Раздел 3
ЯЗЫКОЗНАНИЕ

УДК 81′0:091(054)
DOI: 10.22213/2618-9763-2021-4 82-89

В. А. Баранов, доктор филологических наук, профессор
Ижевский государственный технический университет имени М. Т. Калашникова, Ижевск, Россия
Р. М. Гнутиков
Удмуртский государственный университет, Ижевск, Россия
К. И. Зинатшин, студент
Ижевский государственный технический университет имени М. Т. Калашникова, Ижевск, Россия

КОДИРОВАНИЕ И ПЕРЕКОДИРОВАНИЕ ТРАНСКРИПЦИЙ ИСТОРИЧЕСКОГО КОРПУСА «МАНУСКРИПТ»

Рассматриваются славянские диапазоны стандарта Unicode с точки зрения возможности создания на их основе транскрипций, передающих графику средневековых славянских рукописей. Обращается внимание на наличие в стандарте вариантов кирилловских букв, что позволяет достаточно точно передавать графические особенности рукописей. В связи с тем, что до сегодняшнего дня в стандарте отсутствуют варианты некоторых букв, существует необходимость использования дополнительных соглашений кодирования символов, коды которых размещены не в стандартных, а в специальном и личном диапазонах Юникода.

Примером большого собрания машиночитаемых копий средневековых славянских письменных памятников является исторический корпус «Манускрипт» (manuscripts.ru), созданный на базе СУБД Oracle с помощью специализированной кодово-шрифтовой системы. Миграция корпуса на иные технологические платформы, использование для анализа лингвистических данных (отдельных текстов, подкорпусов, выборок) внешних программных средств возможны только после перекодирования выгружаемых файлов в стандартные диапазоны Unicode.

Сопоставительный анализ использующихся в корпусе наборов символов и в действующей 14-й версии стандарта позволяет сделать вывод: перекодирование или приводит к потере части графических особенностей, или требует использования дополнительного набора вариантных символов с кодами в личном диапазоне Unicode.

Анализируются случаи, когда в стандарте Unicode имеется два или более символов для перекодировки одного символа корпуса «Манускрипт», указывается, что в стандарте и в наборе символов дополнительного личного диапазона отсутствуют многие лигатуры, а также некоторые единичные особые графемы.

Ключевые слова: лингвистический корпус; славянские средневековые рукописи; транскрипция; кодировка.

Введение

Создание исторических корпусов, обеспечивающих демонстрацию лингвистического материала (конкордансов, контекстов, перечней и др.), а также его разметку, обработку, поиск, упорядочение и анализ, предполагает, в первую очередь, кодирование (транскрибирование) средневековых рукописей и старопечатных книг. Понятно, что машиночитаемая транскрипция в таких случаях должна адекватно отражать графику оригинала, что обеспечивается использованием в том числе и таких букв, диакритических символов и знаков оформления текста, которые отсутствуют в современных текстах.

Используемая в настоящее время система кодировки консорциума Юникод [1] включает коды не только современных букв кирилловских алфавитов, но и большое количество исторических символов (букв, диакритических знаков и др.), которые позволяют достаточно точно передать графические особенности средневековых славянских письменных памятников.

Особенности кодирования средневековых славянских письменных памятников

Стандарт Unicode

В актуальной 14-й версии Юникода славянские кириллические символы расположены в пяти диапазонах:

– основном (U+0400–U+04FF)[1],
– дополнительном (U+0500–U+052F)[2],
– расширенном A (U+2DE0–U+2DFF)[3],

[1] https://www.unicode.org/charts/PDF/U0400.pdf (дата обращения: 17.11.2021).
[2] https://www.unicode.org/charts/PDF/U0500.pdf (дата обращения: 17.11.2021).
[3] https://www.unicode.org/charts/PDF/U2DE0.pdf (дата обращения: 17.11.2021).

– расширенном B (U+A640–A69F)[1],
– расширенном C (U+1C80–1C8F)[2].

Историческая часть диапазонов, появившаяся уже в первой версии стандарта и постепенно пополнявшаяся, включает отсутствующие в современном русском алфавите буквы-монографы, например малый и большой нейотированные и йотированные юсы (Ѧ U+0466, ѧ U+0467, Ѩ U+046A, ѩ U+046B, Ѥ̀ U+0468, ѥ̀ U+0469, Ѩ́ U+046C, ѩ́ U+046D), ять (Ѣ U+0462, ѣ U+0463), фиту (Ѳ U+0472, ѳ U+0473), кси (Ѯ U+046E, ѯ U+046F), пси (Ѱ U+0470, ѱ U+0471) и другие, буквы диграфы, например ук (Ꙋ U+0478, ꙋ U+0479), от (Ѹ U+047E, ѹ U+047F), еры (Ъ̀І A650, ъ̀і A651) и другие, лигатуры, например ук (Ꙗ U+A64A, ꙗ U+A64B, ᲈ U+1C88), буквы палатальных согласных (Ґ 04A4, ґ 04A5, Ꙉ A662, ꙉ A663, Ꙟ A664, ꙟ A665, Ԍ A666, ԍ A667), титла (титло ◌̑ U+0483, взмет ◌͡ U+A66F, покрытие ◌͜ U+0487), знаки чисел (тысяча ◌҂ U+0482, сто тысяч ◌⃝ U+0488 и другие), знаки придыхания (◌̔ U+0485, ◌̓ U+0486) и некоторые другие символы.

