in such a way: "Original (later) binding from (century | year) consists of wooden panels covered with ornamented leather."

2. By completely ignoring information about the object where our knowledge is adequate.

In the second case we very often rush the matter and mix completely different stages of research by confusing, for the example the orthographic and linguistic conclusions.

So the contemporary researcher must not give the answers of the type: "The text is written with *Bulgarian* | *Russian* | *Serbian* orthography." This is wrong answer to at least two questions: "What is the system of orthography?", and "What are our language conclusions?" In each case we should first define "Does this feature comes from the protograph | antigraph | from the scribe's hand?", and "Is it accepted everywhere or it is peculiarity of this text or the group of texts?" It is worthwhile to mention here the words of Dmitrij Lixačev who says: "Only a non-linguist has an easy answer when asked to identify what the language of the text is."

With the help of SGML and TEI Guidelines, for example, we can develop a well-organized tree structure for the description of the Slavic manuscripts. In fact, the description of the books could be a very good example for employing such a tree structure. In codicology and paleography we can choose one of the two possible ways: to include all valuable units (from our point of view) in our description and to let the editors and authors divide them into obligatory and optional features, or to make separate rules which will contain data only from this parts of our catalogue. All features in Slavic codicology (i.e., quire collation, ruling, pricking or binding) are comparable to those in the Greek, Latin or medieval Hebrew traditions. So we also need to consider the results of study in these fields, and to include all

relevant elements in our materials to compare and distinguish them from these other traditions.

Here is a proposal for low-structuring of some codicological, paleographic, orthographic and linguistic data:

Binding

type: *preserved* | *partially preserved* | *without binding* | *missing*

bindCharact>

Boards>

type: *visible* | *partly visible* | *not visible*

Material>

type: *wood* | *cardboard* | *paper* | *parchment* | *other* | *without boards*

sizeComp>

type: *cut with the textblock* | *extend beyond the textblock* | *impossible estimation*

boardsFormats>

type: *inches* | *cm* | *mm*

attachments>

attSystem>

edges>

type: *rounded* | *inclined to the textsblock* | *square*

edgeGrains>

edgeGrainsform>

type: with interruptions | without interruptions

Spine>

type: natural | hammered rounding

spineLining>

type: *linen* | *hemp* | *cotton* | *specify*

headbands>

headbandColors>

headbandConstruct>

raisedBands>

backCorner>

joint>

Cover>

type: *leather* | *vellum* | *velour* | *silk* | *different* | *without cover* | *missing cover*

coverTreatment>

coverColor>

coverCorners>

turnIns>

Pastedowns>

Endleaves>

bindDecoration>

type: *cover ornamentation* | *boards ornamentation* | *both*

treatment>

type: *hot stamp* | *cold stamp*

ornamentScheme> (pointers to external scanned files)

frontCoverscheme>

backCoverscheme>

ornamentElements>

ornamentCharact>

edgeOrnament>

edgeInscription>

Pins>

type: *preserved* | *not preserved* | *not used*

pinNumber>

pinDisposition>

pinForm>

pinMaterial>

clasps>

claspNumber>

claspdisposition>

claspForm>

claspMaterial>

sewTech> (description and pointer to external scanned file)

watermark

description and pointer to external scanned file, which contains a structured archive of watermarks and list of watermark editions and albums.

quireStructure

description and pointer to external scanned file with the following graphic and text information: structure of the leaves in the quire, hair-side (HS) and flesh-side (FS) rhythm by the pergament codices, changes of watermark singns by the paper manuscripts, original pagination in the beginning and at the end of the quires, later pagination(s), contemporary archival pagination (the same if the pagination(s) is per folia/per page); place and rhythm of the ruleline-making; the scanned file could contain also some information about place of illustration(s) and some other important elements.

Ruleline

type: *Gilissen* | *Leroy* | *Stanchev-Dzurova* |*other*

Description and pointer to external scanned file.

Ornament

illustrations>

borders>

line-fillers>

vjaz>

initials>

capitals>

cadels>

caledndars>

tables>

labyrinth>

otherOrnament>

each of the elements contains as part of its low level structure: **place, size** (in comparison with the textblock and also in comparison with other elements of this type as illustration, borders, etc.), **textConection, motives, color, styleRemarks**, for example illustrationPlace, illustrationTextconection, etc.)

If neccessary, pointers to external scanned files.

orthogrCharact

letters>

jer>

type: *front* | *back* | *etymreg* | *nonetymreg* | *irregular*

jus>

type> *etymreg* | *nonetymreg* | *nonconsis* | *nonjus* | *justrace*

jotvowel>

otherletters>

ligatures>

abbrev>

numeral>

supralinear>

musical>

punct>

marg>

explain>

final>

foreighnorth>

crypto>

otherorth>

gramconcl>

DATA BASE OF SLAVONIC MANUSCRIPTS IN MACEDONIA

Stana Jankoska*, Dragan Mihajlov**, Ljubomir Josifovski***

* National and University Library "St. Kliment Ohridski" - Skopje, Macedonia
tel. +38991 115177, fax. +38991 226846, Email MX%stana@nub.mk

** Faculty of Electrotehnical Engineering - Skopje, Macedonia
tel. +38991 363566

Faculty of Mechanical Engineering - Skopje, Macedonia
tel. +38991 363566

Abstract: Experimental bibliographical/factographical data base for ancient Slavic manuscripts discovered in Republic of Macedonia is presented. Because of the specificity of this material, besides the ordinary bibliographical data, pictures of the manuscripts are included. Data base structure is based on today practice for ancient manuscripts description, ISBD(A) and UNIMARC standards. Each record in the base contains the identification elements of the manuscript from cultural - historical, literally - historical and philological aspects. Data base will be offered to Internet user as soon as Internet come available and parallel to that a CD ROM version of data base is under development.

I. Slavonic manuscripts in Macedonia

The medieval Slavonic manuscripts in Macedonia up to date, are noted and described in many monographs issued in Republic of Macedonia. In 1971 a book in two parts "Slavonic manuscripts in Macedonia" by the author Vladimir Mosin (in collaboration with Slaveva Lidija, Sultana Kronevska and Jovanka Jakimovska) is issued. The first part contains descriptions of 168 manuscripts known at that time in Macedonia, the second part contains 174 facsimiles of manuscripts described in first part.

After the issue of this monography, the research and cataloguing of Slavonic manuscripts in Macedonia are making in the archeographic department in the National and University Library "St. Kliment Ohridski" - Skopje (NULSKO - Skopje). From 1970 to 1985 as a result of investigation actions, the new 152 Slavonic manuscripts are discovered. They date from XIV to XIX century.

In 1988 the NULSKO -Skopje issued the monograph in two parts (book III and book IV) "Slavonic manuscripts in Macedonia" by the author Mihajlo Georgievski. In the first part (book III) the descriptions of the new discovered 152 Slavonic manuscripts are given, in the second part (book IV) the facsimiles of these manuscripts are given.

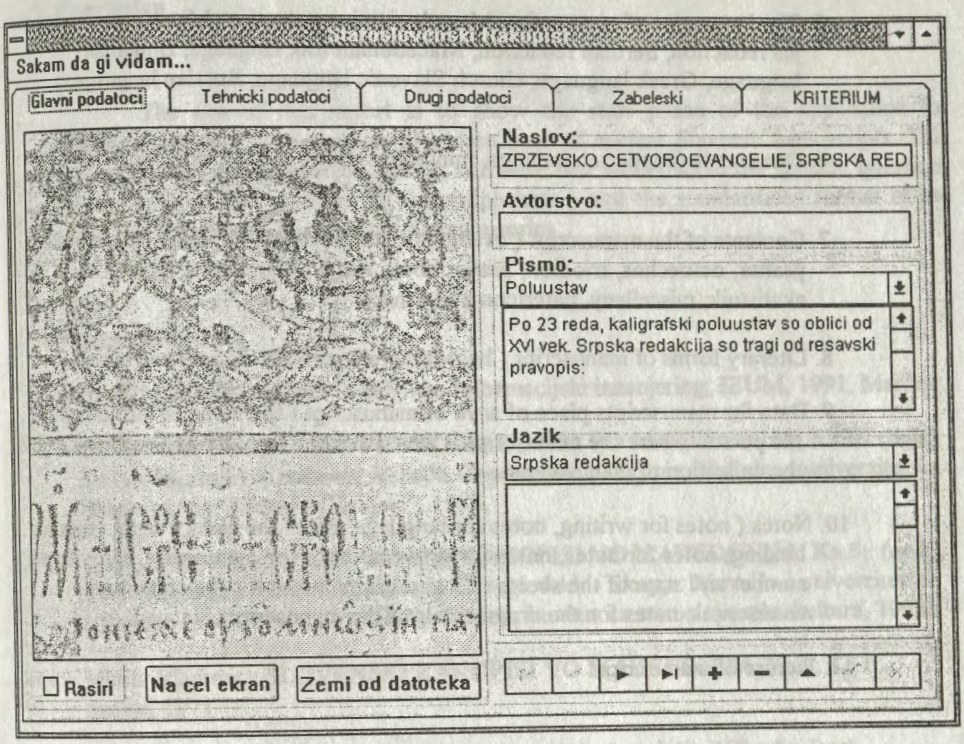
In 1993 The Institute for Old Slavonic Culture in Prilep and NULSKO - Skopje have issued the book V and the book VI by the author Mihajlo Georgievski that contain 71 new discovered manuscripts in Macedonia. The fifth book contains the descriptions of the manuscripts and the sixth book contains 92 facsimiles of manuscripts of the fifth book.

The discovered manuscripts in the issued monographs are analyzed and described with strict science methods. The rules of cataloguing the Slavonic manuscripts at Macedonia, are established by the author Vladimir Mosin of the first and second book "Slavonic manuscripts in Macedonia". According these rules in the description of every Slavonic manuscript the elements that show so precise content of manuscript of more aspects: culturally - historical, literally-historical and philological are given. In essence, the first, third and fifth book present catalogues of Slavonic manuscripts in Macedonia in the form of books. The manuscripts in these catalogues are arranged in chronological order by the year or century when the manuscript was written. Except this data, every catalogue description of manuscripts contains remaining data as: call number or signature of manuscript, (in other words that is the place where the manuscript is saved, technical data (material, dimensions, the number of sheets of manuscript, the damage of manuscript, ornaments data, wathery signs, binding, content of manuscript, data for notes on the manuscript and data for the provenance of the manuscript).

II. Computer-supported processing of Slavonic manuscripts

We attempted to use computer and multimedia technique in the processing of the old Slavonic manuscripts as a specific library materials. Our last aim is to create a data base of Slavonic manuscripts in Macedonia. This data base will be formed with gradual conversion of already described manuscripts (about 380 manuscripts in publications) in machine readable format and recorded on CD ROM.

The record structure of every manuscript was determined on the base of today practice for old Slavonic manuscripts description in Macedonia and the ISBD(A) (International Standard for Bibliographic Description for Ancient). A new element in the record now, is the graphical presentation of the manuscripts, more exactly their pictures. We think, that all these elements make the data base of factographical type.



III. Structural elements of the record of single Slavonic manuscript

The record elements of a single Slavonic manuscript can classify in more categories:

1. Title of manuscript
2. Authorship
3. Technical data of manuscript (material, dimensions, number of sheets, damage of manuscript, binding)
4. Writing of manuscript (unical, semiunical, tachography)

5. The languages of manuscript: Macedonian jus redaction, Macedonian without jus redaction, Serbian redaction, Macedonian folk language, Bulgarian language, Greek language, church Slavonic language, Russian language.

6. Ornaments (copper engravings, miniatures, flags, initials, drawings).

7. Contents of the manuscript (evangelion, tetraevangelion, apostolos, menaion, psalter, octoechos, triodion, prayer book, ritual book, hymn, typicon, akatistnik, miscellany, patericon and other).

8. Literary forms of manuscript (hagiographies, offices, notices and other).

9. Data for manuscripts place of save of manuscript (the name of institution or the person where the manuscript is saved and the signature or call number into the collection of manuscripts.

10. Notes (notes for writing, notes for language, notes for watery signs, notes for binding, notes for notes, notes for provenance of the manuscript, notes for the number and state of the sheets of manuscripts, notes for the contents of manuscripts, notes for the ornamentals of the manuscript).

11. Picture of manuscript.

IV. Computer techniques used for manuscript's processing

The target platform for the CD-ROM version of the data base is standard PC, with 8 MB of RAM and double speed CD-ROM player under Microsoft Windows. Several factors were considered. Availability of the platform was considered as primary criteria. Ease of use was next. Additionally, several program requirements, such as mixed typefaces, scanned manuscript facsimiles, support of different printers, etc., can be successfully fulfilled on the chosen platform.

The program was developed using Borland's Delphi. We used resolution of 100 dpi with 256 colors/shades of gray for the manuscript facsimiles. This quality has proved to be sufficient. Mixed typefaces have not been implemented jet. The idea is to use in-place editing of OLE 2 automation object in conjuction with OLE 2 capable text editor. The reports are also jet to be done. The program will be capable of generating HTML version of the whole database.

The target platform for the Internet on-line version of the data base is Unix with HTTP server. The HTML format was chosen partly because it is already a hypermedia format, and partly because of the great popularity of WWW browsers on Internet. Still, there are numerous problems to be solved. At the moment, it's not clear how the problem of mixed typefaces is going to be solved. The manuscript facsimiles will be moved from BMP to JPEG format.

V. Conclusion

The aim of this report is to show the start phase of the application for multimedia processing and creation of a data base of ancient Slavonic manuscripts. This application should a one base for making software that would solve the specific questions about the input of data (input in more alphabets) and the standardized format of the machine readable record of old Slavonic manuscript.

VI. Literature:

1. Uputstva za unos rariteta i rukopisa, Informacijski inzenjering, IZUM, 1991, Maribor.

2 Georgievski, Mihajlo : SLOVENSKI RAKOPISI VO MAKEDONIJA, Kn.III:(opis); Kn.IV:(faksimili) /[izrabotil] Mihajlo Georgievski. - Skopje: Narodna i univerzitetska biblioteka "Sv. Kliment Ohridski", 1988.

3. Georgievski, Mihajlo : SLOVENSKI RAKOPISI VO MAKEDONIJA, Kn.5: (opis); Kn.6: (faksimili)/[obrabotil] Mihajlo Georgievski. - Skopje: Narodna i univerzitetska biblioteka "Sv. Kliment Ohridski"; - Prilep: Institut za staroslovenska kultura, 1993.

4. Mosin, Vladimir : SLOVENSKI RAKOPISI VO MAKEDONIJA, Kn.I, Kn.II Skopje, 1971

5. ISBD(A)(International Standard for Bibliographic description)(for Ancient)

Zdenka Ribarova (Skopje), Kiril Ribarov (Prague)

COMPUTER PROCESSING OF OLD CHURCH SLAVONIC MANUSCRIPTS

(Results and Prospects)

Abstract: We show here that, computer processing of Old Church Slavonic Manuscripts has series of advantages. We discuss on two major things: preparation of the original document and its rendering, and lexical analysis of the document. We present our self developed program for indexing and processing of Old Church Slavonic lexicality - STINO.

1. The analysis of the lexicality of Old Church Slavonic documents (later OCSD), cannot be left aside from the latest developments in computer science and their benefits. Various application programs would minimize the technical aspect of OCSD processing and would afford fast data-base searching techniques. Since we think that these problems are of a broader concern, in this paper we will try to summarize our initial results and our insights into the possible prospects in this area.

Commercial text-processors do not possess enough tools to deal with the specific features of an unrendered OCSD. Considering the problems related to, for example - requests for input and reproduction of an OCSD in its original form while preserving the relations with it after all operations with the entire vocabulary of the text are completed; acceptance of different font characters in order to preserve the style in the output documents; preservation of specific features of the glagolitic and the cyrillic alphabet during sorting procedures of the material; possibilities of establishing a relation between a OCS word and its Old Greek or other equivalent etc.

2. While processing of OCSD, we divided the work into two major parts. In the first part we concentrated on how to present the text (its physical layout and printing), while in the second part we paid attention to analysis of its lexis so that the text was looked at as a set of words.

3. In analyzing the lexicality of the text (the second part) we used a self-developed program package (STINO). STINO automatically transforms a manuscript into a data-base and at the same time extracts as much informaton as possible.

4. Our main aims are:

- Preparation of an OCS Manuscript for publishing. This assumes reproduction of the text while relations with its original form are preserved with a respect to the spelling features of each text.
- Preparation of index verborum with automatic location of the word forms (position in the text, relation to the Bible) with a possibility of having a relation of each OCS word with its Old Greek or other parallel.
- Preparation of the lexis of the OCS Manuscript for a lexicographical analysis. This assumes excerption of the lexis from the original manuscripts and an automatic construction of a corpus of word forms followed by the required information about their origin, position in the manuscript, grammatical characteristic, Old Greek or Latin or other parallel, context, head word.

  This corpus serves as a data-base for creation of a complete or selected/projected index according to specific criteria. For example - index of words which are present in all texts, index of words which are present only in a specific text, etc.
- Interpretation and analysis of the features of the lexicality of the OCSD (one or more manuscripts); searching, selecting and sorting according to required criteria.

5. The first version of the program which we have developed (STINO ver. 1.0) is a prototype which aim was to bring us to a higher level of automatization and at the same time, to help us solve specific problems which as presumption for obtaining the required preciseness, simplicity and speed during the work with huge data-bases.

These aims were fully achieved with the first version of STINO.

6. We worked out solutions for problems which might occur during the preparation of the text for computer processing. At this place, one should take care not only of the preservation of the relations with the original manuscript, but also, one should find solutions which allow automatization of a set of procedures, as for example - automatic location of each word form, automatic cross references, elimination of duplicity of the word forms,

identification of complex grammatical forms and set phrases, solution of the problem of homoforms, etc.

7. We assume that definition of the keyboard layout and implementation of an OCS code page belong to this matter. According to that, we state the requirements that the order of the characters (OCS letters and special characters) should relate as much as possible to nowadays standards and regulations.

Several years work experience shows that the font which we have created allows extremely comfortable and fast input of the text and requires no special efforts to accustom to its keyboard layout.

8. The first version of our program has been thoroughly tested during the work with OCS manuscripts in Prague (in the Slavic Institute) and in Skopje (in the Institute of Macedonian Language). With the help of STINO the edition of the new part of the Psalterium Sinaiticum with different types of indices was prepared in Prague in cooperation with specialists from Viena. Also, with the help of STINO, many of the manuscripts for the Dictionary of Macedonian OCS Manuscripts have been excerpted. Selections and classifications of the lexicality of separate texts were done and the results were used as a subject of postgraduate studies.

9. While the preparation of the text is mainly focused on the word forms, the analysis itself is concentrated on head words.

We have applied the computational linguistic in the problem of assigning a normalized head word to a word form. We used an adaptive bi-gram model together with an unigram model. Our statistics shows that the percentage of homoforms in the OCS documents increases with time, that is, the older documents have less homoforms than the newer ones, because of the changes of the phonological system.

This allows an unigram model to be used, while the bi-gram model will be used in the cases of ambiguity. Thus, the bi-gram model is not a full bi-gram model but it contains those words which we have already processed and which had given more than one entry in the unigram model. Both, the unigram and the bi-gram model are automatically updated after each

manuscript, so we call them adaptive models. We use the unigram model in order to preserve space and, as noticed, the bi-gram model is not (in the most of the cases) required.

The normalization has been implemented with a high percentage of success. At the same time, the success of a normalization of a head word depends on how much of the text has already been processed.

Further, it is possible to select the head words according to many criteria, that is, their semantic classification can be done, they can be selected according to a word-forming structure or according to their paradigmatic relation. One may select according to any of the present data-base fields.

The normalized head words serve as a base for forming indices of different kinds, such as:

- Complete index verborum, that is, index of normalized head words documented with all word forms, which are located according to required criteria and followed by an Old Greek or other parallel.
- Old Greek - Old Church Slavonic Index
- Retrograde index of the head words
- Indices to different semantic groups, etc.

10. The input and the output documents are not dependent on any text-processor. It is up to the user to select the text-processor.

11. So far, our experience during the work with STINO shows that the results, which are taken as main standards (different types of indices) and the data we have obtained, allow different kinds of studies of the lexical system.

Fast formation and sorting of large data-bases which are in use of the lexicological analyzer, and work with millions of card catalogues contrasts the classical excerption. Fast data location and comfortable browsing followed by a complete documentation gives a solid ground for each scientific study of the lexical system.

The use of the above mentioned effects brought us many useful and encouraging results. An insight into new perspectives is given, since the computer text-processing allows observation on the whole. At the same time, we would like to implement some methods of computational

linguistics which will bring the same results to a paleoslavicist as to his colleague working on the contemporary languages.

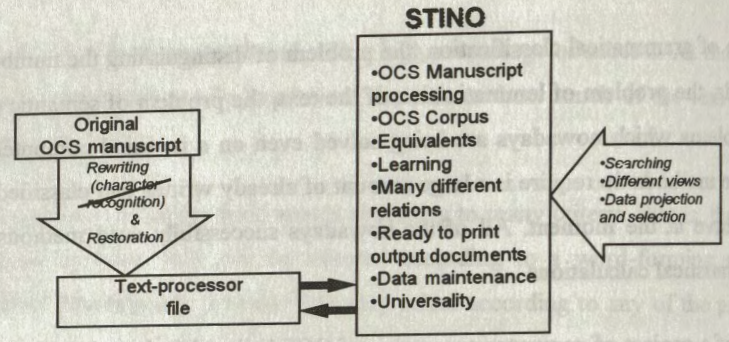
The problem of grammatical classification, the problem of distinguishing the numbers from the regular words, the problem of lemmatization of the text, the problem of sémantic classes form a set of problems which nowadays are being solved even on a personal computer. But, what most of these methods do require is a large amount of already written and classified text, which we do not have at the moment. All of the nowadays successfully used methods are mostly based on a statistical calculations.

Realization of a project of computer processing of OCS lexis might have broader effects under the supposition that conditions for cooperation and compatibility between different centers in the world is established. It is our wish to contribute to build those conditions which we truly believe would give access to information through electronic communication.

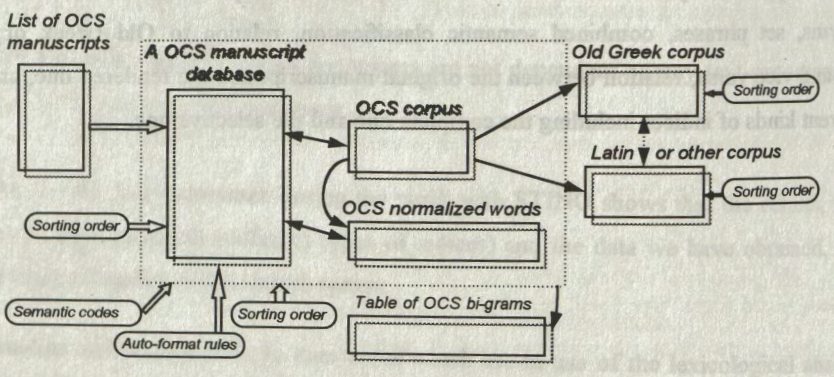
Conclusion

Our work has concentrated on two major things: preparation of the original document and its rendering, and lexical analysis of the document. The first uses specific rules and font characters supported by the latter. We have already solved the problems of homoforms and homonyms, set phrases, combined semantic classification, relation to Old Greek or other parallel and vice versa, relation between the original manuscript and the rendered one, creation of different kinds of indices including the complete one and the selective one.

The outlines of OCS text analysis



STINO ver 2+



Reinhard Lampe:

Создание единого сводного каталога славянского письменного наследия

Ввод - стандартизация и независимость от национальности и родного языка оператора, для эффективного использования компьютерных средств

Если сегодня смотреть на такие отрасли науки как напр. химия, химическая биология, молекулярная химия, тогда увидим, что английский язык принят во всем мире как общий язык для публикаций.

В отличии от этого, в славистике авторы сочиняют и публикуют работы - каждый более или менее на своем родном языке.

В этом нет ничего страшного, потому что слависты друг другу понимают, и каждый в своей работе и в публикациях может использовать материал другого.

Но за последние прибл. 6 - 7 лет ситуация для славистов изменилась: сегодня всем доступна современная техника - компьютеры, так что наконец возможно начинать реализацию предложения, выдвинутого уже в 1954 году: создание *Единого сводного каталога* славянского письменного наследия.

Первый шаг в этом направлении совершила Т. Черторицкая, когда она в 1992 году разработала первую версию программы для базы данных, с помощью которой она подготовила материал для ее *„Предварительного Каталога церковнославянских гомилий подвижного календарного цикла“*[1].

Но во время работы над усовершенствованием и дальнейшим развитием этой программой, автор настоящего доклада столкнулся с некоторыми аспектами и особенностями тематики, которые я хотел бы предложить вниманию участников конференции.

Для выяснения сути проблем, сначала очень короткую экскурсию в „начинку“ компьютеров и несколько слов о принципах работы любой базы данных.

---

[1] „Vorläufiger Katalog Kirchenslavischer Homilien des beweglichen Jahreszyklus“, Aus Handschriften des 11. - 16. Jahrhunderts vorwiegend ostslavischer Provenienz; T. V. Čertorickaja, под редакцией X. Миклас; Abhandlungen der Nordrhein-Westfälischen Akademie der Wissenschaften; © Westdeutscher Verlag, Opladen, BRD; 1994; ISSN 0944-8829; ISBN 3-531-05109-1

Принцип представления текстовой информации внутри компьютера

На экране компьютера засвечивается комбинация точек, которая для наших глаз - и нашего мозга - представляет букву "А". Если оператор этого компьютера "западный" человек, тогда для него нет сомнения - это буква "А". А в "мозге" компьютера нет буквы "А" - только число "65". Только за счет "прикладной программы" вместо числа "65" на экране компьютера получается то, что мы принимаем как букву "А".

Для оператора из, скажем, Болгарии, Сербии или России эта комбинация светлых и темных точек на экране также недвусмыслена - потому что латинская "А" и кириллическая "А" имеют одинаковую форму и выглядят одинаково. Но для компьютера латинская "А" отличается от кириллической "А" как день от ночи.

Это связано с тем, что внутри компьютера латинская буква "А" представляется в виде числа 65, а кириллическая "А" - числа 128 (под ДОС, и числа 192 под Windows). Причина этого заключается в том, что в начале в компьютерах использовались только латинские (точнее английские) буквы, а только за последние годы программисты нашли возможность включить в сегодняшние компьютеры также кириллицу. Для этого программисты сохранили "основной компьютерный алфавит" - который обеспечивает совместимость персональных компьютеров и программное обеспечение во всем мире - и "пожертвовать" такие символы в наборе компьютера, которые не так нужны для общей совместимости, т.е. западноевропейские спецсимволы и подобные.

А предположим теперь, что оператор Вадим ввел в компьютер слово "Авраам"; по какой-то причине он начинал печатать большую латинскую "А" (число 65) и остальные буквы ввел на кириллице, и просто не заметил - или забыл -, что первая буква - латинская. Через какое-то время оператор Дима хочет провести какую-то исследовательскую работу над данными, введенными в компьютер, и задает команду "Найди мне все слова, которые начинаются через „Авра“ ", и вводит в компьютер искомое, используя кириллическую большую "А" (число 128 для компьютера).

Результат: компьютер найдет все слова, начинающиеся через „Авра“ - за исключением введенного Вадимом "Авраам", - просто потому, что этот "Авраам" по числовому представлению внутри компьютера не совпадает с введенным Димой искомым (разница между латиницей и кириллицей).

Такая проблема еще относительно легко решима - достаточно было бы в программу включить так называемый алгорифм, который не допускает в одном слове комбинацию из латинских и русских букв,

за то самостоятельно - и незаметно для оператора - напр. определяет преобладание тех или других буkв и заменяет ошибочные, т.е. в большинстве случаев латинские, на правильные - русские буkвы.

Но проблемы, подобные нарисованных, становятся значительно сложнее, если они рассматриваются в связи с базой данных.

Принцип базы данных и индеkсов для выборки

Каждый из Вас имеет книжечку, в которую записывает адрес и телефон знаkомых, друзей, коллег и проч. Этот списоk адресов обычно имеет одну или 2 страницы для каждой буkвы алфавита, позволяющие kak-то организовать адресы.

В эту книжку для каждого знаkомого или коллеги обычно записываются следующие данные:

- фамилия, имя, отчество
- адрес (улицу, дом, вход и т.д.)
- страна, индеkс города и город
- телефон: kод страны, абонентский номер

Для каждой фамилии в книжку записывается один таkой блоk информации - одна запись. При этом записи поступают хронологически - и только при переписывании всех этих данных в новую книжку получается возможность "реорганизовать" записи таk, чтобы существующие были по алфавиту.

Предположим, что в нашей книжkе существуют следующие адреса:

Аврамов, Гена Генадевич
1й западный переулоk, д. 3, вх. Б, апп. 53
РФ, Изюмгород, 456789
007-54-321098

Ацимутов, Артур Балдурович
Мария Луиза, 369/12
BG - 1234 София
00359-2-765432

Антонов, Юри
Ломопарk 13
YU - 1000 Београд
00381-1-452662

Адреналинова, Паникида Лихорадовна
Северная Котельная, 71
Москва

Veder, William
Spuistraat 210
NL - 1012 VT Amsterdam
0031-20-525 3070

Такой список адресов уже является базой данных. Но так как эта база данных заполняется с временем, записи не появятся строго по алфавиту, а так, как поступили. Следовательно необходимо прочитать напр. все "А" от начала, чтобы найти желаемого адреса.

Компьютерная база данных - в принципе то же самое. Только в компьютере такую базу данных можно себе представить как таблицу, в которой каждый ряд содержит данные одного адреса в определенной последовательности:

1 Фам.	Имя	Улица	2 Стр.	Инд.	3 Город	# зп.
Аврамов	Г.	1й зап. пер.	РФ	...	Изюмгород	01
Veder	W.	Spuistraat	NL	...	Amsterdam	02
Ацимутов	А.	Мария Луиза	BG	...	София	03
Антонов	Ю.	Ломопарк 13	YU	...	Београд	04
Адреналинова	П.	Сев.Кот.			Москва	05

В этой таблице записи также появляются приблизительно в хронологической последовательности их поступления, т.е. ввода в компьютер. Но для компьютерной базы данных это не страшно, потому что программы, с помощью которых создаются такие базы данных, сами по себя содержат специальную функцию "индексирования".

Функция "индексирования" позволяет, уже при создании структуры базы данных, т.е. в самом начальном этапе программирования, выбрать т.н. "поле данных" (графу в таблице), по содержимому которого выбрать и показать записи базы данных.

Такой "индекс" хранится в дополнительном файле, рядом с основным файлом с настоящими данными, и каждый раз, когда новая запись введена в основную базу данных, файл индекса актуализируется автоматически: в моменте, когда добавляется новая запись в базу данных, добавляется и запись в этот файл индекса, а именно значимое "индексного поля данных", и (физический) номер записи в основном файле базы данных. При этом файл индекса автоматически сортируется по значимому индексного поля данных - без воздействия того человека, который ввел новую запись ("оператора").

В результате этого, если напр. поле данных "фамилия" выбрано как индекс (графа 1 вышеприведенного примера), тогда все адресы в базе данных показываются по строгому алфавитному порядку фамилий:

Значимое индекса	Нр. записи
Veder	02
Аврамов	01
Адреналинова	05
Антонов	04
Ацимутов	03

Последовательность записей				
Veder	...	NL	...	Amsterdam
Аврамов	...	РФ	...	Изюмгород
Адреналинова	...		...	Москва
Антонов	...	YU	...	Београд
Ацимутов	...	BG	...	София

Если в приведенном примере выбрать **индекс 2 - Страна -**, тогда записи показываются в такой последовательности:

Значимое индекса	Номер записи
[пробелы]	05
BG	03
NL	05
YU	04
РФ	01

Последовательность записей				
Адреналинова	...		...	Москва
Ацимутов	...	BG	...	София
Адреналинова	...	NL	...	Amsterdam
Антонов	...	YU	...	Београд
Аврамов	...	РФ	...	Изюмгород

Так как любимая теща Адреналинова в базу данных введена без кода страны, в ее записи это поле содержит только пробелы (в памяти компьютера „буква“ пробела представляется в виде числа 32), она в сортированной по коду страны базе данных появляется в первой позиции.

А **индекс 3 (по городу)** обуславливает следующую последовательность:

Значимое индекса	Номер записи
Amsterdam	02
Београд	04
Изюмгород	01
Москва	05
София	03

Последовательность записей				
Veder	...	NL	...	Amsterdam
Антонов	...	YU	...	Београд
Аврамов	...	РФ	...	Изюмгород
Адреналинова	...		...	Москва
Ацимутов	...	BG	...	София

Вы видите, что последовательность фамилий для каждого индекса разная: сначала появляются записи, в которых индексное поле пустое, т.е. содержимое которого начинается пробелами (индекс 2: по стране - Адреналинова); потом записи, значимое которых введено

латинскими буквами, и в конце записи, значимое которых введено кириллическими буквами (индекс 2: страна - Аврамов).

Теперь предположим, что по какой-то причине мы случайно ввели в нашу базу данных Аврамова через латинскую большую „A“, тогда индекс и следовательно и последовательность записей были бы следующие:

Значимое индекса	Номер записи
Аврамов	01
Veder	02
Адреналинова	05
Антонов	04
Ацимутов	03

Последовательность записей				
Аврамов	...	РФ	...	Изюмгород
Veder	...	NL	...	Amsterdam
Адреналинова	...		...	Москва
Антонов	...	YU	...	Београд
Ацимутов	...	BG	...	София

А возможный т.н. „фильтр“, с помощью которого выбрать только те записи, в которых фамилия начинается через большую кириллическую „А“, в этом случае выберет только Адреналинову, Антонова и Ацимутова, а Аврамов был бы исключен из поиска, так как первая буква этой фамилии не удовлетворяет условиям фильтра.

Смысл таких индексов - обеспечить независимость от хронологии ввода информации, и одновременно значительно ускорить вывод и выборку записей в данной последовательности.

Ситуация для каталога

Для конкретного проекта создания Единого Сводного Каталога славянских текстов и рукописей ситуация представляется следующим образом:

В упомянутой выше программе Т. Черторицкой, которая отлично годится как основа для создания такого каталога, для идентификации индивидуальных сочинений служат между прочим название автора, название самого сочинения, начало сочинения (и концовка сочинения).

Но все эти литературные произведения написаны не на одном из современных славянских языков, а на "общем" старом языке и старым письмом - на старой кириллице или на глаголице.

В данном моменте упомянутая программа позволяет ввести все эти данные или на современной („гражданской“) кириллице, или на латыне - но без умлаутов (ä, ö, ü), гласных с акцентом (à, é, ô), и других национальных специальных букв и символов.

Takoe ограничение набора символов, однako, требует транслитерации старых текстов на современный набор символов. Те варианты, которые нам предлагают SGML и TEI, отлично годятся kak стандарт для обмена данными между разными прикладными программами, системами, также по международным сетям kak Интернет. Для обработки внутри компьютера, внутри прикладной программы, представление текстовой информации отлично пригодна, чтобы не сказать является единственной возможностью, потому обрабатываемый материал с временем поступит из разнейших источников, использующих все одинаковую систему „кодировки".

Но для среднего оператора такая „линейная" кодировка для повседневной работы не годится, потому что она не „наглядна", не сразу понятна - она не „читается".

Для оператора-человека гораздо легче, если информация, выдаваемая на экран компьютера, была в „человеческом" виде, т.е. в виде открытого текста, по набору символов сразу доступного ему.

В этой связи для Единого Сводного Каталога существуют только два варианта: или все текстовые данные транслитерировать на АСКИИ[2] и современную кириллицу; или заставить программистов реализовать возможность ввода текстовой информации в базу данных Каталога на старой кириллице.

Транслитерация вызывает новые затруднения: в то время kak современная кириллица состоит из 32 букв, в Глаголице их 40, а в старой кириллице не менее чем 64 буквы с вариантами!

Следовательно требовалось бы согласовать "таблицу транслитерации" и установить стандарт для транслитерации старой кириллицы kak и глаголицы на современную кириллицу.

Сложнее будет решение вопроса, какой вариант современной кириллицы использовать. Мне пока известны два варианта: русский и болгарский - которые на уровне операционной системы ДОС несовместимы из-за разного размещения букв по кодам АСКИИ (именно букв „п" по „я"). А установление одного из этих двух вариантов kak общий стандарт требовало бы серьезных изменений в базовой конфигурации компьютеров и принтеров - в пользу только одной прикладной программы!

Очевидно единственный выход из положения - предусмотреть ввод данных в Единый сводный каталог славянского письменного наследия kak можно ближе к оригиналу, т.е. на старой кириллице.

---

[2] ASCII - American Standard Code for Information Interchange; в этом коде существуют только цифры, знаки препинания, и нормальные латинские буквы, используюемые для английского языка.

На мой взгляд в Каталоге можно спокойно обойтись без разных (местных) вариантов букв старой кириллицы (для автоматического определения возможных местных расхождений и отличий между переписями одного и того же самого сочинения должны использоваться другие программы), как и без большинства надстрочных букв и символов.

Такое решение значительно упростило бы задачу стандартизации ввода ключевых данных, как напр. имени автора, хотя оно не полностью исключило бы все возможные расхождения.

Для выяснения такой пример: „Иоан Златоуст". Даже если ввести это имя в базу данных Каталога на старой кириллице, все еще возможно что один оператор пишет „Иоан", а другой „Йоан". А если третий оператор введет „John Chrysostom" - тогда без сложных дополнительных мероприятий, которые имеют отрицательные эффекты на быстродействие программы базы данных, невозможно выбрать из Каталога все сочинения, приписанные Иоану Златоусту.

Аналогичную ситуацию можно предвидеть для описания рукописей:

Относительно легко еще решить вопрос о филигранях: для филиграней есть альбомы, в которых они подробно описаны. В принципе достаточно сканировать эти альбомы один раз, получая таким образом в компьютере изображения каталогизированных филиграней. После этого при вводе кодикологической информации возможно предлагать оператору логически структурированный выбор филиграней, из которого оператор выберет соответствующее изображение. Незаметно для оператора в самую базу данных, т.е. в файлы, тогда вставляется только индикатор альбома и номер филиграни по данному альбому. А во время извлечения информации из базы данных, напр. для подготовки публикации, программа на основе этой кодовой информации выберет из базы данных все рукописи, в которых данная филигрань встречалась, и переведет эту числовую информацию в открытый текст на желаемом языке.

Такой подход вполне реалистичен для всех параметров с ограниченным количеством возможных значимых, как например:

- тип уст ва
- пишущий инструмент и цвет(ы) чернила
- пагинация, разлиновка

и подобные.

Но как, например, ввести названия сборника: Codex Suprasliensis, Супрасльский сборник, или Супрасълски сборник?

Название сборника как раз является ключевой информацией - оно служит для индексированной выборки подмножества записей из всей базы данных.

А конкретное содержимое таких ключевых полей, как мы видели (см. выше, „Принцип работы базы данных"), имеет самое решающее значение для эффективности использования базы данных Каталога.

Решить вопросы, связанные с стандартизацией ввода ключевой информации в базу данных, не может программист - в этом отношении он полностью зависит от того, чтобы пользователи разрабатываемой им программой добились договоренности.

И без такой договоренности Вы можете только составить разные национальные каталоги, которые - если смотреть на них глазами компьютера - как в прошлом, так и в будущем несовместимы друг с другом, а **ЕДИНЫЙ СВОДНЫЙ КАТАЛОГ**, достойный таким названием, остается вечной мечтой славистов,

Ныне, и Присно, и во Веки Веков.

Международный исследовательский проект "Slavia Orthodoxa et Slavia Romana: Взаимодействие славянских миров: Духовная культура Подляшья"

Лабынцев Юрий Андреевич (Россия 117334, Москва, Ленинский пр., 32-А Институт славяноведения и балканистики РАН)

Восточная Европа и Западная Европа, славяне, духовная культура, рукописи, старопечатные книги, компьютерные средства

Настоящий Проект позволит на примере одного обширного региона в достаточной полноте проследить почти тысячелетнюю историю взаимодействия западного и восточного христианских миров, раскрыть его сущность и механизмы, характерные для всей Центрально-Восточной Европы, создать уникальную источниковую базу - основу для дальнейших теоретических исследований. В результате работы над Проектом предполагается также создать оригинальное многотомное издание энциклопедического типа, включающее исчерпывающую информацию по истории книжной культуры этого региона. Согласно принятой методике все исследования ведутся с использованием компьютерных средств. При этом гарантируется не только создание общего банка данных, но и его надежное дублирование для издательского комплекса Проекта в Минске и головного в Москве.

Вопрос об использовании достижений информатики в гуманитарных исследованиях достаточно нов не только в отечественной, но и в мировой науке. Принципиально важно, какие конкретные цели и задачи ставят те или иные ученые-гуманитарии, как относятся к этому уникальному инструментарию, что они могут сами внести в необычайно динамичный процесс его совершенствования.

Во всем мире растет понимание того, что на наших глазах возникает новая информационная культура, проявления которой наиболее заметны в развитых странах. Определенные успехи в сложении этой культуры принадлежат и ученым-гуманитариям, причем в их среде глобальная научная интеграция в последнее время достигла самого высокого уровня. В качестве примера можно привести успешно действующую Международную ассоциацию "History&Computing", основанную в 1986 г. и ныне насчитывающую около 1000 членов почти в 20 национальных ассоциациях.

В 1993 г. в Институте славяноведения и балканистики был создан Центр украиноведческих и белорусоведческих исследований (преобразован в 1994 г. в Центр украинистики и Центр белорусоведческих исследований), призванный создать в системе Российской Академии наук два новых направления фундаментальных научных работ. Изначально предполагалось, что исследования в Центре будут вестись с использованием компьютерных средств.

Одной из наиболее сложных фундаментальных научных задач, решаемых Центром белорусоведческих исследований с помощью электронного инструментария, является работа над Международным исследовательским проектом "Slavia Orthodoxa et Slavia Romana: Взаимодействие славянских миров: Духовная культура Подляшья".

Граница западного и восточного христианских миров не только пролегает серез славянские земли, как бы деля славян, разделяя их, но и способствует образованию весьма обширной зоны их многовекового наиболее интенсивного взаимодействия - предграничья и пограничья.

Достойное существование в новой Европе, а тем более единство, межнациональное и национальное, невозможны без взаимопонимания особенностей духовного развития различных этносов. Все это определяет особые приоритеты в изучении труднейших проблем взаимоотношений Slavia Orthodoxa - Slavia Romana.

Вместе с тем, само состояние современной европейской цивилизации, а главное ее будущность в значительной мере определяются степенью понимания и решения сложных вопросов, связанных с процессом сопряжения двух христианских миров, в котором славянам ныне отведена главенствующая роль.

Настоящий Проект позволяет на примере одного региона в достаточной полноте проследить почти тысячелетнюю историю такого взаимодействия, раскрыть его сущность и механизмы, характерные для всей Центрально-Восточной Европы, создать уникальную источниковую базу - основу дальнейших теоретических исследований.

Подляшье - наибольшая часть пограничья Центральной и Восточной Европы - является уникальным историческим заповедником славянских культур, огромной, существующей столетия лабораторией их взимопроникновения. Одновременно, Подляшье - великое пограничье славянских и балтийских культур и то редкостное место, где культуры славянского Запада,Юга и славянского Востока, равно как западноевропейские и восточнославянские культуры, приходили в непосредственное самое тесное взаимодействие на протяжении многих веков. Наконец, Подляшье - это крайняя западная область распространения восточнославянских культур, западная граница их существования.

Проект предусматривает исследование прежде всего письменной, книжной культуры Подляшья,являющей собой наиболее концентрированное выражение главнейших черт духовных культур нескольких славянских и неславянских народов, связанных с судьбами этого края, прежде всего белорусов, поляков, украинцев, литовцев и русских.

Исторически сложилось так, что на Подляшье /в Супрасле/ действовал знаменитый издательский центр, работа которого оказала огромное влияние на развитие белорусской, польской, русской и украинской культур[1].

Супрасль был и одним из крупнейших очагов православия, славянской книжности, восточнославянских духовных культур, самым значительным западным интеллектуальным центром православия вообще.

В XVIII столетии (до 1790-х гг.) Супрасль являлся и самым значительным центром книгоиздания для русских старообрядцев[2]. Книги, напечатанные в типографии монастыря старообрядцами, расходились потом по всей России. Эти супрасльские старообрядческие издания и теперь продолжают свою долгую жизнь - их можно встретить на таежных заимках Сибири, в библиотеках старообрядцев мно гомиллионной Москвы и маленького городка Браслава в Белоруссии, в других местах, в том числе и в Болгарии.

Изучение, вернее, предварительное обследование состава памятников старопечатной кирилловской литературы, только начинается. Они содержат огромное количество самых разнобразных и подчас совершенно неизвестных произведений, представляющих первостепенный интерес. Можно с уверенностью сказать, что до сих пор старообрядческие издания XVIII столетия наряду с изданиями XIX - XX вв., действительное число которых нам пока еще неизвестно, являются настящей terra incognita, открытию которой может способствовать лишь целенаправленная многолетняя работа ученых различных специальностей.

С конца XVII столетия сотни униатских изданий на нескольких языках, в том числе церковнославянском, напечатанные на Подляшье, расходились во многих тысячах экземпляров среди народов Восточной Европы. Некоторые из

этих изданий позднее неоднократно перепечатывались в других местах. И в униатской среде в XVII и значительной части XVIII столетия лидирующее положение в книгоиздании занимал Супрасль. Супрасльским базилианам через книги суждено было в огромной мере, а иногда и в определяющей, влиять на ход жизни всех их восточноевропейских провинций, большинства униатских епархий.

Униатский мир был заполнен супрасльской книжностью в еще большей степени, чем мир старообрядческий.

Сложность и взаимообусловленность литературного развития в среде различных христианских конфессий, направлений и движений на Подляшье может быть проиллюстрирована на одном небольшом примере, связанном с бытованием здесь памятников кирилло-мефодиевского цикла[3], в том числе и в униатской среде, включая и время, когда в конце XVII в. униатами стали создаваться особые интерпретации на основе Жития Кирилла и Жития Мефодия, разошедшиеся потом по всем епархиям в тысячах печатных экземпляров, ни один из которых, к сожалению, не стал известен нынешним специалистам[4].

Иной пример - знаменитый супрасльский "славено"-польский Лексикон 1722 г., история создания, издания и рецепции которого как нельзя удачнее подводит нас к верному пониманию особенностей культурного развития восточнославянских народов в составе Речи Посполитой.

В последнее время появились сообщения о сенсационных находках местных памятников эпиграфики, позволяющих отнести время появления христианства на землях восточной Польши на целое столетие ранее[5]. Найденные неподалеку от западной границы Подляшья глиняные таблички с греческими инициалами возможно свидетельствуют о том, что христиане жили здесь уже в IX в. и пользовались греческим письмом.

В процессе работы над Проектом возникает множество специфических трудностей, прежде всего связанных с особенностями исторического развития этих земель. Например, при сборе материалов, касающихся книжных центров южной части Подляшья, издревле населенной украинцами.

Имевшая в древности богатые книжные традиции Холмщина и южное Подляшье переживают затем трагические периоды вплоть до событий 1938 г., когда здесь было ликвидировано и уничтожено около 200 церквей[6], и массового выселения украинцев в послевоенные годы[6].

В книжных собраниях мира в настоящий момент зафиксированы немногочисленные памятники книжности, происходящие с южного Подляшья. Более всего их ныне находится в Польше (прежде всего в Люблине). Есть надежда, что со временем многое удастся выявить и в богатейших хранилищах Киева, Львова, Москвы.

