COMPUTER PROCESSING OF MEDIEVAL SLAVIC SOURCES IN THE INSTITUTE OF LITERATURE AT BAS REPERTORIUM PROJECT (1994–2004)

Anissava Miltenova

Introduction

Mixed-content miscellanies (very frequent in the Byzantine and mediaeval Slavic written heritage) are usually defined as collections of works with non-occupational, non-liturgical application, and texts in them are selected and arranged according to no identifiable principle. It is a "readable" type of miscellanies which were compiled mainly on the basis of the cognitive interests of compilers and readers. Just like the occupational ones, they also appeared to satisfy public needs but were intended for individual usage. My textological comparison had shown that mixed-content miscellanies often showed evidence of a stable content – some of them include the same constituent works in the same order, regardless that the manuscripts had no obvious genetic relationship. These correspondences were sufficiently numerous and distinctive that they could not be merely fortuitous, and the only sensible interpretation was that even when the operative organizational principle was not based on independently identifiable criteria, such as the church calendar, liturgical function, or thematic considerations, mixed-content miscellanies (or, at least, portions of their contents) nonetheless fell into types. In this respect, the apparent free selection and arrangement of texts in mixed-content miscellanies turns out to be illusory.

The problem was – as the corpus of manuscripts that I and my colleagues needed to examine grew – our ability to keep track of the structure of each one, and to identify structural correspondences among manuscripts within the corpus, diminished. So, at the end of 1993 I addressed a letter to Prof. David Birnbaum (University of Pittsburgh, PA) with a request to help me to solve the problem. He and my colleague Andrey Boyadzhiev (Sofia University) pointed out to me that computers are well suited to recording, processing, and analyzing large amounts of data, and to identifying patterns within the data, and their proposal was that we try to develop a computer system for description of manuscripts, for their analysis and of course, for searching the data. Our collaboration in this project is now ten years old, and our talk today presents an overview of that collaboration.

1994–1995

Bulgarian-American project "Computer Supported Processing of Old Slavic Manuscripts" begun in 1994, sponsored by IREX – Washington (1994–1995). A new type of software was built, which was based on the SGML (Standard Generalized Markup Language) accepted by the International Society of standardization (ISO) and especially in its TEI (Text Encoding Initiative) implementation. The goal of the project was to create a sophisticated system of processing Slavonic Manuscripts in the universal format with multiple using.

The system for computer analytical description of medieval Slavic manuscripts on the level of modern archeography, palaeography, codicology and textology (from now on – TSM = Template for Slavonic Manuscripts) was carried out in the process of the teamwork of David Birnbaum, Beirend van Dijk, who was then a post-graduate student in Groningen (The Netherlands) Milena Dobreva, Institute of Mathematics and Computing in BAS and Harry Gaylord, who taught computer systems in Groningen,. The experiments on the program, using tests, continued almost to the end of 1994. In July–August of 1995, the last changes and specifications in the system of document type definition (DTD) were made during the visit of David Birnbaum in Sofia together also with Andrey Bojadziev. The research project was sponsored also by the foundation 'Open Society', Sofia, in the period 1994–1997.

The description used here is specifically intended for the developing of a Repertory of the Old Bulgarian literature and letters and is adopted for Medieval Slavic texts. The development of fonts for writing the original texts in Medieval Cyrillic belongs to research associate Rumyan Lazov from the Institute of mathematics and computing in BAS. The searching programs on the second stage of the project were created by Stanimir Velev. The complex description of Slavic manuscripts is built by the standard of Standard Generalized Markup Language (SGML), which was accepted by the International Society of standardization (ISO). This electronic standard is based on the ability to include special "markings" in the texts of natural languages, so called tags. Tagging circles certain parts of the text and signal what the data represents. It makes very easy to draw out data from the text during its computer processing. This standard was used for the first time, in the description of Medieval Slavic manuscripts and for including an arbitrary (free from limitations) sizes of non-normalized texts from the manuscripts themselves in the

process of description. Our SGML-based undertaking was oriented not only toward preparing manuscript descriptions that might be suitable for printing, electronic rendering, and searching, as was the case with the data-base's approach. Rather, we anticipated even at that stage that the manuscript description files would be suitable for direct analysis, so that we would be able, for example, to identify patterns of structural similarity within a corpus of manuscripts on the basis of the same raw data files that we would also use to generate traditional printed manuscript descriptions.