Согласно принципам консорциума, кодированию подлежат символы письма, включенные в современные национальные алфавиты или исторические системы письма и имеющие собственную функцию – обозначение особой фонемы, использование в качестве модификатора, знака сокращения, ударения и под. Различные начертания (*glyph*) одного символа (например, печатная и скорописная буква «т» – т, *т*) считаются одной буквой (*character*), и им не должны присваиваться индивидуальные коды. Считается, что различия в начертании одного символа, его положении в строке и т. п. должны передаваться или гарнитурой шрифта, или с помощью функций текстового процессора, или дополнительной разметкой – тегами. Несмотря на это славянские диапазоны сегодня содержат варианты одних и тех же символов (как, впрочем, и латинские), что связано, с одной стороны, со сложностью определения функциональной значимости особого начертания того или иного символа в средневековой письменности, с другой – с желанием предоставить создателям транскрипций варианты, регулярно использовавшиеся писцами наряду с основными начертаниями. Так, славянские диапазоны содержат сегодня четыре варианта буквы омега (Ѡ U+0460, ѡ U+0461, O U+047A, o U+047B, Ѿ U+047C, ѿ U+047D, Ꙍ U+A64C, ꙍ U+A64D), три варианта буквы ук (Ꙋ U+0478, ꙋ U+0479, Ꙗ U+A64A, ꙗ U+A64B, ᲈ U+1C88), а также надстрочные варианты этих (ᵚ U+A67B, ᷸ 2DF9) и некоторых других кириллических букв, скорописные варианты букв вѣдѣ (є 1C80), добро (ᲁ 1C81), твьрдо (ᲃ 1C84, ᲅ 1C85), еръ (ᲆ 1C86), ять (ᲇ 1C87) и др.

Безусловно, наличие вариантов позволяет более точно передать в транскрипции рукописный оригинал и тем самым с помощью корпусных методов обеспечить анализ общих и частных графико-орфографических особенностей рукописей. Поэтому можно только приветствовать регулярные дополнения в славянскую часть *Unicode*, даже если при этом нарушается основной принцип формирования стандарта – не кодировать начертательные варианты букв (*character*).

В то же время сегодняшний перечень в *Unicode* славянских букв, необходимых для подготовки максимально точных транскрипций, нельзя считать полным: отсутствует достаточно большое количество вариативных символов, которые логично было ввести в стандарт, точно так же, как это уже сделано с другими вариантными символами, например к зеркальным вариантам букв ци и ю – Ꙇ A660, ꙇ A661, Ꙩ A654, ꙩ A655, добавить зеркальные буквы аз, ук, которые встречаются в рукописях и берестяных грамотах, включить в стандарт недостающие выносные буквы, а также другие символы, в том числе встретившиеся в списках единичное количество раз, например хорошо известный «полуйотированный» азъ, см. строки 1 сн. и 6 сн. на л. 142 об. и строку 19 на л. 143 в Остромировом Евангелии 1056–1057 гг. (РНБ, F.п.1.5), обнаруженный О. С. Пайминой зеркальный йотированный юс малый ѧ̈, см. строки 14 и 17 на л. 67 об. Троицкого сборника XII–XIII вв. (РГБ, Тр. 12) [2, с. 144] и др.

В связи с очень большими трудозатратами на создание (набор, сверку, корректуру) транскрипций средневековых рукописей вопрос о предоставляемых стандартом *Unicode* возможностях кодирования в соответствии с оригиналом и ограничениях при этом становится первоочередным при планировании работ и использовании подготовленных машиночитаемых ресурсов в изданиях, коллекциях и корпусах.

Ситуация осложняется тем, что работы по созданию электронных машиночитаемых ресурсов на основе средневековых славянских рукописей начались давно – более трех-четырех десятилетий

[1] https://www.unicode.org/charts/PDF/UA640.pdf (дата обращения: 17.11.2021).
[2] https://www.unicode.org/charts/PDF/U1C80.pdf (дата обращения: 17.11.2021).