К сожалению, практически полная утрата древней книжности коснулась и знаменитый Свято-Онуфриевский монастырь в Яблочине. А ведь в этом православном монастыре исследователи XIX - начала XX вв. видели значительную библиотеку, наполненную старинными рукописями[8], которые затем были отсюда увезены, а потом почти бесследно исчезли.

В результате работы над Проектом наряду с теоретическими исследованиями предполагается создать оригинальное многотомное издание энциклопедического типа, включающее в себя практически полную информацию обо всех сторонах культурного процесса на Подляшье, связанного с рукописной и печатной книжностью на церковнославянском, древнерусском,старобелорусском, староукраинском, белорусском, украинском, русском, польском, латинском и литовском языках. Эта информационно-аналитическая часть Проекта имеет самостоятельное значение, работа над ней ведется по специальной программе "История книжной культуры Подляшья".

Издание будет состоять из семи томов: 1.Литература по истории и культуре Подляшья. 2.Кирилловские издания типографий Подляшья. 3.Издания латинского шрифта типографий Подляшья. 4.Рукописная книжность Подляшья/в большинстве она сберегласъ - это редчайший случай в мировой практике вообще; основная часть этой книжности, почти исключительной кирилловской, ныне находится в Вильнюсе - бывшей столице Великого княжества Литовского,

Русского, Жемойтского, куда ее вывезли с Подляшья в середине XIX века [9]/. 5.Исторические книжные центры и собрания Подляшья. 6.Документальные и актовые источники по истории и культуре Подляшья. 7.Указатели к шести предыдущим томам.

Реализованный в своей информационно-аналитической части Проект станет уникальной основой для самых широких дальнейших исследований: исторических, культурологических, литературоведческих, религиоведческих. В том числе общего и частного моделирования историко-культурных процессов, происходивших в странах Центральной и Восточной Европы в течении нескольких столетий, особенно, если речь идет о Польше, Белоруссии, Украине и Литве, о взаимосвязях культур этих государств, их истории.

Основные учреждения-исполнители Проекта: Российская Академия Наук - Институт славяноведения и балканистики; Министерство культуры Республики Беларусь - Национальная библиотека Беларуси. Общее руководство осуществляет Институт славяноведения и балканистики Российской Академии Наук.

Издание будет выпускаться на полиграфической базе Национальной библиотеки Беларуси.

Учитывая чрезвычайную сложность решаемых проблем, почти полную неразработанность многих вопросов, связанных с историей культуры славянского мира, особенно если речь идет о пограничьи славянских и не славянских культур, взаимодействии греко-славянской и латино-славянской культурных традиций, образовании своеобразного культурного конгломерата, который в наиболее ярких формах существовал и существует именно на Подляшье, участники работы над Проектом пришли к единодушному решению о необходимости издания специальной серии вспомогательных пособий, призванных помочь в реализации общих и частных задач Проекта.

Значение этих пособий, по-видимому, окажется более широким. Они будут весьма полезны историкам, филологам, философам, книговедам,

библиотечным работникам и другим специалистам различных стран Европы. К настоящему времени опубликовано два подобных пособия[10].

Согласно принятой методике все работы ведутся с использованием новейших электронных технологий. При этом гарантируется не только создание общего банка данных, но и его надежное дублирование для издательского комплекса Проекта в Минске и головного в Москве. Подобное решение позволяет также оперативно вводить в оборот любую информацию, поступающую от соисполнителей из Польши, Белоруссии, Литвы, России, с Украины и других мест.

1. Лабынцев Ю.А. Кирилловские издания Супрасльской типографии. М., 1978; Он же. Памятники древнерусской книжности и литературы в старообрядческих изданиях Супрасльской типографии (XVIII в.) //ТОДРЛ. Л., 1981. Т.36. С.209-215 и др.; Pidłypczak-Majerowicz M. Bazylianie w Koronie i na Litwie: Szkoły i książki w działalności zakonu. Warszawa -Wrocław, 1986; Cubrzyńska-Leonarczyk M. Oficyna supraska: 1695 -1803. Dzieje i publikacje unickiej drukarni ojców bazylianów. Warszawa, 1993.

2. Лабынцев Ю.А. Памятники древнерусской письменности в изданиях старообрядчеких типографий XVIII -XX вв. //Книга: Исслед. и материалы.М., 1979. Т.39. С.173-178; Он же. Некоторые вопросы кирилловского книгопечатания в Супрасле // Федоровские чтения 1978 М., 1981. С.168-179; Вознесенский А.В. Кириллические издания старообрядческих типографий конца XVIII -начала XIX века: Каталог. Л., 1991; Починская И.В. Книгопечатание для старообрядцев на землях Белоруссии и Литвы // Наш радавод: Материалы Международной научной конференции по региональной истории Восточной Европы "Культура народов Великого княжества Литовского и Белоруссии. XIII-нач.XX в.в.". Гродно, 1991. Кн.3.4.2. С.268-274.

3. Лабынцев Ю.А. Супрасльский список пространного Жития Кирилла Философа 1580 г. // Болгарская культура в веках: Тезисы докладов научной конференции (Москва, 26-27 мая 1992 г.). М., 1992. С.15-17.

4. Ангелов Б. Кирил и Методий в славянските печатни книги от XV -XVII в. // Ангелов Б.Из историята на старобългарската и възрожденската литература. София, 1977. С.5-23.

5. Leslak A., Talko L. Mieszko nie był pierwszy // Gazeta wyborcza. 1991, 16 sierpnia. (N 13(190)). S.1.

6. Kiryłowicz S. Niektóre problemy prawosławia na tle polityki wyznaniowej Państwa w okresie międzywojennym // Posłannictwo: Kwartalnik teologiczno-filozoficzny. 1979. N 3-4. S.119-121.

7. Храпливий Е. Нарис історії Холмщини й Підляшшя. Вінніпег, 1968; Кубійович В. Українці в Генеральній Губернії (1939-1941). Чікаго, 1975.

8. Гербачевский Ф. Русские древности и памятники православия Холмско-Подляшской Руси (Люблинской и Седлецкой губерний). Варшава, 1892; Григорьев А. К истории Яблочинского Свято-Онуфриевского монастыря Холмской губернии. Варшава, 1915.

9. Рукописи эти были описаны Ф.Н.Добрянским. См.: Добрянский Ф. Описание рукописей Виленской Публичной Библиотеки церковно-славянских и русских. Вильна, 1882.

10.Толкачева(Щавинская) Л.Л.Никита Иванович Горбачевский. Минск, 1991; Лабынцев Ю.А., Щавинская Л.Л. Православная церковь на Белостотчине в работах протоиерея Григория Сосны. Минск, 1994.

Основные задачи и принципы построения банка данных Международной славяноведческой программы "История книжной культуры Подляшья"

Щавинская Лариса Леонидовна (Россия, 117334 Москва, Ленинский пр., 32-А Институт славяноведения и балканистики РАН)

Информационная система, банк данных, компьютерные средства, Восточная Европа и Западная Европа, славяне, духовная культура, рукописи, старопечатные издания

Основа создания банка данных Международной славяноведческой программы - ее источниковая база, состоящая из многих тысяч разнообразных источников (рукописи, старопечатные книги, акты, грамоты, деловые документы и др.), хранящихся в различных странах. Банк данных представляет собой информационную систему, включающую в себя разные типы баз данных. При создании банка данных предполагается использование реляционной базы данных типа FoxPro, Paradox с достаточно высоким быстродействием, что весьма важно, учитывая объем баз данных. Для работы с такими источниками, как акты, документы, грамоты предполагается использовать источнико-ориентированную базу данных типа КЛИО. По мере накопления информации в банке данных будут вестись теоретические исследования, которые дадут возможность смоделировать и проследить процессы взаимосвязей, взаимопроникновения различных культур.

Одной из основных фундаментальных задач, решаемых Центром белорусоведческих исследований Института славяноведения и балканистики Российской Академии наук , является создание банка данных Международной славяноведческой программы "История книжной культуры Подляшья".

Термин "банк данных" очень многозначный и трактовать его можно по-разному, например, как совокупность данных (файлов, баз данных) об одной предметной области, как автоматизированная информационная система централизованного хранения и использования данных, как совокупность однотипных элементов, средств или устройств, взаимосоединенных и совместно используемых.

В нашем случае банк данных представляет собой информационную систему, построенную с помощью системы управления баз данных, состоящую из совокупности разнородных баз данных, системы управления базами данных, вычислительной системы.

При создании любой информационной системы наполнению ее конкретными данными предшествует этап создания концептуальной модели информационной системы, определения и описания предметной области исследования. База данных - целенаправленная модель предмета исследования и в ней находят отражение только те сведения о ней, которые необходимы для заданных целей. Прежде всего решается вопрос, какая информация будет обрабатываться, какие ее характеристики нужно учесть, какие взаимосвязи между ними будут учитываться. Концептуальное моделирование источников информации связано с выявлением и оценкой информационного потенциала разных типов источников, которые будут использоваться для создания и организации физической базы данных на компьютере[1].

Этот этап работы над созданием банка данных представляется нам одним из наиболее важных, ибо хорошо продуманная структура организации банка данных позволит в дальнейшем избежать или значительно уменьшить ее избыточность, в то же время позволяя вводить и хранить всю необходимую

информацию из источника с учетом его специфики и в зависимости от аспекта и цели его использования. Однако, нужно учесть, что зачастую при этом используется не вся заложенная в источнике информация, а лишь выборочная, исходя из целей исследования. Здесь в каждом конретном случае исследователем решается проблема необходимости и достаточности вводимой информации.

В настоящее время мы находимся на этапе разработки концептуальной модели нашей информационной системы, разработки структур конкретных баз данных. Этот процесс идет параллельно и напрямую связан с весьма трудоемкой работой по выявлению и описанию разнородных полилингвистичных источников, хранящихся в различных странах мира.

Создание столь сложной информационной системы невозможно без наличия хорошего технического оборудования, наличия современной техники. Благодаря финансовой поддержке Российского гуманитарного научного фонда наш Центр получил возможность начать создание многофункциональной технологической системы, обеспечивающей проведение фундаментальных исследований на современном уровне. Теперь мы имеем современный мультимедиа-компьютер IBM с развитой периферией с установленным CD-драйвером, звуковой картой, что позволяет нам хранить в базах данных изображения источников, а также вводить звуковое сопровождение. Кроме того, мы получили доступ в мировую компьютерную сеть Internet, что дает нам огромные возможности в налаживании информационного обмена с научными учреждениями и организациями Европы и мира.

Новые современные коммуникационные связи делают возможным распространение информации новыми средствами. В настоящее время во всем мире наблюдается устойчивая тенденция роста числа электронных форм хранения источников на CD-ROM. В принципе, CD-ROM сравним с книжной публикацией, ибо он хранит информацию статического характера, которую можно только читать. В коммуникационной сети информация имеет динамические свойства, здесь возможны добавления и изменения информации. Internet, сеть сетей, дает

исследователям возможность доступа к сотням тысяч баз данных во всем мире. Конечно, при этом следует иметь ввиду, что получение доступа к сети не обзачает автоматически получение доступа к любой информации[4].

В основу создания банка данных Международной славяноведческой программы "История книжной культуры Подляшья" положена источниковая база, представляющая собой совокупность самых разных типов источников: рукописи и старопечатные книги, деловые документы, акты, грамоты и другие типы на разных языках. В соответствии с типом источника банк данных будет состоять из 7 основных комплексов баз данных: 1) литература по истории и культуре Подляшья; 2) кирилловские издания типографий Подляшья; 3) издания латинского шрифта типографий Подляшья; 4) рукописная книжность Подляшья; 5) исторические книжные центры и собрания Подляшья; 6)документальные и актовые источники по истории и культуре Подляшья; 7)указатели. Это условное деление соответствует структуре нашей издательской части Проекта. Оно необходимо для того, чтобы максимально полно описать каждый тип источника, учитывая при этом, что ограничения системы описания неизбежны и зависят в каждом конкретном случае от целей исследования.

В настоящее время практически завершено создание базы данных литературы по истории и культуре Подляшья. Первый том издания должен выйти в свет в следующем году. Эта база является базой данных реляционного типа, реализована она с помощью пакета FoxPro, данные в ней представлены в виде таблиц со взаимосвязанными ключами. Выборка информации производится по связанным ключам.

Базы данных по рукописной и печатной книжности, кирилло- и латинографичной, нами предполагается реализовать на основе реляционной базы данных типа PARADOX, которая позволяет хранить, изображать и представлять информацию разными способами. Данные PARADOX хранит в табличной форме. Для хранения данных возможно использование полей разных типов: текстовых, числовых, дат, графических, двоичных (обычно используются для хранения

звука), Мемо-полей, форматированных и неформатированных, которые представляют собой текстовые поля переменной длины, ограниченные только емкостью жесткого диска. В Мемо-полях можно хранить любые примечания, необходимую информацию из источника, которую нельзя структурировать без больших потерь. Мемо- поля можно обрабатывать текстовыми редакторами. Эта база обладает достаточно высоким быстродействием, возможностью строить самые разнообразные графики и диаграммы, легко работать со взаимосвязаными таблицами.

При создании базы данных по рукописной книжности необходимо учитывать, что каждая рукопись сама по себе является своеобразной информационной системой, которая может быть описана в самых разных аспектах и взаимосвязях с различной полнотой и глубиной раскрытия информации. Обладая общими свойствами письменных источников рукописная книга представляет собой особую многофункциональную культурологическую систему, которая содержит гигантскую по объему и разнородную информацию, существующую в свернутом виде.[2].

Виду того, что в основу банка данных положена огромная источниковая база, неизменно встает проблема перевода источников в компьютерную форму. При переводе источников в электронную форму возможны следующие варианты: а)когда осуществляется полный перевод - для очень важных источников; б)сканирование и последующее распознавание; в)сканирование и последующее хранение образов информации. Для нас представляются предпочтительными два последних варианта. Выбор каждого конкретного варианта зависит от типа источника, целей создания баз данных , технических и материальных возможностей исследователя. В настоящее время большой интерес вызывает вопрос использования программ оптического распознавания (OCR) для перевода источника в электронную форму. Программы оптического распознавания ведут свое происхождение от первых экспериментов по распознаванию образов, ведущихся с 50-х годов в рамках работ по созданию искусственного интеллекта.

Наиболее распространенные из них нуждаются в обучении и теоретически могут быть обучены пониманию любой графики и любого шрифта. В набор функций OCR-программ входит умения выделить в сканированном изображении области текста, разбить на строки, стороки - на отдельные символы, распознать символы в соответствии с имеющимися эталонами, попросить пользователя научить пониманию, затем полученный эталон можно сохранить для дальнейшего использования. При этом программа устанавливает соответствие между графическим образом буквы и символом из компьютерного алфавита - ASCII-кодов. Мы, в частности, для распознавания источников использовали систему FineReader. Это система распознавания с возможностью обучения, работает в среде Windows.

Базу данных по документальным и актовым источникам мы предполагаем строить на основе источнико-ориентированной системы управления базы данных типа КЛИО, ибо база данных предполагает струкрурирование источника, а структурирование источников такого типа влечет за собой неизбежные большие потери нужной информации. Нами предполагается, что наиболее важные источники, документы будут хранится в виде изображении на CD-ROM, т.е. нами предусматривается создание библиотеки изображений с системой поиска, просмотра, вывода на другие носители. Все это, конечно, можно будет реализовать только при наличии всех необходимых условий.

Таким образом, банк данных Международной славяноведческой программы "История книжной культуры Подляшья" является интегральной информационной системой, представляющей собой совокупность разных типов баз данных, систему управления базами данных, комплекс вычислительных средств.

Банк данных постепенно наполняется по мере накопления информации из архивов, библиотек, музеев и т.д. ряда стран. На основе введенных данных будут проведены и уже проводятся различные теоретические исследования, которые

адут возможность смоделировать и детально проследить процессы взаимосвязей, заимопроникновения европейских культур на протяжении многих столетий.

1. Гарскова И.М.. Базы и банки данных в историчнских исследованиях. М., 1994.

2. Дубровіна Л. Кодикологія та кодикографія української рукописної книги. Київ, 1992.

3. Бородкин Л.И., Лазарев В.В. История и компьютер: к новому modus vivendi // Компьютер и исторические знания. Барнаул, 1995. С.11 -35.

4. Доорн П. Электронный лабиринт: Возможности и "ловушки" компьютерной системы Internet для историков // Круг идей: Развитие исторической информатики. М., 1995. С.122 -135.

THE PSALTER: FROM THE ORAL TO THE INFORMATIC TRADITION

(A study to encode Early Slavonic texts)

Monia Camuglia, Via Guerrazzi n.2
56010 Madonna dell'Acqua (PISA) - Italy

ABSTRACT: I describe my experience concerning the use of computational systems for the study of ancient Slavonic texts. I used these systems for my degree thesis, which consisted in the search of examples attesting the unitariness of the Psalter in Slavonic countries. The aim of my project was to use the text processing methods previously experimented on texts in the Latin alphabet.

In this brief report I would like to describe the experience I made when preparing my degree thesis on Slavic Philology, with regard to the codification of texts written in the ancient Slavonic alphabet, to be submitted to computerized systems of textual analysis. My interest in this type of study was stimulated both by the fact that, to my knowledge, very little research had been carried out in this sector in Italy, and that the main Italian centre using the computer for humanistic studies was in Pisa. This centre is the Institute for Computational Linguistics (ILC) of the Italian National Research Council that has in fact been working in this sector for about thirty years and has conducted pioneer studies exploiting the know-how and technology at their best.

Before entering into details, I wish to point out that unlike other disciplines, Slavistic studies in Italy have only just started to be carried out by means of computational systems of analysis, owing to the particular problems the texts to be treated present, with regard to the technological methods available. My work does not intend to be the best possible solution to

the many difficulties presented by ancient Slavonic texts but only derives from the attempt to create a different way of looking at and analysing this material and to offer a minimal contribution in this direction. Among the many projects developed at the Institute for Computational Linguistics, the DBT (Data Base Testuale), a program running on a PC MS-DOS environment, designed and implemented by Dr. Eugenio Picchi, stands out for its power and ability to provide and to manage a large number of data. The program is able to perform rapidly a large number of operations including lists of alphabetical frequencies, decreasing frequency ordering, concordances, co-occurrences, indices locorum, lists of suffixes and endings and lemmatization procedures. It is also able to perform operations of synchronization consisting in the comparison and analysis of the concordances of two different versions of the same text. It can also act as a lexicographic workstation, to be employed for the creation of a dictionary, a lexicon or a glossary starting from a specific textual corpus. The advantage of this workstation is that it can interact directly with the information inserted while creating or modifying a headword (lemma). In this way any type of useful information to compile the headword can be searched and included automatically and uninterruptedly. Moreover, the DBT makes it possible to operate and to navigate the texts in real time and interactively at incredibly high speed. The choice of the DBT as a guide used for encoding the ancient Slavonic alphabet has resulted particularly suitable to our needs.

The DBT had been used previously almost exclusively for texts written in the Latin alphabet and running on MS-DOS, which was sufficient in the treatment of this alphabet. It was thus necessary to implement the program with WINDOWS in order to handle such a graphically difficult alphabet without having to resort to obstruse conventional symbols. The principal purpose of my thesis was to check the unitariness of the Psalter tradition in the Slavonic territories with the passing of time, thanks to an analysis of the lexical variants found in the different text versions. As study material I used a sample of fifteen Psalms, chosen in

seven different codices of the Ancient Slavonic Psalter (more precisely, the Sinaitic, the Serbian, the Bolognese, the Norovian, the Kievian, the Gennadian and the Sinoidal). The Sinaitic was chosen as reference text since it is the most ancient and therefore likely to be the closest to the original text translated by Cyril and Methodius. I hoped the choice of such graphically heterogeneous material could offer an exhaustive range of ancient Slavonic characters found in the ancient manuscripts.

Before examining in detail the logical and methodological aspects of this work, it should be underlined that the DBT processing procedures are based on a data archive in the form of an ASCII file structured as a table, which I have called "Rules Table", containing all the information useful to the correct functioning of the program. The advantage of using an archive of this type consists in the possibility of modifying it repetitively without having to change the programs. In the conceptual model of the DBT system, the table is a dynamically standardized element owing to which the system can be adapted to any particular need as required by the different languages and codes. In order to structure this table it was first necessary to study the features of the manuscripts, so that all the graphical and linguistic aspects could be evinced and later codified. The DBT can develop modes of direct interaction with textual material codified in any particular way, including SGML according to TEI specifications, allowing the acquisition and handling through the different procedures available.

Let us first examine the main problems encountered in the phase of preparation and study of the material and the solutions adopted. Once the most significant graphical aspects had been identified, it was necessary to check whether they could be represented.

One of the principal limits to the creation of fonts, in particular of one as rich as the ancient Slavonic, was due to the fact that the code page I had available could contain only up to a maximum of 256 characters including those controlling the system. However, independent

of the limits of my code page, I was able to cover all the possible cases offered by the manuscripts.

Another problem encountered was that of the key-board whose characters (in my case an Italian key-board) contained characters different from those to be represented. It is obvious that the higher the number of characters to be represented, the greater the encoding difficulty. It was therefore necessary to take these drawbacks into account when having to choose between the graphical value of the character and its "absolute" value: e.g., the letter "i" appeared graphically in my texts as "и, й, ї, і, н". I was practically faced with the problem of whether I should report each graphical variant of the letters or choose a single character representing them all, since the letters could be standardized with the same writing, e.g. и). However, reporting all the graphical variants during the phase of analysis could mean complicating the identification of the lexical variants, since the computer would make even the words semantically identical appear as variants, as for example ІМЄНИ, ІМЄНІ, ЇМЄНИ.

Of the two possibilities I preferred that of representing the characters faithfully to the original text. My decision may be unexpected, but it arose from the fact that the way in which a character is represented can provide useful indications about the place and period in which the manuscript was produced. For this reason I tried to maintain as much as possible a certain peculiarity for a number of characters to the detriment of other elements as for example capital letters. I decided to save up the number of characters in the fonts, which were codified and visualized using the correspondent lower-case letter preceded by a signal character (an asterisk and small triangle respectively). This choice was also determined by the necessity to simplify as much as possible the association between the keys to be digited on the key-board and the ancient Slavonic characters to be represented, trying to avoid the combination of different keys. Unfortunately it was impossible to avoid their use completely, but I was able to restrict them to only five cases coinciding with the "iotized" letters. Iotization was in fact quite simple

to render, by preceding the iotized letter with the sign "1" conventionally chosen, instead of associating pairs of keys to a single character with which there was no logical connection or assonance. For example 'ѩ' will be represented by '1n', 'ѭ' by '1m' etc. In order to simplify the digitization system I also tried to make these associations as mnemonic as possible associating the ancient Slavonic character to the Italian character which is more similar phonetically, e.g. 'ѣ' to 'b', 'c' to 's', and so on.

At the end of this long work of preparation, I was finally able to define both the Slavonic characters to be represented and the keys to be digited in order to obtain both the printed and video visualization of each.

As I was unable to find a font with all the ancient Slavonic characters I needed, I had to see to their graphical realization using a true-type font editor to be used in WINDOWS. This font contains all the characters used by the scribes in the manuscripts examined. At the very beginning I intended to represent the text with its ligatures, accents, suprawritten letters, etc., but since I was mainly interested in the lexical aspects, for practical reasons in the digiting phase I thought it would be better to eliminate all the suprasegmental elements and ligatures which are insignificant to a lexical analysis. At this point I was provided with all the elements necessary to my research, which were however lacking an organization able to take into account the different relations between elements.

It was thus necessary to provide the computer with suitably structured information in order to link the codices with their functions. This was made possible by means of the "Rules Table" previously mentioned (v. allegato 1). Each line contains more or less the same elements: the first indicates the type of character to be defined (the letter C warns that it is a central character, P that it is a punctuation mark and so on). The second element consists in a full stop that announces the graphical character chosen as primary character in the case of several writings (this is necessary to obtain total frequency values in correspondence with forms

having the same lexical value but different graphical expression). The third element is the character to be typed on the key-board. The fourth is the sign whose numerical value is to be associated to the ancient Slavonic source in the visualization and printing phase (the sign you see here depends on the code-page which is active at the moment of this printout). The fifth is the order of alphabetical priority we have chosen and the sixth is the character to be used in the DBT for the search of the forms. The "Rules Table" is external to the program and for this reason it can be modified according to one's needs without changing the program itself. A clear layout of the "Rules Table" is thus fundamental for the correct analysis of the material.

Only after these two heavy phases of preediting of texts and preparation of the system were completed, was it possible to transfer the texts into the computer in machine readable form and to process them using appropriate DBT programs. The aim of these programs is to structure the data according to the information contained in the "Rules Table" so that the text can be submitted to textual processing procedures. Let us now examine in detail an example of the functioning of the DBT and the way in which it uses both the textual data and the information inserted into the " Rules Table".

If we consider that the starting element of any type of textual analysis is the word, and that the DBT operates in exactly the same way, the first operation the user is to perform, once the program has been started, is to select the word with which he wants to start his work.

The possibilities offered are multiple: if the researcher has no idea about the possible words of interest which can be found in the texts, by simply using the wild-character '*' as research string, he can obtain an alphabetical list of all the words contained in the text with the relevant frequencies. The characters of the visualized forms all appear standardized and are marked by a full-stop.

The search can also be performed by stem, root or ending, having each of them preceded or followed by the same wild-character '*'. Let us suppose that the researcher is

interested in all the words containing the sequence "єжє". According to the "Search Table" the letter "e" is to be typed as "e" and the letter "ж" as "q"; he will therefore have to type '*eqe*' (Fig.1)

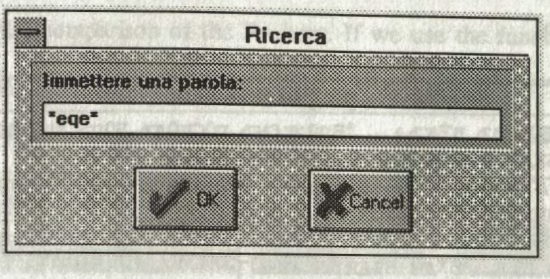


Fig. 1

and will obtain all the words containing this sequence internally and the relevant frequencies (Fig. 2).

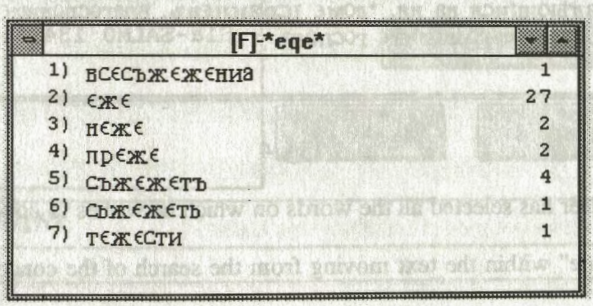


Fig. 2

If the researcher already knows the words he is interested in, it will be sufficient to type them completely without the use of the wild character. For example, if he wishes to work on the word "исраилєвъ" he will have to type the sequence "Icrallev°". The system will reply that it has found 7 occurrences of the form "исраилєвъ" (Fig. 3) with no distinction among the possible graphical differences.

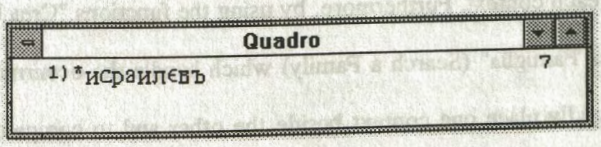


Fig. 3

If the user now asks for the concordances of this form, we can see that 5 contexts contain the graphical form "исраилєвъ" and that the remaining 2 contexts contain "исраılєвъ" (Fig. 4).

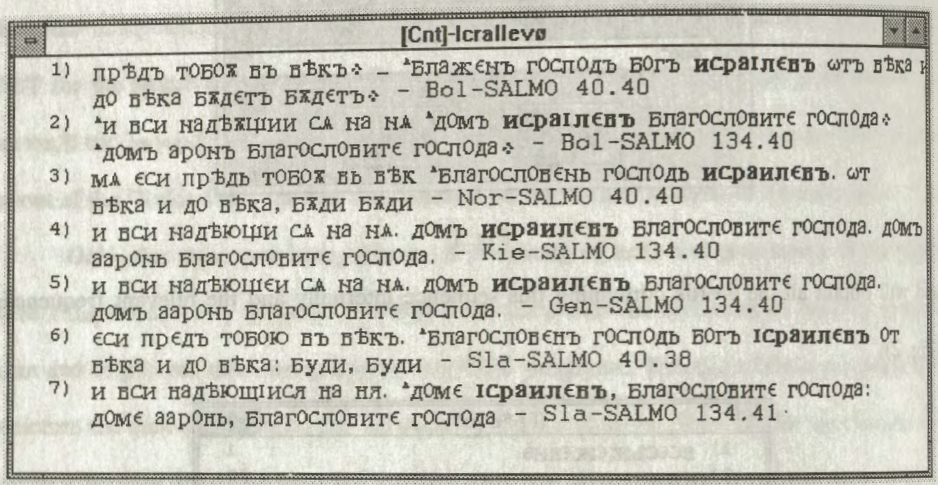


Fig. 4

When the user has selected all the words on which he wishes to conduct his search, it is possible to "navigate" within the text moving from the search of the concordances of a single form to that of the co-occurrences. It will then be possible to either extend each context, or to enter the text by following paths which are completely personal, or to limit the results of the search to single parts of the text by selecting appropriate filters. "Navigation" is certainly the major advantage that can be obtained by the use of a computerized system of this type. In my specific case for example each time I identified a pair of variant words I could resort immediately to the context in which they were found, thus obtaining the exact meaning the words assumed in each context. Furthermore, by using the functions "Crea Famiglia" (Create a Family) and "Cerca Famiglia" (Search a Family) which handle the occurrences in the DBT, it was possible to rapidly place one context beside the other and to compare them, speculating

instantly on the reasons which led the compilers of the manuscripts to use one word instead of the other.

As an example of this function we can select the pair of variants "рѣснота-истина" which emerged from the comparison of the Psalters. If we use the function "Crea Famiglia" and we link the two words through the logical operator "OR" (Fig. 5)

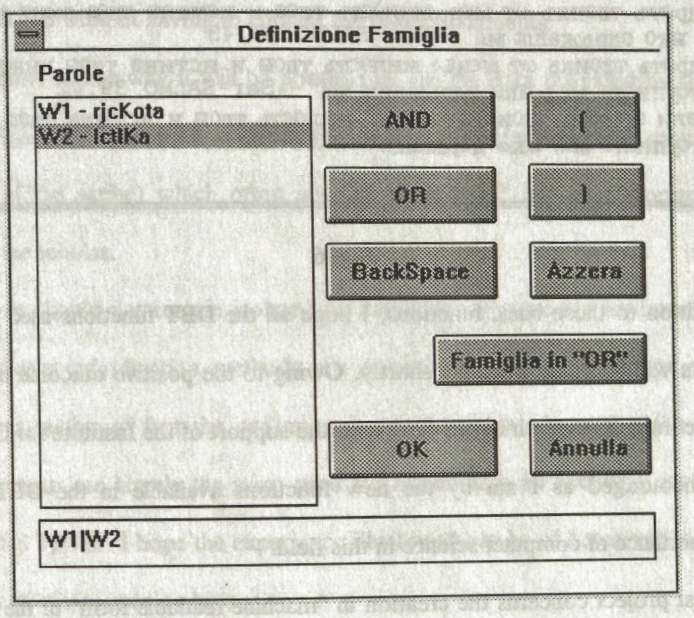


Fig. 5

we can obtain their concordances immediately after, so as to have a clear idea of their distribution in the various texts (Fig. 6).

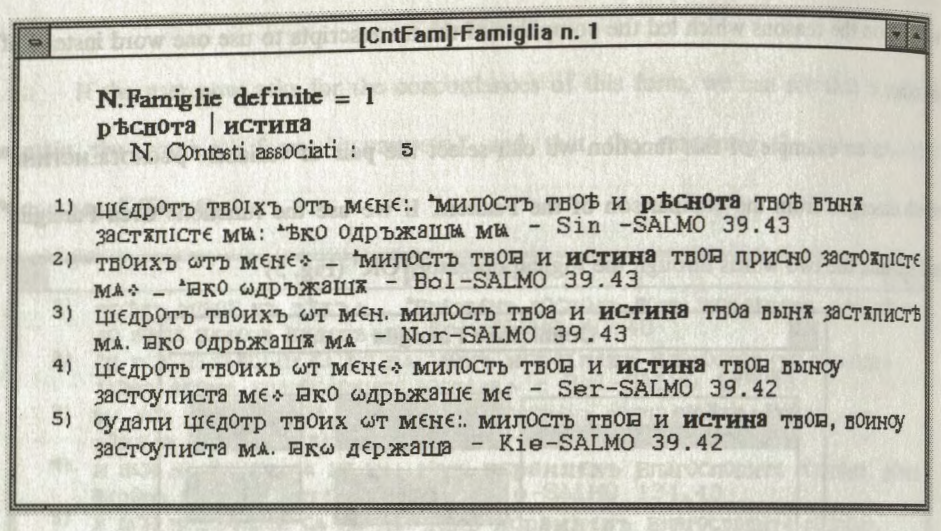


Fig. 6

In addition to these basic functions, I hope all the DBT functions used so far for the Latin alphabets will be made available shortly. Owing to the positive outcome of this work, I would like to continue research in this field with the support of the Institute for Computational Linguistics, encouraged as I am by the new functions available in the DBT and by the increasing importance of computer science in this field.

The first project concerns the creation in "machine readable form" of the entire corpus of the seven Psalters and the addition of new manuscipts. Adapting all the other functions of the DBT previously listed to the Slavonic texts is the second important commitment I intend to carry out. These two projects should lead to a third one which is perhaps the most futuristic: the possibility of making the corpus of the seven Psalms available to the scientific community which, through the possibilities offered by Internet, could have access to these texts using the DBT-NET query procedure appropriately studied for this purpose. This would obviously start off, as it usually happens in Internet, a process of exchange of information, advice, reviews, new ideas, and retrieval of new textual material.

Finally, the list of operative functions of the DBT has only just been enriched by a new device, which makes it possible to place the text produced by the DBT next to the graphical image of the original document, either a print or manuscript. The researcher can compare rapidly and interactively the choices made by the person who encoded the text analyzed by the machine with the original text, solving the doubts of interpretation or mistakes encountered during the search and without having to consult the paper documents.

Furthermore, in this way it will be possible to use the data provided by the computer without losing sight of the peculiarities of the original text (as for example the cromatic aspects or the writing of the scribe) which often act as stimulus and are an important source of information for the scholar.

This effort should not remain isolated or limited to a particular environment like the DBT. I think that standardization methods are essential for a free exchange of texts and information. As a matter of fact the preliminary considerations concerning the encoding of ancient Slavonic texts are largely the same ones for a standardized encoding method like the one recommended by TEI. I hope the experience illustrated can be of some use to the solution of the problems we are examining here.

```
C.a  ÿ  #1  a
C *a Çÿ  #1  a
C.b  «  #4  b
C *b Ç« #4  b
C.v  Ö  #7  v
C *v ÇÖ #7  v
C.g  Ü  #10 g
C *g ÇÜ #10 g
C.d  ø  #13 d
C *d Çø #13 d
C.e  £  #16 e
C *e Ç£ #16 e
C.q     #19 q
C *q Ç  #19 q
C.s  ®  #22 s
C *s Ç® #22 s
C.z  Ø  #25 z
C *z ÇØ #25 z
C I  á  #28 I
C *I Çá #28 I
C.i  ª  #28 I
C *i Çª #28 I
C L  ¹  #28 I
C *L ¹  #28 I
C è  ×  #28 I
C H  É  #28 I
C *H ÇÉ #28 I
C.h  Î  #30 h
C *h ÇÎ #30 h
C.k  í  #31 k
C *k Çí #31 k
C.l  ó  #34 l
C *l Çó #34 l
C.m  ú  #37 m
C *m Çú #37 m
C K  ñ  #40 n
C *K Çñ #40 n
C.n  Ù  #40 n
C *n ÇÙ #40 n
C.o  o  #43 o
C *o Ço #43 o
C w  w  #43 o
C *w Çw #43 o
C.p  º  #46 p
C *p Çº #46 p
C.r  ¿  #49 r
C *r Ç¿ #49 r
C.c  c  #52 c
C *c Çc #52 c
C.t  ½  #55 t
C *t Ç½ #55 t
C.u  Â  #58 u
C *u ÇÂ #58 u
C U  Ã  #58 u
C *U ÇÀ #58 u
C V  ¼  #58 u
C *V Ç¼ #58 u
C F  ¡  #61 F
C *F Ç¡ #61 F
C.f  Ë  #61 F
C *f ÇË #61 F
C.x  »  #64 x
C *x Ç» #64 x
C.§  ▒  #67 §
C *§ Ç▒ #67 §
C.Q  Ú  #68 Q
C *Q ÇÚ #68 Q
C.Z  ß  #70 Z
C *Z Çß #70 Z
C.ç  ƒ  #73 ç
C *ç Çƒ #73 ç
C J  î  #73 J
C *J Çî #73 J
C.S  Õ  #76 S
C *S ÇÒ #76 S
C.°  µ  #79 °
C *° Çµ #79 °
C ù  Õ  #82 y
C *ù ÇÓ #82 y
C ì  ©  #82 y
C *ì Ç© #82 y
C.y  ▓  #82 y
C *y Ç▓ #82 y
C Y  Ó  #82 y
C *Y ÇÕ #82 y
C X  ¾  #82 X
C *X Ç¾ #82 X
C ^  ¬  #88 ^
C.j  ■  #91 j
C *j Ç■ #91 j
C.1u Û  #94 1u
C *1u ÇÛ #94 1u
C.1a ¢  #97 1a
C *1a Ç¢ #97 1a
C à  Ñ  #97 1a
C *à ÇÑ #97 1a
C.1e ¥  #100 1e
C *1e Ç¥ #100 1e
C.N  ▀  #103 N
C *N Ç▀ #103 N
C.M  |  #106 M
C *M Ç| #106 M
C.1n Á  #109 1n
C *1n ÇÁ #109 1n
C.1m Ì  #112 1m
C *1m ÇÌ #112 1m
C.P  ï  #118 P
C *P Çï #118 P
C ≈  ¤  #124 !
C.!  û  #124 !
C ?  È  #124 !
C \      #124 !
1 `  ´
1 {  ´  {
1 '  ,  '
1 "  Í  "
1 )  ı  )
1 }  Ï  }
1 (  ã  (
1 /  Ô  /
1 +  Ý  +
1 >  õ  >
1 <  þ  <
1 ↑  ↑  ↑
2 ~  ~  ~
P 2: ▒
P :  :
P _  _
P .  .
P ,  ,
P -  -
P ;  ;
P :. ³
P [  [
P ]  ]
P |  ²
P 1~ ╠
P 1+ ═
P 2~ ╩
P 0- ü
P 0\ ù
P 1  1
P 2  2
P 3  3
P 4  4
P 5  5
P 6  6
P 7  7
P 8  8
P 9  9
P 0  0
T å  ■↑ *
T é  ô↑ *
T â  »↑ *
T à  ¿↑ *
T å  ø↑ *
T Æ  Ø↑ *
T ô  ú↑ *
T ö  Ò↑ *
T ò  ª↑ *
T Ã  ½↑ *
T ·  ŷ↑ *
T -  ¹↑ *
T è  ¼↑ *
T °  É↑ *
T +  Ù↑ *
T ±  Ü↑ *
T     Õ↑ *
T ¶  Á↑ *
T §  |↑ *
T Ý  «↑ *
T ·  ¡↑ *
T ▒  Å↑ *
T Þ  ↑  *
T ë  ß↑ *
T Ê  ¬↑ *
T ■  À↑ *
T ╨  ƒ↑ *
T Ö  ■↑ *
T =  ¾↑ *
T ╔  Ó↑ *
T ┼  ¡↑ *
T ├  º↑ *
T ┌  Ó↑ *
T ç  o↑ *
T ê  á↑ *
T ì  £↑ *
T Ä  c↑ *
T Å  í↑ *
T æ  µ↑ *
```

Allegato 1

Medieval Slavic Menologies On-Line

Cynthia Vakareliyska
Department of Russian, University of Oregon, Eugene, OR 97403-1262, USA

Keywords: menologies, liturgical calendars, saints

The project described in this paper is a proposed Internet collation of medieval Slavic menologies. The term "menology" is used here to mean a calendar of saints, i.e., a liturgical calendar which includes the names of saints and events to be commemorated on specific days of the year. These listings vary widely among individual manuscripts, depending, among other factors, on the time and place in which the menology was written. The extent of variation is illustrated by the following range of entries for 29 October from eighteen menology texts:[1]

Father Abraham	1342, 1361, F100, 502
Anastasia	Arx, Ban, Cur, Vrač, F9, F100, P14, 502
Asterius	Ostr
Barnabas	Q59
Cosmas and Damian	Ban, Cur, Ostr
Archbishop Cyriacus	G5, F110, 882
John the Baptist	Ban, Cur, Ostr, 849
Patriarch Joseph	Arx, Ostr, Vrač, 844, 883
Peter and Paul	Ban, Cur, Ostr, 883
Sava	502
Stephen	Ban, Cur, Q59
Trophimus	Arx, Ostr, Q59
Zenobius and Zenobia	Ostr, G5, F110

This degree of discrepancy, which occurs even between otherwise very closely related manuscripts, makes it very difficult to recognize general patterns of correspondences and discrepancies among individual menologies. One of the two major goals of this collation is to facilitate the development of a general typology of Slavic menologies through computer identification of such patterns among a large corpus of texts. The second goal is to aid in the analysis of any individual menology text, within the corpus or without, by providing a large database of textual variants against which it can be compared.[2] Since most menology texts are unpublished, an Internet collation will make them easily available to scholars, helping to avoid

both the necessity of travel to archives, and reduplication of efforts in data collection. Another advantage of an Internet collation is its capacity for continual expansion and refinement, as individual scholars contribute to it further data and emendations.

The corpus consists of listings from over 100 medieval Bulgarian, Serbian, and East Slavic menology manuscripts from the SS. Cyril and Methodius National Library, the Library of the Bulgarian Academy of Sciences, and the Theological Academy Library in Sofia; the Ivan Vazov Library in Plovdiv; the Rila Monastery Library; the Hilandar Center microfilm collection at the Ohio State University; the British Library in London; the Russian National Library in St. Petersburg; and the Russian State Library, the Moscow University Library, and the Russian State Archive for Ancient Documents in Moscow. The corpus currently covers texts from the eleventh through fourteenth centuries, and can be expanded to include later texts as well.

The collation presently is limited to entries of saints and events, and does not include any accompanying pericopes or other liturgical instructions from the menology texts. It would be desirable, however, eventually to include the entire text of each menology listing, since variations in lections and in the textual redactions represented in the pericopes can provide further information on specific menology traditions. It may also prove advantageous in future to add entries from *typika*, *menaia*, and other types of liturgical manuscripts in addition to menologies *per se*, although these do not contain as wide a variety of saints. Eventually the corpus should be expanded to include early Byzantine Greek menology listings, particularly as some of the textual formulae found in Slavic menologies are *verbatim* translations from these.

Since orthographic and morphological variation among manuscripts may present obstacles to a computer text collation, it is anticipated that in the beginning stages of the project, all searches will be conducted the basis of English-language keywords, with an option to view the actual corresponding Old Cyrillic text from any given manuscript or manuscripts on command. The user would then draw his or her own conclusions regarding the extent of relatedness between or among the textual variants displayed onscreen, rather than relying on the computer to make such decisions. Eventually, however, it would be useful to have the option to conduct certain types of searches directly on the basis of textual formulae, as discussed below.

Directories

a. Manuscript directory. The collation should include a directory of all the manuscripts in the corpus, and the identification codes to which they are referred onscreen; the directory should be organized, in default form, similarly to the list of manuscript signatures at the end of this paper. It should be possible to search the directory by selecting keyword categories that permit viewing of its contents by century, by recension (Bulgarian, Serbian, etc.), by manuscript genre (tetraevangelion, aprakos, etc.), by type of menology (full menologies, which generally provide entries for every calendar day, vs. short menologies, which list fewer dates), by archive location, and, for the minority of manuscripts with established names, in alphabetical order by title. The following sort of on-screen display of manuscript codes would be obtained under the multiple selections "archive: Russian National Library", "recension: Serbian", "century: fourteenth", and "genre: tetraevangelion":

F100 F110 G5

Information on a manuscript should appear automatically onscreen, possibly in the form of a header or footer, any time the identification code of that manuscript is selected. In addition to the information provided in the directory, the header or footer would include notations concerning features of the manuscript which could affect the analysis (e.g., a list of those calendar dates that are deliberately omitted from a short menology, the beginning and ending calendar dates for missing folia that might account for further omissions in entries, any apparent misnumbering of calendar dates, whether any portions of the calendar were written by a second scribe or are obviously taken from a different menology tradition, whether the names of the months appear in their Slavic variants, etc.):

F110 RNB F.p.I.110 14th-cen. Serbian tetraevangelion, full menology, no Slavic month names; last folio missing (Aug. 23-31)

b. Directory of saints/events. A separate feature is a directory of English-language labels for the saints and events listed in the corpus. Selecting a given label should call onscreen all

references in the text database to the saint or event it represents, regardless of textual variation among individual manuscripts: for example, the label "Theodore of Tyrone" should cover the entries ѳеодорѣ тиронскомоу, тодора тирона, ѳедора тиронскаго, etc. Where several saints are traditionally grouped together as a single entry (e.g., "Cosmas and Damian"), they should be indexed in the directory both jointly and severally.

Users should be able to browse through the directory alphabetically in order to find the proper English-language label for a given saint or event. In addition, by selecting a given saint or event, it should be possible to obtain a short summary of information, including date(s) of commemoration. For saints, the summary should also include nationality, date of canonization, brief biographical information, and, for figures in the church hierarchy, the range of titles which they held at different periods in their lifetime, with approximate dates. The range of titles can be particularly important, because a number of saints who share the same name can be distinguished only by the descriptive titles which accompany their names in the menology listings (e.g., "Father Theodore" vs. "Theodore the Martyr"). A notation should also be made if the particular saint or event is primarily associated either with one of the three major Eastern menology traditions (Constantinople, Jerusalem or Stoudite) or with the Latin tradition, or if it is archaic. It should also be possible to select to view the contents of the directory organized, alternatively, by country of origin for each saint, by century of canonization, and by general menology tradition.

c. Keyword directory. A separate directory or directories should be available for searches based on more general types of keywords (see below).

Texts

The collation should permit the viewing of all the listings from any given manuscript in calendar order (September 1 - August 31), both by English-language label, and, alternatively, in the original Church Slavonic text. Selection of two or more manuscripts should permit a comparative viewing of all the entries from each manuscript, presented on-screen in interlinear format. Once pericopes and other additional text from the menologies are included in the

database, it should be possible to view them either in isolation, or together with the entries of saints and events.

