PRELIMINARY CONSIDERATIONS CONCERNING THE AUTOMATED LEMMATISATION OF MIDDLE BULGARIAN TEXTS

Àâòîð Juergen Fuchsbauer
27.08.2012 ã.

Summary. 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[1] 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 (2 and ¬ to å;[2] s to B; and ¶ to è;[3] w, 3, and 5 to î;[4] Ó and ? to ¹;[5] û to ¥; and v to y.[6]). 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;[7] 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):

B ~ ç / ç ~ B: only a few cases contradict the etymological spelling; most of these deviations seem to be lexicalised (e.g., the adjective ïîëåçí¥è, is always spelt with ç, the noun ïîëBà, by contrast, unexceptionally with B).

ë ~ ø epenthetic
l is comparatively frequently omitted.

ú ~ ø (/ ü / î): weak ú may be skipped, but is usually preserved in spelling;
it is hardly ever replaced by ü; î-vocalism occurs only in a few words (ëþáîâü, íà÷-òîêú) and seems to be lexicalised.

¥
~ è: both are
only exceptionally mistaken for one another; a few cases of regular,
lexicalised commutation occur (íèíý, ïîñëåàòè).

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

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

- ~ ©: as a rule, the choice of one of the nasal graphemes is
influenced, but not strictly determined, by the quality of the preceding sound;
- is preferred at the word onset,
after soft consonants and forward vowels, © 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. çåìëý = 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).

-à nom. sg. fem. and masc. former +-stems,
which were adopted to the ja-stem-paradigm (ìëúíèà, ñ©äèà)

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

  nom. sg. neutr. of the short form of the part. praet.
act. (è äðýâî å-
âåòõî æå è èçãíèâúøå; according to [Diels, 1963: 242], also attested in Supr.)

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

nom. pl. of some masc. jo-stems (êîíå, êîâà÷å, ïðýëþáîäýå)

-åâå nom. pl. of monosyllabic masc. jo-stems (rare! e.g. âðà÷åâå, ïëà÷åâå, êðàââå; cf. [Diels, 1963: 159])

-åè loc. fem. long form of soft adjectives (rare! âú ïîñëýäíåè ñòàðîñòè; âú ïðî÷åè òâàðè)

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

-åìú loc. sg. masc./neutr. of the long form of soft adjectives, comparatives, and part. praes./praet. act. (âú íàñòî©ωåìú æèòèè)

-åõú loc. pl. of masc./neutr. jo-stems (âú àãíüöåõú)

-èå nom. pl. masc. of jo-stems, especially of those ending in -tel’, -ar’, and soft monosyllabic roots (e. g. ðîäèòåëèå, ð¥áàðèå, öàðèå, ì©æèå)

-èè gen. pl. of masc. jo-stems (ì©æèè)

-ì¥ 1. pers. pl. of the athemat. verbs (åñì¥, âýì¥, èìàì¥, äàì¥; according to [Ivanova-Mir eva and Charalampiev, 1999: 134], this ending is already attested in OCS documents)

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

-îìîy dat. sg. masc./neutr. of the long form of hard adjectives, comparatives, part. praes. act., praes. pass., praet. act., praet. pass. (e. g. áîãàòý©ωîìîy)

-îìú instr. sg. and dat. pl. of neutr. jo-stems ending in -ie in nom. sg. (èñêîyøåíèìú çúìèèíN è çàâèñòè- äèàâîëå-); rarely also of with a stem ending in a vowel (after the loss of intervocalic j; e.g. êú ñàäîyêåìú, êú èìyääìú)

loc. sg. masc./neutr. of the long form of hard adjectives (âú ÷åòâðúòîìú ñëîâý) and the part. praes./praet. pass. (âú ... íàñàæä

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

-(ü)ìè instr. pl. of masc. jo-stems (îy÷·òåëìè; according to [Diels, 1963: 157], -úìè is attested with OCS o-stems)

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

-ýìú instr.
(!) sg. masc./neutr. of hard adjectives (ñú øîyìîìú âåëèöýìú; otherwise also as regular loc. form)

-- nom./acc. pl. of r-stems (äúωåð-)

acc. pl. of masc. n-stems (ñòåïåí-)

-©(è) nom. sg. masc. short
(long) form of the part. praes. act. replacing -¥(è)

Many of these morphological innovations,
which affected almost exclusively the nominal and adjectival inflexion, were
caused by inter-paradigmatic equalisation.[8] 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. -üìè in the --stems, -îâå in
the former m-stems). On the other hand, intra-paradigmatic
neutralisation (as in -ýìú 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] I have in mind the OldEd developed at Izhevsk State Technical
University.

[2] The letter 2 is preferred after vowels, at the word onset, and at the
end of lines, but may occur in any position; ¬ appears only in ¬T… (= ¬ñòú) and, occasionally, in ¬„ωå.

[3] The letter is frequently, yet not obligatorily, used in front of
vowels, but may appear in any position; ¶ is restricted to Greek loanwords (¶„íä êòèwí, ¶„2ðåè) and names of Greek or Hebrew origin (¶„ïîîêðàòú, ¶„2„çåê èëú).

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

[5] Digraphic ¹ is by far most common, but may be replaced by Ó in any position; ? (an v set above an î) occurs only exceptionally.

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

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

[8] 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 äüíü at least twice took over
jo-stem endings (gen. sg. è ä‚íý íå âýñè, otherwise: ä‚íå îíîãî; also: dat. sg. ä‚íþ), k-stem öðüê¥ the a-stem
(öðúêâ¥ Bèçäàòè), and ìàòè the dat. pl. ja-stem -ýìú (ìàòåðýìú).