назад. Судя по тому, что первые такие издания и коллекции появились в Интернете в середине 90-х годов прошлого века (см., например, электронные издания Пражских глаголических отрывков[1] и глаголических Киевских листков[2], Супрасльской рукописи[3] и Мариинского Евангелия[4] проекта Titus, Хельсинский корпус старославянских рукописей[5], Санкт-Петербургский корпус агиографических текстов[6] и некоторые другие), начало их подготовки восходит к 70-м и 80-м годам. Первые транскрипции создавались на основе символов латинского алфавита, при этом для передачи собственно славянских фонем и исторических букв использовалась специальная разметка. После 1991 г., года официального введения Юникода как международного стандарта, появилась возможность создавать транскрипции на основе кодировок собственно славянских символов. Но только с той степенью точности, которую предоставлял стандарт в период подготовки транскрипции. Например, йотированный аз (Ѣ̈ U+A656, ѣ̈ U+A657) появился только в 5-й версии стандарта, вышедшей в 2008 году.

Выходов из ситуации, при которой отсутствуют некоторые символы, необходимые для подготовки транскрипции, максимально точно передающей рукописный оригинал, может быть несколько:

1) упрощение транскрипции с последующим ее уточнением после введения в стандарт необходимых символов;

2) использование тегов для модификации основных вариантов;

3) разработка согласованного с точки зрения кодировок перечня символов за пределами стандартных диапазонов *Unicode*;

4) создание собственной кодовой таблицы символов.

Дополнения к стандарту Unicode

В 2010 г. в журнале *Scripta & e-Scripta* была опубликована коллективная статья специалистов в области прикладной славистики из разных стран [3], в которой обосновывалась необходимость кодирования отсутствующих в *Unicode* символов кодами дополнительных областей для специального (*Supplementary Special-purpose Plane*, U+E000–U+EFFF) и личного (*Supplementary Private Use Area planes*, U+F000–U+FFFF) использования. В статье предложена кодировка 177 символов:

38 заглавных,
35 строчных,
88 надстрочных букв (U+F330–U+F3D3),
2 титл (U+EF60–U+EF61),
5 диакритических знаков (U+EF63–U+EF66),
2 знаков порчи (U+EF67–U+EF68),
7 знаков пунктуации (U+EF70–U+EF76).

Опубликованные предложения были разработаны для того, чтобы предоставить авторам создаваемых транскрипций согласованные коды символов, которые отсутствуют в стандартных диапазонах.

В настоящее время в действующей версии *Unicode* зарегистрировано несколько символов, включенных в предложения 2010 г.: широкая строчная буква слово с U+F361 → U+1C83[7], палатальный согласный нашь Ҥ U+F355 → → U+04A4, ҥ U+F356 → U+04A5, ук лигатура ᵹ U+F372 → U+A64A, ᵺ U+F373 → U+1C88, надстрочные буквы широкое есть (украинское е) ᵋ U+F339 → U+A674, восьмеричное и ᵎ U+F342 → U+A675, десятеричное і ᵎ̈ U+F347 → U+A676, ук монограф ᵞ U+F364 → → U+A677, омега ᵚ U+F37C → U+A67B, ер ъ U+F389 → U+A678, один из диграфных еров ы U+F389 → U+A678, ерь ь U+F39E → U+A67A, йотированный есть ᵸ U+F3A8 → U+A69F, всего 14 буквенных символов. (С большой долей вероятности в Юникод постепенно будут добавляться и другие символы предложений 2010 г.).

Еще одна задача может быть решена с помощью предложений 2010 г. – перекодирование существующих электронных машиночитаемых ресурсов, созданных не на основе *Unicode*, в кодировки стандарта без потери точности передачи оригинала.

Кодировка транскрипций исторического корпуса «Манускрипт»

Первые транскрипции будущего исторического корпуса «Манускрипт» (manuscripts.ru) начали создаваться в начале 90-х годов. Для набора текстов использовался текстовый процессор *ChiWriter*, для обработки лингвистических данных – СУБД *Paradox*. С помощью редактора шрифтов *ChiWriter* был разработан шрифт *Putiata* (*Putyata*) в стандартной кодировке

[1] http://www.schaeken.nl/lu/research/online/editions/pragfrag.html (дата обращения: 17.11.2021).
[2] http://www.schaeken.nl/lu/research/online/editions/kievfol.html (дата обращения: 17.11.2021).
[3] https://titus.uni-frankfurt.de/texte/etcs/slav/aksl/suprasl/supra.htm (дата обращения: 17.11.2021).
[4] https://titus.uni-frankfurt.de/texte/etcs/slav/aksl/marianus/maria.htm (дата обращения: 17.11.2021).
[5] https://korp.csc.fi/download/ccmh-src/www/index.html (дата обращения: 17.11.2021).
[6] http://project.phil.spbu.ru/scat/page.php?page=project (дата обращения: 17.11.2021).
[7] Где U+F361 – код предложений 2010 года, U+1C83 – код стандарта Unicode.