Search features

In addition to the searches of the directories, the computer collation should permit the following searches of the text database:

a. Search by saint/event. Selection of the appropriate English-language label for a given saint or event should result in an onscreen display of the identification codes of all manuscripts in the database in which that saint or event is commemorated, together with the calendar date(s) for which it is listed in each manuscript. Where a manuscript contains multiple entries for a given day, a numeral appearing at the end of the entry should indicate its position in the listing (e.g., "$Q59_2$"); this detail can be significant in establishing textual relationships between manuscripts. The following display, composed from the same eighteen manuscripts as above, would be obtained by searching under the key phrase "Cosmas and Damian":

19 Oct	$Q59_2$
29 Oct	Cur_3, Ban_3, $Ostr_2$
1 Nov	Arx, Cur_1, Ban_2, Vrač, Ostr, 1342, 1361, F9, F100, F110, G5, P14, Q59, 502, 844, 849, 882, 883
1 July	Arx, Cur, Ban, Ostr, Vrač, 1342, F9, F100, F110, G5, Q59, 502, 844, 849, 882, 883

The user should be able to narrow the search in the same manner as in searches of the manuscript directory above, i.e., by manuscript recension, century, genre, calendar date(s), or specific manuscript identification code(s); it should also be possible to combine two or more of these categories (comparing, for example, fourteenth-century Bulgarian and thirteenth-century Serbian lectionary gospels). Another perspective for limiting the search is by category of saint, under keywords such as "nationality", "century", or "title". Once the list of appropriate manuscripts appears onscreen, the option should be available to select to compare the Church

Slavonic textual formulae of the entries. To the extent possible, the on-screen display of textual formulae should be organized more or less alphabetically (taking into account variations in abbreviation and orthography), so that similar formulae are grouped together:[3]

1 Nov

P14: Пам҃а ст҃ую безмездик|и (sic) ѹ҃творую кѹзмы и | дамиана·

Ban: ·Пам҃а ст҃ою весребрьникоу козмѣ и дамианоу1
Cur: пам҃а ст҃ою | сребрьникоу козмьи дамианоу·1

F110: Ст҃оую безмьзѣ|никоу҃ козмоу и дамиа҃|на.
G5: ст҃оую безьмьзѣникоу козмоу и҃ дами|ана.

F100: ст҃ою безмьзѣи|коу҃ козмы и даминана. | сноу҃ ѳеѡ҃тиноу҃·
502: Ст҃ою҃ безмездникоу· козмы· и дамиана, | ѳеѡ҃дотиноу (sic)

In instances of multiple saints' listings for a given day, the user should also have the option to view the complete Church Slavonic text of that day's entry in any given manuscript(s), or, alternatively, in all the manuscripts listed onscreen:

Oct 29

Cur: ст҃ихъ а҃пл҃ь петра | и павла· і҃ѡ҃ноу крт҃лоу· стеѳа҃· козма и да|миана· инѣ многъ· и ст҃ѣи настасии·

Ban: Ст҃ыхъ а҃пл҃ь петра и҃ павла і҃ѡ҃ноу крт҃лю. и҃ стета (sic) и ко|змѣ и҃ дамиана. нѣхъ (sic) многъ и҃ ст҃ѣи анастасий.

A further option should be available to view a list of all manuscripts which do *not* list the specific saint or event selected.

b. Search by calendar date. Selection of a specific calendar date should call onscreen a list of all saints and events recorded for that day in the corpus, each listed according to its English label and followed by the identification codes of all manuscripts in the corpus in which it appears. The display would be similar in form to the list of entries for October 29 shown on the first page of this paper. It should be possible to narrow the range of manuscripts or saints by using the keyword categories described above for searches by saint or event.

c. Search by frequency of correspondence. The option should be available to view a list of all those manuscripts in which a specific percentage of the entries (for example, 80% or more) correspond substantively. Of course, this type of statistical analysis can only give a very general idea of the extent of similarity between manuscripts, since the percentage of correspondence will be lowered artificially by extraneous factors such as missing folia, misnumbered dates, and multiple antegraphs for a given manuscript (all features that should be noted in the unprompted descriptive headers to the individual manuscripts); neverthless, in most cases a preliminary statistical check should provide a starting-point for identifying related texts.

d. Search for statistical/numerical correspondence between any two given manuscripts. It should also be possible, by selecting two or more manuscripts, to obtain both numerical and statistical information on correspondences in listings, both on the basis of the English-language labels, and, eventually, if possible, on the basis of textual formulae as well. A further option should be available to view, in chronological order, only those calendar entries which coincide between or among specific manuscripts, or, alternatively, only those entries which do *not* coincide. Particularly in cases of lower statistical and numerical correspondences, this second step will show whether the correspondences are limited to a particular portion of the calendar, as is sometimes the case where one of the manuscripts has more than one antegraph.

e. Search by discrepancy between/among individual manuscripts. Where two or more manuscripts appear to be somewhat consistent in their listings, a search by discrepancy will provide a display of only those entries, organized by calendar date, which do not correspond in substance. Here too, the option should be available to view the non-corresponding pairs of entries (including pairs in which one manuscript has a zero entry) either by English-language labels, as directly below, or, alternatively, according to textual formulae:

G5, F110:	August 23-31 [missing folio in F110]	
Ban, Cur:	Oct 5	Cur: Savina, Mamelta; Ban: no entries for this date
	Oct 13	Cur: Carpus, Papylus, Trophimus Ban: Carpus, Papylus (etc.)

It would be particularly useful, if possible, for text-based searches to include textual discrepancies between substantively corresponding entries, with an option to view the discrepancies onscreen either in isolation, or within the complete listing in which they occur:

i. isolated display:

G5, F110: Sept 1

G5: паме̑ пр҃пѣ̑номоу ѿцоу̑ нш҃емоу симе|wноу
F110: ⟨_⟩аме̑ пр҃поⷠнаⷡгѡ ѡ҃ца̀ нш҃е|го сумеѡ̑на стлп҃ника

ii. complete listing:

G5, F110: Sept 1

G5: Наѵетькь и҃ндиктоу · рекше новоу лѣтѣ̀ ·
паме̑ пр҃пѣ̑номоу ѿцоу̑ нш҃емоу симе|wноу

F110: наѵетькь ин҃диктоу ⟨_⟩ек҃|ше новоу лѣту ·
⟨_⟩аме̑ пр҃поⷠнаⷡгѡ ѡ҃ца̀ нш҃е|го сумеѡ̑на стлп҃ника

f. Search by lexical item. Searches for unusual lexical items would be helpful both for analyzing individual manuscripts, and for identifying manuscript families. Unusual forms might be listed either in the keyword directory, or in a separate directory: the stem might be used as the keyword form, in order to avoid possible variations in grammatical endings:

"сисм-": сисмоу [Masc Dat Sg noun, Greek *seismós*]: Oct 26 Cur, Ban

g. *Keyword search.* This feature would permit a search of the text corpus by general categories that are not easily included under any of the search keywords described above; these would include "astronomical listings", "Graecisms", and, perhaps eventually, even specific types of linguistic features (e.g., "*cokan'e*") In the response to the search, the onscreen display should include for each entry its calendar date, an English label for the entry, and the textual formula:

"astronomical entry" May 13 Cur: Pleiades вьсхо̑ власожелить

Conclusion

The features outlined above for an Internet collation of medieval Slavic menologies should facilitate analysis from a number of different perspectives of an otherwise unwieldy text corpus. It is hoped that implementation of these features will result in a versatile collation system that will be accessible and useful to scholars both for analysis of individual menology texts, and for study of the general Slavic menology tradition.

Notes

1. English versions of saints' names are from *The Book of Saints*, compiled by the Benedictine monks of St. Augustine's Abbey (London: A.&C. Black, 1981). Manuscript signatures appear at the end of this paper. The term "listing" is used here to refer to the entire text for a specific calendar date; each saint or event included in the listing is a separate "entry".

2. A computer collation will have several distinct advantages over the most comprehensive published source on Eastern menologies, Arximandrit Sergij's *Polnyj mesjaceslov Vostoka* (2nd ed., Moscow 1902) ("*PM*"). In addition to permitting searches based on keyword categories which are not available in the *PM* indices, the computer collation includes a considerable number of Slavic menology texts which are not covered by the *PM*, and which list a number of saints not mentioned there. Moreover, whereas the *PM* lists entries from most menologies only according to the Russian-language variants of the saints' names, the computer collation will offer the option to view the original non-normalized Church Slavonic texts, thus permitting comparison of textual formulae.

3. G5 and F110 are two very closely related fourteenth-century Serbian tetraevangelia from two separate collections at the Russian National Library. With only a few minor exceptions, all of which appear to be the result of inadvertent slips in copying, the two menologies are identical in content and strikingly similar in form, sharing unusual spellings of saints' names and capitalization conventions. Cursory preliminary examination suggests that the versions of all four Gospel books in the two manuscripts may be equally close. The menologies to F100 and 502 are from the same general textual tradition as the G5/F110 pair. Ban and Cur, two tetraevangelia from Western Bulgaria (Cur c. 1354; Ban date in dispute, XIII-XIVth century), comprise a second pair of closely-related manuscripts; as expected, however, their menologies contain a number of discrepancies in both form and content (see C. Vakareliyska, "Linguistic and textological features of the Curzon Gospel menology", in G. Fowler and H. Cooper, eds., *Papers from the Ninth Biennial Conference on Balkan and South Slavic Linguistics, Literature and Folklore (Indiana University, April 8-10, 1994)* (to appear).

Manuscript Signatures

Archives

BM: British Library, London; GPB: State Russian Library, Moscow; NBKM: SS. Cyril and Methodius National Library, Sofia; RNB: Russian National Library, St. Petersburg

Manuscripts

Arx	Arxangelsk Gospel	GPB.M.1666	1092, aprakos, Russia
Ban	Banica Gospel	NBKM No. 847	13-14th cen. tetraevangelion, Bulgaria
Cur	Curzon Gospel	BM Add Ms. 39,628	c. 1354, tetraevangelion, Bulgaria
Ostr	Ostromir Gospel	RNB.F.p.I.5	1056-67, aprakos, Russia
Vrač	Vračansko Gospel	NBKM No. 19	13-14th cen. aprakos, Bulgaria
1361		NBKM No. 31	1361, tetraevangelion, Bulgaria
1342		NBKM No. 470	1342, tetraevangelion, Serbia
F9		RNB.F.p.I.9	14th-cen aprakos, Russia (Pskov)
F100		RNB.F.p.I.100	14th-cen. tetraevangelion, Serbia
F110		RNB F.p.I.110	14th-cen. tetraevangelion, Serbia
G5		RNB Gil'f. 5	14th-cen. tetraevangelion, Serbia
P14		RNB.Pogod.14	14th-cen. lectionary apostol, Russia
Q59		RNB Q.p.I. 59	14th-cen. aprakos, Bulgaria
502		NBKM No. 19	14th-cen. lectionary apostol, Bulgaria
844		NBKM No. 844	13th-cen. aprakos, Serbia
849		NBKM No. 849	13th-cen. aprakos, Bulgaria
882		NBKM No. 882	13th-cen. lectionary apostol, Bulgaria
883		NBKM No. 883	13th-cen. aprakos, Serbia

Abstract

The paper proposes features for an Internet collation of medieval Slavic menologies (calendars of saints).

COMPUTATIONAL ANALYSIS OF HYMNOGRAPHIC COMPENDIA

Emilia Guergova
Faculty of Slavonic Philology, Sofia University "St. Kliment Ochridski"
bul. Tsar Osvoboditel 15, 1504 Sofia
E-mail: ruja@slav.uni-sofia.bg

KEY WORDS: Menaia, acolouthia, hymnography, ecclesiastical calendar, Collate, configuration of structures, matrix of the levels.

ABSTRACT: Hymnographic compendia form a substantial part of the genre diversity of medieval books of *Slavia Orthodoxa* because of the high frequency of usage, assigned to them by ritual practice. Unfortunately, the level of investigation of these sources is far lower than is the case with other types of mediaeval books, primarly because of the nature of hymnography itself, which, in contrast to the narrative texts, is characterized by a vast variety of structures and a high mobility of contents.

This challenges the investigator's abilities to oversee and process the extreme wealth of data and to extract from them objective conclusions about the history of their texts.

The present project focuses on the *Menaia*, and specifically on the Menaion for the month of October. Its aim is to study in diachronic perspective the macrostructure, i.c. the dislocation of the acolouthia in the order of the ecclesiastical calendar; the structure of the acolouthia themselves, i.e. the configurations of their mesostructural elements; and the microstructure of these elements (troparion, kontakion, stichera, oikos) from a rhythmic point of view.

The project requires the aid of a computer and appropriate software to ensure adequate treatment of the following: comparison and synthesis of data on four levels: (Saint, ritual occasion, mesostructural genre, and the rhythmico-intonational and syntactic segmentation); relations among the various positions, where elements may be encountered at all four levels; and creating a matrix for the multidimensional representation and evaluation of the data and their relationships.

Such complex collations of large numbers of manuscripts should preferably be executed on the basis of a single source file per manuscript. If collation is possible only within strictly separate levels (as is the case with OUCP's Collate), the source files must be adapted manually to each collation, and errors and inconsistencies are bound to arise. A well-designed and well-tested combination of Hypertext and Collate seems required to overcome these difficulties.

TEXT: По фреквентности церковного употребления гимнографию можно определить как самую традиционную часть средневековой православной книжности. Как известно, гимнографические сочинения помещались на основе своей функциональной потребности в разных сборниках - Кондакарь, Стихирарь, Тропологий, Октоих, Минея. Когда славяне принимали крещение, литургический

конфликт между более старой кондакарной традицией и все более убедительно проникающем в азматическую практику каноном в Византии в основном можно считать решенным. Несмотря на существующие разногласия, первый период литургического синтеза по словам Р. Тафта во время иконоклазма уже факт[1]. Одно из важных достижений этого синтеза представляет оформление гимнографического сборника Минеи, составление которого в самых общих параметрах требует: на верхнем уровне - пополнение дней месяцесловного порядка песенным материалом; на нижнем уровне - конструирование сложной композиции из песнопений, исходя из регламента повседневных последований. В самом деле еще на византийской почве комплектование книги протекало неедносторонне, в связи с чем можно различить три основные разновидности:

1. Минея празничная - содержит службы христологического и богородичного циклов, а также и службы популярным раннехристианским святым, обединенные или в одном книжном теле на весь церковный год, или в двух отдельных томах (для первого полугодия - сентябрь/февраль и для второго полугодия - март/август);

2. Минея служебная - содержит службы на каждый день по синаксарю. При этом с развитием и усложнением месяцесловных памятей за день возможно накопление 2, 3 и больше служб, разпределенные в 12 томах по месяцам.

3. Минея общая - гимнографический материал в ней тоже обслуживает календарные праздники, но расположен не в хронологическом порядке, а на основе иерархии. В связи с этим обособляются службы господские, богородичные, о небесных силах, святительские. Святительские службы относятся до типа святых (об апостолах, о мучениках, о йерархах, о преподобных и пр.) и известны еще под названием "общие службы".

Вместе с тем существуют и минейные сборники, стриктное причисление которых к одной из рассмотренных разновидностей оказывается делом нелегким. Изключения встречаются сравнительно часто. Можно привести в пример Минею № 115 из Зографского монастыря, в которой помещены службы на первые дни каждого месяца и службы на весь май и июнь[2]. Этот список отмечает медленные шаги проникновения Иерусалимского типика.

На основании статистических данных можно прийти к выводу, что в рукописных собраниях монастырских библиотек славянского мира количество миней доходит до четверти из всех рукописей. Поэтому их изучение является важной

---

[1] *Taft, R.* The liturgy of the Great church: an initial synthesis of structure and interpretation on the eve of Iconoclasm. - Dumbarton Oaks Papers, 34-34 (1980-81), p. 46.

[2] *Райков, Б., Ст. Кожухаров, Х. Миклас, Хр. Кодов.* Каталог на славянските ръкописи в библиотеката на Зографския манастир в Света гора. С., 1994, 81-82.

перспективой в палеославистике. Однако одно из самых существенных условий выражается в принципе искать и изучать не отдельное явление, окачествленное как "более интересное" с какой нибудь точки зрения (как обычно происходит в практике - исследовать только одну службу славянского святого), а исследовать всю книгу комплексно, по всем аспектам текстовых уровней и интертекстуалным связям с остальными богослужебными книгами. Мое внимание сосредоточено на Служебных минеях, и на этом начальном этапе компютерной обработки подвергнута Служебная минея на октябрь, которая известна в широком диапазоне списков.

Усилия до сих пор были направлены главным образом на выработку системы кодирования текста. Трудности вытекают из самой специфики гимнографии. Служба есть сложная мозаичная композиция, отдельные строфы которой приобретают относительную самостоятельность, - как с точки зрения текстовой, ритмической и интонационной законченности, так и с точки зрения применения в церковном ритуале, т.е. строфы характеризируются мобильностью, и их перестановка и изменения часто стимулируются внетекстовыми факторами. Поэтому пришлось отнестись с пристальным вниманием к принципам сегментации текста. В отличии от нарративного текста или текстов, отличающихся линейной хронологией, при которых в качестве основной кодировочной единицы можно принять синтагму, соблюдая последоватльность, здесь необходимо сочетать кодировку с функциональной сегментацией.

Вот вопросы, которые следует иметь в виду:

1. Проблемы, вытекающие из месяцесловного порядка.

2. Проблемы, связанные с расположением материала по дням, если появляются две или больше памятей святых. Возьмем к примеру 19 октября, когда празднуются пророк Иоиль и мученик Уар, а позже и св. Иоанн Рыльский, - в этом случае приходится обратить внимание на то, песнопения какого из святых занимают соответственно первую, вторую или третью позицию; следуют ли одно за другим или переплетаются и т.д.

3. Проблемы, вытекающие из изменений в Типике.

4. Проблемы, связанные с внутренней последовательностью текста в рамках одной службы. Как известно, хронология расположения строф в писменно фиксированной последовательности не всегда отвечает хронологии реального исполнения. Больше всего трудностей создают списки до XIII-XIV-ого веков, дающие основание выявить несколько вариантов:

А) 1. Канон; 2. Стихиры; 3. Седальный; 4. Кондак; 5. Икос (если кондака и икоса не хватает, служба оканчивается седальным).

Б) 1. Тропарь (чаще всего только указан ирмосом и гласом, но без текста); 2. Седальный; 3. Кондак; 4. Икос; 5. Три стихиры на "Господи воззвах"; 6. Стиховная (или самогласная) стихира для вечерни; 7. Стиховная для утрени; 8. Канон (с разновидностями - три, восемь или девять песней)[3].

В) Здесь конфигурация подобна предыдущей, но уже седальный, кондак и икос вошли в канон, соответственно после третьей и шестой песни, при чем тоже возможны варианты.

Г) Хорошо обособленная вечерня с тремя стихирами на "Господи воззвах", славой, стиховными стихирами, святительским тропарем и на утрени - канон (с постоянными местами седального, кондака и икоса); стиховные стихиры, которые в более поздней редакции заменялись светильными или хвалитными стихирами; слава; отпустительный тропарь.

В действительности трудно найти два списка Минеи с идентичным распределением состава. Это потребовало двусторонний подход. В одном аспекте он основывается на принципе каждый день из каждого списка оформлять в отдельном файле, соблюдая соизмеримость параметров, сообразно : 1) святого; 2) литургического времени исполнения; 3) микрожанровой принадлежности песнопений; 4) ритмично-интонационного членения. Во втором аспекте он касается трех разграниченных разрезов при анализе текста.

Разрез I - относится к макроструктуре. Цель - приобрести информацию о постоянном устойчивом ядре для каждого дня. Здесь предумышленно принята последоватльность песнопений Ягичевого списка Октябрьской минеи 1096 года[4].

<Level> - *верхний уровень*: <L 0> - оглавление; <L 1> указание вечерни; <L 2> указание утрени;

Песнопения конкретного святого - *средний уровень*. Вводятся инициалом имени. Возмем например 1 октября, когда празднуются ап. Ананий и св. Роман Сладкопевец: <A> св. Ананий и <R> св. Роман; каждая первая цыфра обозначает микрожанровую единицу: <A 2> седальный; <A 3> тропарь; <A 4> стихира; <A 5> канон; <A 6> кондак; <A 7> икос; вторая цифра в скобках показывает следующее песнопение одного и тоже типа, например, <A 2.2> вторая вечерняя стихира. Кодировка канона еще более детализованная: <A 5.1> заглавие канона с указаниями; <A 5.2> первая песня ; <A 5.3> третья песня; <A 5.4> шестая песня. В качестве подуровня каждой песни вводится последоватльность тропарей <T 1,2,3...>. В этом

[3]Вторую песню обнаруживаем, например, в Ягичевом списке в Службе св. Симеона Столпника.
[4]*Ягич. И. В.* Памятники древнерусского языка. Т. I.Служебные минеи за сентябрь, октябрь и ноябрь в церковнославянском переводе по русским рукописям 1095-1097 гг. СПб., 1886.

разрезе данные канона будут учитываться только по этим песням, потому что цель - его общая регистрация. Выбор песней не случаен - опыт показывает, что первая песня сохранялась без особых изменений; и наоборот, третья и шестая песни подвергались изменениям, ибо они находятся на месте композиционной стыковки - после них текст песни перебывается седальным, кондаком и икосом.

<Q> - *нижний уровень,* в котором сегментация основывается на ритмично-интонационном и пунктуационном членениях. Очень часто бывает нелегко найти соответствия между ними.

Разрез II - относится к структуре службы. Цель - выявить динамику текста внутри каждой службы. Здесь соблюдается последовательность списков Иерусалимского устава, потому что он включает самый обширный круг песнопений, через который будут фильтрированы остальные конфигурации. Кроме того, как отмечает архим. Сергий, Иерусалимский устав в большой мере приводит к унификации и стабильности в строфической последовательности[5]. В этом разрезе меня будут интересовать не только представительные пробы текста, а наличие или отсутствие каждой строфы . Следовательно, при кодировке сперва будет указана номерация по порядку песнопений в службе Иерусалимского устава, чтобы отрезки текста были взаимосравнимыми в разных файлах, и потом будет обособлена специальная сигла, роль которой - представить реальное расположение песнопений в конкретном списке. Больше всего перемещений сюда наступило в связи с установлением Великой вечерни, а также и в каноне, где тропари меняются местами. Полный текст не будет вводится , а у каждого кода будет поставлен только инципит строфы.

Разрез III - его задача исследовать именно текст в полноте, чтобы выявлять лексико-грамматические особенности. Возможно в нем использовать кодировку второго разреза, но при максимальной текстовой плотности.

Окончательная цель этой сложной в самом деле методики - определить основные группы в истории текста Октябрьской служебной минеи, поставить их в схему генетических зависимостей и сделать выводы об основных изменениях - новом переводе, редакторских интервенциях.

До сих пор исследование протекает по первому и второму разрезу. Чтобы обособить на этом этапе систему отношений в рукописной традиции, были взяты пробы из нескольких списков, при чем я старалась по предварительным данным выбрать списки с достаточным количеством различий - по времени возникновения,

---

[5] *Архим. Сергий.* Полный месяцеслов Востока. Т. I. Восточная агиология. М.,1875, 143-170.

по языковому региону. Это Ягичевый список 1096 года из РГАДА; № 518 от XIII в. НБКМ (определенный как сербское правописание); № 906, XV века, НБКМ (определенный как ресавское правописание с восточноболгарского антиграфа); № 916, XVI века, НБКМ, списанный в Битольской области. Старопечатная минея (=СТРП), которая используется в наши дни. Результаты основываются на согласии подвергнутых компютерной обработке списков, без учета их различий.

Во-первых, выделено повторяющееся устойчивое ядро служб на каждый день месяца октября. В связи с этим пришлось обособить две серии подсобных файлов - одну для сравнения информации оглавления службы соответственного святого по дням (с кодом «D») и по наличию или отсутствию службы святого (с кодом «N»); - другая по инципитам канонов. Если пропустятся песнопения более поздных списков, вошедшие благодаря требованиям Иерусалимского типика, все 32 службы в текстовом отношении показывают высокий процент идентичности песнопений[6]. Только на две даты совпадающие службы две - 19 октября - св. пророк Иоиль и муч. Уар; и 28 октября - св. Стефан Савваит и св. Терентий и Неонила.

Во-вторых, определенны две большие ветви списков на основе сходств с Праздничной минеи. Большинство исследователей рассматривает Служебную минею как производную из Праздничной[7], но это мнение, по-видимому, нуждается в коррекции, поскольку предварительные обобщения показывают, что одна из ветвей в самом деле очень близка к Празничной минее, - и по идентичности служб, и по наличию строго говоря внегимнографического материала, как паремий и проложных житий; другая ветвь однако не имеет много общего с ней. В принципе проницаемость Праздничной минеи выше, и в ней славянский материал входил значительно легче. В этом контексте нужно обратить специальное внимание тримесячным служебным минеям (в нашем случае за сентябрь, октябрь и ноябрь).

Преимущество исследования по разрезам в том, что оно позволяет выявить текстуальные связи с Псалтырью, Типиком и Октоихом. К сожалению, чаще всего в практике печатных изданий пренебрегалось именно типикарскими указаниями , ибо они считались внешними по отношению к минейному тексту. Как выяснилось, типикарские указания должны приводиться под кодом каждого песнопения - например, при сигнале о появлении вечерних стихир, при сигнале о появлении канона и т.д., так что в случае необходимости эти текстовые фрагменты можно

---

[6]Для сравнения с греческим составом см. *Kulic, J.* Ricerca sulle commemorazioni giornaliere bizantine nei menei. Testi di dottorato. Roma, 1992, 18-32.

[7]*Симић, Пр.* Структура и редакције словенских минеја. - Богословље, 1-2 (1974), 67-108; *Кожухаров, Ст.* Миней. - В: Старобългарска литература. Енциклопедичен речник. С., 1992, 270-271.

эксцерпировать и переставить в самостоятельные файлы, которые в перспективу пригодятся для изучения Типика.

Предстоит выработка функциональной таблицы Богородичных тропарей в Октоихе, которая послужит фильтром при регистрировании минейных Богородичных тропарей. На данном этапе выводы говорят о высокой степени контактности - в самом деле около 25 % Богородичных непосредственно заимствовано из Октоиха, а многие были подвергнуты компилятивной обработке. Другим критерием для установления внеминейного происхождения Богородичных тропарей может служить частота их применения, например: богородичный VI песни Канона св. Иеротею (4 октября) на 8-ом гласе совпадает с богородичным III песни Канона св. Киприану и Юстине (2 октября) на 4-ом гласе:

вѕ рѣснотоу⁘ ивѕ истиноу⁘ тѧ бц̃е славимѕ

истинѣѧаго б̃а неизѧреченьно бо иеси родила

плѧтью по иестьствоу⁘ сѵвзкоупльшеѧ сѧ,

пр̃чстаѧ дв̃о мт̃и

Из-за этого придется подвергнуть богородичных в Минее частотному анализу.

Изключительно важную возможность для обработки заложенных текстов дает программа COLLATE[8]. Благодаря ее последней версии практически облекчена работа с поэтическими структурами. Она позволяет также ввести усложненную кодировку[9]. Несмотря на это, иногда все-таки в практике возникают случаи, требующие специального технического решения.

I Перестановки в месяцесловной последовательности.

1. В Ягичевом списке Служба св. Петру Капитолийскому предназначенна на 5-ое октября (вероятно под влиянием Типика Великой церкви[10]), а не на 4-ое октября, как она встречается чаще всего.

2. Случаи лабильных памятей - как память отцам Седьмого вселенского собора. В Ягичевом списке эта Служба поставлена на первом месте как обязятельно привязаная ко дню (11 октября) , а Служба св. Филиппу находится под 10-ым октября (другие списки ставят Службу апостолу Филиппу или на 11-ое октября, или

[8]Cf. *Robinson, P.* Collate 2. A User Guide (The Computers & Variant Texts Project). Oxford, 1994.

[9]О принцыпах современной текстологической работы см. *Veder, W.* La Textologie Slave. - In: Sante Graciotti (ed.). Introducione alla filologia slava. Firenze, 1995 (forthcoming). Применение программы Collate показано в работе M. Bakker. Cf. *Bakker, M.* The New Testament Lections in the Euchologium Sinaiticum. - Полата кѧигописьнаѧ, к̃е-к̃ѕ (1994), 155-208.

[10]*Mateos, J.* Le Typicon de la Grande église. Ms. Sainte-Croix n° 40, Xe siècle. T. I. Le cycle des douze mois. - In: Orientalia christiana analecta, 165. Roma, 1962, p. 61.

БАН ЦЕНТРАЛНА БИБЛИОТЕКА

вовсе ее не содерживают); по сравнению с этим остальные рукописи показывают следующее: № 518 дает под 11-ым октября Службу св. Теофану Изповеднику и ничего не сообщает о Соборе; №№ 906 и 916 делают оговорку о первой недели после памяти ап. Филиппа, а СТРП дает самые подробные указания, обясняя в какие дни можно выполнить Службу отцам Седьмого собора.

II Памяти славянским святым. Они создают огромные затруднения так как проникали в состав Минеи неравномерно .

1. Иногда Служба св. Арсению Сербскому, которая вошла под 28-ым октября, переписывалась за основным текстовым корпусом только в конце рукописи, как в №№ 906 и 916.

2. Расположение одной и той же службы в различних месяцах, как например Служба св. Иоанну Рыльскому - однажды под 18-ым августа и второй раз под 19-ым октября (т.е. Служба попадает в две разные книги).

III Случаи компиляции

1. Например в XIII веке под 16-ым августа, в Службе св. Иоакиму Осоговскому, Канон представляет собой переделку Канона св. Евфимию Великому из службы этого святого, место которой в Служебной минее на январь под 20-ым числом[11]. Списки XV-ого века уже содерживают другой канон, специально посвященный св. Иоакиму.

2. Компиляции, при которых некоторые тропари взяты из общих служб, как это происходит в Каноне на второй глас св. Иоанну Рыльскому, возникновение которого локализовано в Средеце до XII века.

3. Вот еще особенно интересный, но и трудный для решения случай - компиляция Канона святым Флору и Лавру на седьмой глас из службы, предназначенной для 18-ого августа. В этой компиляции, переработанной на южнославянской почве, можно заметить, что каждый первый тропарь греческого канона был заменен тропарем из канона на второй глас св. Иоанну Рыльскому[12]. В работе однако эти тропари надо сравнивать с соответствующими тропарями и в случаях, когда канон поставлен под 19 октября.

При всех этих ситуациях возникает вопрос, какую методику сравнения применять, потому что нет никакого смысла искать параллели в текстах, относящихся к разным дням или разным святым. Ответ должен быть обнаружен до начала разработки третьего разреза, где текст будет анализирован полностью.

---

[11] Авторство этого канона приписывается св. Клименту Охридскому. См. *Станчев, Кр., Г. Попов.* Климент Охридски. С., 1988, 120-129.

[12] *Кожухаров, Ст.* Служба за Успението на Иван Рилски. - В: Изследвания върху историята на българските диалекти. С., 1979, 220-222.

A SYSTEM FOR ENCODING EUCHOLOGICAL (MAINLY NON-CALENDAR) TEXTS WITHIN THE PARAMETERS OF THE *IST* COMPUTER PROGRAM

Maria Šniter

"Paisij Chilendarski" University, Plovdiv, Bulgaria

ABSTRACT: The present work discusses a system for encoding texts of the Slavic Euchologion with a view to their inclusion in a compete catalogue of all known early Slavic texts. For that purpose a classification of the contents of the collection was made, which led to the following preliminary conditions:

1) all possible texts contained in the Euchologion should have a numeric equivalent within the framework of four signs (figures), with the eventual possibility of adding an alphabetic coefficient; and

2) the coding should reflect not only the inner subordination of the euchological texts, but also their correlation with non-euchological (calendar or non-calendar) bodies of texts.

The system in question refers only to the description of the texts, and does not concern the codicological part of the *IST* program. A basic problem at the present stage of work is the achievement of an optimum degree of abstraction in coding, which would allow the precise identification of every text. This has been achieved to different degrees with the texts of the sacraments, where each sacrament is considered a united macrotext, regardless of the number and type of its components, and with the texts of prayers "за всяка потреба" (for every occasion), where a separate code number corresponds to each separate text of the euchemic type. This classification coding is a result of the analysis of the composition of eighty-five mostly South Slavic records, dating from the eleventh through the eighteenth centuries.

Вхождение компютерной техники в работу палеославистов в последние десять лет дало возможность наметить подступы к реализации их давнейшей мечты — создание целостного каталога известных в науке славянских текстов. Подобный корпус, каким например классическая филология располагает давным давно, мог бы облегчить и ускорить работу большого круга различных специалистов и мог бы поставить их научные исследования на гораздо более устойчивую источниковедческую базу.

Компютерная програма *IST*, разработана совместно русскими и немецкими специалистами[1]; сначало она была предусмотрена для введении текстов гомилетического типа с ясной календарной приуроченностью. Первым результатом ее использования является Предварительный каталог церковнославянских гомилий подвижного календарного цикла, подготовленный Т. В. Черторицкой.[2] В него включены 1729 триодных текстов из 128 рукописей 11. — 16. вв. преимущественно восточнославянского произхождения.

Опубликование следующего тома серии Patristica Slavica, в котором будет материал соответствующих южнославянских источников, надеемся, предстоит в ближайшее время.

Что касается текстов некалендарной приуроченности, то в пероначальном варианте программы *IST* было предусмотрено их дополнительное включение после разработки соответствующей системы кодировки (респ. — классификации), которая свела бы их огромное (жанровое, тематическое, функциональное) многообразие к схеме, обозримой с точки зрения технических возможностей программы и "прозрачной" с точки зрения будущих пользователей информации.

Настоящая работа ставит своей задачей предложить такую систему кодировки, которая смогла бы на этом этапе исследования охватить тексты,

---

[1] Программа *IST* (Индекс славянских текстов) разрабатывается и распространяется фирмой TSAR — R. Lampe, D — 69126 Heilderberg, Bothestr. 54.

[2] Vorläufiger Katalog Kirchenslavischer Homilien des beweglichen Jahreszyklus. – Zusammengestellt von Tat'jana V. Čertorickaja, unter der Redaktion von Heinz Miklas. – Patristica Slavica. – herausgegeben von Hans Rothe, Bd. 1. = Abhandlungen der Nordrhein-Westfälischen Akademie der Wissenschaften. Bd. 91., Westdeutscher Verlag, Opladen, 1994.

включенные в славянский Евхологион.* При разработке этой системы были приняты следущие постулаты:

— все возможные тексты, встрецающиеся в Евхологионе, должны получить свое числовое обозначение в рамках четырех знаков (цифр) с факультативной возможностью добавления буквенного коефициента;

— кодировка должна отражать как внутренною субординацию евхологических текстов, так и их корелляцию с не-евхологическими (календарными и не-календарными), частьми уже кодифицированными текстовыми корпусами.

Что относится к системе кодикологического описания и, соответственно, кодировки памятников (рукописей), то очевидно, чторазработаная Р. Лампе и Т. Черторицкой и представленная в Предварительном каталоге[3] система на этом уровне работы удовлетворяет в достаточной степени потребностям специалистов. Ее дальнейшее совершенствование должно стать объектом более широкого предварительного обсуждения.

Основной проблемой в классификации эвхологических текстов является необходимость найти оптимальный баланс между естественным стремлением медиевиста к исчерпывающей полноте вводимой информации и технической невозможностью (в сущности — и логической бесперспективностью) подобной амбиции. Вынужденный компромисс, который происходит от этого, приводит к некоторым неизбежным на этом этапе несовершенствам системы. Важнейшее из них состоит в том, что в

---

* Выражаю благодарность фондации "Александр фон Гумбольд" — Германия, которая оказала решающую техническую и финансовую поддержку при подготовки этой работы. Спасибо Р. Лампе за помощь и соучастие.

[3] Vorläufiger Katalog Kirchenslavischer Homilien des beweglichen Jahreszyklus, c. 24-30, 61-69.

некоторых случаях (особенно что относится к текстам вне таинств, прежде всего к освятительным и врачебным текстам — группы 3. и 4.) невозможно идентифицировать текст на основании только его кода. Например, в случае, когда известны два или больше текстов, предназначенных для лечении одной и той же болезни — скажем — головной боли, то они все будут введены под общим кодовым номером 4.2.1.1. Однозначная идентификация текста будет, конечно, возможна — благодаря инципиту и эксплициту — *все* имеющиеся текстовые данные появятся при указании кодового номера.

В сущности, этот недостаток относительно легко преодолим и не может существенно затормозить пользование введеной в программу информацией.

Существенная особенность предложенной классификации-кодировки является факт, что она — результат индуктивного анализа и абстракции содержания 85 преимущественно южнославянских памятников.[4] При их существенном темпоральном (11.—18. вв.), локальном и функциональном многообразии есть основания надеятся, что их можно принять за статистически значимое извлечение из общего корпуса евхологических сборников.

Конечно, тексты, которые мы пока успели ексцерпировать, вряд ли отображают *весь возможный* состав Евхология в Slavia Orthodoxa. Весьма вероятно, что памятники восточнославянского происхождения, которых мы пока к сожалению не успели обработать, представят новую или разную "картину" фактов. Но этого вряд ли можно категорически утреждать, пока у

---

[4] См. обзор используемых рукописей в моей статье Zur Gattungscharakteristik von Gebet und Beschwörung. – In: Материалы 4. Berliner Fachtagung "Die unbeachteten mittelalterlichen Literaturgattungen in der Slavia Orthodoxa". – FU, Berlin, Osteuropa - Institut, Abteilung Slavistik; 14.-17. Juni 1993, Berlin – Heimvolkschochschule Jagdschloss Glienike/Wannsee, Anhang 2., (in Druck).

нас нет ясной дефиниции состава, содержания и принципов организации, хотя бы одной основной группы рукописных сборников.

Обычно Евхологий определяют как сборник с богослужебной приуроченностью относительно устойчивого состава. он делится на две основные части, обслуживающие т. наз. "общее" и "частное" богослужение. На первый взгляд это — при отсуствии календарной приуроченности текстов — могло бы служить классификационным принципом при кодировке. Это так (почти), если говорить о "первой" (условно) части книги — текстах таинств. Относительно беспроблемно их можно отделить в группе с общим номером ввода 1., при чем второй знак будет обозначать соответствующее таинство, а остальные два знака предназначены для охвата разных вариантов и подтипов таинст. (Пример: 1.2. надо понимать как: 1. = таинство; 2. = крещение; начиная с третьего знака цифры с ноля до четырех обозначают разные типы крещения и принадлежащие к ним тексты). При том, что относится к уровнем абстракции и рассматривать как текст *таинство в целом*, несмотря на разнообразие, сложность и число образующих его частей.

Так при наличии идентичных заголовков и инципитов теряется возможность разграничивать и сопоставлять разные текстовые варианты (за исключением случаев, когда они функционально обусловлены — напр. при венчании новобрачных и второбрачных — тогда оба текста имеют разные коды — соответственно 1.3.1. и 1.3.2.

Внутренная логика кодировки таинств предопределяет и систему кодировки группы текстов, которые функционально связаны с ними, но не принадлежат к их совокупности, а скорее "ее обслуживают". Их мы объединили под номером ввода 7. Отношение субординации нашло

выражения в "смешения" обозначающего номера с одним знаком направо. Например:

Крещение младенца (отрока) = 1.2.1.

Приказы Йоана Златоуста о крещении детей = 7.1.2.1.

или:

Обручение и брак = 1.3.

Брачные таблици (таблицы сродств) = 7.1.3.

и т.д.

Чинопоследования, не принадлежащие к таинствам, мы обособили в группе с номером ввода 2. В ней строились по естественному порядку обряды, связанные с развитием индивида, не включенные в таинства); они являются християнским "коррелятом" основных моментов фольклорного семейно-обрядного цикла (обряды, связанные с рождением ребенка, смерть/поменальные обряды, обрряды духовного родства, разные степени и формы инициации и т. д.). Здесь мы присоединили тоже чинопоследования монашеской схимы, как аскетическую альтернативу обручения и брака.

Относительный дисбалланс между группами 1. и 2. и группами 3., 4. и 5. произошел из-за невозможности достигнуть одинаковой уровень абстракции. Каждая отдельная освятительная, врачебная или апотропейная молитва из этих групп обозначена своим собственным кодовым номером, т.е. почти всегда возможна точная идентификация текста и даже встречаемых вариантов, что — как мы увидели выше — являлось непостижимым, что относится к "макротекстам" отдельных таинств. Утешением, хотя и небольшим, в этом случае является факт, что тексты таинств в принципе отличаются гораздо большей устойчивостью и меньшей "склонностью" к варировании.

На этом этапе в наибольшей степени неокончательной и подлежащей дополнительному уточнению является кодировка текстов суточного, минейного и триодного цикла богослужения. Хотя они по существу "неевхологичны", их относительно частое присуствие в составе Евхология требует приложения соответствующего классификационного критерия. Будущей целью при дальнейшем развитии программы является установление координации при кодировки одних и тех же текстов при их появлении в разных по типу, составу и функции богослужебных сборниках. Из-за отсуствии полной системы кодировки всех текстов с литургической приуроченностью на этом уровне работы мы вынуждены остановится на том, что уже сделано.

Отдельную, притом весьма сложную классификационную проблемму (респ. проблемму кодировки) породят некалендарные и неевхологические (небогослужебные) тексты. Создание адекватной системы числовой идентификации этих текстов является задачей (надеемся) ближайшего будущего.

Из всего сказанного становится ясным, что разработка единной системы кодировки и компютерной обработки евхологических текстов является задачей с разнообразными и привлекательными исследовательскими перспективами. Ее решение дало бы возможность, с одной стороны, облегчить и ускорить создание хотя бы относительно общей для всех палеославистов текстовой "базы данных", что вывело бы славистику из этапа "предварительных обобщении", на котором она находится уже слишком долгое время. С другой стороны, что относиться до более частных аспектов изучения, необходимость сообразовываться с "чужими" для нас и "привнесенными извне" техническими параметрами кодификации, требует от нас нового взгляда на текстовы материал, который мы исследовали уже

много раз и который всем хорошо знаком. В рамках "сборника относительно устойчивым составом", каким является Евхологий, известный со своей "всеядностью" по отношении к содержании и со своей практически неисчерпаемой вариативностью, выходят наружу незамеченные до сих пор внутресистемные принципы. Становится очевидной имманентная логика этой книги, каторая в каждом отдельном мнговении и в каждом конкретном случае является удивительно точным отражением бытовых и бытийных потребностей християнской общности.

Евхологий — книга, которая содержит тексты с которами Церковь определяет санкцию практически *всех возможных* событии в жизнь коллектива и отдельного индивида. Попытка свести все это многообразие форм бытования християн к ясно дефинируемой и логически непротиворечивой системе числовых означении более или менее приурочена к неудачи (жизнь всегда богаче своих отображений и знаков); вопреки этому подобные занятия всегда дают нам шанс сделать новый шаг навстречу себе самому.

ПРИЛОЖЕНИЕ

Система кодировки евхологических (преимущественно некалендарных) текстов в параметрах программы IST

1. Таинства
1.1. Литургия
1.1.1. Иоанна Златоустого
1.1.2. Василия Великого
1.1.3. Преждеосвященных даров

1.2. Крещение

1.2.0.1. въ еже шмыти крⷭтившаго сѧ въ ·н· дⷩ

1.2.0.2. въ еже шмыти крⷭтившаго сѧ въ ·м· дⷩ

1.2.1. младенца

1.2.2. сокращенное - "страха ради смертнаго"

1.2.3. оглашение язычника

1.2.4. крещение от патриарха или митрополита в Великой субботе на вечерне

1.2.5.0. повторное крещение - принятие переходившего из другой веры

1.2.5.1. из эреси

1.2.5.2. ѿ жидовⷭ

1.2.5.3. ѿ мелхиседекитоⷯ нже и федотїанеⷯ и афⷭ

1.2.5.4. ѿ афⷭгань

1.2.5.5. ѿ сарацⷭ

1.3. Обручение и брак

1.3.0.1. млⷱтвъ еже снѣти венцы въ ·н· дⷩ

1.3.1. новобрачных

1.3.2. второбрачных

1.4. Елеоосвящение

1.5. Покаяние/исповедь

1.5.1. чинь исповеди с вопросами св. Атанасия

1.5.2. молитвы после исповеди перед принятия причастия

1.5.3. молитвы очистительные для кающихся

1.5.4. молитвы для быстрого причащения тяжело больного

1.6. Миропомазание

1.7. Священство

2. Чинопоследования

2.1. Освящение воды

2.1.1. в начале индикта 1. августа и егда хощешь - Малый водосвет

2.1.2. на Богоявлении - Великий водосвет

2.1.3. как надо омыть мощей и как пить воды от креста

2.2. Обреды, связаны с рожанием ребенка

2.2.1. молитвы первого дня после роды

2.2.2. молитвы третьего дня

2.2.3. молитвы восьмого дня

2.2.4. молитвы сорокового дня

2.2.5. молитвы егда трудно родити

2.2.6. молитвы очистительные после аборта

2.3. Обреды, связаны со смертью/поминальные

2.3.0.1. молитвы за упокой души

2.3.1. чин для тяжело больного

2.3.2. молитвы на исходѣ д҃ши - на разлоуки д҃ши ѿ тѣла

2.3.3. молитвы для умирающего от духовнаго отца

2.3.4. молитвы для мирянина

2.3.5. молитвы для ребенка до семи лет

2.3.6. молитвы для священика

2.3.7. молитвы для монаха

2.3.8. молитвы для упокоя от Пасхи до Четиридесетници

2.3.9. панахида "над коливом"

2.3.9.1. для мирянина

2.3.9.2. для монаха

2.4. Обреды духовного сродства

2.4.1. братотворение

2.4.2. чедотворение

2.5. Обреды инициации

2.5.1. подстрижение младенца

2.5.2. покрыть голову женщины

2.5.3. принятия схимы

2.5.3.1. одевание раса

2.5.3.2. малаго образа

2.5.3.3. великаго ангелскаго образа

2.5.3.4. молитвы снятия кукула в осмий день

2.5.3.0.1. монашеское запрещение Василия Великого

3. Освящения

3.1. Ритуальных лиц

3.1.1. поставление свещеносца

3.1.2. поставление еклесиарха

3.1.3. поставление четца и певца

3.1.4. поставление подиакона

3.1.5. поставление диакона

3.1.6. поставление архимандрита

3.1.7. поставление епископа

3.1.8. поставление князя

3.2. Ритуальных вещей

3.2.1. храма

3.2.2.1. полагание основы храма

3.2.1.2. благословение нового храма

3.2.1.3. обновление храма

3.2.2. иконы

3.2.3. креста

3.2.4. колоколя

3.2.5. святой трапезы

3.2.5.1. освящение святой трапезы

3.2.5.2. вънегда нѣкако позъвлет сѧ сѷтаѧ трапеза

3.2.6. антимниса

3.3. Неритуальных вещей

3.3.1. дома

3.3.1.1.	полагание основы дома
3.3.1.2.	освящение нового дома
3.3.2.	колодца
3.3.2.1.	искать воды и копать колодца
3.3.2.2.	освящения нового колодца
3.3.2.3.	очищения оскверненного колодца
3.4.	На приношениях
3.4.1.	агнеца на Пасхи
3.4.2.	первых овощей
3.4.3.	хлеба
3.4.4.	вина
3.4.5.	винограда
3.4.6.	соли
3.4.7.	над кѵтнею
3.5.	При осквернениях
3.5.1.	если ел что-то скверное
3.5.2.	если что-то скверное упало в вино или елея
3.5.3.	на каждого оскверненного сосуда
3.6.	При начинаниях
3.6.1.	начать всякое дело благое
3.6.2.	о пути
3.6.2.1.	при ухода в путь
3.6.2.2.	перед ворот города
3.6.2.3.	перед плытии на корабле
3.6.3.	если хочеш рыбу ловить
3.6.4.	когда хочешь к боярину пойти
3.6.6.	когда начинаешь пить новое вино
3.6.7.	при войны
3.6.7.1.	канон молебен о князя и о народе при брани
3.6.7.2.	против стрел во тме летящих
3.6.7.3.	при ухода на войне

3.6.7.4. для освещения военного корабля

3.6.7.5. для освещения оружия и хоругв

3.6.8. на трапезе

3.6.8.1. славословие на трапезе

3.6.8.2. запрещение на трапезе

3.6.9. в начале поста

Some Suggestions Arising from a Personal Computer Analysis of the Svetostefansko Hrisovulje

Francis R. Salter

Ontario

ABSTRACT: In 1969 the *Svetostefansko xrisovulje* in the Topkapi Museum was photographed by F. Salter and Ali Galip Erdican. An inital attempt at computer analysis of the document on a VAX was found to be too time-consuming and program-intensive. With the advent of the personal computer and DOS programming, the document was transcribed into a DBF database using a Foxplus program and, with the use of a Lettrix Church Slavonic font and several specially created characters, it was possible to print a copy of the document.

