И.С.Маргулис, ИПИ АН РТ, ЦИТ КГУ
ilya_margulis@mail.ru

Заявка на использование материалов проекта «Гиперазбука» в целях подготовки планов корректировки кириллических и глаголического диапазонов Unicode.

Проект
ГИПЕРАЗБУКА
Объединённый алфавит славянских символов

Содержание

1. Спецификация проекта

Цель проекта
Создание упорядоченного набора современных и древних славянских символов (кириллица, глаголица), целесообразного для использования:

1. В кодировке UCS/Unicode в качестве тактичного дополнения славянских страниц;

2. В алгоритмах сортировки в текстовых редакторах, текстовых и табличных процессорах, словарях и СУБД (MS Office, KOffice, ABBYY Lingvo и т.п.).

Задачи проекта

1. Определение набора уникальных символов (базиса), необходимого для их семантической, лингвистической идентификации (детерминация символов с использованием материалов научно-исследовательских работ по лингвистике).

2. Определение порядка размещения символов в объединённом алфавите и в кодировке.

Этапы проекта

1. Создание принципов построения славянских диапазонов кодировки (реализовано).

2. Исходная версия Гиперазбуки на базе символов, уже размещённых в славянских Unicode-диапазонах (реализовано).

3. Версия Гиперазбуки на основе базиса славянских символов (в процессе разработки).

Комментарии

Проект развивается на основе материалов печатных и электронных изданий, существует с 2005 года.

Гиперазбука является открытым проектом, принимающим изменения принципов построения «Гиперазбуки» и рекомендации специалистов.

Стандарт Гиперазбуки

Стандарт Гиперазбуки представляет собой ряд принципов, для последовательного расположения символов, а также ряд исключений.

Версии Гиперазбуки

1. Гиперазбука 1.1 Unicode 4.1.0 (на базе символов Unicode 4.1.0).

2. Гиперазбука 1.1 Unicode 4.1.0 Linguistic (интегрированный кирилло-глаголический алфавит на базе символов Unicode 4.1.0).

3. Гиперазбука 1.2 (кириллица и глаголица дополняются).

4. Гиперазбука 1.2 Linguistic (кириллица и глаголица дополняются интегрировано).

Планы на ближайшее время

1. Представление версий «Гиперазбука 1.2» и «Гиперазбука 1.2 Linguistic».

2. Создание сайта проекта.

3. Вступление проекта в Сообщество славянской типографики (размещение ссылки на сайт проекта, участие в почтовой рассылке «c-slav»).

Некоторые нюансы

В рамках проекта необходимо учитывать, что лигатуры и сочетания букв и диакритических знаков иногда требуют отдельных позиций для каждой пары символов. Недостатки других методов представления лигатур и диакритики изложены в предлагаемом ниже материале.

Анонс

Ниже предлагается обзор проблемной области, материал по стандарту Гиперазбуки и по версии Гиперазбука 1.1 Unicode 4.1.0 Linguistic. В материале приведены аргументы в пользу предлагаемых в рамках проекта подходов: приводятся мотивы создания объединённого алфавита, оптимально упорядоченного для большинства славянских языков, объясняются причины, заставляющие предложить версию Гиперазбуки с интегрированной глаголицей и т.д. Проект Гиперазбуки не подразумевает, что после определения базиса символов не будут обнаружены новые исторические факты, требующие корректировки Unicode. Однако нельзя не принимать во внимание, что по прошествии достаточного времени возможно изменение структуры диапазонов Unicode, которое представит возможность воплотить весь наработанный к тому времени опыт, касающийся как базиса славянских символов, так и их расположения. В связи с известными причинами в Unicode практически невозможно добиться изменения уже позиционированных символов. Однако часть проекта Гиперазбука, отвечающая за упорядоченность расположения символов, на данный момент также применима для осуществления лексикографической сортировки текстов на языках, сортировка для которых не поддерживается программным обеспечением, а также с целью составления специалистами по лингвистике [1-2] смешанных пользовательских словарей.

На сегодня структура кодового пространства Unicode размечена в диапазоне от 0 до $10FFFF_{16}$ символа. Она включает около 800000 свободных, резервных ячеек в зоне неинициированных позиций «Unassigned» и 137468 ячеек в зоне для частного использования «Private Use» (рис. 1), из которых 6400 находятся в диапазоне BMP (Basic Multilingual Plane). По некоторым оценкам для индексирования славянского письма потребуется не более 10000 знакомест, что вполне укладывается и в зону «Unassigned», и в зону «Private Use». В случае если консорциум Unicode не примет расширений, достаточных для корректной работы с древнерусским письмом, некоторые специалисты предлагают разместить дополнительные символы в зоне «Private Use». Возможно, в таком случае более рациональным было бы размещение дополнительных символов в зоне «Unassigned» без официаль-

ного индексирования в Unicode. Согласованное использование такой кодировки в рамках ассоциации «Электронные издания/представления славянских рукописей» или Сообщества славянской типографики [3] может привести к пересмотру консорциумом Unicode разметки кодового пространства. Даже если этого не произойдёт, обращение к единой кодировке для древнеславянских текстов значительно упростит работу членов ассоциации на ближайшие годы. Если же консорциум Unicode примет решение иначе индексировать диапазон, выбранный ассоциацией, это не приведёт в ближайшее время к бóльшим проблемам, чем, если разместить дополнительные славянские символы в зоне «Private Use».