MS DOS CP866, включающий все символы, необходимые для создания максимально приближенной к оригиналу машиночитаемой копии текста. Подготовленные транскрипции были воспроизведены в печатном издании служебной минеи на май («Путятиной минеи»), XI в., 135 л., РНБ, Соф. 202 [4], в котором, помимо двух видов текста, были опубликованы слово- и формоуказатели, созданные с помощью программ обработки лингвистических данных на базе СУБД *Paradox* (программист А. Н. Миронов, конвертирование В. А. Романенко).

Интернет-версия будущего печатного издания появилась в 2001 г. [5]. Для ее демонстрации была разработана и создана кодово-шрифтовая система «Манускрипт» (КШС «Манускрипт») (разработчик В. А. Романенко), включающая перечень кодов для кирилловских символов и семейство шрифтов *Menaion* (дизайнеры А. В. Шарова, В. А. Баранов), что обеспечило точную графическую передачу текста и указателей на сайте (веб-интерфейс А. А. Вотинцев, администратор БД С. В. Ощепков). В 2004 г. был открыт портал «Манускрипт: славянское письменное наследие» [6], на котором в течение нескольких лет, кроме второй версии электронного издания Путятиной минеи, были опубликованы и другие древнейшие славянские рукописи, конвертированные из формата СУБД *Paradox* в формат СУБД *Oracle*, а с помощью специализированного редактора *OldEd*, обеспечивающего не только ввод текстов, но и их корректуру и разметку, продолжилось создание транскрипций, воспроизводящих славянские рукописи XI–XV веков.

В настоящее время исторический корпус «Манускрипт» (manuscripts.ru) представляет собой многофункциональный интернет-ресурс, содержит более 140 полных транскрипций славянских рукописей и отрывков X–XV веков объемом более 3,5 млн словоупотреблений, обеспечен мета-, аналитической и лингвистической разметкой и процедурами и программами ввода, редактирования, обработки, поиска и демонстрации данных, что позволяет решать различные историко-лингвистические задачи.

Основой корпуса являются транскрипции текстов, подготовленные с помощью КШС «Манускрипт», коды символов которой располагаются в дополнительной области для специального использования (U+E000–U+EFFF). Применение веб-шрифтов позволило снять проблему визуализации текстов, указателей и других форм представления данных корпуса при отсутствии на локальном компьютере пользователя шрифта *Menaion* с кодировкой символов в нестандартных диапазонах.

Буквенные символы старославянского кирилловского алфавита в КШС «Манускрипт» располагаются в диапазоне U+E000–U+E0EA, глаголического – в диапазоне U+E700–U+EA51 и содержат заглавные, строчные и надстрочные символы, лигатуры в диапазонах U+E100–U+E14E, U+ED00–U+ED69, вариантные буквы – U+E154–U+E19F, символы чисел – U+E1A0–U+E1A4, диакритические символы – U+E200–U+E342, U+EF63–U+EF66, титла – U+E400–U+E41E, U+EF60–U+EF61, U+FE20–U+FE26, небуквенные знаки в строке – U+E519–U+E5C8, U+EC01–U+EC11, знаки порчи – U+E600–U+E603, U+EF67–U+EF68. Кроме этого, КШС содержит и символы стандартных диапазонов: символы основной и дополнительной латиницы, расширенной латиницы A, расширенной латиницы B, модификаторы, комбинированные диакритические знаки, греческие символы, символы современной и исторической части кириллицы, расширенной кириллицы A, расширенной кириллицы B, диакритические знаки для греческих текстов, знаки пунктуации, надстрочные и подстрочные символы и др., в общей сложности более 1800 символов.

До сих пор использование для транскрибирования текстов символов с нестандартными кодами имело только одно ограничение: при отсутствии на локальном компьютере пользователя установленного шрифта *Menaion* скопированные с сайта тексты или указатели отображались неверно.

В то же время желание анализировать транскрипции рукописей и подготовленные выборки с помощью внешних программ, использовать имеющиеся в корпусе лингвистические ресурсы на других технологических платформах (см., например, [8–11]) побудило начать технологическую подготовку конвертирования данных в стандартные диапазоны *Unicode*.

Цель предпринятой работы – подготовка конвертирования транскрипций исторического корпуса «Манускрипт», созданных в личном диапазоне *Unicode*, в кодировку стандартных диапазонов.