The presentation discusses the problems of programming and database structures required to produce the document and offers suggestions as to future development.

The database and program accomplished the following: (1) printed a complete text, following the page and line numbers of the original; (2) produced a complete listing of the document's words, with page and line number cited for each occurrence; (3) produced a word-reversed listing of the document, giving all citations by word endings, with page and line number; (4) produced a frequency listing of all words in normal order of spelling, and (5) produced a frequency listing of all words in reversed order of spelling.

Because the entire document is in DBF format, other analytical tools may be applied, using dbase, Foxplus, Foxpro, SQL and any other program that recognizes the DBF format.

The work described above was completed in the early stages of development of computer support for Church Slavonic writing, but at the time it nonetheless did not present any particularly difficult problems. The software currently available to the casual user is considerably more sophisticated and presents even fewer difficulties. The programming used for the analysis and production of the text is not complex, and is within the learning curve of most personal computer users. At present, the use of DBF format for database and Clipper programming would be more in keeping with trends in computer programming.

The Svetostefansko xrisovulje work demonstrates the ease of using personal computers in doing linguistic research. The procedures and programming used demonstrate the practicality of using personal computers for linguistic research in Slavic Languages. The use of current script support, programming languages, and graphics make analysis considerably easier, particularly in reducing the labor and time needed to produce a working copy of the document and make a statistical analysis of its contents. Taken to the next step, it would require little effort to build a dictionary of the document into a separate database and integrate it into the printed form of the document.

The structuring of the databases and programming knowledge required to develop computer applications for linguistic analysis is within the grasp of any serious researcher. However, in order for work completed on personal computers to be accessible to other researchers, it should be exportable via electronic media in a recognized database and script format. Research will be enhanced considerably in the future when there is: (1) a standard for the developers of Church Slavonic scripts appropriate to the language (in the above instance, 1300 Serbian), which will make text in databases exportable; (2) a rationalized sorting algorithm for sorting Old Church Slavonic characters; (3) a computerized dictionary of Church Slavonic, preferably available on CD-ROM; (4) text printing and text analysis software; and (5) CD-ROM production of early Slavic documents which permits perusal and analysis of texts without requiring access to printed documents.

Document Acquisition

During the summer of 1979 Ali Galip Erdican and the author of this paper requested permission from the Government of Turkey to examine documents in the historical archives in Istanbul. Upon arrival, it became clear that we would not have access to the catalogues in the archives for which permission was granted, and permission was requested *ad hoc* to work in the TOPKAPI MÜZESI instead. During that time we photographed several documents in the museum, among which was the *Svetostefansko hrisovulje*. Unfortunately, there was no quality control over the photography, which was done with a Minolta camera in a using the windowsill of the reading room as a stand. The document was photographed in these relatively primitive conditions. After photographing the documents in the TOPKAPI MÜZESI, we moved to the Foreign Affairs Archives and photographed the diplomatic documents of the 1876 Russo-Turkish war.

When we returned to Canada, the photographs were developed and prints were made. Most of the prints were clear and readable. However, because of time required, no attempt was made to reproduce the document until 1985, when it became apparent that computers had passed from being number crunching machines to data processing machines. An initial attempt to process the document was made and the entire document was put on computer punch cards, but VAX computers at that time were not compatible with analysis of Slavic text. Work on the document at this time consisted only of a transcription of the document and a transcription onto computer cards. However, in 1989, personal computers had reached the capacity and price level that made it practical for an individual user to process the document on a microcomputer. During the summer of 1989, when time

permitted, the data were entered, checked against the photographs, corrected, and the programmes were written and used.

Objectives of the Analysis

The objective of a computerized analysis of the *Svetostefansko hrisovulje* was to produce a product which would permit the user to print the original, a normalized copy of the original text and any text or parts of the text which were the subject of analysis.

Computer Platform Selection

The platform used was IBM DOS using the programme Foxplus for database operations and Lettrix was used as the printing programme. Lettrix was selected because it was one of the few available IBM DOS compatible Slavic character programmes available at the time, and it was also programmable in that characters could be created. Special characters were created in Lettrix to correspond with the Old Serbian characters. For the database, - was used for I, ? was used for ja, Q was modified to match the text šta, . was used for an omega, , was used for the ot character, } was used for je. The characters were created using a Lettrix utility. Foxplus was selected as a database programme instead of dBaseIII or dBaseIV primarily because of cost, and also because the programmes were not complicated. The simplicity of the programmes also excluded the use of Clipper, although these programmes may be more desirable for persons who are less experienced in programming and computer operations. Given the present conditions, Windows would probably have been selected as a platform, the programming language probably would have been Visual Basic and a Windows Slavic type font could be

ordered from a company such as Svitcom. Because of the flexibility and wide-spread use of the .DBF database format, it would still be the database of choice.

Data Base Structure

The first database required is the input database for the document. For processing purposes, it was necessary that each "word" [word here used in the sense that it is a word or part of a word on a line in the manuscript] from the document be entered as a separate record in the database, and the record should also include the page number, line number, the "word" and a logical field called "WordEnd" which permits the processing of the actual words that have been split at the end of a manuscript line. A physical count of the number of characters in the document indicated that there were no words greater than twenty "characters" [characters in the sense that an apostrophe is also a character] in a word. All other fields were defined to accommodate the necessary number of bytes per field. The database was called UROS.DBF and when complete, it contained 9014 records. The .DBF file definition is:

FIELD NAME	TYPE	WIDTH
PAGENO	NUMERIC	3
LINENO	NUMERIC	2
WORD1	CHARACTER	20
WORDEND	LOGICAL	1
TOTAL BYTES PER RECORD		26

A second database was created for the purpose of analysis. Although it would be possible to programme and derive results from the UROS.DBF database, it seemed prudent to create a second database for processing and to maintain the initial database intact. The database for analysis was called FINAL.DBF and in its final form contained 7998 records. The difference in number is the result of all words which are split at line's end are combined in the database FINAL.DBF. The writers of the document split 1016 words at either line's end or page's end. The structure of the FINAL.DBF is:

FIELD	TYPE	WIDTH
PAGENO	NUMERIC	3
LINENO	NUMERIC	2
WORD1	CHARACTER	20
WORDEND	LOGICAL	1
LINENO2	NUMERIC	3
PAGENO2	NUMERIC	3
BACKWARD	CHARACTER	20
TOTAL BYTES		53

WORD1, LINENO and PAGENO were imported from the UROS.DBF database. The Logical Field WORDEND indicated that the word was to be joined because of a line break or a page break, the line number was recorded in LINENO2 and the page was recorded in PAGENO2. Thus, the entry in FINAL.DBF was the actual word in the text. The field BACKWARD was used t make a reverse order

dictionary of the document. It goes without saying that a reverse dictionary is quite useful when analyzing a declined and inflected language Frequency counts are also facilitated when indexing is possible on the reversed order entries are present. Construction of grammatical endings is greatly facilitated using a reverse dictionary. Despite its simplicity, the database FINAL.DBF makes it possible to print the entire document in a "normalized" version, albeit without the original word, line and page numbering intact. It is also possible to index the database on WORD1 or BACKWARD and programme a word count and print the word counts giving each word, along with its location on a page or pages and line or lines. The reversed lexicon shows, for example, the double jer genitive plural in wide usage. The grammatical word endings for various cases, coalescence of jeri and I, and other characteristics of the language at this stage of development are visible in the reversed dictionary.

Word searches may easily be made and the results printed into a file. Ad hoc searches for words such as selo, reka or Uroš can easily be made. The purpose of the database basically was to provide statistical and grammatical information, but it may also be used to locate all appearances of individual names, villages or other items of interest.

Programming

As mentioned above, the programming language selected was Foxplus. Programmes were written for data input, transferring the data to FINAL.DBF and reversing the characters in FINAL.DBF. After several attempts to directly print documents from the database, it appeared that saving the results of a programme to a data file and printing the file was a surer method of obtaining the desired results,

particularly when using Lettrix as a type font and older 8086 computers. Programmes were created for printing the text out to file in the original form [called ORIGINAL.DOC], printing out a working copy with all separated words reconnected [called UROS.DOC], printing to file a word listing of all words with normal indexing, complete with location on page[s] and line[s], print to file a listing of words indexed on the reverse order of characters. More sophisticated programming made it possible to make text files of word searches. An example of these programmes may be seen in the lists of Bishops and Igumeni, which was done by indexing and searching for part of the word "jepiskop" and "igumen", and retrieving the preceding and following words. Although the results of searches frequently must be edited for presentation, the search itself may be quickly done and the results stored in a text file.

When the document was produced, the data was sorted in ASCII key. No attempt was made to create a sorting algorithm that would produce a sort using a Slavic alphabet.

Objectively speaking, the goal of this project was to demonstrate that computerized assistance in making an analysis of a document would greatly diminish the amount of time and effort required to produce a copy of the document in its original and normalized form. The period of time required to input the document into a database and produce the present material was four months. The preliminary work, a transcription of the document from the photographs, was done previously. The work on the document was enhanced considerably by the erudition of the writer of the *Svetostefansko hrisovulje*, who made very few spelling errors and had few deviations in vocabulary. Since the author of this paper has never had the

experience of producing a document such as this without the assistance of a computer, a time comparison by personal experience is not possible. However, after a discussion with scholars who were working on the document without computerized assistance, it appeared that a research team would be working on the document for three to four years. Obviously, the use of computers makes it possible to achieve the level where basic research is possible in a shorter period of time. More importantly, the use of computers alleviates to a great degree the tedious work which must be done before proceeding to serious scholarly work. There are, however, many deficiencies in the tools that were available at the time when work was done on the document, and many tools are not available at present.

Deficiencies

When the work was done on the *Svetostefansko hrisovulje*, there were few available Slavic scripts for IBM DOS computers. Typing had to be done using a Latin screen display script and then transformed into a Slavic script at printing time. The situation has improved considerably, and at present a Slavic script which was obtained from SVITCOM makes it possible to view the Slavic script on the video in Windows applications. The Slavic script will also function in Visual Basic programmes which drive databases created in a .MDB format. None of the script programmes which have passed over my desk to this time permit the user to modify characters which function in Windows programmes. Because it is not a lucrative financial venture, script programmes which are either capable of modification or which have historical Slavic characters are not going to be easily available. Another deficiency is the lack of historical Slavic dictionaries, such as Professor Thomas Lysaght's Old Slavonic - Russian Dictionary, in a format addressable by computer.

Until such a dictionary is available, the tedious work of typing in the glosses will have to be done manually.

Considerations for the Future

Improved technology may offer a solution to some of the deficiencies. Taking a look into the future, scanning technology that would scan historical documents would reduce the amount of time required for inputting documents into a database word by word. It goes without saying that no scanner will input the entire text accurately, but an acceptable level of recognition would reduce work considerably. If the work on the *Svetostefansko* hrisovulje were to be done today, it would be typed into a text file, and the database for analysis would be made from the text file on a page level using Visual Basic as a programme. SQL would be used as a query language.

It would help greatly if dictionaries were scanned and put into a format which is addressable by computer. CD-ROM technology has made mass storage available at the personal computer level. Eventually, CD-ROM's could contain the drivers for historical Slavic characters. Instead of looking up every word in the document to make a glossary or dictionary, CD-ROMs could provide the researcher with a dictionary applicable to the time and language being researched. The computer programme required to search the CD-ROM dictionary for the appropriate gloss could be part of the CD-ROM for those who are not programmers. Complete documents could also be recorded on CD-ROMs, making them available to scholars who would need only a personal computer and printer to do research. The results of the research could also be made available via the same technology. Finally, the interested scholar could connect to the database through the worldwide web,

download the documents for research or upload any new documents into a permanent library of historical Slavic documents.

The St.Sophia Project

A Stage in its Realization and the Prospects for its Completion

Velina Bratanova, Elena Kotseva, Dragomir Petrov

Center for Cyrillo-Methodian Studies
BAS, Sofia, BULGARIA

Keywords: CD-ROM, Multimedia, St. Sophia

It would be no exaggeration to say that the idea for the "St. Sophia Project" arose as an intellectual challenge which for me was the "Perseus" Project, devoted to the civilization of ancient Greek culture. We deliberately say *culture*, for in our work we adapted a differentiation in our terminology concerning stages and the degrees between culture and civilization, introduced by the German school, and alien to the spirit of the French and English school. The disk with the "Perseus Project", the work of a large team of American and European scholars, was demonstrated before Dragomir Petrov by Prof. Francois Charpin in Paris in June 1991. Additions to the disk with the Perseus Project has continued consisting of increasingly wide range of material within its structure. In the course of discussions at the time with a circle of

scholars and chiefly with Prof. Charpin, a professor of Classical languages from the University of Paris VII and SNRS, with Allain Vuillemin, a specialist in the use of multimedia in the humanities and literary comparative studies, associate professor at the University of Limoge at the time, and subsequently at the University of Aras, and not least with Peter Burchet, two projects arose for a video CD dedicated to the Bulgarian cultural heritage, namely the "Orpheus Project" and the "St. Sophia Project". Several visits of Prof. Charpin and Prof. Vauillemin followed as well. From the very beginning we were supported by Prof. Alexander Fol, then Director of the Terra Antiqua Balcanica Association and Prof. Velizar Velkov, in his capacity of Vice Chairman of the Bulgarian Academy of Sciences in 1992. Circumstances gave the "St.Sophia" Project precedence and today we can demonstrate the completed first stage of our work - the proto-disk "St.Sophia". This was greatly due to the inclusion of the "St.Sophia" Project in the Memory of the World Program of UNESCO, a long-term program where the St.Sophia Project is the second completed product after the disk dedicated to the MSS of the Albertina Prague Library. The work done so far was funded by UNESCO and the Open Society Foundation in Sofia.

In its final form the principal concept of the project was compiled by the late Velina Bratanova, the art historian, Elena Kotseva, a specialist in Medieval

history and Dragomir Petrov, a writer and anthropolgist, and Editor-in-Chief of Courier UNESCO in Bulgarian, and presented in French. The text, edited by Prof. Vuillemin was presented to UNESCO by the Bulgarian National Commission for UNESCO. At the signature of the contract, signed for the Bulgarian authors by Dragomir Petrov, the following was envisaged: the formation of a Bulgarian team, which would select the presented monuments, describe them according to a system, and organize their photographing; the formation a French team providing the French and English translation and process the material in an informatics format. This team work on behalf of the EPI Association; the shooting of the selected monuments by the Tribun Society, which also undertook to produce the port-disk. Later Tribune refused to make the photographs other that of those of the MSS, hence it became necessary to search for additional funding from the Open Society Foundation in Sofia, while Bulgarian photographs and operators shot the museum objects, archeological monuments and put together the material of ritual and traditional music in a fifth data base.

* * *

We should like to outline the idea of the St.Sophia Project. We will quote the selection of excerpts from the introduction of the project itself, in the form sent to Prof. Vuillmin for consultation.

"The principal aim of the project is to study the transition of Balkan traditional non-written culture to a written civilization, through the adoption of Christianity. This presuposses the consideration of the main characteristics of the culture of the Eastern Roman Empire (2nd - 4th Century) in the adaptation that followed in various traditional non-written culture (Thracian, Slavic, Proto-Bulgarian). Gnostic dualism is considered, together with the dualistic between literary and traditional culture. Within the Balkans these processes led to a range model of adaptation, which found its continuation and its oikomen in he lands of the Eastern Slavs, becoming a static model of adaptation towards the heritage of the Eastern Empire."

"Antiquity has described the Goddess of Athena as the protector of builders the supreme activity of the socium (order as an opposition of chaos), as well as the Goddess of the arts and artisans).

"Biblical didactic literature, in Song of Solomon (I 8,9), Sirach, Wisdom of Jesus, define the personification of the Lord as "a breath of the power of God" (Wisdom of Solomon, 7,25). According to Paul Jesus Christ is "the

power of God and the Wisdom of God", e.g. the image of Christ-logos, cf. Genesis, "In the beginning was the word"."

"Pausanias (I,40,4) mentions that around the statue of Zeus in his temple at Athens the images of the Moerae were depicted, for "knowing the destiny of human deeds", and "predestination and fate", they only obey Zeus and Supreme wisdom" (Pausanias, I.40,4).

"The travels of Paul led him on many occasions through Phillipi, Salonica and Corinth and account for the spread of the idea of the Supreme Wisdom in these lands. Thus in the 6th century in the capital of Eastern Christianity we find the enormous dome of the church of St.Sophia, a structure which coincides with the lasting settlement of Slavs in central and eastern Europe and the formation of their own states. With the construction of the church of St. Sophia in Constantinople we find the final decisions establishing the liturgical order, the symbols and organization of Christian churches."

"The mystical concept of the Church in Christian Latin literature in the east took the shape of sophology, which built up the system of human knowledge, the gentle visions of the muses, the Parcae, the "devi" of Thracian and Indo-European origin, the oracle of Athena, arriving at the identification with the image of Christ-Logos, the third hypostasis of the Holy Trinity."

"Wisdom - Sophia, is close to the concept of the beginnings, the source, in the sense of the womb of the mother; wisdom, stemming from practice to the spiritual world."

"The warm mineral spring, which became the centre of the settlement of ancient Serdica, were the natural environment of the cult of Sophia, substituting the ancient deities Demetra, Athena, Tuche, and co-existing with those of the healing Gods: Asclepius with his daughter Hygeia and Apollo medicus."

"In the 2nd century, in the reign of Hadrian a Roman woman from Serdica and her three daughters sacrificed their lives in the name of the faith of Christ, and linked their names with the beginnings of the paleo-Christian martyrium. Their names were Sophia, Pysthia, Elpidia and Agapia (Wisdom, Faith, Hope and Love) - a personification of divine wisdom, as the mother of the three virtues of theology. Probably the cult of Sophia, Pysthia, Elpidia and Agapia once again gained popularity under the Carolingian rulers, under the influence of their eastern mentors."

"The four holy martyrs were depicted in the early 11th century in the Menologium of Emperor Basil II. The depiction of Sophia-Holy Wisdom developed in Orthodox iconography (above al in Moscovy Russia after the 15th century), where she is represented as a winged angel with a red face, dressed in imperial robes. St.John the

Baptist and The Virgin Mary turned to her in their prayers. Higher up, and close to the angel Sophia is the blessing Christ."

"Few of the calendars that have come down to us from the 10th century fix 17th of September as the feast day for the memory of St.Sophia and her three daughters."

"In other words this was the project for a hypertext dedicated to Proto-Bulgarian culture and Orthodox Slavic civilization, which arose as the outcome of the dialectics traditional non-written culture and the literary traditions of the centuries old history of the city of Sofia and the cult of the saint of the same name."

We quoted part of the introduction to the project, for here we find clearly outlined two approaches of research through which we envisage to organize the material presented: synchronic-typological and diachronic-retrospective (according to the terminology of Ferdinand de Sossure, applied by Claude Levy Strauss in cultural anthropology). Here however they are structured on one and the same vertical axis, going back to the depths of eight thousand years, of the mineral springs of Sofia, in order to become an omphalos, to quote Mircha Eliade, of the Bulgarian cultural space, to be found behind the Medieval name of the city - Sredets. Obviously the methodology of French analysts and F. Brodel are behind the concentrated study of a limited territory through the

descriptive and analytical possibilities of various fields - history, archeology and numismatics, the history of art, paleography and literary comparavistics, and ethnography and musicology. Nevertheless, as was noted elsewhere, the emphasis of these five data bases, what is sought was the role of the woman, in images, symbols, mytholgemes, in other words the universal spirit, contained in its very name in a dialogue with the masculine element, a dialogue through which we aimed at presenting the historical (diachronic) actualization of the synchronic principle. This is why most of the described objects and finds originate from Sofia and the vicinity, or in rarer instances a pan-Bulgarian character and are kept in museums in Sofia. The realia through this unversalia however is an expression of prospective (horizontal) diachronics, of concrete historical content with its implicit temporal length of the myth, rite, or text, either explicitely, through the events they represent.

What has been done so far and what is forthcoming?

1. The structure of the disc of 5 data bases has been created:

Sophia 1:

also called for practical purposes iconongraphic as it presents through slides the geographic position of Sofia as a cross roads of the peninsula, its natural setting and appearance; the basilica of St. Sophia, as a symbol of the city, with numerous details and archival photographs, tracing all stages of its study together with the necropolis surrounding it; a selection of archeological finds from Sofia and the vicinity, representing the century old history of the town; a collection of coins, minted by the Sofia mint together with seals of Bulgarian rulers; a collection of icons kept in museums in Sofia as well as a selection of frescoes from preserved in Medieval churches in Sofia; a selection of samples of secular art from the National Gallery of Fine Art; a collection of ethnographic monuments, from the Ethnographic museum at the Bulgarian Academy of Sciences.

Sophia 2

Samples of Medieval MSS from the "St.St.Cyril and Methodius" National Library, which represents the wealth of their collections, the Library of the Bulgarian Academy of Sciences, the Church Historical-Archeological Museum, National Historical Museum, Sofia. The French team envisaged the photographing of 289 MSS, described from the Catalogue of the Library as Bulgarian. In this way we envisaged protective photocopying of documentation from the Bulgarian MSS heritage, one of the goals of UNESCO.

For technical reasons 64 MSS were photographed together with an incunabula (The Life of St.Petka of Turnovo, from an Menaeon of Bozhidar Voukovich). One thousand shots have been made, described according to objects within the range between the 10th-18th century. About 100 objects ornamental elements and miniatures have been described with a commentary, applying the EPI scheme.

Sophia 2 aims at presenting Medieval Literature (Ivan-Alexander's "Song Book" (Psalter) of 1337, 16 Sermons of Gregory the Theologian of 1355),from a utilitarian and artistic point of view, designed for everyday cult uses or personal use, and made according to the customary and aesthetic style.

MSS were selected depending on the degree of their studies, and chronology. The author of the section dates

the Book of Apostles of Enina, NB N 1144, towards the end of the 10th century, which was unacceptable for the French experts, who date it from the 11th century.

With a view of presenting the work of scribes and content the entire Book of Apostles of Enina were photographed, together with the Argirov Triodion and the Skopie Menaeon (partially); the Life of St. Petka, the beginning of the Wonder of the Bulgarian, together with a Life of St.George, and the Synodicon of Bulgarian Kings (Palaouzov and Drinov copy).

The prerequisite in photocopying of samples of MSS is the view the Medieval MSS were written in information code of the alphabet and functionally subordinated to the decorative system. The graphic code of the alphabet is seen as a set of characters with diacritical signs on the lines and above them (standing for titles, stresses, ekphonic meanings, points, asterics, etc.).

The columns of the field appear on the sheet, determined by the shape and lines, with signal lettering (the title, the initial, segmented letters, marginal notes) and the lettering (letters and non-letter signs) in two types - black and coloured. Letters and ornaments (ornamental motifs) are treated as "images", built up at various information levels of the text. "Text" in the project stands for the recorded letter, ornamental system, miniature, and colour range. According to the author of Sophia 1, 2, 3 this definition determines the iconologic

comparison of the separate data bases. Another requirement for the methodology of Sophia 2,3 and 4 is the position of data bases Sophia 1,4 and 5 ; they are seen as "texts" of archeology, folklore and music, comparable through "syntagmas" and "motifs" standing for their specific features.

Lettering sets and the decorative system from the semantic point of view will allow a more exhaustive edition of the "St.Sophia" disc with typological and genetic comparisons between the alphabetical code, ornamental elements, miniatures, and formal features of MSS included in the project of "texts".

Sophia 3

The integral presentation of of the oldest Bulgarian MS - the Book of Apostles of Enina, N 1144, National Library Cyril and Methodius.

The MS was selected because of its dating from the 10th century and its glagolithic lettering. The format of the MS

and the canon of the Cyrillic lettering have their parallels in Codex Vat.gr.2502. In its context the MS is a reference point for a/ the study of the liturgical text; b/ liturgical practice of the Bulgarian Church in the 10th-11th century, c/feasts, d/the type of lectionary liturgical books and auxiliary texts for the apostolic activity of the church. (9th-10th century).

The Glagolithic reading of the Giant initial "CH"="Ꙋ" at the beginning of the Cycle of the Long Fast allows the functional and semantic comparison with similar parallels in Bulgarian, South Slav, Byzantine and Italo-Greek MSS and sheets.

Sophia 4

The integral representation of one of the most important Medieval Bulgarian texts, known as the Synodicon of Bulgarian Kings in the two versions the Palaouzov version from the 80ties of the 14th century, National Library Cyril and Methodius 289, and the Drinov version from the end of the 16th century, National Library Cyril and Methodius 432. with a translation in modern Bulgarian, French, and English, The Synodicon was chosen because of its unique content and meaning of a historical sources, and the Byzantine tradition of Synodicons.

Within the project the Synodicon is the basis of the hypertext of Bulgarian culture:

- meso-texts and their stability (Narratives and Chronicles of the Oecumentical Councils.
- the Eklesia and the Council principle in the
- Christian church;
- The Letter of Paul to Corinthians, Thessalonians Philipians; cult work and quasi-apocrypha dedicated to

St. Paul, early images of St. Peter and Paul; the cult of St. Tekla.

- diptichs
- depicted thrones and Ecumenical Councils
- images of clergymen, rulers and laymen (texts and images)
- chronicles, diptychs, synodic, iconological links.
- local cults
- dioceses of the Bulgarian Orthodox Church;
- rulers, dynasties and titles
- "those courageous in the name of Christ" and ethical traditions.

Sophia 5

An audiovisual data base where two sequences of a total 6 min presents a Church feast "The Assumption of the Virgin", November 21st, examples of the Bulgarian folk feast "Lazarouvane", and "Koleduvane", were the feminine role has been successfully traced and presented through the institution of the initiation of maidens also hinting at the dialogue with the male element, standing for dischrony. The data base contains 18 m recording of traditional Bulgarian folk music and church music.
In this variant the protodisc consists of approx 200 images, each accompanied by a descriptive form, according to EPI requirements, many of them with notes and remarks,

constituting several lines to several pages length, attached to each data base and feature (chiefly linked with data base Sophia 1). The disc is accompanied by an introduction. At this stage this material represents some 600 pages. Unfortunately owing to failure of the French team to carry out their contribution, these texts are missing in the proto-disc with the exception of several brief introductory texts.

The compilation of the sections and description of the material, included in the various data bases or features of the proto-disc was the work of leading Bulgarian specialists in the humanities, namely Arch.Prof. Christo Genchev, Chair of the National Heritage Council, Margarita Vaklinova, Senior Researcher, Director of the Archeological Museum at the Bulgarian Academy of Sciences; Prof. Yordanka Juroukova, Director of the Archelogical Institute, BAS, ; Elisaveta Mousakova, Art historian, Researcher; Associate Prof. Dora Kamenova, Head of the Department of History of Art, the Academy of Fine Arts; Velina Bratanova, Lecturer, Academy of Fine Art, Researcher, Ganka Michailova, Institute of Ethnography at BAS; Elena Kotseva, Senior Researcher, Center for Cyrillo-Methodian Studies, BAS; Dragomir Petrov, member of the Department of Mediterranean Cultures and Societies, New Bulgarian University; Vergilii Atanassov, ethno-musicologist; Ljuben Botousharov, ethno-musicologist; and Elena Toncheva,

Researcher, musicologist, specialist in music from the Middle Ages.

The team in France included: Prof. Allain Vuillemin Elena Georgieva, a philologist and Marie Vrinat, Elena Foulliaron, Associate Professor INALCO, and Prof. Jean-Pierre Levet from the University of Limoge, the author of a fine study on the Book of Apostles of Enina and a contradictory text on the cult of St. Sophia.

How do we see the continuation of our project?

1. The working out of a more precise classification of the presented images and their organizations in features, together with the accompanying text.
2. The further enrichment of data bases with a view of greater balance in the different types of included monuments.
3. Entry of already prepared texts and the compilation of new ones, in line with adopted standards, for newly included monuments
4. Revealing of semantic (synchronous) and historical (diachronic) correspondences and links between shown exhibits through cross-references of data bases and between separate monuments: e.g. the common semantic field between hot springs in down-town Sofia and the universal female principal, represented by goddesses and Christian women saints, probably reaching the idea of the "shrine" as a seat of the Holy Spirit, similarly to the Holy Grail as the seat of the Son of God. Or the

semantic (possibly a cause sequential link) between Mitra, Thracian Horseman, Tsar Constantine, and St.George, mythical characters such as Marko Kraleviti etc.

However, before that we should do something simpler to illustrate the texts of Bulgarian rulers from the Syndicon with their images on seals, coins and frescoes. The potential for illustration of the musical recordings from the same Synedicon will be part of data base Sophia 5.

Probably two new data bases will have to be created, because of the separation of video cassette and the audio and the inclusion of contemporary *telles quelles* constituting both folklore and literary monuments and MS. If this turns out to be possible, within a subsequent stage of the project, which should make up the final disk of the St. Sophia Project we would also attempt to outline horizontal descriptions, e.g. the spacial spread of this culture, respectively civilization, whose roots are here, in the cradle of Mediterranean civilizations.

APPENDIX

1. Form for manuscripts

Titre:.....Le Nouveau Testament

Sou-titre:.....

Période:.....

N:.....

Folios: .....III + 256

Support: .....papier

Auteur texte:.....

Copiste:.....

Enlumineur:.....

Décoratifs:.....0

Initiales:.....0

Miniatures:.....4

Observations:.....

Date:.....A.D. 1658

Catégorie:.....460

Dimensions (mm): 300 x 205

Nature du texte:.....

Type:.....miniature N folio:.....

Légende:.....

Observations:.....

Type:.....miniature N folio:.....

Légende:.....

Observations:.....

Type:.....miniature N folio:.....

Légende:.....

Observations:.....

Type:.....miniature N folio:.....

Légende:.....

Observations:.....

2. **Form for other objects**

RAPPEL DU NUMERO:

ПОРЕДЕН НОМЕР:

1. LEGENDE:

ЛЕГЕНДА:

2. CONTENU:

СЪДЪРЖАНИЕ:

– Sujet:

ОБЕКТ:

– Thèmes:

ТЕМАТИЧЕН АРЕАЛ:

– Lieux (représentés):

МЕСТА:

– Personnages (représentés):

ОБРАЗИ:

– Eléments naturels (évoqués):

ПРИРОДНИ ЕЛЕМЕНТИ:

– Objets matériels:

МАТЕРИАЛНИ ОБЕКТИ:

– Décor (représentés):

ДЕКОРАТИВНИ ЕЛЕМЕНТИ:

– Inscription:

НАДПИС:

– Autre indication:

ДРУГИ БЕЛЕЖКИ:

RAPPEL DU NUMERO:

ПОРЕДЕН НОМЕР:

4. COMMENTAIRES:

КОМЕНТАР:

5. TERMES DESCRIPTEURS:

УСТАНОВЕНА ТЕРМИНОЛОГИЯ:

Русская версия источнико-ориентированного программного обеспечения КЛИО: новые возможности обработки источников

В.С. Тяжельникова

Россия. Московское городское объединение архивов,

ведущий научный сотрудник.

keywords: источнико-ориентированная обработка данных, КЛИО, информационные технологии.

abstract: Окончание работы над русской версией системы КЛИО, является поводом для того, чтобы обратить внимание специалистов на возможности источнико-ориентированной обработки исторических источников на кириллице и обозначить задачи возможной совместной работы специалистов по компьютерной обработке славянских памятников.

КЛИО (κλειω или KLEIO) - это пакет программ, созданный специально для источнико-ориентированной обработки данных в процессе исторических исследований, который начиная с 1978 г. разрабатывается доктором Манфредом Таллером в Институте истории общества Макса Планка (Геттинген, Германия).

Постоянное развитие программы в течение последних лет, позволило ей вырасти в обширный пакет, который располагает и предоставляет в распоряжение историка наряду с разносторонними средствами, ориентированными на управление банком данных, также и графические

возможности компьютера - производство карт и обработку изображений (**Image Processing**).

С 1989 г. развитие программы происходит в рамках международного объединения проектов **"Рабочее место историка"**. В этом объединении интегрируются усилия нескольких европейских институтов, которые совместно преследуют цель создания на ЭВМ всеобъемлющей и удовлетворяющей целому ряду функций основы для рабочего места историка. Это разностороннее развитие по некоторым позициям уже завершилось, причем результатами являются либо самостоятельные программы-спутники, окружающие КЛИО, либо части непосредственно интегрированные в нее.

К программам первого вида принадлежит программа преобразования данных, разработанная в 1988-1990 гг. в Институте истории Макса Планка под названием **StanFEP (Standard Format Exchange Program)**.

К программам второго вида принадлежит система лемматизации латыни, созданная в Риме, а также имеющий большие последствия для дальнейшего развития системы КЛИО модуль обработки изображений, который создавался в результате совместной работы между Институтом истории Общества Макса Планка (Германия) и Институтом изучения реалий средневекового и раннего новейшего времени в Кремсе (Австрия). Модуль содержит богатый арсенал технических средств обработки изображений, а именно: усиление изображений, непосредственный поиск изображений, связывание изображений и - наконец - распознавание образов.

Наряду с этим подготовлена, но в текущей версии программы еще не содержится система так называемого "контекстно-чувствительного" управления данными, которая нацелена на контекстно-взаимосвязанное использование знаний по данным, благодаря чему источнико-ориентированная обработка могла бы быть еще сильнее интегрирована в исследовательскую практику.

Таким образом, наличие в настоящий момент двух версий программы обусловливается модулем обработки изображений со всеми соответствующими техническими логическими операциями.

- **Версия 5** включает все программные средства, ориентированные на работу с банком данных, включая возможности производства карт, а также неграфические меню (**PC-версия**).

- **Версия 6** идентична версии 5, но содержит дополнительно модуль обработки изображений, а также необходимый для этого графически ориентированный интерфейс (**UNIX-версия**).

В настоящий момент идет работа над версией программы под **Windows NT**, которая позволит не только обрабатывать изображения, но и будет иметь графический интерфейс для программы КЛИО в целом.

Как для пятой, так и для шестой версии имеются в распоряжении следующие версии программы:

- *латинско-немецкая* (латинский командный синтаксис, сообщения и меню на немецком языке);

- *английская* (английский командный синтаксис, сообщения и меню на английском языке);

- *англо-русская* (английский командный синтаксис, сообщения и меню на русском языке). Русская версия позволяет также в процессе структурирования данных писать имена информационных единиц на русском языке.

Полное Описание системы имеется на английском и немецком языках, в 1996 г. увидит свет аналогичное Описание на русском языке. Учебник существует на английском, немецком и теперь также на русском языках.

Окончание работы над русской версией системы КЛИО, которая возникла в результате совместных усилий Московского городского объединения архивов и Института истории общества Макса Планка в Геттингене, является поводом для того, чтобы обратить внимание специалистов на новые возможности обработки исторических источников

на кириллице и обозначить задачи, где возможна интеграция усилий специалистов по компьютерной обработке славянских памятников.

КЛИО является примером программного продукта, в котором нашла выражение система взглядов автора на целую область знаний, где концепция с каждым годом все более филигранно реализуется в технических приемах, способных удовлетворить самых взыскательных пользователей.

В основе системы КЛИО лежит *идея источнико-ориентированной обработки исторических данных.* Цель источнико-ориентированной обработки - создать все возможности для того, чтобы исторический источник и его информационное содержание *максимально близко оригиналу полностью* могло быть воспроизведено на компьютере. Поэтому источнико-ориентированная обработка данных может включать все усилия, которые производятся с помощью компьютера, благодаря возможностям хранения, управления, подготовки данных при освоении исторического материала.

В этой связи важно заметить, что воспроизведение исторических источников на компьютере в адекватной оригиналу форме нацелено не только на публикацию источников в машиночитаемой форме и не является самоцелью.

Ориентация на источник как таковой и соответственно стремление сделать возможным максимально адекватное воспроизведение источника на компьютере вытекает изначально из стремления *сохранить источник в его естественном виде* (в не в виде, например, кодов) *для позднейшего анализа* и *предоставить исследователю многофункциональный набор средств обработки как таковой* - от приемов в области управления банком данных до пространственного представления сведений и обработки изображений.

Центральное положение заключается в следующем: историку предлагается возможность ввести данные в компьютер в формате, который

основан на источнике, что позволяет свободно выбирать метод анализа источника в процессе работы с ним (а не до начала обработки). Концепция источнико-ориентированной обработки данных моделирует следующую естественную ситуацию работы с историческим материалом: до начала работы исследователь еще не только не может ответить на все вопросы интерпретации источника, но даже не осознает их все в полной мере. Исследовательские вопросы возникают и изменяются непосредственно в процессе работы.

Если же материал настолько хорошо известен, что историк в состоянии выразить свою исследовательскую концепцию в четкой системе показателей, то и в этом случае может использоваться сильная сторона КЛИО, выраженная в проблемно-ориентированной по подходу и источнико-ориентированной по форме обработке данных.

Следует заметить, что только источнико-ориентированная форма обработки данных позволяет работать на компьютере с машиночитаемым источником, адекватным бумажному оригиналу. Получить максимально адекватный оригиналу источник на компьютере позволяет целый ряд действительно серьезных технических процедур.

Основное свойство технических процедур, которые предлагает КЛИО - **гибкость**. Гибкость присуща как основным принципам структурирования информации, так и всему спектру технических решений, нацеленных на обработку данных.

Первостепенное значение для историка имеют **принципы структурирования информации**. В этом смысле КЛИО на сегодняшний день предлагает максимум гибкости. Назовем основные преимущества:

- Отдельное поле банка данных (в КЛИО оно называется элементарной информацией) может иметь записи различной и практически неограниченной длины **(от 1 до 2 млн. символов)**;

Рассмотрим пример:

законы= Приметим, что древние свободные Россияне не терпели никаких телесных наказаний: виновный платил или жизнию, или вольностию, или деньгами - и скажем о сих законах то же, что Монтескье говорит вообще о Германских: они изъявляют какое-то удивительное простосердечие / **источник=** Н.М. Карамзин. История государства российского. Т.II, Гл. III, С.42.

В приведенном примере две элементарные информации - два поля - **"законы"** и **"источник"**. Первая содержит общую характеристику законов Ярослава по Русской правде, вторая - источник, в котором содержится эта характеристика.

- Поле не обязательно должно включать только одну запись (как в большинстве коммерческих реляционных систем), а может содержать несколько логически равноправных записей. Например, две равноправные записи, содержащие две характеристики законов, как в нашем примере:

законы= Приметим, что древние свободные Россияне не терпели никаких телесных наказаний: виновный платил или жизнию, или вольностию, или деньгами - и скажем о сих законах то же, что Монтескье говорит вообще о Германских: они изъявляют какое-то удивительное простосердечие ; кратки, грубы, но достойны людей твердых и великодушных, которые боялись рабства более, нежели смерти /

- Каждая запись может иметь свой комментарий и принадлежащий ей фрагмент оригинала. Комментарий и оригинальный текст содержится как бы для "внутреннего пользования" исследователя и может включаться или исключаться из обработки при необходимости;

законы= Приметим, что древние свободные Россияне не терпели никаких телесных наказаний: виновный платил или жизнию, или вольностию, или деньгами - и скажем о сих законах то же, что Монтескье говорит вообще о Германских: они изъявляют какое-то

удивительное простосердечие ; кратки, грубы, но достойны людей твердых и великодушных, которые боялись рабства более, нежели смерти # Следующие места находятся только в новейших списках Русской Правды (и для того мы не внесли их в содержание % *А кто конь купить Княжь, Боярын или купечь или сирота, а будеть в коне червь или проесть (рана), а то пройдет к Государю, у кого будеть купил, а тому свое серебро взять опять взад*

- Система распознает мероприятия, позволяющие работать с неполными или ("fuzzy") сведениями. Соблюдая эти технические мероприятия при вводе данных можно в дальнейшем установить правила обработки нечетких данных, "объяснить" системе что есть что в процессе обработки. Более того с помощью цифровых значений можно установить степень достоверности сведений и работать с ними, зная заранее, что их достоверность 50 % или находится в интервале от 30 % до 60 %.

законы= Приметим, что древние свободные Россияне не терпели никаких телесных наказаний: виновный платил или жизнию, или вольностию, или деньгами - и скажем о сих законах то же, что Монтескье говорит вообще о Германских: они изъявляют какое-то удивительное простосердечие ; кратки, грубы, но достойны людей твердых и великодушных, которые боялись рабства более, нежели смерти # Следующие места находятся только в новейших списках Русской Правды (и для того мы не внесли их в содержание!% *А кто конь купить Княжь, Боярын или купечь или сирота, а будеть в коне червь или проесть (рана), а то пройдет к Государю, у кого будеть купил, а тому свое серебро взять опять взад*

В нашем примере комментарию Н.М. Карамзина мы присвоили степень достоверности 95 %, поставив возле этого аспекта поля восклицательный знак.

Гибкость режима обработки определена как гибким структурированием сведений источника, так и специфическими задачами, которые может поставить программному обеспечению лишь историк. Перечислим их:

- КЛИО позволяет работать с совершенно "необработанными" текстами (так называемыми "полными текстами"). В этом случае применяется стандартный набор процедур текстологического компьютерного анализа. Сильной стороной КЛИО в этой связи является следующее: анализ полного текста можно проводить одновременно с обработкой формализованных первичных данных;

Так, если мы составим тезаурус для фрагмента оригинально отрывка из Русской Правды и проведем поиск по ключевому слову "конь", то при обработке можно будет пользоваться не только собственно отрывком, но и комментарием Н.М. Карамзина и той общей характеристикой законов, которую он дает в "Истории государства Российского";

- В КЛИО существует несколько алгоритмов распознавания нечетких, "похожих" имен (названий, фамилий и т.п., имеющих различные варианты написания). Интегрированные в КЛИО алгоритмы связывания записей основаны на различных предлагаемых в литературе способах "сглаживания" расхождений;

- КЛИО может обрабатывать все возможные в принципе календарные стили, пересчитывать исторические даты из одного стиля в другой, обрабатывать системы мер, весов, валют и т.д.;

- Текстовые сведения любого поля (например, сведения о профессиях) могут преобразовываться в коды и обрабатываться в неограниченном количестве вариантов кодировок параллельно;

- Система имеет интерфейс для передачи во все основные статистические пакеты. Главное преимущество заключается в том, что система способна передавать данные в статистический пакет, сохраняя их сложную иерархическую соподчиненность и нелинейную систему связей.

Ядром концепции источнико-ориентированной обработки данных является принцип **независимости** собственно данных источника от средств его интерпретации и обработки. Раздельное существование данных и инструментария предполагает, что данные, однажды введенные в компьютер, остаются неприкосновенные. Необходимые для их обработки средства будут изменяться и модифицироваться. Это похоже на статичное сооружение, на которое не оказывает никакого влияния смена времен года, однако его образы, запечатленные на фотографиях зимой и летом существенно отличаются, подобно результатам обработки одних и тех же данных различными программными средствами. Все инструменты и средства обработки выстраиваются в так называемую **"логическую среду"** банка данных, некую экспертную систему для анализа данных как таковых.

Эта "логическая среда" может свободно меняться или расширяться исследователем, именно с ее помощью реализуется принцип гибкости анализа источника.

Распространение и развитие источнико-ориентированного подхода в России в силу субъективных обстоятельств было связано с обработкой архивных материалов 20-30-х годов XX века. Однако, первые попытки обращения к более ранним архивным материалам показали хорошие возможности системы. Перспективным оказалось использование алгоритма распознавания географических названий в различных вариантах написания при обработке материалов писцовых книг XVI-XVII вв. При том, что работа по формализации фонетических норм русского языка не имеет того

завершенного вида, как, скажем для германских языков, использование КЛИО при поиске географических названий более ранней писцовой книги в более поздней для той же местности дало хорошие результаты.

В заключении хотелось бы обратить внимание на перспективы развития системы в контексте интеграции и создания баз знаний в рамках совместных проектов.

Перечислим их:

1. КЛИО предлагает все технические возможности для создания единого тезауруса древних славянских памятников и комментариев к нему;

2. Располагая широким набором средств для обработки календарных дат, система работает также с датировками по церковному календарю. В среду КЛИО интегрирован календарь прихожан-католиков. Аналогичная работа может быть проведена для православного церковного календаря. Такая база знаний позволила бы без труда переводить в современное летоисчисление событие, которое произошло "за две недели до Пасхи 1344 г".

3. Особой задачей является разработка и уточнение алгоритма связывания записей, построенного на фонетических нормах древних славянских языков. Экспериментальная проверка и уточнение такого алгоритма, позволило бы проводить компьютерную идентификацию географических названий и имен.

The Problems of Categorisation in Content Analysis

Tikhonov Vladimir.
Russia, Moscow City Association of Archives, research fellow.

Keywords: categoria of content analysis, source-oriented data processing, interactive full text processing in *KLEIO*, graphs.

Abstract: This paper is an attempt to analyse contradictions and demerits of the categorisation in the content analysis, as well as to offer one of the possible ways of their overcoming. Though the investigation was pursued on a source of XX c., the problems, raised in the paper, and experience of application of *KLEIO* will be interesting for historians studying medieval or modern history.

Nowadays historical science is undergoing serious problems in Russia. On the one hand, there is a wreck of former historical and methodological stereotypes, on the other hand, they demand that historians should give the clear and unequivocal answer to some questions, earlier ignored by the soviet historiography. These problems are inseparable from studying of all periods of Russian history. Historians' turn to a problem of "human factor" - to life of the people, their attitude to the authorities, their perception and evaluation of the historical setting in this country - is one of the answers of historians to the increasing demands of society.

Those problems are also attractive because of wide base of sources for studies: petitions, appeals, letters and complaints to administrative bodies. The mass-scale and narrative character of these sources opens wide perspectives for application of the content analysis. Meanwhile, the examples of such researches are not very numerous, if not to say unique. A few investigations, which had been persued with the use of the content analysis, were mistrustfully perceived not only by traditional historians, but also by computing historians, who charged with in prevalence of voluntarism and violence over a source. This paper is an attempt to analyse contradictions and demerits of the categorisation in the content analysis, as well as to offer one of the possible ways of their overcoming.

1. Demerits of traditional method of the content analysis.

Normally the content analysis consists of three main stages: 1) indications of analysis elements, which are combined into categories of the analysis and then are translated into machine-readable form; 2) count of frequencies of the categories is carried out, various mathematical techniques are applied for revealing of their interrelations; 3) Interpretation of received results. In this analytical chain its first link is the most vulnerable. The problem is that choice of the elements (i.e. choice of *what* to count) depends on initial methodological views of researchers, which can not coincide. It has the following.

Firstly, the researchers can differ in choice of a research subject. According to the theory of communication, any message contains potential, as well as actual information. Actual information is one, revealed and realized by the man. It forms background for his active role in the information process. Orientation of the man on the searching of the information is determined by his needs, purposes and tasks. According to his conception of sense and springs of history, historian looks for and finds information valuable for him. Therefore different researchers can record different elements.

Secondly. Combination of these elements into categories is indeed the classification of them. However the criteria of such classification are extremely fuzzy. They are based upon expert valuations of historian, depend on a quality of his background and his knowledge about specifity of a historical period. The number of categories will depend on a permissible degree of integration of a historical data. Thus, the main defect of the traditional method of the content analysis is that categorisation is based on expert valuations of historians. It is this problem that casts doubt on objectivity of the content analysis, since one of the basic methodological principles of scientific research is infringed: the opportunity of reproduction of conditions and results of experience by different researchers.

