The Tenth-Century Cyrillic Manuscript *Codex Suprasliensis*: the creation of an electronic corpus UNESCO project (2010–2011)

Hanne Martine Eckhoff
University of Oslo
Kanonhallveien 10e
0585 Oslo
h.m.eckhoff@ifikk.uio.no

David J. Birnbaum
University of Pittsburgh
Department of Slavic Languages and Literatures
1417 Cathedral of Learning
djbpitt@pitt.edu

Anisava Miltenova
Bulgarian Academy of Sciences
Institute for Literature
52 Shipchenski prohod
am-iltenova@gmail.com

Tsvetana Dimitrova
Bulgarian Academy of Sciences
Bulgarian Language Institute
52 Shipchenski prohod
cvetana@dcl.bas.bg

Abstract

This paper presents an overview of principles and problems connected with the preparation of an electronic edition of the largest Old Church Slavonic manuscript, the *Codex Suprasliensis*, in the context of a project funded by UNESCO. Specifications of the manuscript, its history, and previous paper-based and electronic editions are discussed, together with a strategy for the preparation of a complete digital edition, including newly acquired digital images, electronic text, analysis and commentaries, parallel Greek text, and updated bibliography. In particular, our paper sheds light on automating the morphosyntactic annotation of the text and the difficulties that had to be resolved in this part of the project.

1 Introduction

The UNESCO-funded project *The Tenth-Century Cyrillic Manuscript Codex Suprasliensis* aims at digitizing the largest Old Church Slavonic manuscript, the *Codex Suprasliensis* (http://csup.ilit.bas.bg/).

This early Cyrillic manuscript has been dated to the end of the tenth or the beginning of the eleventh century and has been published three times on paper (Miklošič, 1851; Severjanov, 1904; Zaimov and Capaldo, 1982–83). The most recent of these, the two-volume edition by Zaimov and Capaldo (1982, 1983), was published more than two decades ago and contains photographic images of the entire manuscript; a transcription reproduced from Severjanov, 1904 and corrected (not entirely without error) against the facsimile; and a Greek text (compiled from multiple Byzantine sources, which necessarily implies complications in its philological interpretation; see also Abicht and Schmidt, 1896).

In section 2, this paper presents information about the content, condition, and history of the manuscript. Section 3 reviews efforts in digitization of the manuscript, and section 4 discusses previous electronic editions of the deciphered text, reviewing problems with representation and availability and solutions adopted by the editors. Section 5 gives an overview of the principles of application of morphosyntactic annotation conditioned by the chosen annotation tool and strategy. The conclusion in section 6 explores distinctions among the publication of a text, digitization of a manuscript, development of language corpora, and a true electronic edition of the text, which is the goal of the UNESCO project.

2 The Manuscript

The *Codex Suprasliensis* is a Cyrillic manuscript, arguably copied at the end of the tenth or the beginning of the eleventh century (Krăstev and Bojadžiev, 1999). It is the largest extant Old Church Slavonic manuscript and it is associated with the Preslav literary school.

The *Codex* contains twenty-four vitae of Christian saints for the month of March and twenty-three homilies for the triodion cycle of the church year. In content it is a lectionary menaeum (or panaegyricon), combined with homilies from the movable Easter cycle, most of which written by or attributed to John Chrysostom (http://csup.ilit.bas.bg/node/7).

According to most researchers, the Miscellany was not translated as a stable compilation from any single Byzantine menological or hagiographical manuscript. Rather, it was com-

Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop, pages 57–61,
Hissar, Bulgaria, 16 September 2011.

piled from texts translated at different times, long before the compilation of the *Codex Suprasliensis*. Presumably, at least one of the sources was the Glagolitic Epiphanius homily. Folio 104v of the manuscript has a marginal note that reads *g(ospod)i pomilui retъka amin* ('Lord have mercy on Retъk. Amen'), and some researchers have suggested that Retъk is the name of a scribe.

The language of the manuscript follows the Preslav literary norm of the tenth century. It is considered the most representative source of linguistic information about canonic Old Church Slavonic because of its size and because it contains texts otherwise unattested in the early mediaeval Slavic tradition. The codex is, thus, the main source for studying the language, writing, and culture of Bulgaria during the Preslav period.