Для этого необходимо установить соответствия символов в личном и стандартных диапазонах, в случае наличия вариантов перекодировки обосновать предпочтительность одного из них, в случае отсутствия точного соответствия предложить наиболее приемлемый вариант кодировки, создать шрифт для представления текстов в формате стандартных диапазонов *Unicode* на основе шрифта *Menaion*.

Перекодировка транскрипций: соответствия, варианты соответствий, отсутствие соответствий и пути выхода из ситуации

Процедура установления соответствий для большего количества символов проста: соответствуют друг другу символ личного и символ стандартного диапазонов в том случае, если они являются соответствующими друг другу буквами в старославянском и современном алфавитах (имеют одно фонологическое значение), даже если их изображения (глифы) отличаются. Так, соответствуют друг другу инициальная буква азъ **Ⱑ** U+E001 и заглавная **А** U+0410, строчная есть **ⷺ** U+E056 и строчная **е** U+0435, строчная восьмеричная и **ⷭ** U+E05D и строчная **и** U+0438 и др. Использование для визуализации текста фонтов, имеющих в стандартной кирилловской части глифы старославянского алфавита, позволяет представить текст в итоговом файле в соответствии с исходной транскрипцией и оригиналом.

Понятно, что аналогичным образом сопоставляются буквы и символы исторической части, например, строчный ять **ⷦ** U+E086 → U+0463, строчный юс малый **ⷢ** U+E08B → U+0467, строчная пси **ⷲ** U+E092 → U+0471, титло **◌҃** U+E400 → U+0483, знак мягкости **◌҆** U+E242 → U+0484, символ числа «колода» **◌꙱** U+E1A1 → U+A671 и др.

В стандарте *Unicode* в настоящее время размещены несколько кирилловских символов, которые в истории славянской письменности на определенных этапах были вариантами одной буквы и могли использоваться в одном и том же документе. Речь идет о буквах:

зэ **З** U+0417, **з** U+0437 и земля **ⷵ** U+A640, **ⷶ** U+A641,

о **о** U+043E и он узкий **ⷪ** U+1C82,

эс **с** U+0441 и слово широкое **ⷯ** U+1C83,

вэ **в** U+0432 и круглая в**ⷦ**д**ⷦ** **ⷴ** U+1C80,

дэ **д** U+0434 и добро с длинными ножками **ⷣ** U+1C81,

тэ **т** U+0442, твердо с высокой мачтой **ⷮ** U+1C84 и твердо с тремя ножками **ⷰ** U+1C85,

твердый знак **ъ** U+044A и ер с высокой мачтой **ⷱ** U+1C86,

ять **ⷦ** U+0463 и ять с высокой мачтой **ⷷ** 1C87 и др.

Отношения в этих парах (рядах) различны: в одних случаях современная буква была основным вариантом в славянской письменности эс – слово, вэ – в**ⷦ**д**ⷦ**, дэ – добро, тэ – твердо, твердый знак – еръ и др., в других наоборот – современная круглая зэ была вариантом буквы земля, современная о – широким вариантом узкого она. В связи с этим во втором случае возникает ситуация выбора буквы: земля **ⷶ** U+E05C исходной транскрипции может быть конвертирована как в зэ **з** U+0437, что соответствует общему правилу перекодировки символов из личного диапазона в стандартный, т. к. современная буква зэ является функциональной преемницей буквы земля, так и в **ⷶ** U+A641, что дает возможность различать в итоговой транскрипции буквы земля с хвостом и земля с круглыми петлями. Аналогичная ситуация с буквами о и оном узким: перекодирование буквы он **ⷫ** U+E019, **ⷪ** U+E069 исходной транскрипции в **О** 041E, **о** U+043E стандартного диапазона (не в узкий он **ⷪ** U+1C82) приведет к тому, что необходимо будет искать в стандартном диапазоне код для широкого она **О** U+E01A, **о** U+E06A, который, строго говоря, отсутствует (использование **О** U+047A, **о** U+047B проблематично в связи с тем, что в стандарте это варианты буквы омега).

Конвертирование букв **земля** и **он** исходной транскрипции в буквы исторических диапазонов, соответствующих с точки зрения изображения (глифа) оригиналу и позволяющих использовать букву современного алфавита для отображения вариантной буквы **земля округлая** или **он широкий**, как будто представляется предпочтительным. Такое решение ведет и к нежелательным последствиям: кодировка лингвистических единиц с этими буквами в итоговых транскрипциях будет отличаться от кодировки тех же слов в других электронных ресурсах, ср., например, два вида слова зло – **зъло** U+0437 U+044A U+043B U+043E и **ⷶ**ъ**ⷧ**ⷫ U+A641 U+044A U+043B U+1C82. Иначе говоря, необходимость сохранить в итоговой транскрипции графику оригинала вступает в противоречие с возникающими при этом различиями в кодировке одних и тех же лингвистических единиц. Аналогичные отношения существуют между современной у **У** U+0423 и **у** U+0443 и историческим диграфным уком **ⷫ**ⷭ U+0478 и **ⷪ**ⷭ U+0479, для которого монограф ук **ⷭ** был ранним вариантом; между современной ща **Щ** U+0429, **щ** U+0449 и историческим диграфным шта **Ш**ⷮ U+0428 U+0422, **ш**ⷮ U+0448 U+0442; современной буквой ы **Ы** U+042B, **ы** U+044B и историческими ерами **Ъ**І U+A650, **ъ**і U+A651.