The team has followed five main principles, formulated by David J. Birnbaum (see – http://www.slavic.pitt.edu/~djb/): 1. Standardizing of document file formats; 2. Multiple use (data should be separated from processing); 3. Portability of electronic texts (independence of local platforms); 4. Necessity of preservation of manuscripts in electronic form; 5. Orientation to the well-structured divisions of data according to the Slavic traditions of codicology, orthography, paleography, textology, etc.

The system for encoding of medieval Slavic text (TSM) was discussed on an international conference in Blagoevgrad (24th–28th July, 1995). The reports from the conference were published in a separate volume. The philosophy of SGML helped to settle some well known misunderstandings among palaeoslavists concerning philological questions of terminology, inventory of units, character sets and data structure.

1996–1999

During the period from 1995 through 1998, a team of scholars supervised by me based primarily at the Institute of Literature at the Bulgarian Academy of Sciences produced SGML descriptions of some 200 medieval Slavic manuscripts of all types.

At the same time, the Institute of Literature entered into a project with Ralph Cleminson at the Central European University entitled "Computer-Supported Processing of Slavonic Manuscripts and Early Printed Books", which led to the encoding of additional manuscript descriptions and the publication of several articles addressing the technology underlying the project. Ralph Cleminson, David Birnbaum, and others presented the results of their research at the Twelfth International Congress of Slavists in Kraków in 1998, where the International Committee of Slavists established a Special Commission to the Executive Council of the Committee for the Computer-Supported Processing of Slavic Manuscripts and Early Printed Books, with David, Ralph, Andrey, and me as officers. The Commission's authorization was renewed at the Thirteenth International Congress of Slavists in Ljubljana in 2003. Participants from Belorussia, Bulgaria, Chech Republic, Finland, Italy, Macedonia, Great Britain, the US, etc. put on discussion some mainstream questions in the field.

The other principal achievement of this stage was the development by Stanimir Velev of a query interface for the manuscript descriptions that had been prepared within the Repertorium project. Stanimir's interface was an interim solution that has now been superseded by XSLT scripting, but for several years it served as the principal query engine for scholars at the Institute of Literature who were conducting philological research on the basis of our manuscript descriptions.

2000–2003

For three years amount of analytically described manuscripts increased to three hundred. They were processed by using TSM system in the SGML environment with the corresponding interface A/E (Author/Editor, SoftQuad, Canada) software package. Members of the team were: Anna Stoykova, Nina Georgieva, Elena Tomova, Adelina Angusheva, Andrey Boyajiev, Margaret Dimitrova, Dimitrinka Dimitrova, Desislava Athanasova, Maya Petrova, Radoslava Stankova, Marina Jordanova, Dilyana Radoslavova, and Anissava Miltenova. The book under the title: "Medieval Slavic Manuscripts and SGML: Problems and Perspectives" (Sofia, 2000) is sponsored by IREX and Central European University). The articles in the book not only put into scientific circulation the achieved results from the analysis of the manuscripts, but also mark the problems that are waiting to be solved.

In general, the description in Repertorium is much more detailed in comparison with all other projects in the field, especially in such areas as orthography and the description of the texts in the manuscripts. The textological part includes information on the level of the whole manuscript (<manuscriptContentDesc> element) and on the level of each text (an element <articleContentDesc>). Because the intention was to provide research results from philological text investigations, there are elements such as <source>, <translation>, <protograph>, <antigraph>, and <litRedaction> on the level of the manuscript and on the level of each of the texts. A special element <neighbour> was introduced in the DTD to facilitate describing the organization of the component texts in mixed-content miscellanies, where the texts are not arranged according to ecclesiastical feasts or medieval typicons.