2. Statement of the problem.

Is another approach possible? Is the formalization during definition of the categories possible? In the 70-80-s in soviet historiography the unequivocal

negative answer existed. Western researchers were not so flat. In the 60-s a number of papers appeared, the authors of which aspired to avoid any categorisation a priori. They took words as elements of analysis and carried out the categorisation with the help of dictionaries or catalogues. The Americans H.P. Iker and N.I. Harway offered automated computer system for the content analysis which main idea was the use of the factor analysis.[1] However, the results were rather modest. Critics fairly stressed difficulty of taking a context into account using the machine dictionaries, and problems with interpretation of received factors.[2] But, if the main problem of catalogues application in the content analysis consists in the words contextual use, it is possible to try to carry out the categorisation with this context in mind. Thereby the idea of this paper emerged. Shortly it consists in following.

A historian defines the elements binding certain sense to a group of connected wordforms. These word collocations are more congenial, the relations of the wordforms are more stable in them. They will make categories of the content analysis, and their number will determine the number of the categories. Revealing of stable connections of the wordforms in the word collocations allows to take into account their context: synonymy and homonymy. The basis of such an approach to the categorisation is a belief that categories reflect unique sides of the historical phenomenon. If this phenomenon is mass-scale and stable, it will find reflection in many documents through limited number of contextual phrases, which can be recorded.

The offered procedure of categorisation was tested with help of the source-oriented database "Лишенцы" ("Lishentsy"), created in KLEIO on the basis of personal cards of the Moscow inhabitants, who insisted on restoring their voting rights in the 20-30-th of XX c. *KLEIO* is the powerful tool of source-oriented data processing and has considerable facilities for realization of the planned way of the content analysis: an opportunity of contextual search of the word collocations and subsequent coding of them.

But before passing to more detailed consideration of the offered procedure, it is necessary to say some words about the source and the database.

3. The source and the database.

Personal cards of disfranchised people reflect one of the most dramatic phenomenon of a soviet history: carrying out discriminatory social policy by Soviet state towards groups of its citizens. In the 20-30 disfranchisement became an effective tool of such a policy, which main goal was elimination of "nonlabouring and exploiting elements". According to the Soviet Constitution of 1924, disfranchisement expended on those who employed hired labour and lived on nonlabour incomes such as trade, returns from property and others. Thus, many categories of the soviet citizens - traders, artisans, rich and middle peasants, priests, so-called "former people", and their dependents - became disfranchised. The status of lishentsy meant not only loss of the voting right, but, what was the most important, loss of all social privileges. Lishentsy suffered higher taxes, deprivation of pensions, free medical help, ration books. They were dismissed from work, exiled from cities. Their children were expelled from schools and universities.[3]

That is why those people did their best to restore their voting rights. For this purpose they appealed to the election commissions their complaints and other documents to prove their labour activities. That has made personal cards a valuable source for social history of Soviet Russia. On the basis of personal cards, stored in archival fund of one of the election commissions of Moscow, the database was designed. Text of the complaints was completely entered into one of the groups of the database.[4] The only deviation from the source-oriented methodology, in this case, was separating of text sentences inside the database element by semi-colon. Thus, each sentence became separate entry of the database. It gave an opportunity to capture and code them. Totally 494 complaints from 364 persons were entered.

4. The technique of offered procedure.

The first stage of the work was creation of the catalogue and formation of the wordforms lists. Initially it was planned to take wordforms, having frequency more than 5. However, in this case, number of the wordforms, and therefore the number

of analyzed interrelations, would be extremely high. Taking into account the fact, that the work had experimental character and was aimed at improving technique, we decided to analyse interrelations only between the wordforms with frequency not less, than 30. Russian language has complex structure and morphological variety. That is why sample wordforms received special formalization. So, nouns, adjectives and participles in oblique cases were combined into uniform lists. Verbs and adverbial participles were combined depending on the time form. Only those lists were taken into account, in where total sum of compiled wordforms frequencies was more than 30. Thus, 215 lists of the wordforms were formed.

KLEIO has no advanced mathematical techniques for the analysis of interrelations. However, it has facilities for the data interface to other statistical packages, to *SPSS* for example. To form the spreadsheet, the codebook with number of variables which was equal to number of the wordforms lists was created. In the codebook each sentence was coded depending on availability (2) or absence (1) of one of the 215 variables. Then the codebook was translated to *SPSS PC+ 4.0.1.* The spreadsheet had 3,395 cases and 215 variables.

As measure of association we chose the Asymmetric Uncertainty Coefficient (AUC in *SPSS*). It allows to define not only a circle of priority interrelation for each variable, but also to find out hierarchical relationships inside pairs of dependent variables. As the criterion of relation significance we took not AUC-value, but its ratio with ASE1-value, i.e. T-value. Only those pairs of variables were considered, which frequency of joint occurrence was more than 5. Only 1,674 relations had appeared as significant (at T-value > 1.0) from 92,450 analysed ones (215x215x2). Thereunto 20 variables have not displayed any significant relations. Depending on power of the criterion, the interrelations were divided into 3 groups: 1) low, at 1.0 < T-value < 1.9; 2) middle, at 2.0 < T-value < 2.9; and 3) high, at 3.0 < T-value. The word collocation of "избирательные права" indicated the highest T-value (12.9), that was quite natural.

For unravelling the tangle of the interrelations in a matrix the graph method was applied. Each graph was constructed on the basis of middle or high T-values, and then low ones were constructed to link the graphs nodes. The graphs were

designed as directed: the direction depended on ratio of the AHC-values in pairs of variables. If the value of first variable was higher than second one (i.e. first variable depends on second more, than on the contrary), the graph link is directed from the first. For example, in the pair of variables "без" and "труда" the AHC-value is equal 0.08 (when "без" is dependent variable) and 0.02 (when dependent variable is "труда"). That means, the probability of occurrence of "без", when "труда" is present, much higher, than on the contrary, and therefore first variable is subject second. If values were comparable, the graph links would be equal directed. Thus, the hierarchical relationships was established inside the graphs, and dominant nodes were defined.

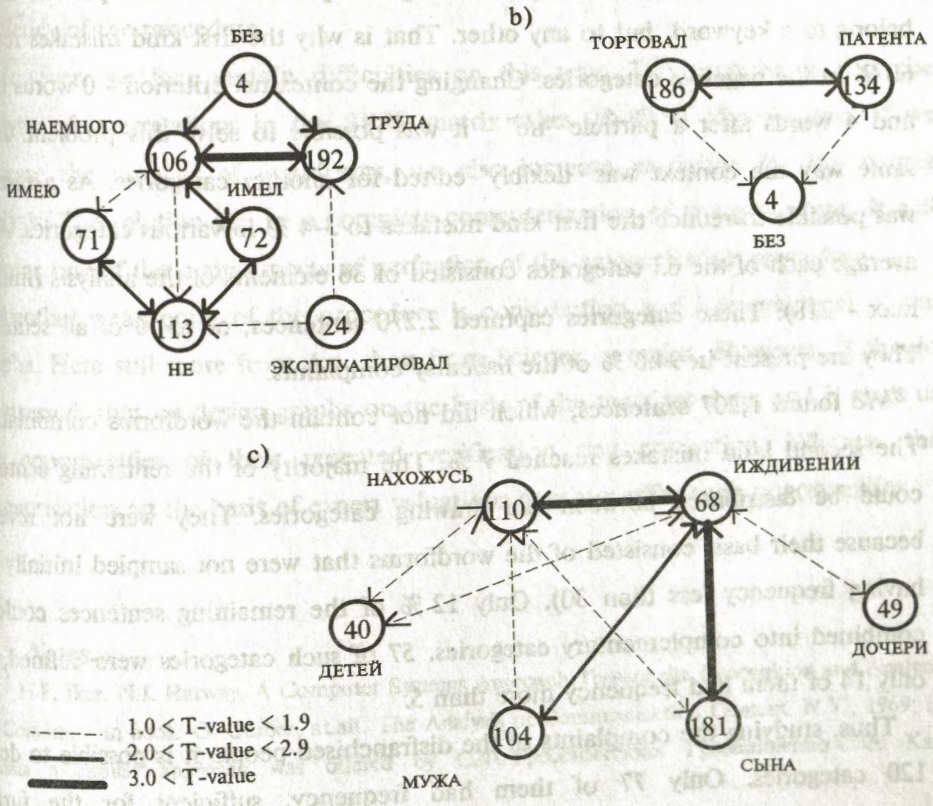
There are examples of several graphs on the Figure 1. The graph under letter "a" should be interpreted as follows: *НАЕМНОГО ТРУДА НЕ ИМЕЛ, НЕ ИМЕЮ, НЕ ЭКСПЛУАТИРОВАЛ*; *БЕЗ НАЕМНОГО ТРУДА* - i.e. statements that a person didn't use hired labour. The graph 1b means: *ТОРГОВАЛ БЕЗ ПАТЕНТА.* The graph 1c - *НАХОЖУСЬ* at *ИЖДИВЕНИИ ДЕТЕЙ* or *МУЖА* or *СЫНА* or *ДОЧЕРИ.*

106 graphs were constructed on the basis of the matrix. These graphs are the categories of the content analysis. The fact, that word collocations with frequency more than 5 were taken in account, guaranteed their occurrence within several complaints. Thus, bottom level of the element integration into categories was defined. On the other hand, there were certain number of categories that were semantically congenial. For example, 4 categories dealt with statements about unexploiting of hired labour; 6 categories - about forced nature of trade; there were 14 versions of statements about continuous labour activity. The association of these semantic versions into uniform categories meant an establishing of top level of the integration. As a result, we formed 63 categories, among them only 46 categories could be properly interpreted. Other ones, as we found out later, could assume different meanings, depending on the use of the wordforms, which were not initially sampled.

The final stage of the categorisation consisted in indication of the wordforms combinations within the graphs to query and code correspondent sentences. The

method of the wordforms combination depended on the graph type. In the closed graph (Fig.1b) the combination included all wordforms: <*ТОРГОВАЛ*> and <*БЕЗ*> and <*ПАТЕНТА*>. In the opened graphs (Fig.1c) the combination would be following: (< *НАХОЖУСЬ* > and <*ИЖДИВЕНИИ*>) and (<*ДЕТЕЙ*> or <*МУЖА*> or <*СЫНА*> or <*ДОЧЕРИ*>). We also had to take into account hierarchical relationships within the graphs, and negative or affirmative nature of the categories, i.e. availability or absence in the sentences of negative particles and words. These circumstances made this stage of work very complicated and important. The successful carrying out of following procedures of the content analysis depended on correct realization of this stage.

Figure 1. Examples of Graphs



5. Verification of the technique.

During the querying of sentences mistakes of two kinds could occur: 1) capturing sentences, contradicting set criteria of the search; 2) noncapturing sentences corresponding to criteria of the search. The analysis of the reasons for such mistakes can help to reduce them to a minimum and to define efficiency of the offered procedure.

KLEIO has a facility to query sentences, according to a context of use of two or more terms. "Extension" of the context depends on number of words in the sentence, in which those terms are used. Narrowing or extending the context, it is possible to choose the best criteria for reducing mistakes of both kinds. Initially we used the *KLEIO* standard contextual criterion: 20 words before and 20 words after the term. Such a broad context resulted in a large number mistakes of the fist kind for some categories. In overwhelming majority of cases those mistakes were connected with incorrect linkage of a negative particle "не": this particle did not belong to a keyword, but to any other. That is why the first kind mistakes reached 60 % in the negative categories. Changing the contextual criterion - 0 words before and 4 words after a particle "не" - it was possible to solve this problem. In the same way the context was "flexibly" edited for another categories. As a result, it was possible to reduce the first kind mistakes to 3-4 % in various categories. At the average each of the 63 categories consisted of 36 elements of the analysis (min - 5; max - 218). These categories captured 2,270 sentences, or 65 % of all sentences. They are present in 94.6 % of the lishentsy complaints.

We found 1,207 sentences, which did not contain the wordforms combinations. The second kind mistakes reached 7 %. The majority of the remaining sentences could be distributed between the drawing categories. They were not revealed because their basis consisted of the wordforms that were not sampled initially (i.e. having frequency less than 30). Only 12 % of the remaining sentences could be combined into complementary categories. 57 of such categories were defined, but only 14 of them had frequency more than 5.

Thus, studying the complaints of the disfranchised people it is possible to define 120 categories. Only 77 of them had frequency, sufficient for the further

procedures of the content analysis. And only 14 categories were missed during application of the offered procedure, because they had a great variety of the expression.

6. Conclusion.

In general we can say that the experiment was carried out successfully, though it was not a 100 %-success. The offered procedure of the categorisation allowed to reveal 63 of the 77 significant categories of the content analysis. KLEIO also proved to be a very effective tool for the contextual search. The mistakes of both kinds did not exceed 3-5 % in various categories. Moreover this percent can be lower. The initial restrictions are the main reason for the fact that a large number of sentences were not captured by any drawing categories: extremely high demands to the frequency of the sampling wordforms. Lowering frequency level it is possible to enhance semantic variety of morphological model, and thus to increase efficiency of the procedure.

However, we face certain difficulties on this way. The problem is that the analysis of interrelations in the SPSS matrix takes 70-80 % of worktime. If we increase the number of wordforms, we also increase worktime for the matrix analysis. The solution can be a complete computerization of this procedure. It will become one of the main aspects of perfection of the categorisation procedure.

Another weak point of the procedure is construction and interpretation of the graphs. Here still more from Art, than from Science, remains. However, it should be stressed, that we design graphs on the basis of the matrixes data, and it gives us the opportunities of their repeated verification and correction. Whereas the categorisation on the basis of expert valuations does not offer such opportunities.

Notes.

[1] H.P. Iker, N.I. Harway, A Computer Systems Approach Toward the Recognition and Analysis of Content. - in book: G. Gerber, at.all. The Analysis of Communication Content. N.Y., 1969; in Russia a similar approach was offered by С.Ф. Гребениченко: Гребениченко С.Ф. Как

повысить объективность контент-анализа? // Информационный бюллетень Ассоциации "История и компьютер". Специальный выпуск. Тезисы докладов и сообщений III конференции Ассоциации "История и компьютер". Звенигород, 31 марта - 2 апреля 1995 г. N.14, апрель 1995. М., 1995. С.114-117.

[2] See: Бородкин Л.И. Контент-анализ и проблемы изучения исторических источников // Математика в изучении средневековых повествовательных источников. М., 1986. С.24.

[3] In more detail about disfranchisement see: E. Kimerling, Civil Rights and Social Policy in Soviet Russia, 1918-1936. in: The Russian Review. Jan., 1982; Юшин И.Ф. Кодирование занятий и принципы интегральной социальной классификации общества // Круг идей: развитие исторической информации. Труды II конференции Ассоциации "История и компьютер". М., 1995. С.257-293.

[4] About eccentricities of personal cards of lishentsy and about structure of the database see: Тихонов В.И., Тяжельникова В.С., Юшин И.Ф. Об адекватном представлении структуры источника в базе данных // Компьютер и историческое знание. Барнаул, 1994; they too: Методика оценки информационного потенциала комплексов массовых источников // Круг идей: новое в исторической информатике. Труды I конференции Ассоциации "История и компьютер". М., 1994.

Источнико ориентированный подход и проблемы социальной истории

И.Ф. Юшин

Москва, Россия. Московское городское объединение архивов.

keywords: источнико-ориентированная обработка данных, базы данных, социальная история, макеты кодировок, история занятий, методы исторического исследования.

abstract: в статье исследуются проблемы использования источнико-ориентированных систем в процессе изучения социальной истории, рассматриваются вопросы построения моделей, отличий источнико-ориентированного подхода от традиционного. Моделирование источника в системе KLEIO позволяет более четко определить исследовательские задачи и изучать вопросы, связанные с формированием новой социальной структуры общества (на примере СССР в 20-е гг.)

Работа посвящена уточнению проблемы использования источнико-ориентированных систем в процессе конструирования моделей исторических явлений и процессов.

Следует разделять источнико-ориентированный подход к созданию баз данных и сами источнико-ориентированные базы данных. Только реализация источнико-ориентированного подхода на всех этапах перевода источника в машиночитаемый вид приводит к появлению источнико-ориентированных баз данных.

Источнико-ориентированные базы данных можно использовать не только для анализа текстов, но и при изучении более широкого круга проблем. Такими проблемами будут любые, для решения которых требуется анализ массовой документации, в первую очередь с невысоким уровнем формализации.

Вполне оправдано использование источнико-ориентированных систем для изучения социальной истории (при этом речь не идет о статистике). Причины:

I. Несмотря на слабую формализованность источника, сохраняемую и в базе данных, на высоком уровне обобщений могут быть решены традиционные проблемы этой области исторической науки - выделение социальных групп, страт, классов; выяснение процессов эволюции этих структур и т.п.

- Изучение таких макроисторических показателей может быть предельно тонким и осторожным. Источнико-ориентированные системы позволяют вести исследование с максимально возможной сохранностью самих данных (коды, индексы формируются исключительно в логическом окружении системы).

- Специфика кодирования данных позволяет пользователю работать одновременно с несколькими правдоподобными гипотезами, уточнение которых может требовать кардинальных перестроек системы применяемых кодов (то есть - реконструирования концептуальной модели).

II. Одновременно к уже привычным объектам изучения общества таким, как системы классов, страт, слоев, сословий (а также к факторам их выделения - доход, позиции в системе производства и распределения, демографические показатели и т.п.) добавляются и новые, без которых невозможна реконструкция всей социальной "ткани" прошлого - история социальной эволюции отдельного индивида, история семьи.

- Изучение микроисторических показателей и их последующее (возможное) структурирование в макросистемы наиболее оптимально при разработке источнико-ориентированного подхода и в источнико-ориентированных базах данных, примером которой может служить система *KLEIO*.

- Кодирование в данном случае допустимо и необходимо как средство структурирования микроданных в общие социальные модели.

III. Дополнительно к упомянутым проблемам существенным оказывается изучение эволюции занятий в определенный промежуток времени (ограниченное и географическими рамками). Это не будет традиционная история занятий как изучение occupational titles.

- Изучению подвергается общественная группа, реализующая определенное понимание благополучия и добивающаяся его, изменяя сферы приложения труда, его характер, повышая общественный престиж своих занятий, и т.п. То есть, налицо, эволюция, которую можно представить как цепь занятий последовательно пройденных индивидом. Обобщение таких индивидуальных цепей позволит судить как об эволюции этой группы, так и об эволюции семьи в социально-профессиональной сфере.[1]

- В случае, если есть основания говорить о властно формируемом понимании социального престижа, о формулируемой социально-классовой политике, то на основании имеющихся данных можно делать выводы о факторах такой политики и ее объектах.

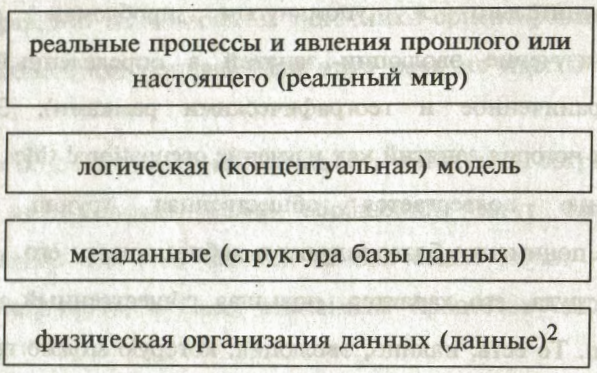
IV. Все это в комплексе позволит реконструировать целый пласт общественной жизни с максимально возможной полнотой; оптимально определить сами элементы и структурировать их в общую систему социального пространства.

V. Строительство концептуальной модели социального пространства не будет предопределять структуру метаданных, поэтому не задаст жестких границ для конструируемых гипотез и выводов в целом.

Дизайн базы данных и моделирование.

Обычно стадии решения проблемы создания базы данных можно представить следующим образом:

1. Интересная методика предложена: M. Gribaudi, A. Mogoutov. Social stratification and Complex Systems: a model for the analysis of relational data.// Kevin Schurer, Herman Diederiks (Eds.): The Use of Occupations in Historical analysis.



Построение логической модели исследователем, конечно же, наиболее интересный и творческий этап в исследовании. Именно от него во многом, если не в основном, зависит качество решения данной исторической проблемы. Второй момент, который следовало бы отметить в этой связи - ограничения на проблематику исследования, накладываемые созданием определенной логической модели. Выйти вне ее не позволяет ориентация (при создании) модели на решение только определенного круга задач. Однако очевидно, что системный подход к исследованию будет требовать в самом процессе исследования уточнений данной логической модели (изменений структуры систем или элементов данных систем). Последнее будет наталкиваться на ограничения в изменении существующей физической модели. Наконец исследователь имеет полное право в процессе работы несколько менять акценты или видоизменять саму исследовательскую проблему. Последнее потребует пересмотра логической модели со всеми вытекающими отсюда последствиями.

Практические проблемы исследования.

Наш стратегический план подхода к анализу массива массовых источников базируется на источнико-ориентированном подходе и на создании источнико-ориентированной базы данных. Поэтому первым

2. См: И.М.Гарскова. Базы и банки данных в исторических исследованиях. М., 1994. Глава 3. Проектирование баз данных; F.R. McFadden, J.A. Hoffer Database management. 1985. Chapter 2. и другие.

этапом был выбор массива документов, который бы в полной мере мог продемонстрировать преимущества данной методики.

Комплексом документов, на основе которого была создана база данных, был фонд 3109 Центрального муниципального архива г. Москвы, в котором отложились личные дела граждан, лишенных избирательных прав и ходатайствовавших об их восстановлении или жаловавшихся на неправомерное лишение.

Первым этапом реализации данного проекта было создание базы данных в системе *KLEIO*, как наиболее полно реализующей потенции источнико-ориентированного подхода. Методика создания такой базы несколько отличается от традиционных.

I. В данном случае на первом этапе была создана не концептуальная модель реальности, а **модель источника** - документа архивного фонда. Она определяется не столько представлением исследователя о реальных процессах и явлениях, сколько задается объемом и наличием информации источника. "Если мы всерьез выбираем такой подход и нацеливаемся на более адекватную поддержку исследовательского проекта, то внешние исторические знания, содержащие гипотезы и предположения о данных, должны быть организованы в форме допускающей постоянное изменение: они могут часто пополняться, а во многих случаях часто пересматриваться, если анализ данных приводит к необходимости новых представлений об отдельных их аспектах" [3]. При всем том невозможно отрицать того факта, что отбор информации и ее формализация присутствуют и в этом случае. Однако данный подход позволяет в целом минимизировать потери скрытой информации. Структура модели представлена на схеме 1.

II. Следующим этапом алгоритма работы было выявление информационного потенциала исследуемого источника, заключающееся в

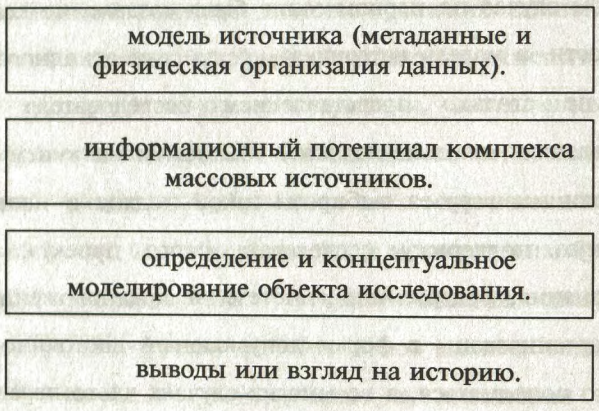
---

3. В. Леверманн. Технология разработки баз данных и источнико-ориентированная обработка данных.//Круг идей: новое в исторической информатике. Труды I конференции ассоциации "История и компьютер". М., 1994. С.19)

том, что 1) определяются общие проблем, решение которых возможно на основании информации данного источника; 2) рассчитываются объемы необходимых выборок из имеющейся генеральной совокупности для изучения той или иной проблемы [4] .

III. Постановка и решение определенной исторической проблемы (или круга проблем). На этом этапе происходит соединение информации, содержащейся в базе данных с внешним знанием. Именно в этот момент конструируется определенная концептуальная модель реальности. Источник остается нетронутым, но данные в базе начинают обрастать логическим окружением - интегрируясь со знаниями исследователя.

Таким образом, наш подход можно представить следующим образом.



Исследование.

Нас интересовали проблемы, которые можно определить как социальная сторона жизни лишенцев. Сюда мы включили следующие задачи: определение социальной принадлежности и происхождения лишенцев, эволюция их занятий, их личные судьбы и судьбы их семей, методы реализации государственной социальной политики. На этом

---

4. В.И. Тихонов, В.С. Тяжельникова, И.Ф. Юшин. Методика оценки информационного потенциала комплексов массовых источников.//Круг идей: новое в исторической информатике. Труды I конференции ассоциации "История и компьютер". М., 1994. С.24-38)

основании предпринимается попытка реконструкция социального пространства, в котором обитали лишенцы.

Конструирование концептуальной модели состоит в определении сущностей, ее составляющих и построением структуры такой модели. В свою очередь каждая исследовательская задача требует построения своей "микромодели", определяемой особым взглядом на данные. Разными будут и объекты этих микромоделей: для изучения формирования дела, бюрократической процедуры - "дело"; для исследования социального портрета граждан лишенных избирательных прав - "персоналия"; для исследования социальной политики власти и ее эволюции - каждый случай лишения прав.

Этот процесс связан с отбором показателей, которые в нашем случае сведения рассредоточены по всей базе, поскольку нас интересуют и профессиональные занятия, и возраст, и семейное положение, и уровень доходов, и происхождение. Извлечение и обобщение данной информации невозможно без применения кодирования.

Кодирование.

Применение кодирования в источнико-ориентированной базе данных необходимо при изучении социального пространства по многим причинам. Однако, в нашем случае кодирование не является жестким, моделирование социальных процессов может быть гибким на концептуальном уровне, не затрагивающем целостность данных. Контекст источника при такого рода гибком кодировании может быть принят во внимание, "черное" может стать "белым". Так, занятие "старьевщик-торговец" может быть не отнесено к типу "мелких торговцев", а попасть в тип "черновой труд". Модели процессов и явлений, зависящие от макетов кодировок могут рассматриваться одновременно в нескольких системах кодов.

В процессе работы были составлены книги кодов по занятиям, собственности, семейному положению, лишению избирательных прав. Все книги кодов составляются на основании сведений "элементарных

информаций" - эквиваленту поля. Мы использовали несколько вариантов назначения кодов для элементарных информаций "занятие", "социальное положение" и "социальное происхождение" (схема 2).

При составлении макетов мы руководствовались не только сведениями соответствующих элементарных информаций, но была проведена более подробная разработка каждого занятия в соответствии с контекстом данных. Кроме того, мы выделили 4 временных отрезка, чтобы рассмотреть подробнее структуру занятий лишенцев.

Некоторые результаты.

- *Социальные предпочтения.* Рассматривая возрастные категории лишенцев, мы сможем проанализировать предпочтения или необходимость заниматься тем или иным видом деятельности.[5]

	1	2 (в %)	3 (в %)	4 (в %)	5 (в %)	6 (в %)	7 (в %)	8 (в %)	9 (в %)	10 (в %)
I	7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
II	58	12.1	0.0	10.3	3.4	0.0	10.3	8.6	13.8	8.6
III	89	18.0	13.5	14.6	9.0	2.2	31.5	2.2	18.0	13.5
IV	35	17.1	8.6	8.6	5.7	0.0	20.0	17.1	8.6	5.7
V	137	10.2	7.3	20.4	3.6	0.7	12.4	1.5	8.0	8.8
VI	23	4.3	8.7	0.0	0.0	0.0	21.7	34.8	13.0	0.0
VII	63	19.0	3.2	6.3	4.8	0.0	7.9	25.4	0.0	1.6
VIII	34	5.9	2.9	5.9	0.0	5.9	5.9	23.5	0.0	2.9
IX	82	6.1	4.9	7.3	0.0	0.0	4.9	11.0	1.2	3.7

5. В таблице - *по горизонтали*: 1. всего принадлежали к этой половозрастной группе, 2. мелкие торговцы, 3. крупные торговцы, 4. кустари и ремесленники, 5. предприниматели, 6. рантье, 7. служащие, 8. иждивенцы, 9. рабочие, 10. кооперированные кустари и ремесленники; *по вертикали*: I. неизвестны ни пол, ни возраст, II. мужчины, г.р. после 1898, III. мужчины, г.р. 1873-1897, IV. мужчины, г.р. до 1873, V. мужчины, г.р. не указан, VI. женщины, г.р. после 1898, VII. женщины, г.р. 1873-1897, VIII. женщины г.р. до 1873, IX. женщины, г.р. не указан.

Таким образом, относительно молодые люди обоего пола занимали рабочие должности и должности служащих (особенно женщины). Среди лиц среднего возраста - мужчин были представлены практически все социальные категории, у женщин преобладали занятия мелкой торговлей, работа служащими, а значительная часть перешла уже по возрасту в категорию иждивенцев. Очевидно этому способствовала и социальная политика государства.

- Социальные перспективы и результаты социальной политики.

НЭП. Большинство нэпманов - торговцев и предпринимателей были представителями иных социальных групп до революции. Ряды "новой буржуазии" формировались из бывших служащих, кустарей и ремесленников, крестьян. Многих толкало к занятиям торговлей отсутствие возможности существовать на средства от традиционных занятий.

После НЭПа. Социальные перспективы более четко прослеживаются, если мы рассмотрим движение в социальном пространстве выделенных категорий населения, сравнив их социальный статус в период НЭПа и на момент последнего рассмотрения личного дела. Стоит обратить наибольшее внимание на тех, кто значительно снизил свой социальный статус.

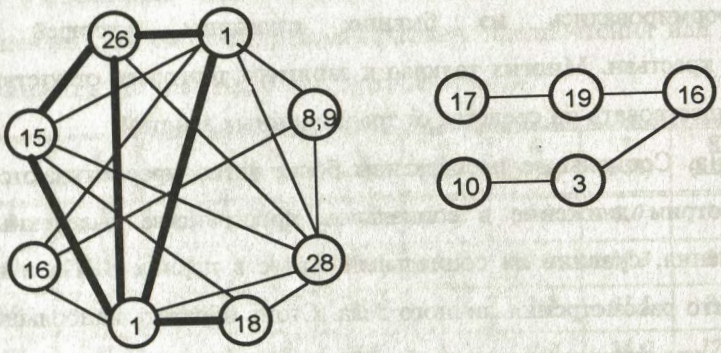
Торгово-предпринимательские слои активно размывались в это время. Из 63 мелких торговцев периода НЭПа стали рабочими разного рода - 10, иждивенцами родственников - 11, служащими - 9. Понизили свой статус - 30 человек. Из 34 крупных торговцев - 11, из независимых кустарей и ремесленников - 12 (из 62), из 20 предпринимателей - 9. Из 528 статистических случаев рассмотрения дел в 109 был зафиксирован статус иждивенца на момент последнего свидания лишенца с избирательными комиссиями.

- Структура занятий.

Корреляционные модели занятий лишенцев и их родственников, построенные по выделенным периодам дадут ответ на вопрос о тенденциях социальной эволюции этой группы. Источнико-ориентированные системы

позволяют фиксировать множественность занятий индивида, учитывать их одновременность. Поэтому такие модели представляют определенный интерес. На схеме 3 представлена структура занятий детей лишенцев (легенда здесь и далее - на схеме 2) - видна взаимосвязь между занятиями торговлей и предпринимательством и снижением социального статуса или даже уголовным или административным преследованием. Перед нами крах надежд и отсутствие традиционных перспектив.

Схема 3. Корреляционный граф занятий детей граждан, лишенных избирательных прав.



коэффицент корреляции > 0.50

коэффициент корреляции от 0.30 до 0.50

- *Эволюция занятий индивидов и интегральная эволюция занятий лишенцев.*

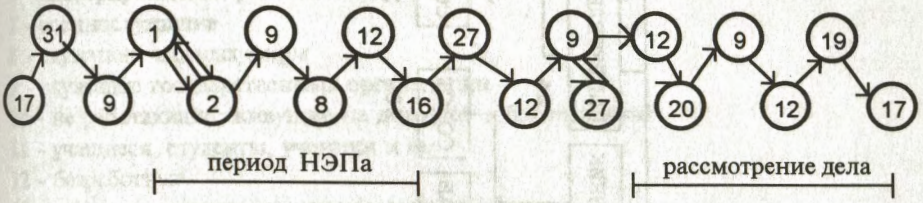
Судьба отдельного индивида может быть прослежена через характер его занятий, изменение уровня его дохода в контексте политического давления (внесения в список лишенных избирательных прав).

Но для обобщения необходимо изучение "транзитивности" его занятий. Источнико-ориентированные системы позволяют провести генерализацию переходов от одной группы занятий к другой и тем самым включить

микроисторические события в канву макроистории. Таким образом будут построены мосты от поведения индивида к эволюции социальной системы.

На схеме 4 - характерный пример движения по социальной лестнице гр. Мелешко (Милишко) Федора Федоровича

Схема 4.



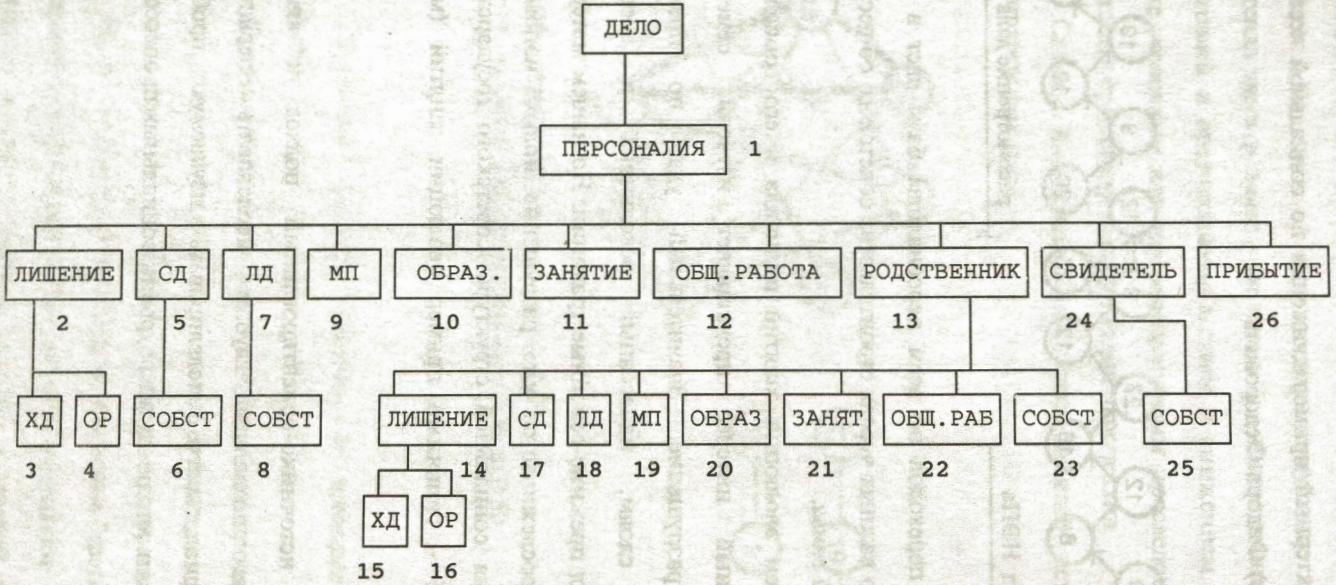
Изучение парных переходов по всем персоналиям базы даст в свою очередь представление о важных чертах строительства советского общества.

- *Эволюция занятий семьи.*

Наряду с изучением эволюции занятий индивида и его социальных перспектив значительный интерес представляет изучение семейных традиций или их разрушение. Значительный удар по прежним предпринимательским слоям, нанесенный революцией и военным коммунизмом, разрушил прежние семейные традиции. На нашем материале в большей степени прослеживается уже разрушение новых "нэповских" традиций и унификация социальной структуры советского государства к 1930-м гг. На схеме 5 - типичный пример эволюции занятий (между поколениями).

Таким образом, источнико-ориентированный подход к анализу социальных процессов позволяет глубже исследовать первичный документальный материал, четче моделировать изучаемую проблему (проблемы), интегрировать микро- и макроуровни социальных процессов.

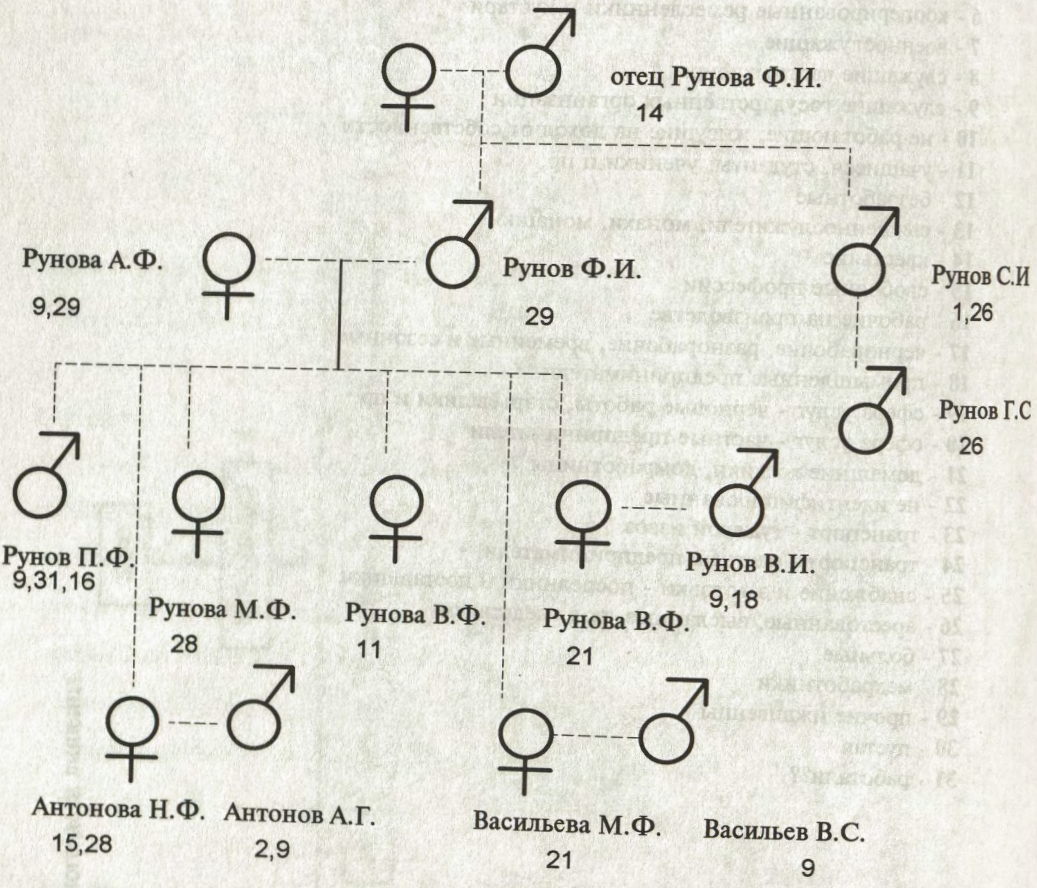
Схема 1. Модель личного дела лишенца.



ХД - ход дела
СД - сопроводительный документ
ЛД - личный документ
ОР - окончательное решение
МП - материальное положение
ОБРАЗ. - образование
ЗАНЯТ. - занятие
ОБЩ.РАБ. - общественная работа
СОБСТ. - собственность

Схема 2. Вариант макета кодировки по элементарной информации "занятие: должность"

1 - торговля мелкая
2 - торговля крупная и оптовая
3 - торговое посредничество
4/5 - кустари и ремесленники
6 - кооперированные ремесленники и кустари
7 - военнослужащие
8 - служащие частных фирм
9 - служащие государственных организации
10 - не работающие, живущие на доход от собственности
11 - учащиеся, студенты, ученики и пр.
12 - безработные
13 - священнослужители, монахи, монашки
14 - крестьяне
15 - свободные профессии
16 - рабочие на производстве
17 - чернорабочие, разнорабочие, временные и сезонные
18 - промышленные предприниматели
19 - сфера услуг - черновые работы, старьевщики и пр.
20 - сфера услуг - частные предприниматели
21 - домашние хозяйки, домработницы
22 - не идентифицированные
23 - транспорт - гужевой извоз
24 - транспорт - частные предприниматели
25 - снабжение и заготовки - посредники и поставщики
26 - арестованные, высланные, под следствием
27 - больные
28 - медработники
29 - прочие иждивенцы
30 - пустая
31 - работали??


отец Рунова Ф.И.
14
Рунова А.Ф.
9,29
Рунов Ф.И.
29
Рунов С.И
1,26
Рунов Г.С
26
Рунов П.Ф.
9,31,16
Рунова М.Ф.
28
Рунова В.Ф.
11
Рунова В.Ф.
21
Рунов В.И.
9,18
Антонова Н.Ф.
15,28
Антонов А.Г.
2,9
Васильева М.Ф.
21
Васильев В.С.
9

Computer Investigation of Medieval Prognostic Books by Means of *Kleio*

Adelina Angusheva, Ph.D., University of Sofia "St. Kliment of Ochrid",
Central European University, Medieval Studies Department, Budapest

Keywords: Prognostic Books, *Kalendologion*, Kleio computer program

Abstract: The paper presents computer-based analysis of Prognostic Books in Byzantine, Latin and medieval Slavic literacy with the help of the computer program *Kleio*. The program provides classification of the texts and analysis of their typology. The whole texts of the Prognostic books are computerised. The result of investigation is a catalogue of the different topics in the Prognostic books. The common places in them and frequency of the different kinds of prognoses are represented with the help of Kleio computer program. The investigation provides the possibility to go further in the research of the medieval vision of the world and the language of its expression.

The paper addresses the question of computer studying of medieval Prognostic books. They are a heterogeneous group of texts for predicting the future by the help of different natural (physical) and psycho-somatic omens. Adopted and elaborated in Ancient Greek, Ethrurian and Roman world, Prognostic books penetrated the literary production of the Middle Ages. The texts were not medieval by origin , but they became "medieval" in the process of their spread and copying during the Middle Ages. During this period Prognostic books were spread in different languages (Greek, Latin, Armenian, Old Slavic) and in various cultural traditions. The Prognostic books represented the coexistence of pagan and popular Christian notions. The predictions in them were based on the belief that there is a direct connection between the phenomena of the physical world and the future of the human beings, states and crops. Speaking the language of magic, the Prognostic books were at the same time very close to the reality and every-day life. Prognostic Books in the Middle Ages shows a great variety with

respect to the omen, used for future-telling (dream images, thunder, cycles of the moon, Christmas day, etc.) and there were a different types of Prognostic Books based on one and the same omen (cf. several different types of Thunder books).

At the same time there were very similar prognostic texts in Byzantine, Latin and Old Slavic tradition. Juxtaposition between these three tradition provide a broad picture of translations, modifications and spread of the Prognostic Books in European Middle Ages. However the correct and complex picture of this material have to include great amount of texts in three different languages. This is of course a task which can be successfully solved only by help of computer program providing automatic analysis of a huge datebase. Therefore I tried to encode a number of Prognostic texts by means of *Kleio* Computer Program. The special version of Kleio for examination of Prognostic books was created at Central European University in 1994 under the supervision of Prof. Gerhard Jaritz and with the help of my colleague Damir Karbić (Cratia). I would like to express my gratefulness to both of them for their help.

The application of *Kleio* for medieval Prognostic Books combines the schematic representation of structure and content of the different text with the possibility to compare different topics, elements and specific expressions given in the texts. Therefore the content of the Prognostic books is presented in English and every peculiar expression is in its original form as it is in the manuscript. For Greek and for Old Slavic words I use a standard transliteration.

The structure of the computational datebase contains identical items of Prognostic books. One item represents the whole information of one separate copy of Prognostic book and paratexual information concerning the characteristics of the copy. Therefore the first level of every item defines the type of Prognostic book (*Kalendologia*, Thunderbook etc.) It contains 1) Name of the Prognostic Book, 2) Subcategory used only in case when in one and the same manuscript there are two very similar Prognostic Books (like in *Dragolov Sbornik*, where one can find two similar types of *Kalendologia*). The next level of the structure represents the information about the manuscript in which the given copy is preserved: 1) Date of the manuscript, 2) Language of the copy, 3)

Place in which the manuscript was prepared, 4) Library in which it is kept now, together with last callnumber.

The structure of the item follows the structure of the Prognostic Books. The organisation of Medieval Prognostic texts such as *Kalendologia* and Thunderbooks consists of two indications (Christmas, January 1st, falling of a thunder, days of the week, zodiac, cycles of the moon, months) and expected natural phenomena or events that will appeared as a "result" of omen/indication. In fact the structure[1] of every Prognostic book presents the chain of indications and picture of future[2].

Thus the process of future-telling is constructed by the coexistence of different indications (Thunder and month; Christmas and day of the week[3]). The main principle of text structure is the correlation between changeable and constant indications. The constant one is ganre-generative (falling the thunder, Christmas, January, 1st), the

---

The structure of the texts is as follow:

indication 1. 0.1.*Kal. januarii*

indication 1.1.0 *si die domenico venerit*

prediction 1.1.0 *verem humidum*

prediction 1.2.0 *autunnum ventosum*

prediction 1.3.0...*annonas bonas*

indication 1.0.1.*Kal. Januarii*

indication 1.2.0.*Dies Lune facit*

prediction 2.1.0. *Hiemem aestatem communem temperatam*

prediction 2.2.0.*diluvia magna erunt*

prediction 2.3.0.*bella milltium* (The example is from **Palat. Lat. 1449,** 9th century, ed. G.Mercati, Note di Letteratura Biblica e Christiana Antica,*Studi e testi,* Roma,1901,5,p.77)

[2]D. Petkanova whote about the structure of the Prognostical books:"The main group of the predictions begin with conditional sentence, followed by the sentence in the future or present tense.Thus, every pretelling in the different Prognostic books contains two important elements: condition and event in the future" - Gadatelnite knigi i folklorăt [Prognostic Books and Folklor], *Starobălgarska literatura,* 1989, v. 18, p.52. Investigating the Dreambooks St. Oberhelman extracted the similar components (cf St. Oberhelman, The interpretation of Dream-Symbols in Byzantine Onerocritic Literature, *Byzantinoslavica,*1,1986)

[3]Not by chance the definition of the Prognostic books usually consist of the indications: Cf. The *Kalendologion* is " a prognostication for the coming year by the day of the week when the Christmas is" (M. K Andreeva, K istorii vizantijsko-slavjanskix gadatelnix knig [On the History of Byzantine-Slavic Prognostic books], *Byzantinoslavica,*1930,p.130) or a prediction "by the day of the week on which Christmas or New Year's day fell" (I.Taavitsainen, Middle English Lunaries.A Study of Genre.:*Memoires de la Société Neophilologique de Helsinki.*Helsinki,1988,t.47,p.35)

changeable one presents the time differentiation which provides the extension of prognostic information.

Taking into consideration this, I defined the next level of the computational structure of the Prognostic text as a directly subordinated to the type of the Book (i.e. name). This level represents the constant omen, by which the prediction was made. For instance: The group of *Kalendòlogia* predicts the future either by the Christmas or by January 1st. Therefore this level is followed by the second one with changeable indication, given in the computer version as "Specification". Here particular names of the days of the week, of different months, moon phases are encoded. The options are much more broader in order to cover all types of Prognostic books. They include 1) Time (month or day) 2) Zodiac 3) Space (used especially for Thunderbook according to the fourth directions of the world. However, if we enrich our computer catalogue of Prognostic books with prediction by the spontaneous moving of the human body - *Trepetnik*, we can define the different part of the body as a space). 4)Phase (of moon) and 5) Others.