Решение возникающей при перекодировке некоторых символов проблемы видится в создании дополнительных процедур и программ, обеспечивающих обработку лингвистических данных с учетом различий в их кодировке,

в создании электронных словарей, унифицирующих графику и орфографию одних и тех же лингвистических единиц.

Предложения 2010 г. [7] содержат большое количество символов, встречающихся в рукописях, но отсутствующих в стандарте *Unicode*. При их перекодировке из внутреннего формата корпуса «Манускрипт» во внешний существует два решения: кодировать символы, используя коды предложений 2010 г., или использовать символы стандартных диапазонов. Так, зеркальное йотированное есть **ⷉ** U+E191, **ⷊ** U+E192 можно кодировать или **Ҿ** U+0464, **ҿ** U+0465, или **ⷉ** U+F3A9, **ⷊ** U+F3AA, в первом случае потеряв зеркальность, во втором – требуя от пользователя использовать фонт, поддерживающий символы предложений 2010 г. Подобная ситуация и во многих других случаях: приведения к основному варианту буквы или использования символов предложений 2010 г., требуют зеркальный **аз** **ⷡ**, мягкий глаголь **ᲂ**, мягкий како **ᲄ**, мягкий хер **ᲅ**, переставленный ук **ᲆ**, переставленный зеркальный ук **ᲇ**, перевернутая шта **ᲈ**, лигатурные еры **ъы**, полумягкий йотированный аз **ⷡ**, переставленный йотированный есть **ⷤ**, зеркальный йотированный юс большой **ⷬ**, закрытая омега **ꙍ** и другие, а также надстрочные буквы, отсутствующие в стандартных диапазонах, – мягкое добро **ᲀ**, широкое есть (украинское) **ᲁ**, зело **ᲃ** и зеркальное зело **Ᲊ**, мягкие глаголь **ᲂ**, како **ᲄ**, людие **ᲊ**, хер **ᲅ**, наш **ᲇ**, очное о **ᲈ**, двуочное о **Ᲊ**, еры **ᲊ** и др. Все эти буквы имеют собственный код в предложениях 2010 г., и при конвертировании соответствующих символов в эти коды транскрипция сохранит свое соответствие оригиналу.

Особой задачей является конвертирование во внешний файл многочисленных лигатур, представляющих собой объединение в одном символе двух букв: **ав** **ⷢ** U+ED01, **аг** **ⷣ** U+ED03, **лу** **ⷥ** U+ED05, **нг** **ⷮ** U+ED43, **нк** **ⷯ** U+ED49, **пн** **ⷱ** U+ED4D и многих др.

В отличие от лигатур ук **ꙋ** U+A64A, **ꙋ** U+A64B, **ᱸ** U+1C88, представленных в *Unicode* и имеющих диграфные соответствия **оу** U+0478, **оу** U+0479, лигатуры **ав** **ⷢ** U+ED01 и подобные отсутствуют в стандарте, и, насколько нам известно, их кодирование не планируется. Это понятно: полный перечень лигатур в средневековых текстах не установлен, а кроме того, современные типографские возможности шрифтов семейства *OpenType* позволяют создавать необходимые для отображения на печати или на экране лигатурные объединения букв.

В простом варианте (без использования тегов) перекодирования символов из внутреннего формата корпуса «Манускрипт» во внешний неизбежна потеря информации о таких лигатурах.

К сожалению, информация может быть потеряна и в случае наличия в рукописи символов, не зарегистрированных ни в стандарте *Unicode*, ни в Предложениях 2010 г. В корпусе «Манускрипт» это два символа – полуйотированный азъ **ⷠ** Остромирова Евангелия 1056–1057 гг. и зеркальный йотированный юс малый **ⷭ** Троицкого сборника XII–XIII вв.

Есть еще одна сторона проблемы перекодирования в стандартные диапазоны Юникода. Понятно, что правильное отображение всех символов итоговой транскрипции возможно только в случае их наличия в выбранном для визуализации текста фонте. В неспециализированных фонтах могут отсутствовать некоторые исторические славянские символы, их изображение (глиф) может не соответствовать изображению буквы в средневековом славянском тексте. Для точной визуализации транскрипций можно рекомендовать шрифты, размещенные на сайтах проектов «Пономарь» [8], «Кодекс» [9], Манускрипт» [10].