The *Codex Suprasliensis* is written on parchment and shows careful writing and craftsmanship. It was discovered in 1823 in a Uniate Basilian monastery in Supraśl (then in Lithuania, now in Northeastern Poland in the Podlaskie Voivodeship) by Canon Michał Bobrowski. Bobrowski sent it for study to the Slovenian scholar Jernej Kopitar. After Kopitar's death, the first 118 folios were donated to the University Library in Ljubljana, where they are still kept. The following 16 leaves were purchased by A. F. Byčkov in 1856 and are now kept in the Russian National Library in St. Petersburg. The remaining 151 leaves were part of the collection of the Counts Zamoyski. The last, so-called Warsaw part had disappeared during World War II and were long considered lost until re-emerging in the US. In 1968, those folios were returned to Poland, where they are now part of the manuscript collection of the National Library in Warsaw.

The *Codex Suprasliensis* has been listed in UNESCO's Memory of the World Register since 2007.

3 Digitization

In the present project, digital images of all three parts of the *Codex Suprasliensis*, currently located in repositories in three different countries (the National Library in Warsaw, Poland; the National Library of Russia in St. Petersburg; and the National University Library in Ljubljana, Slovenia), were reunited for publication in a single electronic edition. The digital images are already available at http://csup.ilit.bas.bg/galleries. The separate publication of the photographic facsimile is an interim stage in the project, and the photographs will eventually be republished together with a transcription that will be fully annotated, accompanied by commentary and updated bibliography.

Some previously unknown source materials, including some Byzantine originals identified only after the publication of the Zaimov and Capaldo edition in the early 1980s, have been used in the preparation of the Greek text of the new edition.

Eventually a diplomatic transcription of the text of the *Codex Suprasliensis* will be published together with critical apparatus, parallel Greek text, vocabulary, and grammatical analysis (in the form of corpora annotation). The annotation of the electronic corpus is at initial stage, with only one piece, namely the Life of St. Paul the Simple, completely annotated, and another (the Life of St. Paul and St. Juliana) under active preparation.

4 Electronic text

The principles of manuscript description follow a proposal developed in the context of *The Repertorium of Old Bulgarian Literature and Letters,* which includes descriptions, in both English and Bulgarian, of some 350 mediaeval Slavic manuscripts dated from the eleventh to the beginning of the eighteenth century. The *Repertorium* was designed in conformity with important standards and guidelines in humanities computing (Miltenova, Boyadzhiev, and Velev, 2000; Birnbaum, 1996). The description and analysis of the Cyrillic manuscripts contain comprehensive data drawn *de visu* from old texts (http://clover.slavic.pitt.edu/repertorium/).

The first electronic version of the *Codex Suprasliensis* was a 7-bit ASCII transliteration prepared under the direction of Jouko Lindstedt and distributed by the *Corpus Cyrillo-Methodianum Helsingiense: An Electronic Corpus of Old Church Slavonic Texts* (CCMH, http://www.helsinki.fi/slaavilaiset/ccmh/) and the TITUS project (http://titus.uni-frankfurt.de/texte/etcs/slav/aksl/suprasl/supra.htm). These transcriptions contain numerous errors and come completely without context and critical apparatus (no images, Greek text, commentary, grammatical annotation or analysis, etc.). The new edition under development takes the Helsinki transcriptions as a starting point, converts the text from ASCII to Unicode, corrects the er-

rors, and includes the full range of supporting materials listed above.

A pilot model of an electronic edition of a small part of the *Codex Suprasliensis* with a search program was developed in 2008 (Birnbaum, 2008) at the University of Pittsburgh (http://paul.obdurodon.org). This electronic edition of the Life of St. Paul the Simple was developed in accord with the procedures and priorities described above: it is based on a corrected version of the text published by the CCMH, accompanied by parallel Greek (from the Zaimov/Capaldo edition), a new English translation, detailed linguistic commentary, and photographic facsimiles. Linguistic analysis in the commentary conforms to notation developed in Oscar Swan's *Old Church Slavic Inflectional Morphology* (2008).

There are many collections and editions of classical and mediaeval texts (such as the Perseus Project, http://www.perseus.tufts.edu/hopper/), but most of them are manually annotated. No rule-based morphological guesser is currently available for Old Church Slavonic, partially because of troublesome orthography, although there is preliminary finite-state morphology under development by Roland Meyer (http://rhssl1.uni-regensburg.de:8080/OCS/).

