Информационные технологии и письменное наследие

El'Manuscript–2012

Материалы IV международной научной конференции
Петрозаводск, 3–8 сентября 2012 года

Петрозаводск, Ижевск
2012

УДК 004.9
ББК 81.11+81.2-0
  И741

И741   **Информационные технологии и письменное наследие** [Текст] : материалы IV междунар. науч. конф. (Петрозаводск, 3–8 сентября 2012 г.) / отв. ред. В. А. Баранов, А. Г. Варфоломеев. — Петрозаводск ; Ижевск, 2012. — 328 с.

ISBN 978–5–8021–1402–5

Сборник содержит материалы конференции, посвященной современным электронным средствам хранения, описания, обработки, исследования и публикации памятников письменности и исторических источников.

УДК 004.9
ББК 81.11+81.2-0

ISBN 978–5–8021–1402–5

PRELIMINARY CONSIDERATIONS CONCERNING THE AUTO-MATED LEMMATISATION OF MIDDLE BULGARIAN TEXTS[1]

Juergen Fuchsbauer

University of Vienna, Austria

The present paper attempts to define the philological preconditions for the digital processing of texts written in Middle Bulgarian with the help of software applicable for other recensions of Church Slavonic. With the Slavonic Dioptra as an example, required adaptations on the levels of graphetics, graphematics, and morphology are discussed.

The Dioptra is a voluminous Greek didactic poem composed as a dialogue of body and soul, which was translated into Middle Bulgarian Church Slavonic around the middle of the fourteenth century. As was first noted by Franz von Miklosich, it contains an abundance of remarkable lexical material, which until now has not been analysed conclusively. Therefore, the bilingual critical edition being currently prepared at Vienna University shall be completed by a dictionary eventually disclosing the lexicon of the poem. In view of its considerable length—the Dioptra consists of approx. 62.000 words—a largely automated lemmatisation appears highly desirable. This requires a device for approximate string matching directly applicable to Middle Bulgarian texts, which, as to my knowledge, for now does not exist. The present paper lists the deviations of the Dioptra from Old Church Slavonic relevant to the automated processing of the text. Its goal is to outline from a philological point of view the prerequisites for an adaptation of approximate string matching techniques developed for other variants of Slavonic[2] to the Dioptra. At that, OCS is unquestionably a more natural point of reference than Old Russian. The results can be expected to be applicable for other Middle Bulgarian texts as well.

Our edition relies on the L'viv manuscript of the Dioptra (LNB NAN imeni Stefanyka MV-418), as this is the only completely preserved Middle Bulgarian testimony of the poem. First of all, in order to allow fuzzy string matching, the software processing the text should be capable of abstracting from certain graphic peculiarities of the ms represented in the print version. Thus, the 12 letters (out of a total of 51 used in our edition) representing positional or arbitrary allographs should be assigned to the superordinate graphemes (є and ꙉ to е;[3] ꙁ to ꙃ; ї and ι to и;[4] ѡ, о,

---

[2] I have in mind the OldEd developed at Izhevsk State Technical University.

[3] The letter ꙁ is preferred after vowels, at the word onset, and at the end of lines, but may occur in any position; ¬ appears only in ¬Tˆ (= ¬стъ) and, occasionally, in ¬„ѡе.

[4] The letter · is frequently, yet not obligatorily, used in front of vowels, but may appear in any position; ¶ is restricted to Greek loanwords (¶„нд·ктиꙟн, ¶„2реи) and names of Greek or Hebrew origin (¶„ппократъ, ¶„2„зек·илъ).

and ѡ to o;[1] ꙋ and ꙋ to оу;[2] ꙑ to ы; and ѵ to у.[3]). Additionally, the lemmata in the dictionary should appear in a corresponding "abstract" form, relieving the reader of some time-consuming guesswork. Of course, the actual spelling is to be preserved in the single entries listed under the respective headwords.