A current continuation of the original project, “Electronic Description and Edition of Slavic Sources” (2002–2003, sponsored by UNESCO), is in a transitional stage of migrating from SGML to XML technology. In 1994–1995, when the SGML DTD for the project was first constructed, Extensible Markup Language had not yet been conceived. Since then electronic and web technologies have changed very rapidly, and now we have tools that are very convenient for direct browsing and editing the markup files. Direct access to XML documents from such popular browsers as Internet Explorer, Opera, or the Gecko-engine powered ones, as Mozilla, Doczilla, and Netscape, provide more control and efficiency. This fact, together with the development of special recommendations for the markup languages produced under the auspices of the W3 consortium, Unicode, and other institutions and international initiatives, has led to a rapid growth of academic applications based on XML technology. So, this stage is characterized not only by the accumulation of still more manuscript descriptions, but also by the conversion of our materials from SGML to XML. The transition to XML was dictated by the remarkably broad acceptance of XML within the electronic-text community, and particularly by its adoption by the TEI, initially as an alternative to SGML, but ultimately as a replacement for it. We have currently converted over one hundred manuscript descriptions from our initial corpus of three hundred; the rest will be converted in time, and all new descriptions are being created directly in XML. Contributions of David Birnbaum to the project are enormous. His presentation at the Thirteenth International Congress of Slavists in Ljubljana, demonstrated a new level of document multipurposing: the generation directly from TEI XML manuscript descriptions of dendrograms that illustrated the degree of structural similarity among miscellany manuscripts and SVG plectograms showing the item-by-item correspondences in the contents of pairs of manuscripts. Andrey Boyadzhiev’s presentation at the Ljubljana Congress illustrated the use of the same files to produce prose descriptions suitable for publication electronically or on paper.

At the beginning of Repertorium project we had concentrated on the production of manuscript descriptions based on the promise that one would be able to employ them directly in computer-assisted analysis at some point in the future. Last years had shown that the descriptions were suitable for use in a range of analytical applications, but primarily within a fairly low-level query framework that did not take full advantage of the hierarchical XML structure. David’s work showed that radically new non-textual representations of manuscript structures were available essentially for free from the same files that were used to produce formatted descriptions. This development demonstrated that computers did more than provide a new way of performing such traditional tasks as producing manuscript descriptions. Rather, the production of electronic manuscript descriptions enabled new and innovative philological perspectives on the data. Not only did it make traditional activities easier and more reliable, but it also created opportunities for radically new philological research.

At the end of 2003 the project obtained a membership in Text Encoding Initiative consortium.

Ongoing Projects of the Department of Old Bulgarian Literature, Institute of Literature

1. Joint projects with British Library, London, for analytical description of Slavic manuscripts (Working team: Anissava Miltenova, Andrey Boyadzhiev, Dilyana Radsolavova, Christine Thomas, Ralph Cleminson, with cooperation of Central Library of BAS – Dincho Kr"stev and Sabina Aneva).
2. Joint project with Gothenburg University, for implementation of computer tools in the study of late mediaeval Slavic manuscripts (working team from the Department of Slavic Languages and from the Institute of Literature BAS).
3. Collaboration with the Institute of Russian language, RAS, Moscow, directed to build a network for exchange of language and contents data of Slavic manuscripts in XML and to realize joint electronic editions.
4. The national project “Metadata and electronic catalogues” (Institute of Literature, Institute of Bulgarian language, Sofia university) is concentrated on the terminology in palaeoslavistic. The project includes something about 2000 denotations and keywords in the field of archeography, palaeography, textology, etc. which will be take into an electronic dictionary.
5. An ongoing project is an updating of my 1982 dissertation on the structure of mixed-content miscellanies. Our collaborative application of computer technology has already led us to revise some of my earlier conclusions about structural and textological similarities among manuscripts, and David Birnbaum and I are currently preparing a monograph that takes my dissertation as its starting point, but extends the data and the reevaluates it with the aid of computational tools.

Author Information

Anissava Miltenova – Senior Researcher Dr., Institute of Bulgarian Literature, BAS, 1113 Sofia, Shipchenski prohod str. 52, Bulgaria; e-mail: anmilten@bas.bg