Выводы

Обеспечение миграции лингвистических ресурсов между разными технологическими платформами, которая предоставляет возможность использования различных инструментов для их анализа и увеличивает вероятность долговременного их сохранения, возможно только при создании транскрипций на основе стандартов кодирования и разметки.

При подготовке машиночитаемых копий средневековых славянских письменных памятников, начавшейся более 30–40 лет назад, использовались различные форматы кодирования. Часть созданных ресурсов переведена в кодировку стандартных диапазонов Юникода, часть, например, транскрипции корпуса «Манускрипт», продолжает использоваться в нестандартной кодировке.

Анализ показал, что перекодирование таких ресурсов представляет собой нетривиальную задачу, решение которой затрудняется, в первую очередь, несовпадением состава символов исходных транскрипций, максимально приближенных к рукописным оригиналам, и перечнем кирилловских символов, включенных в стандарт Юникод. Одним из возможных решений является использование при перекодировке личного диапазона Юникода, кодирование символов в котором предложено группой специали-

стов в области прикладной палеославистики в 2010 г.

Библиографические ссылки

1. Unicode // The Unicode Consortium. URL: https://home.unicode.org/ (дата обращения: 03.11.2021).

2. *Паймина О. С.* Языковые особенности Троицкого сборника XII–XIII вв. : дис. … канд. наук: 10.02.01 – Русский язык. Казань : КГУ, 2012. 326 с.

3. Proposal for a unified encoding of Early Cyrillic glyphs in the Unicode Private Use Area / Victor Baranov, David J. Birnbaum, Ralph Cleminson, Heinz Miklas, Achim Rabus // Scripta & e-Scripta: The Journal of Interdisciplinary Mediaeval Studies. Vol. 8-9. Sofia : “Boyan Penev” Publishing Center ; Institute of Literature, BAS, 2010. S. 9–26. URL: https://clck.ru/YeZyU (дата обращения: 03.11.2021).

4. Новгородская служебная минея на май (Путятина минея). XI век: Текст, исследования, указатели / подг. В. А. Баранов, В. М. Марков. Ижевск : Издат. дом «Удмуртский университет», 2003. 788 с.

5. Путятина минея / подг. В. А. Баранов, В. М. Марков; ЛАФИ УдГУ. 2001. 2001. URL: http://manuscripts.ru/ptm/; http://manuscripts.ru/mns/portal.main?p1=19&p_lid=1 (дата обращения: 03.11.2021).

6. Манускрипт: славянское письменное наследие / ИжГТУ имени М. Т. Калашникова, УдГУ ; коллектив авторов. URL: http://manuscripts.ru/ (дата обращения: 03.11.2021).

7. Proposal for a unified encoding of Early Cyrillic glyphs in the Unicode Private Use Area / Victor Baranov, David J. Birnbaum, Ralph Cleminson, Heinz Miklas, Achim Rabus // Scripta & e-Scripta: The Journal of Interdisciplinary Mediaeval Studies. Vol. 8-9. Sofia : “Boyan Penev” Publishing Center ; Institute of Literature, BAS, 2010. S. 9–26. URL: https://clck.ru/YeZyU (дата обращения: 03.11.2021).

8. Ponomar Project. URL: https://ponomar.net/ (дата обращения: 03.11.2021).

9. Kodeks Project / Sebastian Kempgen. URL: https://kodeks.uni-bamberg.de/AKSL/AKSL.Schrift.htm (дата обращения: 03.11.2021).

10. Манускрипт: славянское письменное наследие / ИжГТУ имени М. Т. Калашникова, УдГУ; коллектив авторов. URL: http://manuscripts.ru/ (дата обращения: 03.11.2021).

References

1. Unicode. The Unicode Consortium. Available at: https://home.unicode.org/ (accessed 03.11.2021).

2. Pajmina O. S. *Jazykovye osobennosti Troickogo sbornika XII–XIII vv., dissertacija na soiskanie uchenoj stepeni kandidata filoligicheskih nauk: 10.02.01 – Russkij jazyk* [Linguistic features of the Trinity collection of the XII-XIII centuries. , dissertation for the degree of candidate of philological sciences: 10.02.01 - Russian language]. Kazan, KSU Publ., 2012, 326 p. (in Russ.).

3. Victor Baranov, David J. Birnbaum, Ralph Cleminson, Heinz Miklas, and Achim Rabus. Proposal for a unified encoding of Early Cyrillic glyphs in the Unicode Private Use Area. Scripta & e-Scripta: The Journal of Interdisciplinary Mediaeval Studies. Vol. 8-9. Sofia : “Boyan Penev” Publishing Center ; Institute of Literature, BAS, 2010, pp. 9-26. Available at: https://clck.ru/YeZyU (accessed 03.11.2021).