I do not expect the operations necessary for a simplification of that kind to cause much trouble. By contrast, the frequent alternations of graphemes resulting from phonetic shift can be assumed to pose a much bigger challenge both to computational scientists entrusted with the task of adapting existing software to the requirements of Middle Bulgarian, and to philologists processing the data thus gained. I examined the spelling principles of the Middle Bulgarian Dioptra mss in a recent paper in detail;[4] therefore, I shall only give a brief overview here.

Following graphematic alternations appear regularly in the L'viv ms of the Dioptra (and, of course, in many other Middle Bulgarian mss):

ꙃ ~ ꙁ / ꙁ ~ ꙃ: only a few cases contradict the etymological spelling; most of these deviations seem to be lexicalised (e.g., the adjective полеꙁныи, is always spelt with ꙁ, the noun польꙃа, by contrast, unexceptionally with ꙃ).

л ~ ø epenthetic l is comparatively frequently omitted.

ъ ~ ø (/ ь / о): weak ъ may be skipped, but is usually preserved in spelling; it is hardly ever replaced by ь; o-vocalism occurs only in a few words (любовь, начатокъ) and seems to be lexicalised.

ы ~ и: both are only exceptionally mistaken for one another; a few cases of regular, lexicalised commutation occur (нинѣ, посилати).

ь ~ ø / є / ъ: weak ь may be skipped or replaced by ъ, but is usually preserved; strong historic ь appears as є, weak ь vocalised in order to split consonant clusters either as ь or ъ.

ѣ ~ ꙗ: a complementary distribution prevails; ѣ is used after soft consonants, ꙗ at the word onset and at morpheme boundaries; after vowels only ѧ appears.

ѧ ~ ѫ: as a rule, the choice of one of the nasal graphemes is influenced, but not strictly determined, by the quality of the preceding sound; ѧ is preferred at the

---

[1] Both ѡ and 3 may appear in any position; ѡ is clearly preferred at the word onset; 5 is notoriously restricted to the word *oko*.

[2] Digraphic № is by far most common, but may be replaced by ꙋ in any position; ? (an ѵ set above an о) occurs only exceptionally.

[3] The other letters, ѕ, û, and v, exceptionally replace their more frequent counterparts.

[4] Remarks on the Grammar of the Slavonic *Dioptra*. Part I: Orthography and Phonetics" (submitted for the 2012 issue of *Scripta & e-Scripta*).

word onset, after soft consonants and forward vowels, ѫ after hard consonants with a more ambiguous distribution after sibilants and non-forward vowels.

In general, the spelling of the L'viv ms of the Dioptra seems to be fairly consistent and highly lexicalised. Words deviating from a presupposed OCS standard are likely to be spelt in the same way in other occurrences as well—though the total number of possible variations is rather high, only a limited set is realised. This can, once appropriate parameters were defined, be expected to facilitate approximate string matching significantly.

A pivotal point in the automated processing of a text is evidently the correct assignment of inflexion forms. In the following, I give an overview of the desinences present in the Dioptra which do not or not regularly occur in OCS (merely graphematic phenomena covered above are not quoted expressly; e.g. ꙁемлѣ = nom. sg. fem. ja-stem). For comparison I used [Diels, 1963]. Most of these endings are all but uncommon in Middle Bulgarian; not a few occur even sporadically in OCS (those mentioned by Diels are given in italics).

-ѧ   nom. sg. fem. and masc. former ī-stems, which were adopted to the ja-stem-paradigm (млъниѧ, сѫдиѧ)

-є   nom. sg. masc. jo-stems: proper names ending in -ιος in Greek (e.g. григорие)

nom. sg. neutr. of the short form of the part. praet. act. (и дрѣво еѧ ветхо же и иꙁгнивъше; according to [Diels, 1963: 242], also attested in Supr.)

acc. sg. of r-st. (матере, дъщере; according to [Diels, 1963: 178], also in Sav. and Supr.)

nom. pl. of some masc. jo-stems (коне, коваче, прѣлюбодѣе)