The research project *Pragmatic Resources in Old Indo-European Languages* (PROIEL), which aims at developing morphosyntactic means for the annotation of and research into the information structure in Ancient and Hellenistic Greek, Latin, Gothic, Classical Armenian, and Old Church Slavonic (Haug and Jøhndal, 2008), has developed a statistical morphological guesser and a semi-manual syntactic annotation tool supported by a set of morphology-based rules. The corpus to be built for the electronic edition of the *Codex Suprasliensis* will be annotated manually, but with the assistance of the morphological guesser already developed by the PROIEL project and trained for Old Church Slavonic morphology on the *Codex Marianus* (Haug et al., 2009). Thus, the *Codex Suprasliensis* will be annotated for morphology, syntax, and other features in the PROIEL annotation interface, and the information will be exported in XML for incorporation into the projected electronic edition.

5 Morphosyntactic Annotation

The morphosyntactic annotation tool to be used in the *Codex Suprasliensis* project is an integrated part of the PROIEL parallel treebank of ancient Indo-European languages. The core of the treebank is the New Testament in its Greek original and its earliest translations into each of the other project languages. PROIEL features an electronic version of the *Codex Marianus* fully annotated for morphology, syntax, and various other linguistic features. It has also been automatically aligned with the Greek Gospels at token level (Eckhoff and Haug, 2010).

Test annotation of the *Codex Suprasliensis* is currently in progress. Observations and solutions discussed in this section of the paper were drawn from the process of annotating of the Life of St Paul the Simple and the Life of St. Paul and St. Juliana (the annotated text is currently available at: http://foni.uio.no:3000/).

The PROIEL annotation tool (available at the same site) was developed with certain needs in mind:

When confronted with novel text styles and orthographical conventions (different from the already annotated *Codex Marianus*), annotation initially is primarily manual, but it becomes increasingly automatic as the tool learns from operator input. Because the annotation for some languages, including Old Church Slavonic, is being performed on a diplomatic transcription of a text with substantial orthographic variation (rather than on the normalized texts that are used more commonly in other disciplinary philological traditions), morphological analyzers and syntactic parsers are not available for all of the project languages.

Annotators had to be recruited internationally due to the specialized knowledge required. The application was, therefore, built to work with standards-compliant browsers, which did not require the annotators to perform any extra installation. For the annotators of Old Church Slavonic texts, the tool supports transliterated input, obviating the need for a specialized keyboard layout interface.

Texts are imported in a simple XML format, where they are split into tokens (words) based on spacing, and roughly into sentences based on punctuation. After the import and coarse automatic segmentation, the annotation proceeds as follows:

First, there is adjustment of sentence division. Since punctuation is not a reliable guide to sentence division in Old Church Slavonic, sentences must often be split or merged.

Second, the imported tokenization must be checked and corrected manually. A linguistic

analysis of the text may need to normalize the word boundaries of the edition. In particular, contractions of prepositions and nouns may need to be dissolved.

Third, morphological annotation and lemmatization are implemented. The PROIEL annotation tool provides guesses for morphological features and lemmata based on previous reviewed annotations (Haug et al., 2009).

In the first stages of the annotation of the initial samples from the *Codex Suprasliensis*, the guesser recognized only 15% of the words on the basis of its prior annotation of the *Codex Marianus*. After annotating 2000 tokens of the *Codex Suprasliensis*, the accuracy of the guessing more than tripled, to approximately 50%. The low initial result and rapid improvement is mostly due to the use of diacritics in the *Codex Suprasliensis*, and we are developing an orthographic normalizer that will temporarily strip diacritics to facilitate recognition and automated linguistic tagging.

The lemmata were entered with support from a transliteration device, which also provides guesses based on extant lemmata. The lemmatization follows part-of-speech classification. A single form may, therefore, belong to several lemmata. For example, there are no fewer than four lemmata with the form *jako*: a subjunction, a relative adverb, and two regular adverbs that are deemed to have sufficiently different functions to be separated (one meaning 'as, like' and the other serving as an introductory 'for'). Morphological analysis disambiguates the morphological features as far as possible based on syntax and context, and the information is further stored in the database as a positional tag in the form of a string of symbols where each morphological feature represented by a given symbol has a fixed slot (for positional tags, see also Hajič, 2004).

Fourth, the annotators apply syntactic annotation in an enriched variety of dependency grammar (Haug, 2010). This level relies on overt elements and makes it possible to keep word order information and syntactic analysis in separate layers, which is essential in dealing with free-word-order languages such as Old Church Slavonic and Greek. The syntactic annotation is performed with a simple tool that provides good guesses from a set of morphologically based rules.

Fifth comes the review stage, where the morphological and syntactic analysis is reviewed by project members, and, when found correct, published on the PROIEL website.