4. Victor A. Baranov, Vitaliy M. Markov. *Novgorodskaja sluzhebnaja mineja na maj (Putjatina mineja). XI vek: Tekst, issledovanija, ukazateli* [Novgorod service Menaion for May (Putyata Menaion). XI century: Text, research, indexes]. Izhevsk, Izdatel'skij dom “Udmurtskij universitet”, 2003, 788 p. (in Russ.).

5. Victor A. Baranov, Vitaliy M. Markov. *Putjatina mineja* [Putyata Menaion], 2001. Available at: http://manuscripts.ru/ptm/ (accessed 03.11.2021). (in Russ.).

6. Kalashnikov ISTU, UdSU. *Manuskript: slavjanskoe pis'mennoe nasledie* [Manuscript: Slavonic Written Heritage]. Available at: http://manuscripts.ru/ (accessed 03.11.2021). (in Russ.).

7. Victor Baranov, David J. Birnbaum, Ralph Cleminson, Heinz Miklas, and Achim Rabus. Proposal for a unified encoding of Early Cyrillic glyphs in the Unicode Private Use Area. Scripta & e-Scripta: The Journal of Interdisciplinary Mediaeval Studies. Vol. 8-9. Sofia : “Boyan Penev” Publishing Center ; Institute of Literature, BAS, 2010, pp. 9-26. Available at: https://clck.ru/YeZyU (accessed 03.11.2021).

8. *Ponomar Project*. Available at: https://ponomar.net/ (accessed 03.11.2021).

9. Kempgen Sebastian. *Kodeks Project*. Available at: https://kodeks.uni-bamberg.de/AKSL/AKSL.Schrift.htm (accessed 03.11.2021).

10. Kalashnikov ISTU, UdSU. *Manuskript: slavjanskoe pis'mennoe nasledie* [Manuscript: Slavonic Written Heritage]. Available at: http://manuscripts.ru/ (accessed 03.11.2021) (in Russ.).

V. A. Baranov, Doctor of Philology, Professor
Kalashnikov Izhevsk State Technical University, Izhevsk, Russia
R. M. Gnutikov
Udmurt State University, Izhevsk, Russia
K. I. Zinatshin, Student
Kalashnikov Izhevsk State Technical University, Izhevsk, Russia

ENCODING AND TRANSCODING OF TRANSCRIPTIONS OF THE HISTORICAL CORPUS “MANUSCRIPT”

The article considers capabilities of using Cyrillic blocks of the Unicode Standard for the purpose of creating transcriptions, which would represent graphics of medieval Slavonic manuscripts. In addition, much attention is given to the fact that the Unicode Standard provides variants of Cyrillic letters, which means that one can accurately enough record graphic features of manuscripts. However, some variants of certain letters are still missing, and that is why there exists a need to use additional agreements of character encoding, which code points are placed in special blocks and Private Use Areas and not in standard ranges of Unicode.

The Manuscript – a historical corpus – is the example of a big machine-readable collection of medieval Slavonic manuscripts. It was created on the base of Oracle DBMS with the use of a specialized system of codes and fonts. Transference of the corpus to other technological platforms or usage of external software (including separate texts, parts of corpora, selections) for analysis of linguistic data would be possible only after downloaded files are recoded to the Unicode Standard.

A comparative analysis of the character blocks used in the corpus and in the current version 14.0 of the Unicode Standard leads to the conclusion that recoding either results in losses of graphic features or requires usage of a supplementary set of varying characters with code points of Private Use Areas.

Instances when there are two or more characters of the Unicode Standard that correspond to one recoded character of the Manuscript are analyzed. It is also stated that numerous ligatures and certain singular graphemes are missing in the standard blocks and in the blocks of Private Use Areas.

Keywords: text corpus; Slavonic medieval manuscripts; transcription; encoding.

Получено: 15.11.21

Образец цитирования

Баранов В. А., *Гнутиков Р. М.*, *Зинатшин К. И.* Кодирование и перекодирование транскрипций исторического корпуса «Манускрипт» // Социально-экономическое управление: теория и практика. 2021. Т. 17, № 4. С. 82–89. DOI: 10.22213/2618-9763-2021-4-82-89.

For Citation

Baranov V. A., Gnutikov R. M., Zinatshin K. I. [Encoding and Transcoding of Transcriptions of the Historical Corpus “Manuscript”]. *Social'no-jekonomicheskoe upravlenie: teorija i praktika*, 2021, vol. 17, no. 4, pp. 82-89 (in Russ.). DOI: 10.22213/2618-9763-2021-4-82-89.