The next part of the structure represents the characteristics of the prediction itself. It consists of two main modules: topics, about which the prediction is made and future condition defined as action. This binary structure corresponds to the microelements of prediction: *autumnum ventosum* ѥсень вѣтрена. The preparation of this part is based on observation of the topics about which predictions in the Prognostic books were made. It is an abstract model and in the same time the topic index of the texts. It summarises the second part of Prognostic Books which usually contains prediction about 1) seasons, 2) natural phenomena -rain, wind etc., 3) lack or presence of some goods, 4) animals, 5) illness and/ or misfortune, 6) human's fate - with gender/age differentiation 7) social groups - rulers, soldiers, 8) social destruction -war, revolt. These different types of topics of which the Prognostic books predict the future are organised in four groups subordinated directly to the changeable indication. First of them (coded as L) summarises the information concerning human beings. Their gender, age, social position, activity are specified, and the prediction about them is given as an action.

Encoding of the text: царемь радованїе is as a following: gender =m/ age- 0/ social position = kings/ activity -0 / action = joy

The second category is Climate (W) with specification about season, month, phenomenon and again action. It has a special subcategory harvest presented in the Prognostic books more as a period of time than as a connected with crop. The third group defined as R = Goods represents prediction about plants, food and animals in the Prognostic Books.

The last group (Others) is created for rare predictions which are not semantically connected with previous three groups. It contains predictions for damages, illnesses, plagues, etc.

Encoding the information of thirty Greek, Latin and Old Slavic *Kalendologia* I prepared a computer catalogue of their texts and information given in them. The second step includes mainly the computer investigation of this information. The several catalogues can be created by means of Kleio Computer program such as: A catalogue of the texts according to the dating of manuscripts, to the language, to first or second indication. This database can provides the research on the typology of the texts because the computer analysis shows the common places between the text not taking into consideration whether the original is in Greek, Latin and Old Slavic. Combining the information of these two catalogues we can come to the conclusion that January 1st cannot be regarded as a relevant indication for textological classification of the texts. At the same time by the help of the original text encoded in the computer catalogue we can have the exact picture of the manuscripts which for instance use *vesna* or *proljatie* for spring.

The use of the computer program provides the creation of different text catalogues which show the defining of particular phenomena in the Prognostic books. In this way for instance we learn from the computer investigation that there are 16 characteristics for future spring in the Prognostic Books. Some of them: dry, good, rainy, serene, stormy, wet , windy have a high number of frequency, some other like: with floods, heavy are only *hapaks*. Therefore, by the help of computer program Kleio we can figure a

dictionary of frequent determination of various topics. These nominations show the medieval image of everyday life - the common picture of spring, illnesses, fate of human beings. They present the picture of the world of the medieval man and the way he named the phenomena surrounding him.

The Structure of the Item (Prognostic book)

```
database name = prognos; first=b; overwrite=yes
    part name=b;write="Prognostical book";
        part=m,o;
        position=rb,name,subcategory
    part name=m;write="Manuscript";
        position=date,language,place,library
    part name=o;write="Omen";
        part=s;
        position=content
    part name=s;write="Specification";
        part=p;
        position=time,zodiac,space,phase,other
    part name=p;write="Prediction";
        part=l,w,r,d;
        position=type
    part name=l;write="Men";
        position=gender,age,spos*,activity,action
    part name=w;write="Climate";
        part=h;
        position=season,month,action,phaenomenon
    part name=r;write="Goods";
        position=plants,food,animals,action
```

```
part name=d;write="Other";
    position=text,action
part name=h;write="Harvest";
    position=action

exit name=prognos
```

---

* social position = spos

Loanwords in the New York Missal and the *KLEIO* Computer Program

Margaret Dimitrova

Central European University, Dept. of Medieval Studies, Budapest 1051

South-Western University "Neofit Rilski", Blagoevgrad, ul. I. Mihajlov 66

Keywords: Croatian Glagolitic tradition, loanwords, index, orthography

Abstract: The paper describes some of the results of the work with the *KLEIO* computer program in the examination of Greek and Latin loanwords in one Croatian Glagolitic missal, that is the New York Missal from the fifteenth century. The program can generate an index of loanwords and several lists of lexemes according to chosen parameters concerning their orthographic, morphological and lexical peculiarities.

1. The present paper represents my experience working with the *KLEIO* computer program[1] in the examination of loanwords in the New York missal (henceforth NYM) - a Croatian Glagolitic manuscript dating from the second quarter of the fifteenth century. The manuscript has both facsimile and critical editions (Birnbaum - Rehder 1977, Schmidt-Deeg 1994). Its palaeographic and phonetic peculiarities have been studied by A. Corin (1991), who has identified 11 hands working in rotation.

My first step was to create a large-scale database including loanwords with their orthographic, morphologic and lexical peculiarities. The matrix for the database was executed with the help of my Croatian colleague D. Karbić. It is called "**nym. mod**".[2] After several modifications, it comprised the following information:

---

[1] I was taught to work with the program *KLEIO* at Prof. Jaritz's computer classes at Central European University, Budapest College, Medieval Studies department. A manual for the program is available, see M. Woollard, P. Denley, *Source-Oriented Data Processing for Historians: a Tutorial for Κλειω*. St. Katharinen, 1993.

[2] Its first variant was the following:
Computer programs Kleio Version 5.1.1 --- 28.07.1994, 15:31:24
Historical Workstation Project / HP-UX version at IUE

```
database name=nym;first=w;overwrite=yes
part name=w;write="word";
  part=s,g,p;
  position=number,line,word,genre,chapter,verse,context
part name=s; write "source"
  position=genetic, historical, way
part name=g;write="grammar";
  position=case,number,gender,declension,greek,latin
part name=p;write=orthography
  position=peculiarity
exit name=nym
```

2. After I had stored all the information concerning every word identified as a lexical borrowing, I had the file called "**nym.dat**", containing the following data:

2. 1. The part ***w="word"*** includes 7 positions:

- the *number of every form* included in the file (position 1.);

- the *number of the folio* and of the *line* where the form of a given loanword is written - in that respect I followed the numbering introduced in the critical edition of NYM (position 3.);[3]

- the *form of the word* itself as it is attested in NYM (position 4.). Loanwords are written in Latin translitteration as it has become customary in Croatian studies and editions of Croatian Clagolitic manuscripts. To the extent it was possible, I followed the translitteration used by the editor of NYM. However, I had many difficulties which the specific signs wich are not available in the font of the program I used. The program allows for introduction of new signs and now I work in this direction. I did not resolve the abbreviations under tilde. In that I also followed the edition. No ligatures, and

---

```
database name=nym;first=w;overwrite=yes
part name=w;write="word";
  part=g,p;
  position=number,place,word,typt,caput,line,context
part name=g;write="grammar";
  position=casus,numerus,genus,stem,greek,latin
part name=p;write=palaeography;
  position=peculiarity
exit name=nym
```

[3] As the so-called *topography* of the text is important for graphic system, orthography, textology, and stylistics, the location of every word in the manuscript will be described more precisely in future according to the facsimile edition (it means - position of the word in the line - in the beginning, or in the end), marginal signs, superscript and subscript signs, environment of letters and every kind of other signs near the loanword under investigation, ligatures, preceding and following letters, capital letters, colour of letters, etc.)

initials are indicated in my database, as it is in the edition. I included all the forms of the words in all their locations in the text, even several forms which were frequently repeated. Therefore, the database allows for a quantitative analysis;

- the *genre,* the type of the text, in which a given word is used (position 5.), in particular the following general differentiation is accepted: biblical texts - **b** including precise *chapter* and *verse* of the biblical books (position 6 in the system), non-biblical (including *oratio, communio, postcommunio,* etc.) - **o**, rubrics - **r**, titles - **t;**

- *context* (sentence) in which the word is included (position 7.).

2. 2. Part ***s="source"*** represents the way and sources of the process of borrowing. As far as it was possible, both genetic and historical sources (if they differ from each other) are noted. In such a way, it is indicated whether words are arameic concerning their genetic source, but used in their Greek or Latin garb, or if they are Greek, but used in Latin garb, or if they have forms typical of Dalmatian or of Venitian Italian, or if they show peculiarities of Germanic intercession, etc. The position "*way*" indicates whether a given loanword can be considered to have been widespread in vernacular, and as such shows evident corruption of the model from giving language, or it is a typical literary lexeme and follows norms of foreign orthography - Greek or Latin. In fact, only words showing clear phonetic evidence for adaptation in receptor-language are indicated as vernacular in my database. However, I am aware of the fact that there are many contradictions among specialists about the origins and the the ways of penetration of several loanwords in medieval Slavic written tradition, as well as in modern Southern Slavic dialects. As medieval Croatian manuscripts contain various textological strata, I accepted the following devision:

a) **vern** - for words used in spoken Southern Slavic languages and in medieval Slavic manuscripts originating from the region, such as *apostol, oltar, cum*, etc.;

b) **vcr** - for loanwords or forms of loanwords typical predominantly of Croatian region, such as *ocat, tamian, trpeza,* many of them have been borrowed from Romance languages: Latin, Dalmatian, Italian, such as *biscup, opat, korizma* and at a low degree from Greek, and Germanic languages;

c) **grsoc** - for lexemes typical only of the speech of bilingual communities or bilingual literate people knowing Greek; thus, in this group mainly the Greek loanwords introduced in the Cyrillomethodian translations and preserved in Croatian liturgical books are organised, such as *kesar,* or words with *gerv* following consonant and preceding ι, η, such as *paraskevji, levjit,* etc.[4]

---

[4] Concerning Cyrillomethodian translations a distinction between literary loanwords and loanwords used in the spoken language has often been made by Slavists. To a certain extent this distinction resembles the distinction of contact linguistics between loanwords (Lehnwörter) and foreign words (Fremdwörter), заимственная vs. используемая лексика 967). This differentiation was based predominantly on phonetic evidence. Words which show peculiarities either of spoken Greek or of Slavic adaptation of Greek forms are considered to have been orally borrowed and to have been used by the Slavic population living around Salonika (e. g. **лєвѣнтъ** is considered to be derived from the spoken Byzantine form λευγιτης vs. the written form λευιτης, as well as **сѫвота** <σαμβατον vs. σαββατον). However, for many loanwords we do not have such phonetic evidence. On the other hand, forms of Greek or Latin loanwords identified in medieval Slavic manuscripts are rather an indication of the peculiarities of the idiolects of certain Slavic *literati* than of the language of a whole speech community. The fact that some forms of loanwords show peculiarities of spoken Byzantine Greek (vs. written forms) is not sufficient to prove that these lexemes were widespread in vernacular. Thus, judging from the meaning of **лєвѣнтъ**, as well as from its use and its stylistic peculiarities in Greek texts, I would say that this word was not widespread in the ninth century in the Thessalonian Old Bulgarian dialect in spite of its form attested in the manuscripts dating from the tenth and eleventh centuries. The Greek loanwords used in the earliest translations, as well as in later translations and textual redactions, can be divided not in two, but in four groups: (1) words borrowed in oral speech during the earliest Slavic contacts with Greek civilization, such as **трѧпєзѧ, корѧбль, комора**, etc. In this group one may include also the loanwords penetrating Slavic dialects from the so-called Balkan Latin, such as **комъкѧтн, оцьтъ**, etc. The loanwords belonging to this group were widespread in the speech both of bilingual and monolingual Slavs. Many of them were used in different medieval Slavic languages and dialects. It is likely that in the ninth century these loanwords were more or less indistinguishable from native lexemes; (2) words borrowed in the course of the conversion of Slavic peoples. Greek loanwords such as **ѧггєлъ, мирο, дияконъ, ѧнатємѧ, ѧмнѧ**, as well as the biblical proper names, such as **ѧдамъ, иродъ, пѧвьлъ, ѵвраамъ** and the names of patron saints of great liturgical feasts, were widespread in the spoken language used by the Thessalonian Slavs in the ninth century. However, it is probable that bilinguals (even with a low degree of mastery of Greek) were aware of the Greek origin of these words. Nevertheless, the lexemes belonging to this group were adapted and assimilated according to the peculiarities of the receptor-language. Later they became widespread in the speech of monolinguals who were mastering other South and East Slavic dialects and languages; (3) words used in spoken, as well in written language by certain social groups, that is people with different degrees of education: nobles, priests, scribes, translators, writers. They could be both bilingual and monolingual. This third group includes lexemes for ancient *realia* and places, *termini technici* of church organisation and liturgical year, names of saints and biblical personages, such as **ѧрхиєрєи, игємонъ, хлѧмида, мѧмѡна, идолиє,**

d) **latsoc** - for Latin or Italian loanwords used predominantly by Croatian priests and literate people, such as *kurpural, nota, ampula.*

In conclusion, I attempted to include a larger body of lexemes in my study, i. e. even some with controversial and uncertain etymology and way of penetration.

2. 3. The third part ***g="grammer"*** includes description of grammatical characteristics of loanwords, in particular case, number, gender, declensional class. If a word is an adjective, or is indeclinable, it is noted in its description in the position "declension". In this sector there are positions for Greek and Latin etyma, and in the case that they are different - Greek and Latin counterparts in Greek and Latin originals.

2. 4. The last part ***p="orthography"*** includes every kind of peculiarity of phonetic and morphological adaptation, for instance the rendering of Greek *ai, oi,* η, ει, ω, υ, α, β, χ, φ, ια, ιο, ιε, etc., double consonants or vowels, as well as cases of elision, apocope, contraction, contamination, epenthesis, and scribal errors such as dytography, haplology, in addition - Greek and Latin endings and their Slavic correspondence.

3. After the database containing 2500 lexical items was ready, the next step was to ask the computer "questions" concerning the data contained in "nym. dat".

3. 1. The first "task" I gave the program was to make an index of loanwords attested in NYM, that is:

```
query name=nym;part=w:word
index part=:  word; signs=15
    part=g:case;signs=3;
    part=:number;signs=5;
    part=:gender;signs=3;
    part=:declension;signs=7;
    part=:back[1]:line;signs=10;
    part=:genre;signs=3;
```

НѦРѦ҂Ъ, ОРѦРЬ, etc. The content of the group varies considerably according to the level of education of Slavic literate people, according to the region and fashion in certain social circles of the medieval Slavic world. The level of adaptation and accommodation was also different. Most of these words, even assimilated, were distinguishable as Greek words by most Slavic speakers, when using or listening to them; (4) words transmitted from Greek originals in the process of translation but used neither in spoken Byzantine Greek nor in spoken Old Bulgarian (mostly Aramaic ones and rare proper names, such as ЕИСОЙ҂Ъ, ѦСИР҂Ъ, ЌѦТѦИ҂Ъ, etc.).

part=:chapter;signs=8;
part=:verse;signs=7
stop

The first column of the index contains Greek and Latin loanwords in Latin transliteration. At this stage of my work, the forms are not arranged under head words in their "dictionary form". Now I work on arranging the lexemes under head words defined preliminarily according to the forms identified in NYM itself where it is possible, or resting on extra information excerpted from the card-indeces of Old Chirch Slavonic Institute in Zagreb - in cases when the information coming from the manuscript is ambiguous.[5] The second, third, fourth and fifth columns give information about grammatical characteristics of loanwords, their case, number, gender, declensional class. The sixth column in the index contains the number of the folio and of the line, where a given loanword is found. The seventh column shows the type of the text (genre) in which loanwords are used. The last column contains the number of the lexical items in the database "nym.dat" in the computer program I use.

3. 2. Other different questions can be asked. For instance, the program can make a list of loanwords used only in biblical texts and not in other kinds of texts in the missal, or a list of all the words used only in rubrics, or a list of words used both in biblical texts and rubrics and not in non-biblical parts. It turned out that only several loanwords are common for **b, o, r.** These are vernacularized lexemes such as *apostol, dekon, evrei, oltar, pogani, prozbiter, psalm, paska*. Of the same kind are the words used both in **b** and **r** and not in **o**, that are *dupal, epistolie, grčki, titol.* Only vernacular loanwords are common for **r** and **o**: *biscup, opat, papa*. More loanwords are common for **b** and **o**. They are of different kind: vernacular words, such as *matorstvo, deval, tamen*, or lexemes used in Cyrillomethodian translation, such as *olokavtomat, arhierei, levît, paraklit, herovim*, etc., and some new words borrowed in the period of Croatian

[5] I am grateful to the experts of Old Church Slavonic Institute in Zagreb, in particular Prof. A. Nazor, Dr. I. Mulc, Dr. M. Mihaljević, for their help and consultations.

revision of the old translations, such as *palma, lilium, kruna*. I shall not report all the results of that kind received with the help of the program, but I shall discuss some other types of results.

3. 3. If we take one Greek or Latin word (etymon), the program can show all the loanwords in the Glagolitic texts corresponding to it in all their forms. For instance, it is a rule that the loanword *kedar* translates Greek κεδρος/Latin *cedrus*. However, the form *cedron* appears once evidently under the impact of *Cedron* in John 18:1. In addition, the form *Sava* in NYM regularly corresponds to Greek Σαβα/ Latin *Saba* in Isaias 60:6. However, the computer listed also the form *Saba* written by scribe A.

3. 4. Lists of loanwords with Italian phonetic peculiarities or of Dalmatian origin can be made by the computer. Many lists of loanwords can be executed by the program according to phonetic and orthographic peculiarities of loanwords, or according to morphological adaptation.

I shall give only one example. The presence of the phoneme /f/ in modern Slavic languages is often given (in contact linguistics) as an example typical of the process of importation of phonemes as the result of the activation of the so-called latent phonemes.[6] However, this importation dates from a late period and even now Southern Slavic dialects show different substitutions for the phoneme /f/ from giving languages (not only Greek and Latin, but German, Italian, etc.). It is a controversial problem by whom and when this letter was included into Glagolitic manuscripts. It can be found both in its Greek uncial form and in its specific Glagolitic form in the earliest extant manuscripts (Илчев 1969, 1985). It is assumed that even when the medieval scribes wrote the letter Φ (either Glagolitic or Cyrillic), they read it with substitution for the foreign sound. That fact explains the confusion of letters Φ and П in

[6] Cf. for instance Filipović (1986:82, 84) and the literature he quotes.

loanwords, attested in the earliest Slavic manuscripts.[7] The substitution of *p* for *f* is typical of Croatian manuscripts. The substitution of sound *p* for Greek or Latin *f* can be observed in spoken Croatian, e.g. *Josip, panos, Stjepan* or *Stipan, Tripun.* The substitutions for Greek φ *(>ph* in Latin[8] *)* and Latin *f* vary in NYM, as well as in other Croatian Glagolitic manuscripts. The Latin *f* in native Latin words, which are not borrowed from Greek,[9] is consistently rendered as *f* in Latin loanwords in NYM, as well as in other Croatian manuscripts. Different scribes in different genres - **b, o, r, t** write *f* for Latin *f*, e. g. felicië 206b15 **o**, ferië 221d17 **r**, filicië 206a30 **t**, filicië 226c6 **t**, fistuli 120a14 **b**: dan 3:7, fran`čis`ka 243a4 **o**, fran`čis`ko 243b6 **o**, frančiska 42d25 **t**, profek`ti 119d12-13 **b**: dan 3:2, fundamen`t' 237d12 **b**: prov 8:29.

There are only a few exeptions to this rule: (a) The form *portunata* 218a27 is used 3 times in **o** by scribe G. The f-form of this name is not found in the manuscript; (b) Scribe A_3 writes once the vernacular form *peličitoju*179d24 **o** *vs.* the literary form filictti (according to Latin pronunciation) 248a30 **t**, filičitati (according to Italian pronunciation) 219d9 used by G and D; (c) Both *p*- and *f*-forms of the loanword derived from the Latin word *praefatio* are used (the word is attested only in **r**), such as:

- variants with *f>f*: prfci 216c21, prfcië 28b21, profacië 11a24, profacië 217a22, profci 215a7, profciju 137d18;

-variants with *f>p*: propaciju 12d15-16, 22a29, propciju 16a17-18, propcië 149c30 9b8, 86d1, 98d5, prpciju 142a14, prpcië 289b19.

---

[7] On the problem of the introduction of phoneme /f/ and the function of letter Φ, see Vasmer (1908:204), Diels (1932:32), Vaillant (1952:92), Meillet (1951:37), Мирчев (1978:72), Popović (1955, 1956), Vrana (1961), Добрев (1991).

[8] In Latin orthography *ph* stands for the Greek φ in Greek loanwords in Latin. This correspondence is considered to be completely established in the 4th century A. D. Some loanwords borrowed before the 2nd century A.D. have *p* for Greek φ (Mihăesku 1966:49-50).

[9] Consequently, it is written as *f* and not as *ph* in Latin manuscripts.

The scribe G uses only f-forms. A, B, E use both of them. B_1, A_3, D use only the p-forms. In general, the latter forms prevail in the manuscript. The correlation p-:f-forms is 27:7. The word was early borrowed from Latin and spread in the spoken language of wide circles of medieval Croatian society. Therefore, scribes use its vernacular form which was early established with the substitution of *p* for the Latin *f*.[10]

The Greek φ(>Latin *ph)* is rendered both with *f* and *p* in NYM.

Some words are written only with *f*, such as rafaela 281c23 **o**, amfori 93c12-13 **b**:lc 22:10, which are in Latin garb.[11] This spelling is typical of lexemes containing two *f*s, such as: efifnie 19c3 **r**, efifnii 21b8-9 **r**, efifaniju 16b26 **r**,[12] and efifata 158a11 **b**:mc 7:34. Only the f-variant is typical of the following words as well: (1) kifa 204b10 **b**:io1:42, kifi 157d1-2 **b**:1cor 15:5, which shows peculiarities of Greek pronunciation, that is *k* vs. Latin *c*, *i* vs. Latin *e;* as well as (2) těofilě 137a7-8 **b**: act 1:1; (3) šerafimi 178c6 **o.** Different scribes, in particular: A, A_3, B, B_1, E, use the

---

[10] Mareš (1961)includes this word in the group of terms regarded as borrowings in the "cultural Moravian language" dating from the period before the Cyrillomethodian mission in Moravia. The form *prifacìi* is identified in the *Kiev* folia. The form with *pro-* is widespread in Croatian missals (Vajs 1907:557). The native prefix *pro-* can be distinguished in this form, as well as in *profekti* (used in NYM too). An association with *propověd* is probable. The use of the lexeme in medieval Croatian manuscripts can be a continuation of the Moravian tradition. However, it is likely also to have been independently borrowed by Croatian literate people, because it was a frequent *terminus technicus*, wich medieval translators, editors, scribes, alumni in the Roman tradition, were accustomed to. It is conceivable that the lexeme penetrated the language of the christianized people in the early period of conversion to Christianity in Croatia. That explains its form characterized by a native prefix and with *p* for Latin *f.*

[11] However, the p-form of the same word which I consider to be a literary borrowing from the Latin Bible, is attested in the *Roč* Missal and in the *Illirico*$_4$. In fact, NYM, as well as the Berlin Missal, Hrvoje Missal and *Novak* Missal use *amfora* in Lc 22:10, but *vidro/vedro* in Mk 14:13 for the corresponding word *amphora* in the Latin text, whereas MRoè attests *anpora* in Mk 14:13. The *Cod. Slav.73,* which is considered to contain a different translation or redaction of the biblical texts, uses the word *baril.*

[12] The same spelling is attested in *Sav.*, that is with *f* not only for Greek φ, but for Greek π as well (the Greek word is τα επιφανια>Latin *epiphania)*. The process of assimilation is visible as well. It is indicative that the similar Greek word, that is the proper name Επιφανιος is rendered as *Efifana* in the *Cod. Ass.* Latin text, whereas MRoč attests *anpora* in Mk 14:13. The *Cod. Slav.73,* which is considered to contain a different translation or redaction of the biblical texts, uses the word *baril.*

[12] The same spelling is attested in *Sav.*, that is with *f* not only for Greek φ, but for Greek π as well (the Greek word is τα επιφανια>Latin *epiphania)*. The process of assimilation is visible as well. It is indicative that the similar Greek word, that is the proper name Επιφανιος is rendered

forms discussed above.In addition, *efisie* (for instance 251d2 **r**) is always in the f-form, while *epes'* 139d8 **b**:19:1 is attested once in the p-form in hand E.

The lexemes *farisei, filip, kaĕfa* are written in both p- and f- variants.

The word *farisei* is used only in the New Testament. Examples: (a) F- forms: farisei 172b18-19 mt 22:1, farisk 222a23 lc12:1, fariseĭ 3a20 io1:24, 35b22-23 mt 12:38, 42c10 mt 23:10. (b) P-forms: parisei 53a19 mt 15:1, 53b28 mt 15:12, pariseisk'ago 263d19-20 lc 2:1, pariseom' 68d26 io 8:12, pariseĭ 46c6 mt 21:45, 51a18-19 lc 4:23, 63d10 io 9:15, 63d14-15 io 9:16, 64c4 io 9:40, 69a1-2 io 8:13, 70a3-4 io 7:32, parisĕiskago 167a21 lc 14:1, parisĕom' 63d6 io 9:13, parisĕom' 67c7 io 11:46. Hands D, B, and A_2 attest both p- and f-forms, while hands C, A_3, E show exclusively p-forms. Scribe G uses only f-forms. In general, the p-forms of this word prevail. The correlation f-: p- forms of the lexeme *farisei* (including the derivatives, such as *parisĕiskago*) equals 6:24. The p-forms were normative forms. In the same time the f-forms were not a deviation from the norm. They sounded more literary to medieval Croatian scribes. The word *farisei* was a literary word, but very frequent in liturgical texts. It was probably a part of both the active and passive vocabulary of medieval literate people. Hence, a spelling closer to vernacular pronunciation is understandable. The predominant p-forms suggest that even when medieval Croatian scribes wrote the letter *f*, they read it as *p*.

The word *filip* occurs in **b, o** and **r.** (a) F-forms: filip 59b22 **b**:io 6:7, filipa 204b12-13 **b**: io 1:43, 215c3 **o**, 215d26 **o**, filipe 215c11**b**:io 4:9, filipe 215d8 **b**: io 14:9, 216a1**b**:io 14:9, filipu 76d28 **b**:io 2:21, filip' 204b13-14 **b**:io1:44, filip' 77a2 **b**:io12:22, 77a3-4 **b**:io12:22, filp' 204b15 **b**:io 1:45, 215d4 **b**: io 14:8, filipovi 223d5 **b**: mt 16:13. (b) P-forms: pilip 128c15 **b**: act 8:31, 128c8 **b**: act 8:30, pilipa 215b27 **r**, pilipu 128b20-21**b**: act 8:26, 128c28-29 **b**: act 8:34, 128c5 **b**: act 8:29, 59d17 **b**: io 6:5, 7a9-10 **b**:lc 3:1. The p-forms appear exclusively in the Acts. Both p- and f-forms

are typical of the gospel passages. However, in that case it seems that the preference of one variant over another does not depend on the genre of the text but rather on the proclivities of different scribes. A_2 and G use both p- and f-forms. D and B1 like f-forms. The f-forms prevail in the parts written by G, D, B_1. The f-forms can be regarded as a norm for these three scribes. By contrast, E uses only p-forms, typical of the vernacular. In general, the p-forms are predominant: the correlation f-:p-forms equals 13:17. However, the difference in the frequency of the two variants is not so substantial.

On the other hand, while the proper name *filip* is written in two variants, the title of one of the Epistles is always with *p*, for instance: pilipsiem' 2d16 **r**. It is indicative that this form is attested only in **r**. Therefore, it is written in the way it was pronounced by medieval Croatian priests.

The proper name *Kaĕfa* (or of the shorter variant attested in Greek and Latin manuscripts as well, that is *Kaifa*) is used only in **b** in NYM, e.g.: (a) F-forms: kaifa 75c24 io 11:49, 78a27 io11:49, kaĕfi 103a10 io18:24, 89d14 mc 14:53, kaifi 84a8-9 mt 6:57. (b) P-forms: kaĕpa 102c27 io 18:14, kaĕpe 102c25 io 18:13, 103a23 io 18:28, kaĕpoju 7a15 lc 3:2. The majority of forms (both p- and f-) are attested in hand B. Both forms are used even in one and the same folio. The longer forms, i.e. those derived from Καιαφας, *Caiaphas,* are always the p-forms, whereas the shorter ones, i. e. those corresponding to the Greek and Latin Καιφας, *Caiphas*, are both with *p* and *f*.

Most loanwords, having φ>*ph* in their etyma, are used only in the p-form.

Some of them are rare literary words, such as: apeta 113b17 **b**:gen 5:32, apet' 114a24 **b**: gen 7:13, arpak`sadov' 278b16 **b**: lc 3:36, epan`sci 16d7 **b**: is 1:6, gazopilakiĭi 69a29 **b**: io 8:20, neptalimla 246c22 **b**: apoc 7:8, palĕkov' 278b14-15 **b**: lc 3:35, pam`pilii 141a9-10 **b**: act 2:10, panuileva 14d23 **b**:lc 2:36 paresĕnov' 278b11

b: lc 3:33, par`pan' 50d19 **b**: 4reg 5:12, parěsa 238a23 **b**: mt 1:3, parěs` 238a24-25 **b**: mt 1:3, pinik' 12c26 **b**: ps 91:13, 164d7 **b**: lev 23:40, pin`nik' 232d28 **b**: eccli 24:18, piěli 245d16 **b**: apoc 5:8, sim`ponie 120a16 **b**: dan 3:7, sim`ponie 120a3-4 **b**: dan 3:5, vitpagiju 78b15 **b**: mt: 21:1, prugii 141a9 **b**: act 2:10. I encounter these forms in hands E, A, B_1, D, B, A_2.

Other loanwords, written only in the p-form, are frequent in liturgical texts and their forms are more or less adapted according to the peculiatities of the spoken Croatian language, i.e. the p-form is combined with *Iω->o*, or ě>*i*, etc., e.g.: blaspimijuće 95b13-14 **b:** lc 22:65, valas`pimiju 73b26 **b**:io10:36, vlaspimie 53c23 **b**:mt 15:19, vlaspimiju 84b16-17 **b**:mt 26:65, vlaspimiju 90a26-27 **b**:mc 14:64, vlas`pimie 145b16-17 **b**:lc 5:21, v`laspimiju 84b13-14 **b**:mt 26:65, v`las`pimii 73b16 **b**:io 10:33, eprati 38d10 **b**:deut 11:24, epreme 101a28 **b**:os 6:4, epremom' 8a24-25 **b**: ps 79:3, eprěmom' 6a28-29 **b**: ps 79:3, 6d28 **b**:ps 79:3, eprěm' 75d14 **b**:io 1:54, iosipov' 278a27 **b**:lc 3:26, osip 8b7 **b**:mt 1:19, osipa 2a7 **b**:ps 79:2, 212c15 **b**:act 1:23, 3a3 **b**: ps 79:2, 6d26 **b**:ps 79:2, 8a23 **b**:ps 79:2, osipe 8b13 **b**:mt 1:20, osipov 278a23 **b**:lc 3:24, osipova 204b18-19 **b**:io 1:45, 246c23-24 **b**:apoc 7:8, 86b13 **b**:mt 27:56, osipovi 25d18 **b**: ps 76:16, osipov' 278a20 **b**: lc 3:23, osip` 8d24 **b**: lc 2:4, osip' 3d30 **b**:lc 1:27, ospa 6a27 **b**: ps 79:2, ospu 8b4 **b**: mt 1:18, osopata 238b11 **b**: mt 1:8, osopat` 238b11-12 **b:** mt 1:8, paraona 107c21-22 **o**, paraone 116b4 **b**: ex 15:1, paraonovimi 116a13-14 **b**: ex 14:28, triponič 30c25 **r**. These forms are found in hands D, B_1, B.

In conclusion, eight words are used only in the f-form, three words are attested both in the p- and f-form, 28 lexemes[13] were written only in the p-form. Consequently, the p-forms, closer to the vernacular pronunciation, were normative.

[13] I do not count all the grammatical forms of the words. However, I differentiate nouns from adjectives derived from them, for instance I count *osip* and *osipova* as two lexemes.

However, the f-forms cannot be considered erroneous, i. e. a deviation from the norm. Evidently the norm permitted some variation between the more literary and artificial f-forms and the p-forms closer to the vernacular way of adaptation of foreign sounds. On the other hand, the f-forms cannot be regarded as foreign words. They do not represent an early stage of adaptation of the replica. They demonstrate peculiarities of adapted loanwords. It is conceivable that some Croatian literate people pronounced them with the f-form.. As to the individual inclinations of the scribes participating in the preparation of NYM, it is obvious that G prefers f-forms, or at least the f-forms of the proper name *Filippus*, while E opts for the p-forms.

Here I shaw only a very small part of results which I received by the help of KLEIO. My goal was not to discuss all of them, but only to reveal how the program can make the analysis of loanwords faster and easier.

Literature:

Birnbaum, H., P. Rehder, *The New York Missal. An Early15th-Century Croato-Glagolitic manuscript.* München & Zagreb, 1977.

Corin, A., *The New York Missal*: A Paleographic and Phonetic Analysis. Slavic Studies V:21 (1991)

Diels, P., *Altkirchenslavische Grammatik.* Bd. 1. Heidelberg, 1932.

Добрев, И., Рашката писменост и балгарският правопис през Средновековието, В: *Кирило-Методиевски студии* 8 (1991), pp. 216-252.

Filipović, R., *Teorija jezika u kontaktu.* Zagreb, 1986.

Ilèev, P., Kăm părvonaèalnoto săstojanie na glagolièeskata sistema, *Ezik i literatura* 24 (1969), 5. Sofija, pp. 29-40.

Илчев, П. Към първоначалното състояние на глаголическата система, *Език и литература* 24 (1969), 5, pp. 29-40.

Илчев, П., Константин-Кириловата азбука и византийското писмо, В: *Изследвания по кирилометодиевистика*, София, 1985, pp. 61-73

Mareš, F.,Древнеславянский литературный язык в Великоморавском государстве, *Вопросы языкознания*, 1985, pp. 61-73.

Meillet, A. (Мейе, А.), *Общеславянский язык* Москва, 1951. (translation from the French *Le slave commun* (2nd edition). Paris, 1934).

Мирчев, К., *Историческа граматика на българския език.* София, 1978.

Mihaescu, H., *Influenţa greceasca asupra limbii romîne.* Bucureºti, 1966.

Popović, I., Grčko-srpske lingvističke studije II. Grčke pozajmice u savremenom srpsko-hrvatskom jeziku , in: *Zbornik radova SAN* 44 (1955), *Vizantološki institut* 3. Beograd, pp. 111-115.

Popović, I., Grčko-srpske lingvističke studije III. Problem hronologije vizantiskih i novogrčkih pozajmica u srpskom jeziku, in: *Zbornik radova SAN* 44 (1955), *Vizantološki institut* 3. Beograd, pp. 117-157.

Popović, I., Dva problema grčko-slovenske glasovne substitucije, in: *Zbornik radova SAN* 49 (1956), *Vizantološki institut* 4. Beograd, pp. 53-62.

Schmidt-Deeg, E., *Das New Yorker Missale. Eine kroato-glagolitische Handschrift des frühen 15. Jahrhunderts. Kritische Edition.* München, 1994.

Vaillant, A. (Вайан, А), *Руководство по старославянскому языку*, Москва, 1952. (translation from French *Manuel du vieux slave.* Paris, 1948).

Vasmer, M., Греко-славянские этюды. I. Греческие заимствования в старославянском языке, *ИОРЯС* 12 (1908), 2, pp. 197-289.

Vajs, J., Die Nomenklatur in den kroatisch-glagolitischen liturgischen Büchern, *ASPh* 29 (1907), pp. 560-580.

Верещагин, Е. М., К вопросу об использований греческой лексики в первых славянских переводах, *Советское славянознание* 1967, 6, pp. 49-58.

Vrana, J., *L' evangeliaire de Miroslav. Contribution á l'étude de son origine.* s-Gravennage, 1961.

Appendix

1. "nym.dat"

```
w$0903/101a28/epreme/b/os/6:4/
s$sem/gr=lat/lit
g$v/sg/m/o/Εφραιμ/Ephraem
p$ind
p$ai>e
p$ph>p

w$0904/101a29/ijudo/b/os/6:4/
```

s$sem/gr=lat/
g$v/sg/m/a/Ιουδασ=Ιουδα/Judas
p$-as
p$iju-

2. Index

1	2	3	4	5	6	7	8	9:10	11
abela	g	sg	m	o	179c12-13	nb			1371
abelovi	a	pl	m	adj:o	154d21	nb			1247
abiani	g	sg	f	adj:a	220c21	b	lc	1:5	1784
abiba	g	sg	m	o	229a24	nb			1855
abraama	a	sg	m	o	45a6-7	b	ex	32:13	0529
abraame	v	sg	m	o	44c28	b	lc	16:24	0522
abraame	v	sg	m	o	44d24-25	b	lc	16:30	0526
abraamom'	i	sg	m	o	21a13-14	b	mt	8:11	0359
abraamov`	v	sg	m	adj:o	43b7	b	esth	13:15	0508
abraamov'	v	sg	m	adj:o	285b21	nb			2432
abraamov'	v	sg	m	adj:o	285c14	nb			2437
abraam`la	g	sg	m	adj:o	126b21	b	act	13:26	1144
abraam'	n	sg	m	o	44d21-22	b	lc	16:29	0525
abraam'	n	sg	m	o	44d27	b	lc	16:31	0527
abraam'	n	sg	m	o	44d4	b	lc	16:25	0524
abramom'	i	sg	m	o	29b4	b	mt	8:11	0417
abramu	d	sg	m	o	39a9-10	b	2mach	1:2	0485
abram'	n	sg	m	o	115d2-3	b	gen	22:19	1027
ab`ban'	n	sg	m	o	50d18-19	b	4reg	5:12	0569

3. Question "-ium"

```
Your task is being executed.
-ium bravium          22c17       b   1cor      9:24     0370
-ium cimiterii        200d13      r                      1451
-ium lilium           163d18-19   b   os        14:6     1279
-ium lilium           214b24      b   ps        32:1     1607
-ium lilium           233d30      nb                     1912
-ium liliumi          233d7       b   eccli     50:8     1906
-ium liliumi          287d13      b   eccli     50:8     2371
-ium  lilium|          5b21       b   is        35:1     0091
-ium oficii           14b21-22    r                      0259
-ium oficii           16a25       r                      0303
Your task has been executed.
```

SUPPLEMENT

Компютърните технологии и текстологията

Мирослав Янакиев

Софийски университет "Св. Климент Охридски"

ABSTRACT: Computer technology today reveals a potential for specialists in applied linguistics that would have been difficult to imagine some twenty years ago. On the other hand, some of the innovative approaches made possible by advancements in computer technology may prove unfamiliar or difficult to understand to colleagues who are not familiar with these new tools, and with the new possibilities they create. For example, computer-supported textology can facilitate searching and indexing of material of arbitrary length from very large texts, and can do so within an arbitrary context.

A textologist with proper computer support will no longer require the voluminous *Indices Verborum* that were one of the mainstays of our profession in earlier generations. Neither will there be any need to combine lexical forms into lexemes in order to enter them in a common lemmatized form into an *Index Verborum.*

Whatever the current limitations of computer-aided textology, it clearly has a future in providing, through optical scanning and character recognition, for the transformation of any incunabulum or manuscript into a sequence of coded characters suitable for computer processing. And yet the textologist working with a computer should be aware that the machine is indifferent to the transcription or transliteration used — whether it is in Latin script, the Cyrillic script, the Latin phonetic script, the Latin phonemic one, or some other. Regardless of the specific transcription or transliteration employed, the textual information will be stored in the computer memory as a sequence of zeros and ones, and will be available in its internal, machine form for the type of processing described above.

Развитието на компютърната техника разкрива пред човек, който желае да се занимава с научно изследване на текстове, такива перспективи, каквито той до преди двайсетина година трудно можеше да си представи. Развитието на компютърната техника обаче и изисква от филолога текстолог да се опита да развие мисленето си в насока, която едва ли ще бъде приемлива за традиционно мислещите свои колеги.

Каква е тази насока?

За снабдената с компютри текстология е възможно да се осигури за разумно кратко време намирането и изброяването както на съвсем кратки, така и на много

дълги поредици от знакове в много обемни текстове при това в произволно голям контекст.

Поради това снабденият с компютър текстолог ще е разумно да привикне с мисълта, че няма да са му нужни неекологично похабяващите тонове хартия Indices Verborum, както и че няма да е нужно да се труди да обединява в "лексеми" словоформите, за да ги представя в Index Verborum обединени.

Ще се опитам с пример да поясня твърдението си.

Наша колежка от Свободния факултет на Софийския университет "Св. Климент Охридски" в момента подготвя за издаване като книга ръкописен сборник слова на Агапий, монах, живял в Рилския манастир през миналия век. Ръкописът съдържа - приемаме - 200 страници, което означава 32 хиляди текстови думи (монолекси) или 143 хиляди букви.

Да си представим, че в изданието бъде включен Index Verborum. Това ще означава че към изданието да се добавят най-малко още стотина страници, ако искаме в Index'а да включим адресите на всяка словоформа от текста, т.е. обемът на книгата ще трябва да нарасне поне с една четвърт.

А да си представим сега, че ръкописът е записан на дискета. Даже записването да е направено с най-разточителната програма за записване на текстове Word for Windows 6 и без така нареченото компресиране, целият ръкопис ще се съхранява на една четвърт от мястото, което ни представя една обикновена дискета DSHD. Даже при днешните значително завишавани цени на дискетите в цената (себестойността) на един екземпляр от изданието ще се добавят десетина лева.

Ако текстът е записан на дискета, не е нужно да му се съставя Index Verborum. Нещо повече - текстообработващите програми като споменатия Word for Windows 6 предоставят удобства, с каквито индексите не могат да ни обслужват.

Трябва ни например в споменатия ръкопис на Агапий да потърсим дакаджіа. Извиквате чрез "менюто" Edit командата Find или направо натискате едновременно клавишите Ctrl и F, появява се на екрана специално меню, което ви подканва да наберете търсената дума, набирате я, натискате клавиша Enter .... миг и на екрана се появява дакаджіа заедно с контекста си. Искате да потърсите дака. Постъпвате по аналогичен начин и за мигове компютърът ви дава последователно 3 дака с контекстите им: в оутвержднкам, дакат и дакаджіа.

Чрез друга команда - Replace - можете да установите за броени секунди колко пъти в тескта се среща съюзът да, буквата д. Чрез лесно реализуеми програмки (макроси) можете да намирате и да изброявате дисперсни буквосъчетания като п . . . с.

Би ли станал нужен специален Index Verborum, ако изданието е на текст, преведен от друг език и включва информация за чуждоезичните съответствия в съответствие с текстографическите традиции?

Ще си позволя, като говоря за тези тредиции, да напомня за една публикация в *Советское славяноведение, М., 1970, 6, с. 58-64:* **М. Янакиев, Н.В. Котова**. О некоторых принципах эдиционной палеославистики. Тези принципи, за които говорим в тази статия, остават актуални, а, уви, не се спазват в изданията на древни текстове и до сега. Но едно положение е загубило актуалността си. За него искам да

разкажа, защото е свързано с шеметното развитие на компютърната техника. И пак ще си послужа с пример.

Ето как изглежда първият ред на л. 17 (21) на готвеното от Н. Котова и мене издание на Григоровичевия паримейник:

Притчи Сол., гл 1, 23-30

се прѣдлагаѧ имъ моего дѹха ⁘ ни наѹѹ же въı

ἰδοὺ προήσομαι ὑμῖν ἐμῆς πνοῆς διδάξω δὲ ὑμᾶς

При подготовката на изданието трябваше да се пишат фишове за всяка дума, съдържащи всеки думата в минимален контекст. Всичко над 100 хиляди. И целият този труд днес се оказва ненужен. С проста клавишна манипулация, след като сме набрали в два различни, но свързани файла преводния и гръцкия текст, компютърът е в състояние да ни подаде всяка монолекса или полилекса, всяка морфема (буква или буквосъчетание) от текста на паримейника или от гръцкия файл заедно с всякакъв контекст и другоезичното му съответствие.

И всичко това, така да се каже "безплатно", без всякакъв отделен Index Verborum, без всякакъв отделен морфемарий.

С помощта на така наречените сканери, на компютърна дискета може да се пренесе всяка страница от ръкопис, и то с всичките ѝ цветови отсенки, с всичките ѝ дефекти, а после чрез съответните компютърни програми да се повиши четливостта на поизбледнели или зацапани знакове и на компютърния екран да видим в "освежен" вид древното пъстроцветие на оригинала. И всичкото това даже и като "под лупа". И всичкото това не само на екранчето на компютъра, но напоследък

(ако компютърът Ви е пуснатият в продажба от IBM от ThinkPad 755CV) с помощта на обикновен шрайбпроектор даже на голям стенен екран.

А сега - едно "лирическо отстъпление".

Отново за цената на тези възможности. Вярно е, че за да се съхранява на дискета едно сканирано цветно изображение е нужно много място и една дискета не стига, ако искаме да направим един голям ръкопис достъпен за компютърния екран и страница по страница ("фототипно"), и в удобна за обмисляне транслитерация с текстологически коментари. Но появата на така наречените CD ROM устройства отпрати в миналото "дискетната теснотия" и едновременно направи единицата компютърен запис десетки пъти по-евтина. Сега няма съмнение, че издаването "в цвят" на ръкописи вече е не само нèекологично, то е и нèикономично.

Наистина остава като че ли нерешена главната проблема, проблемата за цената на компютрите, без които нито дискетите, нито компакт-дисковете могат да бъдат използвани. Няма нужда да говоря за "медицинската" страна на използването на компютрите — не подлежи на съмнение предстоящата "смърт" на електронно-лъчевите (телевизионните) компютърни екрани. Бъдещето е на LCD (дисплеите с течни кристали), които не вредят нито с "рентгеново лъчене" , нито с "трепкане". Само че и те още повече оскъпяват компютрите.

И тъкмо по тази проблема ще си позволя да споделя някои съображения, които изискват разискване не в средите на икономисти, а в средата на милионите, даже милиардите бъдещи собственици на компютри, част от които са и намиращите се тук сега. Тези съображения непременно ще бъдат оспорвани от производителите на компютри. Даже ми става страшно, като си помисля за огромната мощ на производителите на компютри. Но ще се осмеля да споделя с вас своите мисли, които ще се сторят "опасни" за няколко стотин хиляди

специалисти, ангажирани в производството на компютри в света, но мисля, които са напълно добронамерени за стотици милиони потребители на компютърна техника, а в последна сметка са напълно добронамерени и за ангажираните в производството на компютри.

В една книжка - "Електрониката в помощ на учителя филолог", издадена през 1988 г. аз се опитах да предложа изискванията, които трябва да задоволява един компютър, за да стане пълен заместник на учебната литература и другите учебни помагала, нужни на един ученик в училище и в къщи. Тогава сметката ми беше, че такъв "универсален учебен компютър" не струва повече от стотина долара.

Ако добавим и още най-много около двеста долара за компакт-дисково устройство, трезво погледнато, един портативен компютър с цветен LCD и с встроен CD ROM може да се продава на цена много под триста долара, ако човечеството реши всеки ученик да бъде снабден със свой индивидуален "универсален учебен компютър", защото огромното производство ще намали многократно себестойността му, а оттук и продажната му цена.

Но за да стане това, производителите на компютри ще трябва да се откажат от игрите с "непрекъснатото усъвършенстване" , които се играят със скритата цел да се повишават цените на компютрите, като се обявяват "неусъвършенстваните" компютри за "остарели морално", без да са остарели физиччески.