In addition to the morphosyntactic annotation, there is an interface for annotating information status and anaphoric relations. There is also an option for customized tagging at the token, lemma, and sentence level. This option has been used to tag semantic features (such as animacy), derivational morphology (such as prefixation), and textual features (such as direct speech).

The annotations are all stored in a relational database, but may be exported in various XML formats. The rich linguistic information provided by the PROIEL-style annotation may, thus, be interwoven in XML format into an electronic text edition that also takes the many textological concerns implicit in the Suprasliensis project into account. The resulting edition will thus be one that can serve a very wide audience with different needs and interests.

6 Conclusion

The paper outlines the stages in creating an electronic edition of the *Codex Suprasliensis*: the digitization of the manuscript, preparation of the electronic text, and application of morphosyntactic annotation. All of these tasks can constitute objectives of separate projects (manuscript digitization; electronic text publication; language corpora compilation), but none of them alone would be sufficient to produce an electronic edition of the manuscript. Such an edition depends on all of these products, as well as the publication and annotation of the Byzantine sources, and the development of indices, a lexicon, glossary, bibliography, and others. The project therefore unites the efforts of an international working team with members with different but complimentary qualifications for the joint work on the edition. The electronic version of the *Codex Suprasliensis* will be freely available under a Creative Commons BY-NC-SA license.

Reference

Hajič, Jan. 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum Charles University Press, Prague.

Severjanov, S. 1904. *Suprasl'skaja rukopis' [Codex Suprasliensis, vol. 1-2]*. Pamjatniki staroslavjanskago jazyka, volume 1, 1-2. Sanktpeterburg.

Zaimov, Jordan and Mario Capaldo. 1982. *Suprasŭlski ili Retkov sbornik (Codex Supraslinesis or the Retkov Manuscript),* volume 1, Bulgarian Academy of Sciences Press, Sofia.

Zaimov, Jordan and Mario Capaldo. 1983. *Suprasŭlski ili Retkov sbornik (Codex Supraslinesis or the Retkov Manuscript),* volume 2, Bulgarian Academy of Sciences Press, Sofia.

Swan, Oscar. 2008. *Old Church Slavic. Inflectional Morphology*, volume 1, Berkeley Slavic Specialties, Berkeley.

Miltenova, Anissava, Andrei Boyadzhiev, and Stanimir Velev. 2000. Computerized Manuscript Corpus Data: Results and Further Development. *Bulgarian Studies at the Dawn of the 21st Century: a Bulgarian-American Perspective. Sixth Joint Meeting of Bulgarian and North American Scholars. Blagoevgrad, Bulgaria, May 30–June 2, 1999*. Gutenberg, Sofia, 237–243.

Birnbaum, David J. 1996. Standardizing Characters, Glyphs, and SGML Entities for Encoding Early Cyrillic Writing. *Computer Standards and Interfaces,* 18: 201–52.

Birnbaum, David J. 2008. Paul the Not-So-Simple. *Scripta & e-Scripta*, 6: 23–45

Krăstev, Georgi and Andrej Bojadžiev. 1999. Supras′lski sbornik: problemi na xronologijata i kompozicijata v iztočnoto-pravoslavie i v evropejskata kultura. *Materiali ot meždunarodnata naučna srešta, posvetena na 1100 godišninata ot načaloto na Zlatnija vek v bălgarskata kultura*. *Varna, 2–3 Juli 1993*. Guturanov, Sofia, 192–197.

Miklosich, Fr. 1851. *Monumenta linguae palaeoslovenicae e codice Suprasliensis*. Vindobonae, 456.

Abicht, R. and H. Schmidt. 1896. Quellennachweise zum Codex Suprasliensis. *Archiv für slavische Philologie*, 18: 138-155.

Haug, Dag. T. T. 2010. *PROIEL Guidelines for Annotation*: http://folk.uio.no/daghaug/syntactic_guidelines.pdf

Eckhoff, Hanne Martine and Dag Trygve Truslew Haug. 2010. Aligning Syntax in Early New Testament Texts: the PROIEL Corpus. *Wiener Slavistischer Almanach.*

Haug, Dag T. T., Marius Jøhndal, Hanne Martine Eckhoff, Eirik Welo, Mari J. B. Hertzenberg, and Angelika Muth. 2009. Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages. *Traitement Automatique des Langues*, volum 50.

Haug, Dag T. T., and Marius Jøhndal. 2008. *Creating a Parallel Treebank of the Old Indo-European Bible Translations.* http://www.hf.uio.no/ifikk/english/research/projects/proiel/Activities/proiel/publications/marrakech.pdf