-еве   nom. pl. of monosyllabic masc. jo-stems (rare! e.g. врачеве, плачеве, краеве; cf. [Diels, 1963: 159])

-еи   loc. fem. long form of soft adjectives (rare! въ послѣднеи старости; въ прочеи твари)

gen. pl. of masc. jo-stems (e.g. мѫжеи; cf. [Diels, 1963: 159])

-емъ   loc. sg. masc./neutr. of the long form of soft adjectives, comparatives, and part. praes./praet. act. (въ настоѧщемъ житии)

-ехъ   loc. pl. of masc./neutr. jo-stems (въ агньцехъ)

-ие   nom. pl. masc. of jo-stems, especially of those ending in -tel', -ar', and soft monosyllabic roots (e. g. родителие, рыбарие, царие, мѫжие)

-ии   gen. pl. of masc. jo-stems (мѫжии)

-мы   1. pers. pl. of the athemat. verbs (єсмы, вѣмы, имамы, дамы; according to [Ivanova-Mirčeva and Charalampiev, 1999: 134], this ending is already attested in OCS documents)

-ове   nom. pl. of monosyllabic masc. o-stems (e.g. родове; cf. [Diels, 1963: 156]) -

омоу   dat. sg. masc./neutr. of the long form of hard adjectives, comparatives, part. praes. act., praes. pass., praet. act., praet. pass. (e. g. богатѣжцомоу)

-омъ   instr. sg. and dat. pl. of neutr. jo-stems ending in -ie in nom. sg. (искоушениомъ ꙁъмиино и ꙁавистиꙗ диаволеꙗ); rarely also of masc. with a stem ending in a vowel (after the loss of intervocalic j; e.g. къ садоукеомъ, къ иоудеомъ)

loc. sg. masc./neutr. of the long form of hard adjectives (въ четврътомъ сло-

въ) and the part. praes./praet. pass. (въ ... насажденомъ раи)

-охъ   loc. pl. of masc./neutr. o-stems (въ нѣдрохъ; masc. already in OCS, cf. [Diels, 1963: 157])

-(ь)ми   instr. pl. of masc. jo-stems (оучїтелми; according to [Diels, 1963: 157], -ъми is attested with OCS o-stems)

instr. pl. of the neutr. jo-stems ending in -ie in nom. sg. (ѿ вѣщанми)

-ѣмъ   instr. (!) sg. masc./neutr. of hard adjectives (съ шоумомъ велицѣмъ; other wise also as regular loc. form)

-ꙗ   nom./acc. pl. of r-stems (дъщерꙗ)

acc. pl. of masc. n-stems (степенꙗ)

-ѫ(и)   nom. sg. masc. short (long) form of the part. praes. act. replacing -ы(и)

      Many of these morphological innovations, which affected almost exclusively the nominal and adjectival inflexion, were caused by inter-paradigmatic equalisation.[1] Therefore, most of the respective desinences should be readily identifiable for software applicable to OCS as they appear in an either identical or similar form in at least one other paradigm (e.g. -ьми in the ĭ-stems, -ове in the former ŭ-stems). On the other hand, intra-paradigmatic neutralisation (as in -ѣмъ for the instr. sg. of masculine and neuter adjectives) is not common enough to seriously aggravate the problem of homonymy, which can be expected to leave the editor with a lot of manual work anyway. All in all, despite the loss of the casus in the contemporary vernacular, in respect to morphology the Dioptra preserved an artificial standard close to OCS. Therefore a digital processing of the poem does not seem less promising than the processing of OCS or Old Russian texts.

----

[1] We detect a few more isolated instances of unproductive stems adopting desinences of their productive counterparts, that were not incorporated in the list above: n-stem дйнй at least twice took over jo-stem endings (gen. sg. и д,нэ не вэси, otherwise: д,не оного; also: dat. sg. д,нꙗ), ū-stem црйкѵ the a-stem acc. pl. -ѵ (цръквѵ ꙃиздати), and мати the dat. pl. ja-stem -эмъ (матерэмъ).