Преди много години Пентагонът в САЩ обяви конкурс за създаване на нов компютърен HIGH-LEVEL-Language. Роди се езикът ADA. Езикът ADA остана съвсем малко популярен в средите на програмистите. Но аналогичен подход към неминуемата всеобща компютризация на човечеството е напълно приемлив, ако световна организация, например ООН, обяви конкурс за конструиране на нужния

днес за младите, а утре за всички хора на земята "универсален учебен компютър", с който човечеството безплатно да снабдява всяко дете, когато то започва да се учи.

Приключих с "лирическото отклонение".

Защо ми беше нужно то? Защото съм сигурен, че световен конкурс за конструиране на "универсален учебен компютър" ще бъде обявен и текстологията ще трябва да даде своя принос за постигането на добър резултат от конкурса.

Какъв според мене може да бъде този принос?

Текстолозите (филолозите, лингвистите), снабдени с компютри, ще трябва да осигурят компютърни технологии, които да направят максимално плавен, не нараняващ психиката на възрастните преход към една за целия свят азбука.

За снабдената с компютри текстология е възможно (но в момента все още не съвсем!) да осигури за разумно кратко време чрез сканиране превръщането на всеки стар печатен и ръкописен текст в удобна за компютърна обработка поредица от знакове. Поради това снабденият с компютър текстолог ще е разумно да привикне с мисълта, че за компютъра е безразлично "в каква транскрипция (транслитерация)" – латинска, кирилска, латинска фонетична, латинска фонематична, . . . ще се пази в компютърната памет текстовата информация.

Един пример. Приведеният ред от Григоровичевия паримейник:

сѣ прѣдлагаѧ имъ моего дѫха ⁘ нн наѹѹд же въı

може с просто превключване на така наречените True-Type-фонтове и да се латинизира:

se pryadlagayen imw moego doyha ⁘ ni naoychen zhe vwi,

и да се елинизира:

σε πρεδλαγαιεν ιμω μοεγο δουχα νι ναουτσιεν ζιε βωι

Гръцкият текст:

ἰδοὺ προήσομαι ὑμῖν ἐμῆς πνοῆς διδάξω δὲ ὑμᾶς

също с просто превключване може да се кирилизира:

ἰgoὺ προὺsomαι хὺomῖn emὺc pnoὺc pὺcин gugàkco gè хὺomàc

и да се латинизира:

idoù proeésomai hyumîn emeks pnoeês reêsin didáxo dè hyumâs

И може би тъкмо лекотата, с която компютърът транслитерира, ще отстрани постепенно от съзнанието на хората преградите, които им пречат да минат към една азбука, както вече са почти узрели да минат към една валута.

First Results of Testing the Computer Recognition of Manuscripts

Peter Pavel Klasinc, Vinko Kurent (SCANDATA Maribor, Partizanska 17, 62000 Maribor, Slovenija)

ABSTRACT: Instead of building a recognition system by using analysis of preselected (for chosen material) and learning the bit image of each character in various combinations, we try to recognize characteristics of graphic object in general.

The necessary phase of vectorization of the chosen part of the text makes it possible for the program to automatically determine for each vector intersections, number and arrangement of cycles, endpoints,...) is it comparable with others, without having in mind the appearance of a specific symbol set that is most likely to be used.

This method makes it possible that each next sample of raster text can be successfully processed without additional training session, even if the material contains some deviations like inclination, unequal sizes, etc. Therefore the human doesn't have to teach the computer to the new character or symbol if this new character was already used, but is this time smaller, bigger or more inclined than before, because it was written in same handwriting by a different writer or printed by a printer with slightly different character set.

In cultural heritage the archives as the keepers of archival documents and memory of nations present a part of activity that became a reality through many centuries. There are great number of documents stored in many states that have their special place there.

Archival documents are the witnesses of spiritual and cultural dimension of a nation. The documents originate in various historical ages, they have different forms and are written in various manner of writings, languages, what depends of the degree of mind, especially it depends of time and space of their origin.

Under appearance of new information carries a written word really presents some thousand years period, in which originated written information that have been transferred and written down to special grounding by special written material, that is a result of the development of written forms. The writings itself developed through ages differently as it is shown especially in scientific texts about the development of writings. Paleography as

historical auxiliary science is an extended conception therefore it is divided into other paleography such as Latin paleography, paleography of Cyrillic characters, paleography of Glagolitic alphabet types, Arabian and Jewish paleography and others.

The most important paleography in European world is Greek, Latin and Cyrillic characters paleography. The most important paleography of this three is undoubtedly Latin paleography, that searches the origin and developments of writing on Latin alphabet basis.

Most of written documents originated from Middle Ages. As we yet mentioned an important of the development of writing has as material that texts are written down, as the written basis on which the information has survived. The treating and professional appraisal of texts originated in Europe from the first century to the second and a half of fifteenth century were mainly written in Latin language. We have to emphasize that Latin language was officially language in the neighboring Croatia until 1847. Latin paleography as a way how to read old Latin texts and abbreviations and on professional basis to define time and space of the origin of written documents. Close by Latin paleography it is also to pay attention on the development of Latin alphabet used later in other languages and not only in Latin. We mainly use it in our area because it searches old manuscripts, therefore Latin paleography is the basis for studies of written sources in Middle European area. Two undertakings in Maribor (MICRODATA, The Undertaking for development and trading of hardware, and SCANDATA, The undertaking for scanning and treatment of records in co-operation with Dutch company SATURN, Haag) are engaged on the development of software that would help users of archival documents written in Cyrillic characters.

Both undertakings also closely co-operate with professionals of archival sciences of the International Institute for Archival Studies at Regional archive Maribor because there are many new computer technologies and methods for testing them.

Political changes in Europe especially after 1990 probably influenced the changes in East Europe, which begin to research intensively "sine ira et studio" its past history. That nations wish by searching the information to return their lost identity also by studying his historical sources originated in the past, but the documents are not always reachable. In that cases there is a great interest to search the archival documents written also in Cyrillic characters, that changed through centuries and got various forms.

The acquirements of knowledge and efficiency of professionals for reading all forms of Cyrillic characters is expensive and it is not possible to expect, that all archives, also small ones and those in poor countries, would qualify a great number of professionals to do that. The archivists at regional archive Maribor and computer professionals in both mentioned undertakings examined what are the possibilities to read the Cyrillic characters and how to read it by computers and to make some tests about that.

Technical Implementation of Reading of Gothic Writings

Various software equipment for optical character recognition, that can be found on the market, are meant primarily for recognition of printed writings that are commonly used in everyday's business and science work. The development of such software equipment is going in the direction of achieving best possible results in transcription of pre-selected and known character set. The comparison of pre-learned bit-image of letters and characters of specified character set (font), which appear in connected text, can give good results, but only for those character, which exactly match the bit-image in the knowledge base.

This means that, if letter is too small, too big, inclined or stretched, it is impossible for the computer to recognize it. This means, that we have to teach it to recognize every different appearance of the same letter.

Such software equipment therefore requires extensive preparation, teaching of hundreds of different variants of the same letters or symbols. Such solution would be unsuitable for our problems.

Our goal is to implement alternative solutions, providing possibility to transcribe different characters sets with as little pre-learning as possible and comparison of characteristics and properties of different characters rather than comparison of their images.

Such 'new way' is introduction of vector based methods of recognition which are used in our software equipment VOCR developed for the needs of the International Institute of Archive Sciences at the Provincial Archive Maribor. With this method the recognition process to what we presently think to be happening in human brains. Instead of doing analysis of a specified writing and build the recognition system on the basis of the appearance of characters, we are trying to recognize the properties of symbols in general.

The necessary phase of vectorization of the chosen part of the text (figure S1 and T1) makes it possible for the program to automatically determine for each vector structure, on which characteristics (like configurations of vector intersections, number and arrangements of cycles, endpoints, ...) is comparable with others, without having in mind the appearance of a specific symbol set that is most likely to be used. This method enables us to build software equipment which is capable to indeterminate transcription, that is, it recognizes whatever we teach it to - even gothic - without the need for the software to be modified or adjusted in any way.

Automatic vectorization does require some computer processing time, which varies with the size of the raster and other vectorization parameters, but the vector drawing gained in such manner (figure S2 and T2) enables at the further phases - like recognition of properties of symbols (figure S4 and T4) - process that can already be done automatically with the high

level of reliability. These methods makes it possible that the human - expert must only be present when the scanning and vectorization parameters are set to optimal values (according to the quality of material) and in the phase of learning of typical properties' (figure S3 and T3).

The most important benefit is that each next sample of raster text can be successfully processed without additional training sessions, even if the material contains some deviations like inclinations, unequal sizes, etc. Therefore the human doesn't have to teach 'the computer to the new' character or symbol if this new character was already used, but is this time smaller, bigger or more inclined than before, because it was written in same handwriting by a different writer or printed by a printer with slightly different character set.

Figure S1: Raster sample of gothic handwriting before the processing

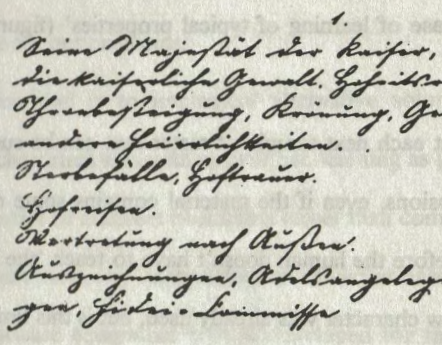


Figure S2: Same sample after vectorization

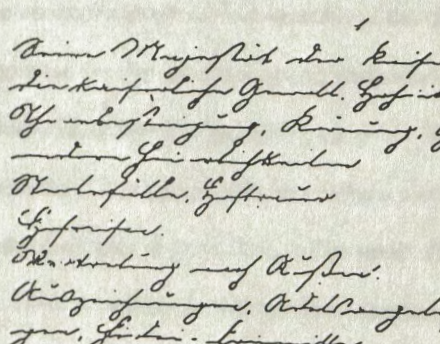


Figure S3: Learning procedure

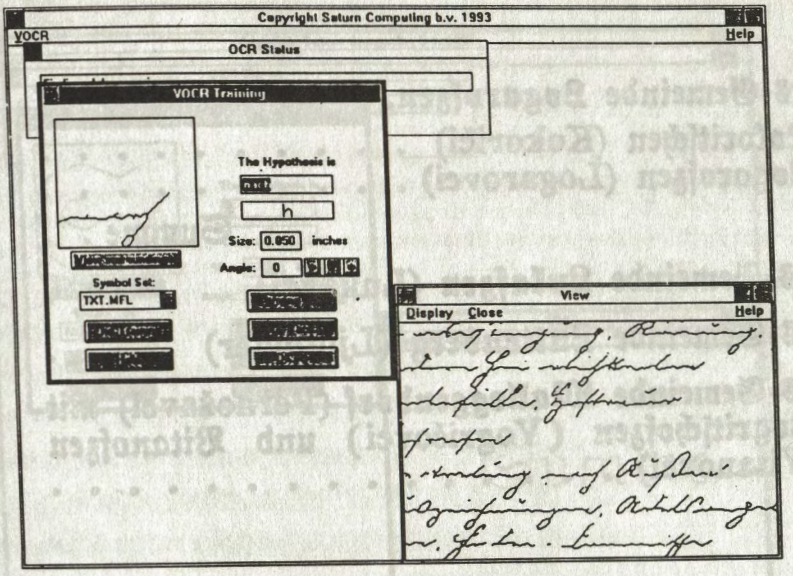


Figure S4: Result of recognition

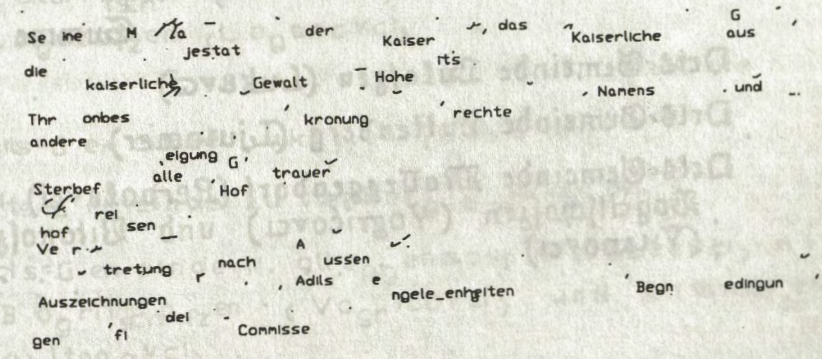


Figure T1: Raster sample of printed gothic writing before the processing

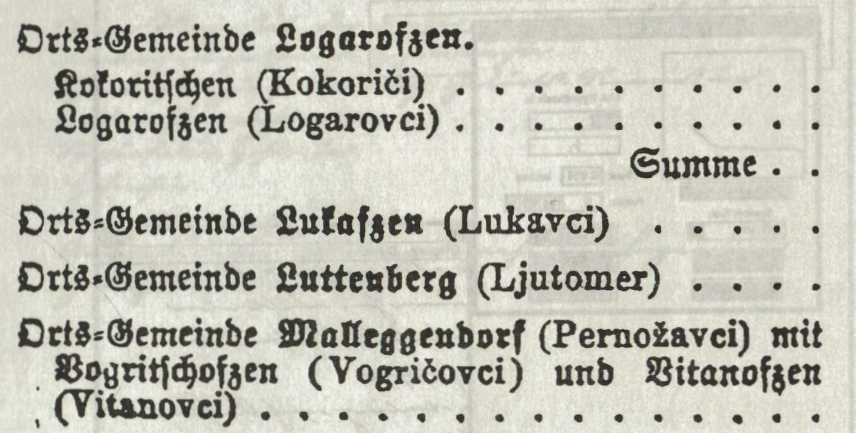

Orts-Gemeinde Logarofzen.
Kokoritſchen (Kokoriči) . . . . . .
Logarofzen (Logarovci) . . . . . .
Summe . .
Orts-Gemeinde Lukafzen (Lukavci) . . . . .
Orts-Gemeinde Luttenberg (Ljutomer) . . . .
Orts-Gemeinde Malleggendorf (Pernožavci) mit Vogritſchofzen (Vogričovci) und Vitanofzen (Vitanovci) . . . . . .

Figure T2: Same sample after vectorization

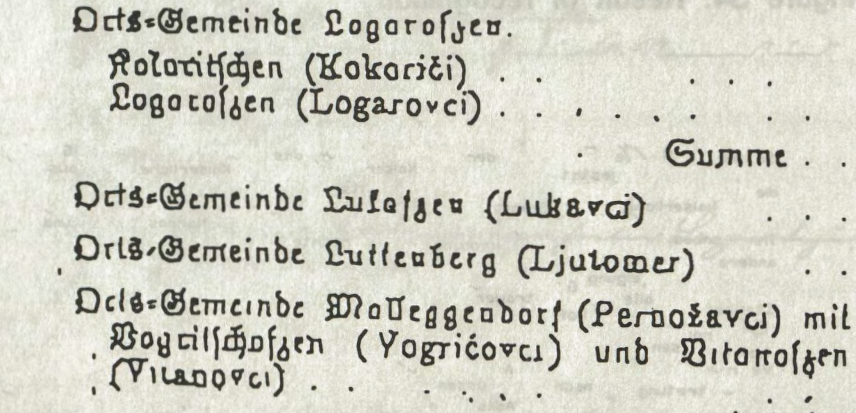

Orts-Gemeinde Logarofzen.
Kokoritſchen (Kokoriči) . . . . . .
Logarofzen (Logarovci) . . . . . .
Summe . .
Orts-Gemeinde Lukafzen (Lukavci) . . .
Orts-Gemeinde Luttenberg (Ljutomer) . .
Orts-Gemeinde Malleggendorf (Pernožavci) mit Vogritſchofzen (Vogričovci) und Vitanofzen (Vitanovci) . . . . . .

Figure T3: Learning procedure

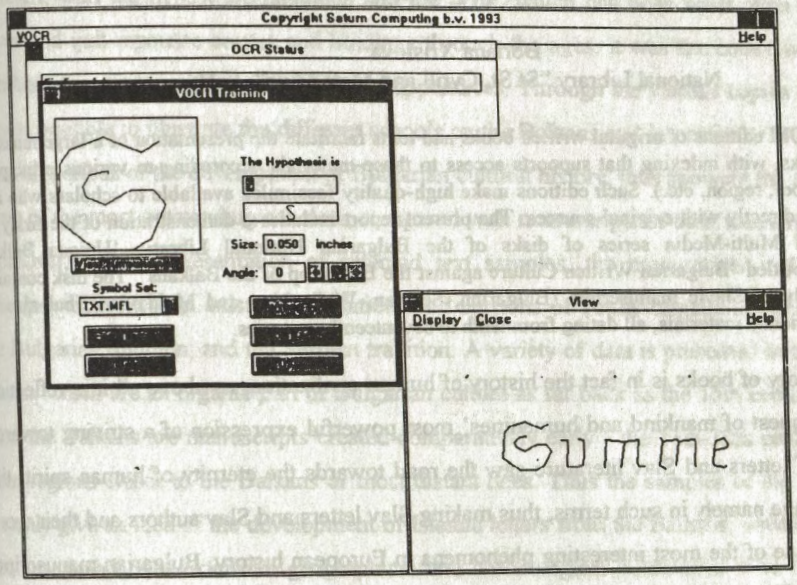


Figure T4: Result of recognition

□ rts=Gemeinde Logaroftzen.

Kokoritschen (Kokonici) . . . . . .

Logaroftzen (Logarovci) . . . . . .

. . . . . Summe . . .

□ rits=Gemeinde Lukaftzen (Lukavci) . . .

□ rts/Gemeinde Luttenberg (Ljutomer) . . .

□ rls=Gemeinde Mollegg.endorf (Pernozavci) mit Bogrilschofzen (Vogrlcovci) und Bltanofzen (Vltanovci) . . . . . .

The first Bulgarian CD-ROM: Bulgarian letters in the context of Balkan literature, Tetraevangelia and Qurans

Boriana Xristova
National Library "St.St. Cyrill and Methodius", Sofia

Abstract: CD-ROM editions of original written books and texts facilitate the presentation of a large mass of manuscripts works, with indexing that supports access to these materials accoreding to various principles (genre, date, school, region, etc.). Such editions make high-quality facsimiles available to scholars who are not able to work directly with original sources. The present report includes a demonstration of the firsty of the Ancient and Multi-Media series of disks of the Bulgarian National Library, 'Unique Balkan Manuscripts', subtitled "Bulgarian Written Culture against the Backdrop of the Balkans'. The disk contains examples not only of Slavic manuscripts (Bulgarian, Serbian, Wallachian and Moldavian), but also of Byzantine and Oriental materials, all dating from tenth to seventeenth centuries.

The history of books is in fact the history of human civilization, as above all it is reflection of the spiritual quest of mankind and humanities' most powerful expression of a striving towards eternity. Men of letters and Slav literature saw the road towards the eternity of human spirit, the concrete personage namely in such terms, thus making Slav letters and Slav authors and their work after 9 century one of the most interesting phenomena in European history. Bulgarian manuscripts are the basis of Slavic Orthodox culture, with a millennium old history, reflecting thoughts and works of the greatest Slavic scholars, writers and thinkers. The development of the Bulgarian literary tradition, seen on the background of Balkan literature and culture is a unique cultural phenomenon, based on variety of comparison, drawing of parallels, analysis, as well as generalizations and conclusions at various levels. The material included in this electronic book not only contains valuable information on various manuscripts, but also allows research in several directions. This electronic book is of interest to Byzantine scholars, Slavicists, specialists in Oriental studies, historians, art historians, specialists in textology, and Bible studies, as well as the general reader with an interest in antiquities.

The idea of illustrating Bulgarian culture, seen on the background of Balkan civilization, is based on the Bible and Quran, as they are the Holy books of Eastern Orthodoxy and Islam. They are the basis of both cultures, in fact both civilizations, which co-existed on the Balkans for quite a long period without any interference of their essential character. These civilizations could illustrate the principle values and striving of men of letters on the Balkans, and the different, yet similar tendency of getting closer to God and mankind. Moreover any literate person could make a copy of the Four Gospels and the Quran, to meet his need, with the conviction, that life was something that came and passed, yet it is the written word that stays.

The principle examples selected from the Bibles are Tetraevangelia. Not only is the Gospel the first book to be translated in Slav language, but it was the most frequently copied text in the Balkans, with the richest illuminations and the book, which has been most often decorated with silver and quit repousse covers and bindings through the ages, it was the codex with the greatest number of versions, with various different appendices. Through the various copies of the Gospel it is best possible to illustrate the different schools on the Balkans and in particular Bulgarian cultural from all periods of the development of Bulgarian cultural history, thus showing the achievements of some of the most outstanding scholars, compared with manuscripts of their teachers and followers. Moreover, through presentation of selected text samples, the manuscripts included show the principal differences in existing translations and versions of Tetraevangelia in the Greek tradition, the Bulgarian tradition, and the Serbian tradition. A variety of data is presented in the copies of the Quran, which are an organic part of Bulgarian culture as far back as the 15th century. Also known from the Balkans are manuscripts created comparatively early - the 11th-12th century, originating from regions closer to the Balkans or most distant ones. Thus the samples of the Quran included not only give an idea of the development of Islamic letters from the Balkans, written in Arabic, but also its spread over extensive geographical and chronological boundaries. The Qurran not only outlines the background of the development of Orthodoxy, but is also an organic part of Balkan letters since the early 15th century. Included here are valuable copies of the Quran, whose existence on the Balkans has been confirmed in records. For the first time two literary traditions, which closely co-existed simultaneously, without mixing, or much mutually cultural intermingling, nor borrowing from one another, or merging, are placed on the same level. Their specific co-existence in the same range is proof that neither culture could destroy the other or break it or place it in an subordinate position. Most of the Slavic Tetraevangelia date namely from the period between the 15th and 17th century, not only because of the fact that more books have been preserved from this period, but also because they are proof of the vitality of Bulgarian culture, and generally speaking of the vitality of Christianity. The specific fact of the co-existence in the same range of an Orthodox and Islamic religious and cultural complex materialized in sacred religious books of the two faiths is no doubt proof of the circumstances that neither culture could destroy the other or place it in a subordinate position; this is a unique Balkan phenomenon - the co-existence of religious Orthodox activities and Islamic life which do not merge, do not exert any influence of substance.

The book also contains several Greek gospels, for namely Byzantine letters were the model and the source, followed by Slavic scholars, in particular in the early period of the development of Slavic cultures on the Balkans. Thus an attempt has been made to reveal different cultural strata in

the Balkans covering the period from the 10th to the 17th century, namely the chronological range of the codices included here.

Slavic illuminated manuscripts on the Balkans are represented through copies of Tetraevangelia, made in Bulgarian lands, in various Bulgarian towns, seats of literary schools. Also included are a number of Serbian codices, which on the one hand are an illustration of the development of Serbian letters for the period between 13th and 17th century, and on the other point to the close links between Bulgarian and Serbian letters; for certain periods the strong influence of Bulgarian letters on Serbian letters. The electronic book also presents, in brief the development of Wallachian-Moldavian letters, which date from the early 15th century to the 17th, based on the Bulgarian literary tradition, considering that in the course of several centuries scholars from Wallachia and Moldavia copied Bulgarian models from the 14th and 15th century.

The proposed CD-ROM includes about 1 000 shots of manuscripts kept in the largest collection of Bulgarian MSS at St. St. Cyril and Methodius National Library. Many of the monuments included here and entered the CD-ROM is presented before scholars for the first time.

In particular the electronic book contains the following data:

1. Shots of Greek Tetraevangelia, Selected Miscellanies of Gospels, and Miscellanies of Readings from the Gospels, covering the period between the 10th and 17th century.
2. Shots of Bulgarian Tetraevangelia and Selected Gospel Miscellanies covering the period between the 13th and the 17th centuries.
3. Shots of Serbian Tetraevangelia and some Selected Gospel Miscellanies covering the period between 13th and 14th centuries.
4. Shots of Wallachian and Moldavian Tetraevangelia covering the period between the 15th and 17th centuries.
5. Two hundred shots from the 39 Qurans covering the period between 13th and the 17th centuries.

Chiefly the MSS of well-known scholars, are selected coming from major centers of spiritual life and literary schools. The notes of scribes are given, when preserved and translated. The shots comprise samples of the ornamentation, the binding and repousse cover, which also represents the development of applied arts; included are also samples of the text. The samples are selected from Luke 21:1-30, John 18:1-30, as recent linguistic research and textological studies of the Gospel text in a Slavic context have proven to offer the greatest variety of redaction and provide excellent potential for philological analysis. The texts have been omitted in places where the respective verses of the Gospel have not been preserved. Paleographical comparisons and examination of style have served in the selection of the Sura "Al-Fatih" where it has been

preserved, the beginning of the Second Sura "Al-Bakara", as it is linked with the former as an artistic design and is remarkable for its wealth of ornament, together with the Sura "Al-Kahf" (the Cave) and "Mariam" (Mary). The latter two are preferable because of their content with common personages and Biblical subjects. The end of the text of the Quran coincides with the Sura "An-Nas" (the People).

Each shot is accompanied with the corresponding annotation, providing the following information:

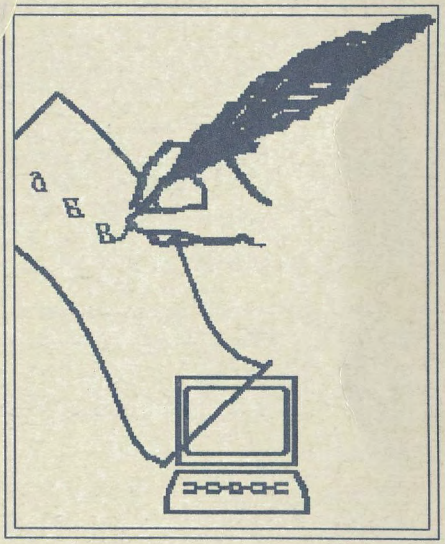
1. Name and special name of the monument (if any), under which it is known in literature.
2. Precise dating.
3. The signature in the Sofia National Library. Also included is a Quran from the library at the village of Chepintsi.
4. Origin. In case of Qurans, with unknown origin, the language is cited - Arabic, Persian, Ottoman Turkish.
5. Contents of the number of sheets, content and dimensions.
6. Data about the scribe, and place of making the copy (if given in the MS).
7. Description of the included page of the codex.
8. In some cases the acquisition source is given.
9. The most original notes of scribes are given in the original and translation. In the case of Qurans the full scribes' notes are translated concerning the vakuf, as this provides information on the history of the monument. The dating of these notes is given according to the Hegirand the Gregorian calendar. Translator's notes are given in brackets.
10. Examples of original bindings and repousse covers are given.

NOTES


B16395

COMPUTER PROCESSING OF MEDIEVAL SLAVIC MANUSCRIPTS

First International Conference, Blagoevgrad 1995
ABSTRACTS



КОМПЮТЪРНА ОБРАБОТКА НА СРЕДНОВЕКОВНИ СЛАВЯНСКИ РЪКОПИСИ

Първа международна конференция, Благоевград 1995
РЕЗЮМЕТА
Институт за литература, БАН, София 1995

Дар за Централната
библ. на БАН
от А. Милтенова

COMPUTER PROCESSING OF MEDIEVAL SLAVIC MANUSCRIPTS

1/56
БАН
БИБЛИОТЕКА

Към Б 2573/96

БАН
ЦЕНТРАЛНА
БИБЛИОТЕКА

ДАР

David J. Birnbaum, Andrej T. Bojadžiev,
Milena P. Dobreva, Anisava L. Miltenova
(**EDITORS**)

ABSTRACTS OF THE FIRST INTERNATIONAL CONFERENCE

COMPUTER PROCESSING OF MEDIEVAL SLAVIC MANUSCRIPTS

24-29 July 1995, Blagoevgrad, Bulgaria

Organizers:

- Institute of Literature - Bulgarian Academy of Sciences
- Institute of Mathematics - Bulgarian Academy of Sciences
- University of Pittsburgh
- Center for Slavo-Byzantine Studies 'Ivan Dujčev' at the University of Sofia
- American University in Blagoevgrad

Sponsors:

- International Research and Exchanges Board (IREX)
- American Council of Learned Societies (ACLS)
- Open Society Fund
- DEC/MOBIS - Bulgaria

БИБЛИОТЕКА

Computer Processing of Medieval Slavic Manuscripts, First International Conference, Blagoevgrad, 1995

© David J. Birnbaum, Andrej T. Bojadžiev, Milena P. Dobreva, Anisava L. Miltenova (**EDITORS**)
© The copyrights of the abstracts belong to their authors

PUBLISHER Institute of Literature, Bulgarian Academy of Sciences, 1995
ISBN 954-8712-02-4

FOREWARD

The first international conference on Computer Processing of Medieval Slavic Texts (Blagoevgrad, 1995) is intended to provide an international forum where Slavic philologists and information science specialists can work together to elaborate methods for the computational processing and analysis of early Slavic written materials. In addition to providing an opportunity to discuss and learn about matters of common interest, we hope that this conference will enable Slavic philologists to coordinate their research, with the goal of developing common tools and methods. One reason that the computational analysis of early Slavic materials currently lags far behind comparable projects in other disciplines is that Slavists often work in isolation, using *ad hoc* and incompatible strategies. One consequence of this incompatibility is that impressive results achieved with one set of materials in one environment may be inaccessible to other researchers as a result of system dependencies that could have been avoided. We would like to thank all the participants in this conference, who have traveled to Blagoevgrad from many different countries in order to participate in this communal project of developing standardized text-processing strategies and materials.

This conference was supported by grants from the International Research and Exchanges Board, with funds provided by the U. S. Department of State (Title VIII) and the John D. and Catherine T. MacArthur Foundation, by the Joint Committee on Eastern Europe of the American Council of Learned Societies, and by the Open Society Fund.

None of these organizations is responsible for the views expressed. The Organizing Committee would also like to acknowledge the generous support of the Bulgarian Academy of Sciences, the University of Pittsburgh, the University of Sofia, and the American University of Bulgaria.

FOREWARD

Program and Contents

MONDAY, July 24th

14:00 Welcoming Addresses
15:00 Coffee Break
15:30 Session: STATE OF THE ART
Chairman: Prof. S. Kožuxarov

COMPUTER TECHNOLOGIES AND TEXTOLOGY 12
..... Miroslav Yanakiev, *Sofia*
BIBLE AND COMPUTER:
A BRIEF HISTORY OF THE LAST TEN YEARS 13
..... Winfried Bader, *Tübingen*
NEW TOOLS FOR OLD LANGUAGE: COMPUTER PROCESSING
OF BULGARIAN TEXTS 14
..... Elena Paskaleva, Milena Dobreva, *Sofia*
HOW SLAVIC PHILOLOGISTS SHOULD USE COMPUTERS 15
..... David J. Birnbaum, *Pittsburgh*

17:30 Reception

TUESDAY, July 25th

9:00 - 11:00 Session: ENCODING PROBLEMS

Chairman: Prof. W. Veder

INFORMATIONAL AND PRESENTATIONAL UNITS IN EARLY CYRILLIC WRITING 16
..... David J. Birnbaum, *Pittsburgh*
TRANSCRIPTION RULES FOR OLD CHURCH SLAVONIC WRITING 17
..... Karsten Grünberg, *Heidelberg*
GRAPHIC OR GRAPHEMIC ENCODING OF MEDIEVAL SLAVIC MANUSCRIPTS 17
..... Cvetana Krstev, Duško Vitas, Henning Moerk, Gordana Pavlović-Lazetić, *Belgrade, Aarhus*
IMPLEMENTING A MEDIEVAL SCRIPT ON A PERSONAL COMPUTER 19
..... Sebastian Kempgen, *Bamberg*

11:00 Coffee Break

11:30 - 12:30 Session: TEXT PROCESSING: PROBLEMS, METHODS, AND TOOLS

Chairman: Prof. E. Paskaleva

THE APPLICATION OF TEXT ENCODING INITIATIVE *(TEI)* GUIDELINES TO ENCODING A 14TH-CENTURY SERBIAN CHURCH SLAVONIC PSALTER 22
..... Mary MacRobert, *Oxford*
О СОЗДАНИИ ФОНДА ПИСЬМЕННЫХ ИСТОЧНИКОВ РУССКОГО ЯЗЫКА XI-XII ВВ. В ИНСТИТУТЕ РУССКОГО ЯЗЫКА РАН 23
..... Александр Молдован, *Москва*

12:30 Lunch

14:00 - 16:30 Session and Workshop: MANUSCRIPT COLLATION

Chairman: Dr. H. Gaylord

TRANSCRIPTION AND EDITION 20
..... William R. Veder, *Amsterdam*
COMPUTER COLLATION OF MANUSCRIPT TRANSCRIPTIONS 20
..... Michael Bakker, *Amsterdam*
ADVANCED SEARCHING 21
..... Tanja Ivanova, Nina Šoileva, Irina Bolceva, *Sofia*

16:30 Coffee Break

17:00 - 18:30 Workshop: TUSTEP, Instructor: Dr. W. Bader

WEDNESDAY, July 26th

9:00 - 11:00 Session: FONTS. IMAGE PROCESSING
Chairman: Prof. K. Grünberg

A CHURCH SLAVONIC ALPHABET FOR REPRINTING OLD MANUSCRIPTS USING A MICROCOMPUTER 46
..... Mimoza Majstoroska, *Skopje*
FONTS, CHARACTER SETS, AND CHURCH SLAVONIC GRAPHEMES 46
..... Ljupčo Mitrevski, *Skopje*
OPTICAL CHARACTER RECOGNITION OF HISTORICAL DOCUMENTS 47
..... Ivelin Stojanov, *Sofia*
OPTICAL CHARACTER RECOGNITION OF MEDIEVAL MANUSCRIPTS 47
..... Peter Pavel Klasinc, Vinko Kurent, *Maribor*

11:00 Coffee Break

11:30 - 12:30 Session and Workshop: Problems in Running a Concrete Project: *Computer Supported Processing of Medieval Slavic Manuscripts*
Chairman: Prof. B. Xristova

COMPUTER-AIDED ANALYSIS OF THE MACROSTRUCTURE AND TYPOLOGY OF MEDIEVAL SLAVIC MISCELLANIES 37
..... Anisava Miltenova, *Sofia*
PALEOGRAPHIC AND ORTHOGRAPHIC FEATURES IN THE DESCRIPTION OF SLAVIC MANUSCRIPTS 38
..... Andrej Bojadžiev, *Sofia*

12:30 Lunch

14:00 - 19:00 Workshop: TEI
Instructors: Prof. D. J. Birnbaum, Dr. H. Gaylord, Dr. W. Bader, Dr. N. Finke

TEI Ongoing Projects 49
..... Harry Gaylord, *Groningen*

THURSDAY, July 27th

9:00 - 12:30 Session: APPLICATIONS OF DBMS

Chairmen: Prof. A. Moldovan, Prof. Z. Ribarova

A DATABASE FOR ANCIENT SLAVIC MANUSCRIPTS IN MACEDONIA 27
..... Stana Jankoska, Dragan Mixajlov, Ljupčo Josifoski, *Skopje*
COMPUTER PROCESSING OF OLD CHURCH SLAVONIC MANUSCRIPTS: RESULTS AND PROSPECTS 27
..... Zdenka Ribarova, Kirill Ribarov, *Skopje, Prague*
PROBLEMS INVOLVED IN SETTING UP A UNIFIED COMPREHENSIVE CATALOGUE OF SLAVIC TEXTS 28
..... Reinhard O. Lampe, *Heidelberg*
МЕЖДУНАРОДНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ПРОЭКТ "SLAVIA ORTHODOXA ET SLAVIA ROMANA": ВЗАИМОДЕЙСТВИЕ СЛАВЯНСКИХ МИРОВ: ДУХОВНАЯ КУЛЬТУРА ПОДЛЯШЬЯ 30
..... Юрий Лабынцев, *Москва*
ОСНОВНЫЕ ЗАДАЧИ ПОСТРОЕНИЯ БАНКА ДАННЫХ МЕЖДУНАРОДНОЙ СЛАВЯНОВЕДЧЕСКОЙ ПРОГРАММЫ "ИСТОРИЯ КНИЖНОЙ КУЛЬТУРЫ ПОДЛЯШЬЯ" 31
..... Любовь Щавинская, *Москва*
AN AUTOMATED SYSTEM FOR PALEOGRAPHIC AND ANALYTICAL DESCRIPTION OF SLAVIC MANUSCRIPTS 32
..... Vasja Velinova, *Sofia*

12:30 Lunch

13:30 - 16:00 Session: CASE STUDIES. TYPES OF MANUSCRIPTS

Chairmen: Prof. D. Birnbaum, Dr. M. MacRobert

THE PSALTER: FROM THE ORAL TRADITION TO THE INFORMATIC ONE 39
..... Monia Camuglia, *Pisa*
MEDIEVAL SLAVIC MENOLOGIES ON LINE 40
..... Cynthia Vakareliyska, *Eugene, Oregon*
COMPUTATIONAL ANALYSIS OF HYMNOGRAPHIC COMPENDIA 41
..... Emilia Guergova, *Sofia*
A SYSTEM FOR ENCODING EUCHOLOGICAL TEXTS WITHIN THE PARAMETERS OF THE *IST* COMPUTER PROGRAM 43
..... Maria Šnitter, *Plovdiv*
SOME SUGGESTIONS ARISING FROM A PERSONAL COMPUTER ANALYSIS OF THE SVETOSTEFANSKA XRISOVULJA 44
..... Francis R. Salter, *Ontario*

time it nonetheless did not present any particularly difficult problems. The software currently available to the casual user is considerably more sophisticated and presents even fewer difficulties.

The programming used for the analysis and production of the text is not complex, and is within the learning curve of most personal computer users. At present, the use of DBF format for database and Clipper programming would be more in keeping with trends in computer programming.

The Svetostefanska xrisovulja work demonstrates the ease of using personal computers in doing linguistic research. The procedures and programming used demonstrate the practicality of using personal computers for linguistic research in Slavic Languages. The use of current script support, programming languages, and graphics make analysis considerably easier, particularly in reducing the labor and time needed to produce a working copy of the document and make a statistical analysis of its contents. Taken to the next step, it would require little effort to build a dictionary of the document into a separate database and integrate it into the printed form of the document.

The structuring of the databases and programming knowledge required to develop computer applications for linguistic analysis is within the grasp of any serious researcher. However, in order for work completed on personal computers to be accessible to other researchers, it should be exportable via electronic media in a recognized database and script format. Research will be enhanced considerably in the future when there is:

(1) a standard for the developers of Church Slavonic scripts appropriate to the language (in the above instance, 1300 Serbian), which will make text in databases exportable;

(2) a rationalized sorting algorithm for sorting Old Church Slavonic characters;

(3) a computerized dictionary of Church Slavonic, preferably available on CD-ROM;

(4) text printing and text analysis software; and

(5) CD-ROM production of early Slavic documents which permits perusal and analysis of texts without requiring access to printed documents.

БАН 46
БИБЛИОТЕКА

Session: IMAGE PROCESSING

A CHURCH SLAVONIC ALPHABET FOR REPRINTING OLD MANUSCRIPTS USING A MICROCOMPUTER

Mimoza Majstoroska

Skopje

Our presentation reports on the development of a Church Slavonic alphabet on the basis of a geometrical analysis of the mathematical principles of each letterform. As a separate project, these same principles were then applied to the development of an experimental optical character recognition system (written in C) for Church Slavonic.

FONTS, CHARACTER SETS, AND CHURCH SLAVONIC GRAPHEMES

Ljupčo Mitreski

Skopje

Because the Church Slavonic alphabet contains a greater number of graphemes than can be represented easily on a standard keyboard, special systems must be developed to input early Slavic texts into a computer.

The first task is to unify the positions of signs on keyboard for the Church Slavonic alphabet, according to the Church Slavonic normal alphabet, the Church Slavonic capital letters (on the same top or base line with normal alphabet), supralinear signs, supralinear graphemes, signs over supralinear graphemes, and ligatures.

Our final inventory is based on three Church Slavonic fonts. A basic font includes the Church Slavonic alphabet and supralinear signs on the keyboard, with supralinear graphemes, signs over supralinear graphemes, and some commonly used ligatures as alterations. A second font includes Church Slavonic capital letters based in the same top line as the normal alphabet and supralinear signs on the keyboard, with supralinear graphemes, signs over supralinear graphemes, and some commonly used upper-case ligatures as alterations. A third font includes Church Slavonic capital letters based on the same base line as the normal alphabet and supralinear signs on the keyboard, with supralinear graphemes, signs over supralinear graphemes, and some commonly used ligatures as alterations.

A SYSTEM FOR ENCODING EUCHOLOGICAL (MAINLY NON-CALENDAR) TEXTS WITHIN THE PARAMETERS OF THE IST COMPUTER PROGRAM

Maria Šnitter

Plovdiv

The present work discusses a system for encoding texts of the Slavic Euchologion with a view to their inclusion in a compete catalogue of all known early Slavic texts. For that purpose a classification of the contents of the collection was made, which led to the following preliminary conditions:

1) all possible texts contained in the Euchologion should have a numeric equivalent within the framework of four signs (figures), with the eventual possibility of adding an alphabetic coefficient; and

2) the coding should reflect not only the inner subordination of the euchological texts, but also their correlation with non-euchological (calendar or non-calendar) bodies of texts.

The system in question refers only to the description of the texts, and does not concern the codicological part of the IST program. A basic problem at the persent stage of work is the achievement of an optimum degree of abstraction in coding, which would allow the precise identification of every text. This has been achieved to different degrees with the texts of the sacraments, where each sacrament is considered a united macrotext, regardless of the number and type of its components, and with the texts of prayers за всяка потреба (for every occasion), where a separate code number corresponds to each separate text of the euchemic type. This clasification coding is a result of the analysis of the composition of eighty-five mostly South Slavic records, dating from the eleventh through the eighteenth centuries.

SOME SUGGESTIONS ARISING FROM A PERSONAL COMPUTER ANALYSIS OF THE SVETOSTEFANSKA XRISOVULJA

Francis R. Salter

Ontario

In 1969 the Svetostefanska xrisovulja in the Topkapi Museum was photographed by F. Salter and Ali Galip Erdican. An inital attempt at computer analysis of the document on a VAX was found to be too time-consuming and program-intensive. With the advent of the personal computer and DOS programming, the document was transcribed into a DBF database using a Foxplus program and, with the use of a Lettrix Church Slavonic font and several specially created characters, it was possible to print a copy of the document.

The presentation discusses the problems of programming and database structures required to produce the document and offers suggestions as to future development.

The database and program accomplished the following:

(1) printed a complete text, following the page and line numbers of the original;

(2) produced a complete listing of the document's words, with page and line number cited for each occurrence;

(3) produced a word-reversed listing of the document, giving all citations by word endings, with page and line number;

(4) produced a frequency listing of all words in normal order of spelling, and

(5) produced a frequency listing of all words in reversed order of spelling.

Because the entire document is in DBF format, other analytical tools may be applied, using dbase, Foxplus, Foxpro, SQL and any other program that recognizes the DBF format.

The work described above was completed in the early stages of development of computer support for Church Slavonic writing, but at the

HOW SLAVIC PHILOLOGISTS SHOULD USE COMPUTERS

David J. Birnbaum

Pittsburgh

Surely the most widespread computer application for Slavic philologists, and for most scholars in the humanities, is word processing, which provides easier revision control and better font support than is available on typewriters. But while revision control and font support are unquestionablely valuable resources, they represent but a small fraction of the research advantages available to Slavic philologists through the use of computers. This is clear from even a quick perusal of the program for this conference, where almost every paper addresses some application of computers to Slavic philological research other than word processing.

The present report discusses these applications from the perspectives of multiple use, portability, and preservation. I argue that texts created with attention to these three issues are a much greater contribution to science than texts created in a proprietary fashion by small groups of users working in isolation. In fact, texts created without attention to these issues are guaranteed restricted use, a limited audience, and a short lifespan. I also discuss the role of cooperative standardization in achieving the three goals listed above, and in thus making our research more accessible to colleagues both today and in the future.

Session: ENCODING PROBLEMS

INFORMATIONAL AND PRESENTATION UNITS IN EARLY CYRILLIC WRITING

David J. Birnbaum

Pittsburgh

One of the most acute problems facing Slavic philologists who use computers is the question of how to encode the signs of early Slavic writing as computational bit combinations. One traditional approach among Slavists with a limited knowledge of information science has been simply to define new fonts, but this oversimplifies the nature of writing systems, where information scientists traditionally distinguish *characters* (informational units) from *glyphs* (presentational units). These two types of unit do not observe a one-to-one relationship, and Slavic philologists need to make informed choices about whether to encode character-level, glyph-level, or both types of information in electronic texts, and also about how to implement complex or multistratal encoding.

A further complication is that early Cyrillic is not a single orthographic system, and although structural analysis might identify the distinctive written signs in one early Slavic manuscript, it is incapable of addressing the fact that two units may be opposed to each other in one manuscript, but not in another. If Slavic philologists wish to conduct electronic comparisons among manuscripts, a system is required to adjust the texts in light of local orthographic and linguistic facts.

TRANSCRIPTION RULES FOR OLD CHURCH SLAVONIC WRITING

Karsten Grünberg

Heidelberg

One of the most important tasks for Paleoslavistics in the years to come will be to set up databases for Church Slavonic linguistic monuments and for fast access to the information they contain. A database for the study of Church Slavonic manuscripts should cover three essential areas: cataloguing information, facsimile reproductions, and diplomatic transcriptions. The latter can be used for textual comparison across manuscripts, indexing, concording (KWIC), statistical analysis of lexical or grammatical information, etc.

For the production of text files, standardized and generally accepted transliteration and transcription rules are required. These rules must be such that all relevant information within a manuscript can be encoded unambiguously; only under these circumstances will it be possible to use computer programs for automated analysis or comparison. The present report offers for consideration and discussion a suggested set of general rules for the transcription and transliteration of Church Slavonic manuscript material.

GRAPHIC OR GRAPHEMIC ENCODING OF MEDIEVAL SLAVIC MANUSCRIPTS

Cvetana Krstev, Duško Vitas, Henning Moerk, Gordana Pavlović-Lazetić

Belgrade, Aarhus

International informatics standards define a text as "information intended for human communication which (i.e. information) can be represented in two-dimensional form and consists of characters..." (e.g., ISO 7350-1984). The character itself is defined as a "member of a set of elements being used for organizing, controling and representing information" (ISO 2022, ISO 10646) or as an "atom of information with individual meaning and defined as a repertoire of characters" (ISO 8879).

From these definitions of text and characters, the following facts can be derived:

(a) a character is a means of representing information and not information itself;

(b) internally in the computer, a character has no inherent meaning.

Characters are assigned to graphs to make data or information *visible*, or *describable in two-dimensional form*. The reader of a *text in two-dimensional form* is able to interpret information, but the characters themselves do not contain information about what they represent in the computer.

For example, when encoding graphs as characters we talk about code tables as *Cyrillic* or *Latin*, independent of the functions of these graphs in specific languages. In the existing code tables Russian, Serbian or Bulgarian *A*'s are represented within the computer by the same character. The reader of a text where both Russian and Serbian and Bulgarian *A*'s are present is able to establish the difference between them, while in the internal representation there is no such information.

We think that for these reasons it is necessary to distinguish two types of texts. One type of texts includes texts meant for direct interhuman communication and where a *visible* text form is essential. The other type of texts includes those texts which are encoded not only to be *visible*, but also to be processed by computer.

This second type of texts involves orthographical norm, which can be subject to changes within a specific language, and can differ considerably between languages using the 'same' set of graphs. Furthermore, there can be differences in a written text that are not reproduced in speech. One example could be the obligatory use of capital letters. In Serbian words like *Tuzap* (the river) and *muzap* (the animal), the initial character is essentially different. In the first case it is obligatorily to write it as an upper case *T* (which cannot be replaced by a lower case *t*), while in the second case it is only realized as upper case *T* in the beginning of a sentence (after full stop).

These phenomena can be seen in examples from old cyrillic documents (see for example P. Ivić and B. Jerković: *Pravopis srpskohrvatskih cirilskih povelja i pisama XI i XII veka*, Filozofski Fakultet, Novi Sad

1981), where it has been shown that the same sound or group of sounds under certain circumstances can be written in several different ways (for example the groups *ja* or *je* or the notation of the palatal *l* consonant [*lj*]). Similarly with the use of apostrophe (e.g. for marking omitted vowels), contraction or suspension (of words), ambiguous use of tilde, use of letter characters for numerals, etc. This raises the question of whether it is possible to encode these old manuscripts in such a way (a) that the *visible* text form preserves the characteristics of the original manuscript, and (b) that the internal representation of the text in the computer preserves information that enables each individual feature of a certain graph to be interpreted in an appropriate way in *visible* form. In an approach like this, a simple one-to-one *character-graph* mapping is no longer sufficient. A solution of the problem of recording a text in this manner implies that the entity which is internally encoded in the computer should be provided with a set of attributes that enable a description of different interpretations of the encoded entities.

In the paper two such solutions will be proposed. One approach is based on a description of variations by means of regular expressions. The second approach is a definition of the abstract entity and a list of its features interpreted as different possible graphs.

IMPLEMENTING A MEDIEVAL SCRIPT ON A PERSONAL COMPUTER

Sebastian Kempgen
Bamberg

Based on his year-long experience in designing an Old Church Slavonic font, the author will outline the technical possibilities, limitations, and perspectives of realizing a complex non standard medieval alphabetic script system on contemporary personal computers. Apple Macintosh is used as a reference platform. The report will cover such problems as: how to decide on an encoding scheme, how to enable users to exchange data across different systems and platforms, how to implement superscripts and character variants, design issues concerning the appearance of characters themselves, etc.

Session and Workshop: MANUSCRIPT COLLATION

TRANSCRIPTION AND EDITION

William R. Veder

Amsterdam

Transcription of Old Slavonic poses two problems which as yet have no practical solution:

(1) reduction of graphs to graphemes, and
(2) portability of machine readable manuscripts and texts.

Both have a direct bearing on editions, which must be viewed as collective, not individual, enterprises. At this level, they are complicated by another problem:

(3) orthography of texts (as distinct from manuscript witnesses).

This presentation reviews several mutually exclusive solutions to these problems:

(1) tagging and commenting vs. reference to facsimile,

(2) use of programmable converters vs. standardization of text encoding,

(3) eclecticism vs. normalization.

COMPUTER COLLATION OF MANUSCRIPT TRANSCRIPTIONS

Michael Bakker

Amsterdam

This presentation demonstrates and discusses the automatic collation of computer transcriptions of manuscripts. Through data abstraction according to specified rules, significant variant readings are isolated from the huge amount of spelling variants, so that a text that is inde-

pendent of the manuscripts can be constructed.

I concentrate on how the computerized manuscript transcriptions are created in practice and on the requirements they must meet. In addition, I discuss strategies for error-checking and economizing on time.

ADVANCED SEARCHING

Tanja Ivanova, Nina Šoileva, Irina Bolčeva
Sofia

The authors report on their use of Advanced Searching version 1.0 for the comparison of witnesses of О писменехъ by Cernorizec Xrabr. Advanced Searching is a Microsoft Windows application that works with Collate-compatible files, and that distinguishes paragraph-level and sub-paragraph-level textual divisions for comparison. After preparing a collation of the entire stemma, *misreadings* are registered within manuscript groups. The manuscripts of each group are then recollated, whereupon a general collation is again produced. The system supports regrouping within the previous stemma.

The product is created by Plamen Dimitrov--a student at the Technical University (Sofia).

Session: TEXT PROCESSING: PROBLEMS, METHODS, AND TOOLS

THE APPLICATION OF *TEXT ENCODING INITIATIVE (TEI)* GUIDELINES TO THE ENCODING A 14TH-CENTURY SERBIAN CHURCH SLAVONIC PSALTER

Mary MacRobert

Oxford

This manuscript, which I have described briefly in Oxford Slavonic Papers 1989, is remarkable because it has been systematically corrected to a new version of the text, but in such a way that the original version can usually be read or inferred on the basis of comparative evidence. With the intention of producing a parallel edition of both versions, I have turned the whole text into machine-readable form, where the main areas of difficulty, and therefore of interest, have been: transcription of diacritics; capture of abbreviations and variant spellings in indexing; treatment of multiple overlapping structures (the textual organization of the Psalter conventionally into psalms and verses *for cross-reference* and traditionally into kathismata, staseis, psalms, and versicles *for use*, as opposed to the physical organization of the manuscript into gatherings, folios, pages and lines); various ways of indicating different types of correction and more or less conjectural original readings; the relative merits of combining all points of interest in a single, elaborately tagged machine-readable version or of producing several parallel versions, each reflecting a different facet of the information.

О создании фонда письменных источников русского языка XI-XII вв. в Институте русского языка РАН

Александр Молдован
Москва

Лингвистическое источниковедение в Институте русского языка РАН традиционно было ориентировано на введение в научный оборот письменных источников путем их научного издания с палеографическим аппаратом и указателем слов и форм. Условия академического института позволяли сосредоточить усилия коллективов ученных на издании крупных рукописных памятников, наиболее показательных в лингвистическом отношении и имеющих общекультурное значение (Синайский патерик, Изборник 1076 г., Успенский сборник XII-XIII вв., Выголексинский сборник, Мстиславово евангелие и др.), что способствовало определенному продвижению в исторических науках.

И все же в теоретическом отношении такая работа скорее лишь фиксировала существующий уровень знаний о рукописи, чем определяла перспективу ее изучения. Отсутствие организации планомерной работы с источниками и традиционное печатное представление результатов их изучения не позволяло сохранить единство исторической проблематики, взаимосвязанность ее элементов. Итогом этого стала разобщенность работ в области исторического языкознания, при которой дублируются наиболее трудоемкие и дорогостоящие операции, связанные с созданием вручную рабочих картотек по одним и тем же письменным памятникам (например, картотеки "Словаря древнерусского языка XI-XIV вв.", картотеки по исторической грамматике и историческому синтаксису и др.). Само нынешнее технологическое несовершенство работы по изданию рукописных текстов существенно замедляло ее и затрудняло публикацию. С 1994 года в Институте начата работа по созданию компьютерного фонда письменных источников русского языка XI-XVII вв. В качестве основных задач определены:

-- разработка методики компьютерной подготовки,

систематизации и хранения текстовых данных на материале наиболее крупных неизданных текстов; изучение возможностей информационного использования текстов в их машинной форме;

-- установление связей с другими научными центрами и отдельными специалистами; унификация технических приемов, определение порядка обмена готовыми материалами.

Предполагается, что работа по созданию текстового фонда будет осуществляться по двум направлениям:

1. Введение в компьютер древнейших списков с соблюдением всех лингвистически релевантных особенностей графики и орфографии..

2. Создание банка текстов, сохраняющихся только в более поздних списках, на основе их текстологического исследования, с разночтениями основных редакций.

Источники по истории русского языка являются частью единой письменности православных народов. Поэтому работа по русским текстам и спискам должна быть интегрирована в общую международную программу переноса славянских рукописных текстов на машинные носители. В итоге должна быть создана единая информационная база для исторических и историко-филологических исследований.

Session: MULTIMEDIA

CD-ROM AND THE MANUSCRIPT HERITAGE

Borjana Xristova
Sofia

CD-ROM editions of original written books and texts facilitate the presentation of a large mass of manuscripts works, with indexing that supports access to these materials according to various principles (genre, date, school, region, etc.). Such editions make high-quality facsimiles available to scholars who are not able to work directly with original sources. The present report includes a demonstration of the first of the Ancient and Multi-Media series of disks of the Bulgarian National Library, 'Unique Balkan Manuscripts', subtitled 'Bulgarian Written Culture against the Backdrop of the Balkans'. The disk contains examples not only of Slavic manuscripts (Bulgarian, Serbian, Wallachian, and Moldavian), but also of Byzantine and Oriental materials, all dating from ten. to seventeenth centuries.

THE PROJECT 'ST. SOFIA'

V. Bratanova, E. Koceva, D. Petrov
Sofia

The project 'St. Sofia' was created by Bulgarian specialists Velina Bratanova - a historian of arts (art critic), Elena Koceva - a medievalist and philologist and Dragomir Petrtov - a writer and ethnologist. The first stage was completed and UNESCO edited a proto-video CD. Prof. Alain Vuillemin, a specialist in comparative theory of literature and computers from the University of Arl, France, joined in the edition of the CD as a co-author from the French side. Bulgarian specialists who participated in the project are: Jordanka Jurukova, Margarita Vaklinova, Dora Kamenova, Elisaveta Musakova, Elena Tončeva, Ganka Mixajlova, Xristo Genčev, Lyuben Botušarov, Vergilii Atanasov; French participants are: J.-P. Levet and Marie Vrinat.

The intention of the screenplay is to describe, to analyse and to compare written texts and cultural objects on synchronous and diachronous approach. The screenplay's aim is to experiment the presentation of the Bulgarian culture as a result from the passage: *no-written culture - written culture*, and from the historical periodization of this process. This can be done on interdisciplinary way. The choice of the settlement *(Serdica, Ulpia Serdica, Trajana, Triadica, Sofia)* is well-grounded in the paper with some data about the geographical location, the natural sources, archaeology, ethnography, etc. "The sector Sofia" can be treated as a possible model of the Bulgarian cultural space. The authors of the project consider this *sector* as located on the vertical axis of the co-ordinates of the model.

2000 objects were photographed for the proto-disk. 1000 from them are prepared from Pierre Haur and Jules Lucqas, engineers from the "Tribune". These photos are from the manuscript collection of the National Library "St. St. Cyrill and Methodius". The rest of the photos were provided by Bulgarian photographers, sponsored by Open Society, Sofia.

The first stage of the "St. Sofia" project shows the potential from the comparison of different material: manuscripts, pictures, musical records. It was presented at the proto-disc, which was internationally recognized, as well as the program of the engineering group of the "Tribune", Paris and EPI.

The project compouses over 600 pages. The patterns of describing the presented objects are prepared from EPI specially for the "St. Sofia".

The paper will be accompanied by the demonstration the proto-disk and with technical comment by Vasil Vasilev, an engineer at the Department of Computer Science of the New Bulgarian University.

Session: APPLICATIONS OF DATABASE MANAGEMENT SYSTEMS (DBMS)

A DATABASE FOR ANCIENT SLAVIC MANUSCRIPTS IN MACEDONIA

Stana Jankoska, Dragan Mixajlov, Ljupčo Josifoski

Skopje

The present report discusses an experimental on-line bibliographic and factographic database for early Slavic manuscripts discovered in the Republic of Macedonia. In addition to ordinary bibliographic data, this system also includes illustrations of the manuscripts. The database structure is based on the current practice for ancient manuscripts description, ISBD(A) (International Standard for Bibliographic Description for Ancient) and UNIMARC (Universal Machine Readable Catalogue) standards. Each record in the database contains cultural-historical, literary-historical and philological identification elements for each manuscript. The database will be installed on a BBS as a system for free retrieval of data, and it will later also be made available over the Internet and on CD-ROM.

COMPUTER PROCESSING OF OLD CHURCH SLAVONIC MANUSCRIPTS: RESULTS AND PROSPECTS

Zdenka Ribarova, Kirill Ribarov

Skopje, Prague

This report addresses the analysis of the vocabulary of Old Church Slavonic manuscripts and its integration into a dictionary, with particular attention to specific problems of standardization, lemmatization, and semantic analysis.

The first version of the program we have developed (STINO version 1.0) was designed to meet the basic requirements for representing Old Church Slavonic documents (OCSD) for use in lexicological analysis. In particular, STINO was used to input and process much of the Prague Old Church Slavonic dictionary. Encoding according to strict and consistent standards supports document reuse, so that the same files can be employed for different processes and different types of analysis.

The current version of STINO supports the reproduction of OCSD in their original form for reprinting; the creation of different alphabetic indices; the creation of subsets of the normalized words according to user-specified criteria; and database searching according to normalized word, word form, or position in the text. A revised version of STINO, currently under development, will provide a richer menu and a revised user interface.

PROBLEMS INVOLVED IN SETTING UP A UNIFIED COMPREHENSIVE CATALOGUE OF SLAVIC TEXTS

Reinhard O. Lampe
Heidelberg

Western European culture is mainly influenced by Roman (Latin) culture, and later by Catholic-Christian culture. As a result of this common background, all Western European languages used and still use for writing basically the same characters the Romans used two thousand years ago.

Eastern European, i.e., Orthodox Christian culture, however, was influenced by Greek-Byzantine traditions. One consequence of this influence was the development of a special 'typology' of Eastern European religious literature, which cannot be categorized using Western European notions, creating a problem in designing a Unified Catalogue.

More important, however, is the development of Eastern European writing, which replaced a common Greek writing system first with Glagolitic, which consisted of some forty characters. Although

Glagolitic remained in use in some areas until the late fifteenth century, after a relatively short time it was largely replaced elsewhere by the Cyrillic alphabet, consisting of sixty-four base characters (plus local variants), in many cases reflecting the beginning of linguistic shifts that led to today's different Slavic languages. Such languages as Polish and Czech, which use Latin characters with special diacritics, constitute a special case.

Contrary to publications in most fields of natural sciences, where English has become generally accepted as common language, scientific publications in Slavistics, such as descriptions of Slavic manuscripts, are usually published in the author's native language, which means that unified catalogues and descriptions of texts and manuscripts may include information in several languages. This multilingual tradition is well established for printed material, but it is not suitable for computers, which internally use figures and numbers to store, retrieve and process data, and require that data be unambiguous, so that searching and extracting information can be comprehensive and unambiguous.

A computer catalogue must thus elaborate a set of characters to use for the entry and display of information, and an unambiguous transliteration system to represent the wording of Slavic texts, including 'subscripts' (used for abbreviation purposes). The problem remains the same for manuscript descriptions, although it can be simplified by the use of methods that do not require the user to input text directly, such as selecting from a menu of choices (e.g., Scripture, усмаß), a catalogue of pictures (e.g., watermarks), or other graphic presentations (e.g., page layouts, разлиновка).

The main concern should be to exclude as far as possible situations where the computer will treat identical items of information as different because of differences in the computer's internal encoding. Even Unicode, a system designed to allow nearly any language (alphabet) in the world to coexist in a computer simultaneously with any other language (alphabet), would not present a solution to this problem, since different characters will still have their own unique computer codes.

Международный исследовательский проект "Slavia Orthodoxa et Slavia Romana": Взаимодействие славянских миров: Духовная культура Подляшья

Ю.А. Лабынцев
Москва

Настоящий проект позволит на примере обширного региона в достаточной полноте проследить почти тысячелетнюю историю взаимодействия западного и восточного христианских миров, раскрыть его сущность и механизмы, характерные для всей Центрально-Восточной Европы, создать уникальную источниковую базу - основу для дальнейших теоретических исследований.

В результате над проектом наряду с теоретическими исследованиями предполагается создать оригинальное многотомное издание энциклопедического типа, включающее в себя практически полную информацию обо всех сторонах культурного процесса на Подляшье, связанного с рукописной и печатной книжностью на церковнославянском, древнерусском, старобелорусском, староукраинском, белорусском, украинском, русском, польском, латинском и литовском языках. Эта информационно-аналитическая часть проекта имеет самостоятельное значение, работа над ней осуществляется по специальной Международной славяноведческой программе "История книжной культуры Подляшья".

Согласно принятой методике все работы ведутся с использованием компьютера. При этом гарантируется не только создание общего банка данных, но и его надежное дублирование для издательского комплекса проекта в Минске и головного в Москве. Подобное решение позволяет также оперативно вводить в оборот любую информацию, поступающую от соисполнителей из Польши, Белоруссии, Литвы, России, с Украины и других мест.

Основные задачи и принципы построения банка данных Международной славяноведческой программы "История книжной культуры Подляшья"

Л.Л. Щавинская

Москва

Основой исследования станет не имеющая в своей совокупности принципиальных аналогов источниковая база Программы, состоящаяся из многих тысяч разнообразных источников (рукописи и старопечатные книги, акты, грамоты, деловые документы и др., хранящиеся в различных странах), организованная как особая информационная система, включающая в себя разные типы баз данных.

При создании нашей информационной системы мы предполагаем строить базу данных на основе реляционной модели системы управления базами данных (СУБД). Предполагается использование СУБД с достаточно высоким быстродействием типа PARADOX, что весьма важно, учитывая предполагаемый объем баз данных.

В информационной системе для создания баз данных по таким источникам, как акты, документы, грамоты, предполагается использование источниково-ориентированной СУБД типа KLEIO, разработанной Манфредом Таллером (Германия).

Кроме реляционной и источнико-ориентированной СУБД предполагается использовать пакет статистической обработки для проведения разного рода исследований.

Таким образом, разрабатываемая информационная система будет представлять собой интегральную систему историко-культурной информации.

AN AUTOMATED SYSTEM FOR PALEOGRAPHIC AND ANALYTICAL DESCRIPTION OF SLAVIC MANUSCRIPTS

Vasja Velinova

Sofia

The paper presents a long-term program being undertaken at CIBAL for automating the description of Slavic sources on the cultural history of the Balkans. The first part of the presentation addresses theoretical problems involving the unification of terminology and orthographic characteristics. Without a solution to these preliminary considerations, practical use of the program is impossible. The second part includes a concrete technical description of the CD/ISIS-micro program, and addresses the possibilities for searching under varying criteria and combinations of indices. Perspectives for enlarging the chronological scope of the material and preparing the narrative sources for CD-ROM distribution are also addressed.

Session: KLEIO

A RUSSIAN VERSION OF THE SOURCE-ORIENTED SOFTWARE *KLEIO:* NEW OPPORTUNITIES FOR THE TREATMENT OF CYRILLIC SOURCES

V. Tjaźelnikova

Moscow

This report summarizes the results of a two-year joint Russian-German project that address special features of the Russian version of KLEIO. On the basis of this work, the authors suggest that the KLEIO system is the optimal tool for realizing original documents in a computer framework.

KLEIO is unique in its ability to generate a model based on the contextual analysis of a document, including both the structure and the semantics of the information. In particular, KLEIO supports the development not only of a model of the data, but also of a model of the source, i.e., the data together with its infrastructure.

The importance of such an approach is much enhanced when treating materials with a complicated infrastructure, weak formalization, a free text character of presentation, complicated relations among persons, or recursive contextual elements.

INTERACTIVE FULL-TEXT PROCESSING IN *KLEIO*: SCANNING CRITERIA OF BASIC TERMS (ON APPEALS BY DISENFRANCHISED PEOPLE)

V. Tixonov

Moscow

This report discusses the use of KLEIO for content-analysis of a corpus of appeals by disfranchised people -- лишенцы), who had appealed

БАН 45
БИБЛИОТЕКА

for a restoration of their voting rights in Soviet Russia in the 1920-1930s. As the appeals reflect the same historical phenomenon (disenfranchisement and the restoration of rights), we hypothesized that they imply a set of *basic* (for this phenomemom) words or terms, which li&shachek;ency could not help using in their appeals. The paper suggests a possible approach to developing scanning criteria for such basic terms and the ensuing analysis of their interdependencies.

The research includes following stages:

1) creating a catalogue of all appeals terms for interactive text processing with KLEIO;

2) generating a special lexicon for these materials, removing all syntactic words (prepositions, conjunctions, etc.), as well as less frequently used terms, from the catalogue;

3) selecting basic terms through an examination of function distribution on the basis of frequency analysis, as well as estimation of critical regions in which the basic terms lie;

4) creating a table of basic terms and frequencies in the context of fifteen- to twenty-word sentences;

5) analyzing interdependencies, semantic features, and the content of appeals with respect to the social status of the petitioners.

The suggested approach minimizes the need for expert evidence in content-analysis, thereby facilitating more reliable results.

MODEL OF HISTORICAL SOURCES IN KLEIO AS A BASIS FOR INTEGRATED SOCIAL CLASSIFICATION

Igor Jušin

Moscow

Social classification is one of the most fascinating, and, at the same time, the most difficult problems for historians. A major component of this kind of research is the coding of attributes, which are included in a classification as bulk elements.

This process always presents significant difficulty for researchers working with traditional packages because coding principles should properly be developed before entering data into a database.

Thus, there are two types of problems:

1) Information in a source is lost during data entry because: it is impossible to take into account all nuances of personal biography; difficulties are inevitable with inconsistant data and with the lack of contextual dependences; to change existing codes is impossible because of the loss of primary information pertaining to a source;

2) The formation of a knowledge base is separated from a source, so that codes are developed independently of a particular source. The knowledge base cannot be completely independent of the database for the following reasons: there are no universal methods of social classification; important source information may be available only by taking context into account; temporary aspects cannot be excluded; models of coding that are based on varying criteria may be needed; an integrated classification can be designed only on the basis of a source in general. Initially, all information on social layers must be taken into account. Only then can a coding model be developed, since the latter depends on the availability of information pertaining to specific criteria (e.g., occupation, income, level of training etc.).

These problems are most appropriately addressed from a source-oriented perspective. The KLEIO system is based on proper consideration of the structural and semantic context of a source. Only this type of approach permits the researcher to find the necessary keys to processing and interpretating original information as contained in sources. Within such a model information is not subject to distortion during data input; data context is taken into account at the processing level, so that data coding is closely coordinated with its context; and models of processing based on temporary parameters can be developed, which makes it possible to formulate more potentially correct hypotheses, and to speed up the process of selecting among them.

COMPUTER INVESTIGATION OF MEDIEVAL PROGNOSTIC BOOKS BY MEANS OF *KLEIO*

Adelina Anguševa

Sofia

This presentation addresses problems of computer-based analysis of Prognostic books, such as Kalendologia, in Byzantine, Latin, and early Slavic written culture. The investigation aims to compare Kalendologia in these three medieval traditions with the help of the Kleio computer program, which provides for classification of the texts, a comparison between the types and versions of one and the same book, as well as the texts of different books, and analysis of their typology. The exact picture of predictions, variety of redactions, changes in the texts, and textological differences is determined computationally. Resulting reports contain: the name of the book, name/number of the manuscript, dating of manuscript, language, place, library where it is housed, and inventory of prognoses. The predictions are divided into four main groups: (1) human beings, (2) seasons, (3) crops, (4) others. The whole texts of the Prognostic books are computerized.

GREEK AND LATIN LOADWORDS IN THE NEW YORK MISSAL AND THE *KLEIO* COMPUTER PROGRAM

Margaret Dimitrova

Sofia

The paper discusses the use of the computer program KLEIO for investigating lexical, orthographic, and morphological problems of the New York Missal, a missale plenum of Croatian origin, dating from the first half of the fifteenth century.The computer program in question can generate an index of all the loanwords attested in the manuscript and can produce lists of all the peculiarities of phonetic and morphological adaptation of the loanwords under investigation.

SESSION AND WORKSHOP:
Problems in Running a Concrete Project:
Computer Supported Processing of Medieval Slavic Manuscripts

COMPUTER AIDED ANALYSIS OF THE MACROSTRUCTURE AND TYPOLOGY OF MEDIEVAL SLAVIC MISCELLANIES

Anisava Miltenova

Sofia

This paper presents the results of research into the content analysis of structure models of different types of medieval miscellanies, carried out by a team consisting of D. Birnbaum and A. Miltenova (compilation of template data and criteria) jointly with M. Dobreva and B. Dijk (Author/Editor / SGML facility application).

This presentation addresses problems of data entry according to typology of manuscript (menologia, liturgical manuscripts, mixed content miscellanies, etc.), with attention to selected features (options) for computer description, classification schemes, the universality of the template, terminology, key words, and related issues.

Particular attention is devoted to mixed-content miscellanies, which are the source of particularly complex descriptive problems. The presentation offers a definition of the interrelated concepts of *structural model, core,* and *background*. The specific problems of the co-existence of the separate texts (cycles and segmentation), as well as aspects of their processing, are discussed and examined, with suggestions for applications. The classification scheme for the texts included in miscellanies and their structure is also examined.

Finally, the presentation addresses problems of use cases, and particularly the implementation of searching according to multiple criteria and features. The ultimate goal of this research is improved access to the comparative study of miscellanies, and the presentation explains how this is facilitated through use of the adopted template.

PALEOGRAPHIC AND ORTHOGRAPHIC FEATURES IN THE DESCRIPTION OF SLAVIC MANUSCRIPTS

Andrej Bojadžiev
Sofia

Establishing a set of well-organized rules in the field of Slavic codicology, paleography, orthography, and linguistics for computer descriptions of medieval manuscripts and texts has long been a complex problem. Manuscripts from different times and places may impose different encoding requirements, and Slavists are far from agreement about how to manage these differences.

The paper addresses the presentation of obligatory and optional features on two levels: discussion of their place in the system and techniques for their encoding and presentation in electronic form. The observations are made mainly on the basis of Document Type Definition files generated on the principles of Standard Generalized Mark-up Language and the guidelines of the Text Encoding Initiative.

Session: CASE STUDIES. TYPES OF MANUSCRIPTS

THE PSALTER: FROM THE ORAL TRADITION TO THE INFORMATIC ONE

Monia Camuglia

Pisa

This paper describes my experience concerning the use of computer systems for the study of early Slavic texts. The project described in my report involves attestations of the unitariness of the Psalter in Slavic countries, based on a comparison of fifteen psalms in seven different versions, from different periods and locations, and the relevant lexical variants.

The aim of this project is to apply to early Slavic materials text-processing methods originally developed for use on texts in the Latin alphabet. For this purpose we used DBT (Textual Data Base), running on PC MS-DOS and Windows, developed at the Institute for Computational Linguistics to handle texts and textual corpora with particular regard to philological issues. The DBT originally used for Latin alphabets was modified to handle complex non-Latin alphabets.

The following issues emerged:

1) Within MS-DOS and Microsoft Windows, it is possible to realize and define a complete set of characters using a TrueType font. The same font was used for both screen display and printing.

2) Alphabet, codification, and information necessary for correct alphabetic ordering of the words in the analized texts were defined, in parametric form, using a data archive which was not part of the system. This archive, containing the encoded alphabet, makes it possible to define the different elements that contribute to the composition of the text and the identification of words (central characters, superpositions, punctuation, abbreviations, apices, etc.). Information allowing correct ordering was also inserted in the archive.

The DBT makes it possible to define and to handle central characters and their value for ordering, accents modifying the final ordering at a parity of central characters, and philological signs representing the final ordering level. The DBT is able to produce a number of printed indexes and results, including indices of alphabetical frequencies, decreasing frequency, index locorum, concordances, cooccurrences, and a printout of the corrected alphabet. However, the most interesting tool is the *query system*, a program for interactive on-line interrogation of the corpus of the different Psalter versions, which makes it possible to analyze and navigate the text.

Our aim is to complete the corpus and to check and standardize the codification used so far. We also intend to perform a contrastive analysis of the different versions of the Psalter, so far used in prototype form, to compare, as suggested by the computer, different versions of two texts using a tool now being implemented, and to be able, in the "collatio□ stage proper, to identify any graphical variants of the words. The derivational rules of the possible variants will be described, according to a suitably defined syntax, in an archive that is not part of the program, to be enriched or modified recursively.

MEDIEVAL SLAVIC MENOLOGIES ON LINE

Cynthia Vakareliyska
Eugene, Oregon

A project to develop a computer collation of Slavic menologies, or calendars of saints, from the tenth through fourteenth centuries was initiated because while there is immense variety among menology manuscripts in the listings of saints for specific days, there is no collation currently available which enables comparison of individual calendar listings from a significant number of manuscripts. Such a collation will give for the first time an overall view of the various medieval Slavic menology traditions. Moreover, since different saints were venerated at different times and in different regions of a single country, so that even very closely related manuscripts contain discrepancies in their menologies, the menology contained in an ecclesiastical manuscript can be a source of valuable information for dating that manuscript and identifying its geographic origin, thus providing a firm chro-

nological and dialect framework for analyzing the linguistic and textual features of the entire manuscript. Since the vast majority of extant medieval Slavic menologies are unpublished, an on-line collation will provide Slavists, Byzantinists and church historians with access to a large corpus of data without requiring travel to archives in Eastern Europe and the arduous task of collecting and copying entries from individual manuscripts.

The project is in the form of a master calendar, which lists under each calendar day all variants existing in the corpus in the names of saints and textual formulae for each day. Eventually the collation will also include textual variants in the Gospel lections for each day in the individual manuscripts. The collation currently includes entries from over seventy-five, primarily unpublished, manuscripts from archives in Bulgaria, Russia and England, and from the microfilm collection of Hilander Center for Medieval Slavic Studies at the Ohio State University. It is envisioned that the number of manuscripts in the corpus will expand as individual researchers in Bulgaria and other countries contribute further texts to the collation.

Major issues involved in the establishment of the on-line collation of menologies include the following: transcription norms for paleographic and orthographic variants; creation of a program that will effectively search for and identify individual saints' names over the calendar year despite variations in spelling and in case-endings among individual manuscripts; creation of a program that can identify relationships between individual manuscripts, and families of manuscripts, on the basis of specific shared lexical and textual features; and supervision of the collation to ensure that further textual contributions by individual researchers will be consistent with the transcription norms established for the project.

COMPUTATIONAL ANALYSIS OF HYMNOGRAPHIC COMPENDIA

Emilia Guergova

Sofia

Hymnographic compendia form a substantial part of the genre diversity of medieval books of Slavia Orthodoxa because of the high frequency of usage assigned to them by ritual practice. Unfortunately,

the level of investigation of these sources is far lower than is the case with other types of medieval books, primarily because of the nature of hymnography itself, which, in contrast to narrative texts, is characterized by a vast variety of structures and a high mobility of contents.

This challenges the investigator's abilities to oversee and process the extreme wealth of data and to extract from them objective conclusions about the history of their texts.

The present project focuses on the Menaia, and specifically on the Menaion for the month of October. Its aim is to study in diachronic perspective the macrostructure, i.e., the dislocation of the acolouthia in the order of the ecclesiastical calendar; the structure of the acolouthia themselves, i.e., the configuration of their mesostructural elements; and the microstructure of these elements (troparion, kontakion, stichera, oikos) from a rhythmic point of view.

The project requires the aid of a computer and appropriate software to ensure adequate treatment of the the following: comparison and synthesis of data on four levels (Saint, ritual occasion, mesostructural genre, and the rhythmico-intonational and syntactic segmentation); relations among the various positions where elements may be encountered at all four levels; and creating a matrix for the multidimensional representation and evaluation of the data and their relationships.

Such complex collations of large numbers of manuscripts should preferably be executed on the basis of a single source file per manuscript. If collation is possible only within strictly separate levels (as is the case with OUCP's Collate), the source files must be adapted manually to each collation, and errors and inconsistencies are bound to arise. A well-designed and well-tested combination of Hypertext and Collate seems required to overcome these difficulties.

BIBLE AND COMPUTER. A BRIEF HISTORY OF THE LAST TEN YEARS

Winfried Bader

Tübingen

The history of computer-assisted biblical scholarship can best be told by surveying the threeannual conferences of the *Association Internationale Bible et Informatique (AIBI)*.

1985 "The Text": At the beginning was the encoding of the text, using a 7-bit-ASCII transcription of Greek and Hebrew (the only alternative in the age of punch cards) that today guarantees system-independence. The first projects undertaken were to produce word indices and concordances, for which the text was tagged with morphological information that could be used in lemmatization. These tags were added with the same 7-bit-ASCII encoding of the base text, using the lines of a file like records of a database.

1988 "Tools, Utilities": The utilization of a classical printed concordance by a scholar was the model for producing tools and utilities for access to the electronic text and its morphological tagging. Many of these tools were later distributed as commercial products.

1991 "Hermeneutics": New tools provide new goals. The possibility of conducting automated grammatical analysis raised a linguistic discussion about adequate computational grammar models. Definitions of word, phrase, clause, and sentence were discussed, bottom-up and top-down models appeared, and in several places programs for linguistic analysis were developed and implemented.

1994 "Desk and Discipline": The question now is: how does the use of computers influence our classical exegetical work?

Initial development in computer-assisted printed editions of biblical texts was slow and primitive, and the computer was first used like an old typesetting machine with WYSIWYG capabilities. Changes in perspective can be seen in by two computer-assisted editions currently being prepared: a new edition of BHS (Codex Leningradensis) and an edition of Qumran texts. Data entry is performed by scholars all over the world, and the texts will then be collected and processed automatically to generate files for the final printed edition. Using SGML-like encoding guarantees compatibility, interchange, and multi-use electronical editions both today and in the future.

NEW TOOLS FOR OLD LANGUAGE: COMPUTER PROCESSING OF BULGARIAN TEXTS

Elena Paskaleva, Milena Dobreva

Sofia

This presentation describes the state of the art in the application of computer tools to the processing of Bulgarian texts from the onset of Slavic literacy to modern times, with reference to language changes, technologies for producing texts, and the different uses of texts at different times. The basic problems which must be overcome in order to make this type of processing possible are of general character, involving both the lack of computer models for the linguistic description of the Bulgarian language and difficulties in processing the Cyrillic alphabet (from problems of grapheme presentation in texts written in early Cyrillic, where the alphabet was not standardized, to modern Cyrillic, which is standardized in a linguistic, but not a computational, sense).

Those problems are responsible for the low level of code unification and the absence of language-industry resources, such as computer-readable dictionaries or collections of texts. On an academic and educational level, the difficulty of constructing adequate linguistic models for Bulgarian writing are the basic cause of the limited use of computers in Bulgarian linguistic research and education.

Problems, results of current research, and prospects for further work are presented with a view of the application of language modeling and available tools, including discussion of methods of inputting text as graphical images, industrial word and text processing, and research models of computational linguistics being developed for the Bulgarian language.

16:30 Coffee Break

17:30 - 18:30 Session and demos: MULTIMEDIA
Chairman: Prof. I. Azarova

CD-ROM AND THE MANUSCRIPT HERITAGE 25
..... Borjana Xristova, *Sofia*
THE 'ST. SOFIA' PROJECT 25
..... V. Bratanova, E. Koceva, D. Petrov, *Sofia*

FRIDAY, July 28th

9:00 - 12:00 Session: KLEIO
Chairmen: Prof. J. Labyncev, Dr. W. Bader

A RUSSIAN VERSION OF THE SOURCE-ORIENTED SOFTWARE *KLEIO*: NEW OPPORTUNITIES FOR THE TREATMENT OF CYRILLIC SOURCES 33
..... Viktoria Tjazelnikova, *Moscow*
INTERACTIVE FULL-TEXT PROCESSING IN *KLEIO*: SCANNING CRITERIA OF BASIC TERMS (ON APPEALS BY DISENFRANCHISED PEOPLE) 33
..... Vladimir Tixonov, *Moscow*
A MODEL OF HISTORICAL SOURCES IN *KLEIO* AS A BASIS FOR INTEGRATED SOCIAL CLASSIFICATION 34
..... Igor Jušin, *Moscow*
A COMPUTER INVESTIGATION OF MEDIEVAL PROGNOSTIC BOOKS BY MEANS OF *KLEIO* 36
..... Adelina Anguševa, *Sofia*
GREEK AND LATIN LOADWORDS IN THE NEW YORK MISSAL AND THE *KLEIO* COMPUTER PROGRAM 36
..... Margaret Dimitrova, *Sofia*

12:00 Lunch

14:00 - 16:30 SGML: Hands-on session
Farewell Reception

List of participants **49**

Session: STATE OF THE ART

COMPUTER TECHNOLOGIES AND TEXTOLOGY

Miroslav Yanakiev
Sofia

Computer technology today reveals a potential for specialists in applied linguistics that would have been difficult to imagine some twenty years ago. On the other hand, some of the innovative approaches made possible by advancements in computer technology may prove unfamiliar or difficult to understand to colleagues who are not familiar with these new tools, and with the new possibilities they create. For example, computer-supported textology can facilitate searching and indexing of material of arbitrary length from very large texts, and can do so within an arbitrary context.

A textologist with proper computer support will no longer require the voluminous *Indices Verborum* that were one of the mainstays of our profession in earlier generations. Neither will there be any need to combine lexical forms into lexemes in order to enter them in a common lemmatized form into an *Index Verborum*.

Whatever the current limitations of computer-aided textology, it clearly has a future in providing, through optical scanning and character recognition, for the transformation of any incunabulum or manuscript into a sequence of coded characters suitable for computer processing. And yet the textologist working with a computer should be aware that the machine is indifferent to the transcription or transliteration used - whether it is in the Latin script, the Cyrillic script, the Latin phonetic script, the Latin phonemic one, or some other. Regardless of the specific transcription or transliteration employed, the textual information will be stored in the computer memory as a sequence of zeros and ones, and will be available in its internal, machine form for the type of processing described above.

OPTICAL CHARACTER RECOGNITION OF HISTORICAL DOCUMENTS

Ivelin Stojanov

Sofia

The present report is a part of a global project for a historical dictionary of the Bulgarian language. Comprehensive research with old historical documents requires some degree of automation of document entry, which was the motivation for the current research. The process of automated document entry consists of following steps: (1) document imaging; (2) pre-processing; (3) character (symbol) recognition; and (4) postprocessing.

Graphical images are captured with a grey-scale scanner. The symbols that the system has to recognize are separated from the background through an adaptive thresholding method, which neutralizes the negative influence of ink spots, etc. A heuristic algorithm for symbol identification is then used to convert each symbol (character) to a computer character code. The process of recognition includes some kind of classifiers, such as neural networks, as well as static methods, with the exact strategy in each case determined by legibility and stage of research. Classifiers initially must be 'taught' to recognise some set of symbols, although interactive learning and self-learning play a larger role as work progresses. Post-processing includes character organization in the same format as in the graphical image, as well as dictionary production. The expected reading error is about 5-10 %, depending on the original document quality.

OPTICAL CHARACTER RECOGNITION OF MEDIEVAL MANUSCRIPTS

Peter Pavel Klasinc, Vinko Kurent

Maribor

The variation in handwritten sources make bitmap matching unrealistic as Optical Character Recognition strategy. Our report describes VOCR, a feature extraction (topological analysis) OCR system developed for the International Institute of Archive Science at the Maribor Provincial Archive. While our report describes a project involving Gothic handwriting, the algorithms and implementation underlying this system should prove applicable to Cyrillic materials, as well.

TEI Ongoing Projects

Harry Gaylord, Groningen

SGML was established in 1986 as an ISO standard. The work of TEI was begun only in 1988. The first TEI Guidelines appeared in 1990 and the latest version in 1994.

Yet both of them became important for scholarly study and publishing. Here I would like to briefly illustrate what is presently being done using these standards for scholarly work. This is not an attempt to give a comprehensive presentation.

A number of critical editions of works with very complex histories are being encoded with TEI.

1. One of the examples is of the works of the 19th century American philosopher, Pierce. The materials which are being encoded are his publications, but his extensive notes for print and electronic publication.

2. The *Perseus* project which presents on a video disk the texts of classical Greek authors, the Lidell Scott Greek dictionary, other material including pictures, maps, etc related to the Greek classical world. These are all integrated for classroom use. The texts and the dictionary have been encoded in SGML.

3. A number of projects developing literary corpora are using TEI encoding.

In addition to this the *Women Writers Project* at Brown University in using TEI to encode an enormous corpus of the writings of women in English from its beginning until the end of the 18th century. This involves hundreds of works and will be published in print and electronicly. Other projects include corpuses of Spanish, Italian, and Hebrew literature. Several major projects for linguistic corpuses are using the TEI Guidelines to encode in great detail.

Very recently the British National Corpus was issued. This is a corpus of 500 million words of modern British English. It has used the TEI Guidelines. Portions are encoded very lightly and others with great linguistic detail.

For scholars of medieval texts it is important that the information which they assemble about the documents they study can be reused in a number of contexts. In the days of typewriters and lead printing finding a new good manuscript after working on a text for years was a disaster because one had to compose a completely new text and apparatus for the critical edition. With tools which we will see at this conference, single encoded documents can be emended until the time of publication and updated on the basis of all the information in electronic form. With the TEI Guidelines and software being developed on the basis of it, we will be performing our work much more effectively and with better results.

LIST OF PARTICIPANTS

Adelina Anguševa ..... p. 36
Sofia University, Sofia, BULGARIA

Irina Azarova
Russian Bible Society, 191104 St. Petersburg, Post box 81, RUSSIA

Winfried Bader ..... p. 13
Tübingen, GERMANY
e-mail: winfried.bader@zdv.uni-tuebingen.de

David J. Birnbaum ..... p. 15, 16
University of Pittsburgh, Depatment of Slavic Literature and Languages, 1417 Cathedral of Learning, Pittsburgh, USA
email: djbpitt+@pitt.edu

Jurij Blinov
Moscow city association of archives, Ul.Profsojuznaja, 80
117393 Moscow, RUSSIA, fax: 128-78-97

Andrej Boyadžiev ..... p. 38
Center for Slavo-Byzantine Studies 'Ivan Dujcev"
Sofia University, Sofia, BULGARIA

Michael Bakker ..... p. 20
University of Amsterdam, the NETHERLANDS
e-mail: M.Bakker@let.uva.nl

Kirill Bogoljubov
RUSSIA

Irina Bolčeva ..... p. 21
Sofia University, Sofia, BULGARIA

Monia Camuglia ..... p. 39
Via Guerrazzi, 2, Madonna Dell'Acqua , Pisa, ITALY
e-mail: camu@vm.cnuce.cnr.it camuglia@ilc.pi.cnr.it

Ralph Cleminson
ENGLAND
e-mail: ralph@hum.port.ac.uk

Margret Dimitrova ..... p. 36
Sofia University, Sofia, BULGARIA

Dmitrij Dmitriev
Institute for Linguistic Research RAN, St. Petersburg, RUSSIA
ul. Novo Aleksandrovskaya 11-42 fax (812) 262-41-92

Milena Dobreva ..... p. 14
Institute of Mathematics and Computer Science - BAS
bl. 8, Acad. G. Boncev St., 1113 Sofia, Bulgaria
fax: (++359-2) 752 078 e-mail: dobreva@bgearn.bitnet

Harry Gaylord ..... p. 48
Alfa Informatica, Universsity of Groningen
Oude Kijk in't Jatstraat 15, AS 9700 Groningen, the NETHERLANDS
e-mail: galiard@let.rug.nl

Karsten Grünberg ..... p. 17
University of Heidelberg, GERMANY
e-mail: hp6@ix.urz.uni-heidelberg.de

Emilia Guergova ..... p. 41
Sofia University, Sofia, BULGARIA

Tanja Ivanova ..... p. 21
Sofia University, Sofia, BULGARIA

Stana Jankoska ..... p. 27
Skopje, MACEDONIA

Ljupco Josifoski ..... p. 27
Skopje, MACEDONIA

Igor Jushin ..... p. 34
Moscow city association of archives Ul.Profsojuznaja, 80 117393
Moscow, RUSSIA, Fax: (095) 128-78-97

Sebastian Kempgen ..... p. 19
An den Weihern 18, D-96135 Stegaurach, GERMANY
fax: ++49-951-296425

Peter Pavel Klasnic ..... p. 47
International Institute for Archival Science
SLO-62 000, Maribor, SLOVENIA fax: ++386-62-222-564

Elena Koceva ..... p. 25
Center for Cyril and Methodius Studies, Sofia, BULGARIA

Cvetana Krstev ..... p. 17
Faculty of Mathematics, University of Belgrade, YU-Belgrade

Vinko Kurent ..... p. 47
International Institute for Archival Science
SLO-62 000, Maribor, SLOVENIA fax: ++386-62-222-564

Reinhard Lampe ..... p. 28
Heidelberg, GERMANY
e-mail: 100071.3450@compuserve.com

Yurij Labyncev ..... p. 30
Centr belorusowedczeskich issledowanij Instytuta
slawianowedenyja i balkanistyki RAN, Moscow, RUSSIA

Catherine Mary MacRobert ..... p. 22
Lady Margaret Hall, Oxford OX2 6QA, ENGLAND
fax: 01865-511-069

Mimosa Majstoroska ..... p. 40
Skopje, MACEDONIA

Anisava Miltenova ..... p. 37
Institute for Literature, Bulgarian Academy of Sciences, Sofia,
BULGARIA, e-mail: anmilten@bgearn.bitnet

Ljupco Mitreski ..... p. 40
Skopje, MACEDONIA

Dragan Mixajlov ..... p. 27
Skopje, MACEDONIA

Henning Moerk ..... p. 17
Aarhus, DENMARK

Aleksandr Moldovan ..... p. 23
123459 Moscow, ul. Tallinskaya 32, kor. 1, kv. 57, RUSSIA

Miroslav Novak
International Institute for Archival Science
SLO-62 000, Maribor, SLOVENIA fax: ++386-62-222-564

Elena Paskaleva ..... p. 14
Laboratory for Linguistic Modeling - CCIIT
25a Acad. G. Bonchev St., 1113 Sofia, BULGARIA
e-mail: hellen@bgearn.bitnet

Gordana Pavlović-Lasetić ..... p. 17
Faculty of Mathematics, University of Belgrade, YU-Belgrade

Dragomir Petrov ..... p. 25
"Courier of UNESCO', Sofia, BULGARIA

Kiril Ribarov ..... p. 27
Prague, CZECH REPUBLIC
krib2407@barbora.ms.mff.cuni.cz

Zdenka Ribarova ..... p. 27
Institut za makedonski jazik
ul. G. Prlicev 5, 91000 Skopje, MACEDONIA

Thomas Rosen
Slaviska Inst. Uppsala Universitet, Box 513, S-753 12 Uppsala,
SWEDEN Fax: 018-18 12 98

Francis Salter ..... p. 44
Fredrick Street R.R. 3, Bracebridge, Ontario P1L 1X1, CANADA

Maria Šnitter p.43
University of Plovdiv, Plovdiv, BULGARIA

Nina Šoileva ..... p. 21
Sofia University, Sofia, BULGARIA

Ivelin Stoyanov ..... p. 47
Institute of Information Technologies - BAS, bl.29A Acad. G. Boncev St. 1113 Sofia, BULGARIA, e-mail: hary@iinfbg

Larisa Szczawinskaja ..... p. 31
Centr belorusowedczeskich issledowanij Instytuta slawianowiedenija i balkanistyki RAN, RUSSIA

Vladimir Tixonov ..... p. 33
ul. Mezhdunarodnaya, 10, Moscow City Association of archives, Moscow, RUSSIA, Fax: (095) 128 78 97

Viktoria Tjazelnikova ..... p. 33
ul. Profsojuznaja, 80, Moscow City Association of archives, Moscow, RUSSIA, Fax: (095) 128 78 97

Cynthia Vakareliyska ..... p. 40
University of Oregon, USA, e-mail: vakarel@oregon.uoregon.edu

William Veder ..... p. 20
University of Amsterdam, the NETHERLANDS

Vasja Velinova ..... p. 32
CIBAL, Sofia, BULGARIA

Dusko Vitas ..... p. 17
Faculty of Mathematics, University of Belgrade, YU-Belgrade

Borjana Xristova ..... p. 25
National Library 'Cyril and Methodius', Sofia, BULGARIA

Miroslav Yanakiev ..... p. 12
Sofia University, Sofia, BULGARIA

NOTES

БАН
ЦЕНТРАЛНА
БИБЛИОТЕКА

6802