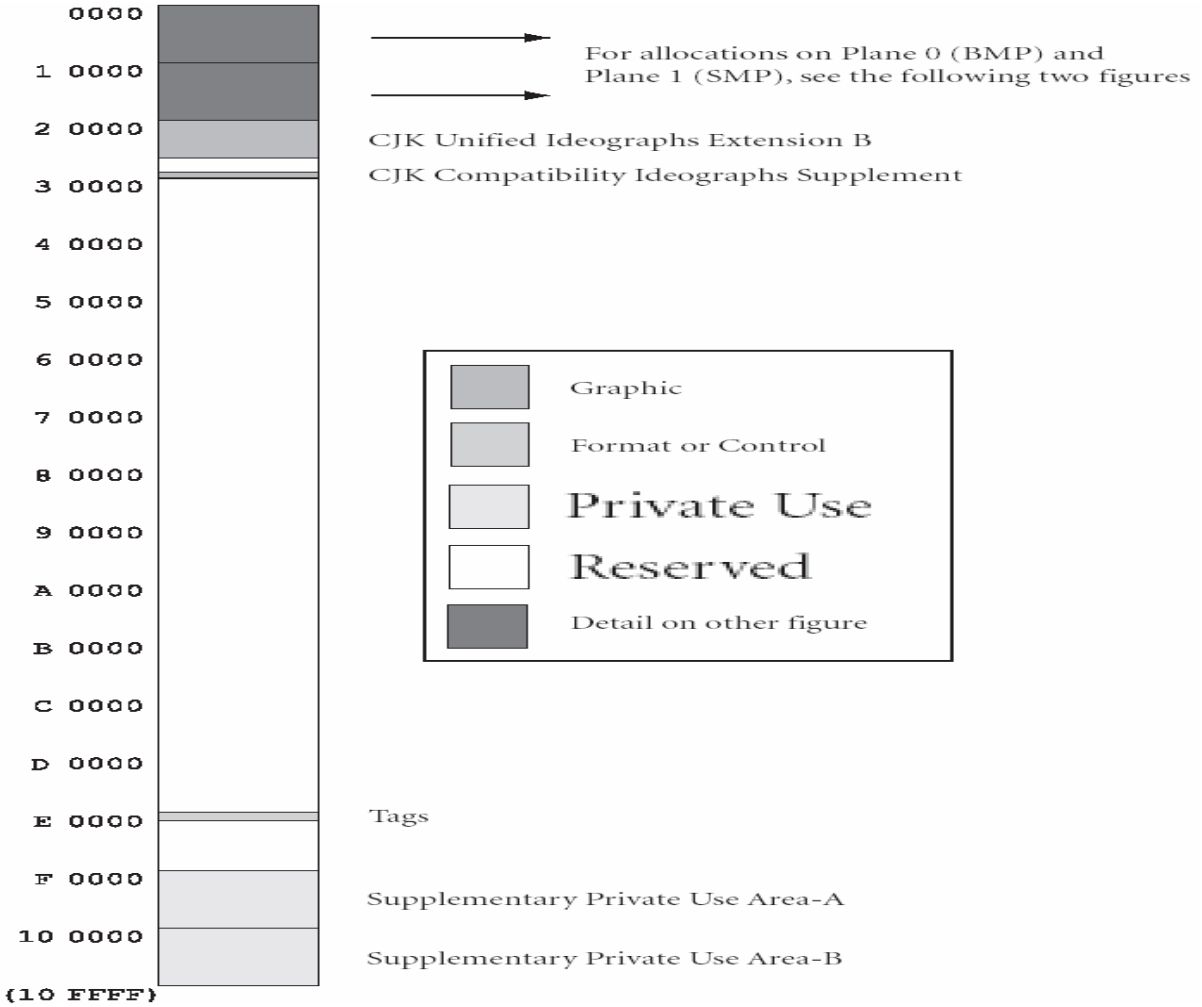


Рис. 1. Зоны, не индексированные в Unicode 4.1.0 (показаны наиболее светлыми цветами). Изображение взято с официального сайта The Unicode Consortium (http://www.unicode.org) и изменено.

Для размещения небольшого числа символов можно также воспользоваться свободными ячейками кириллического и глаголического диапазонов Unicode в BMP. В диапазоне «Cyrillic» (рис. 2) имеется 8 свободных ячеек, в диапазоне «Cyrillic Supplement» – 32, в диапазоне «Glagolitic» – 2, в следующем свободном диапазоне, которым потенциально может быть

расширен диапазон «Glagolitic», – 16 ячеек. Однако, скорее всего, 40 свободных ячеек из BMP для расширения кириллицы и 18 ячеек из BMP для расширения глаголицы не будет достаточно.

Ниже более подробно изложены основные материалы проекта.

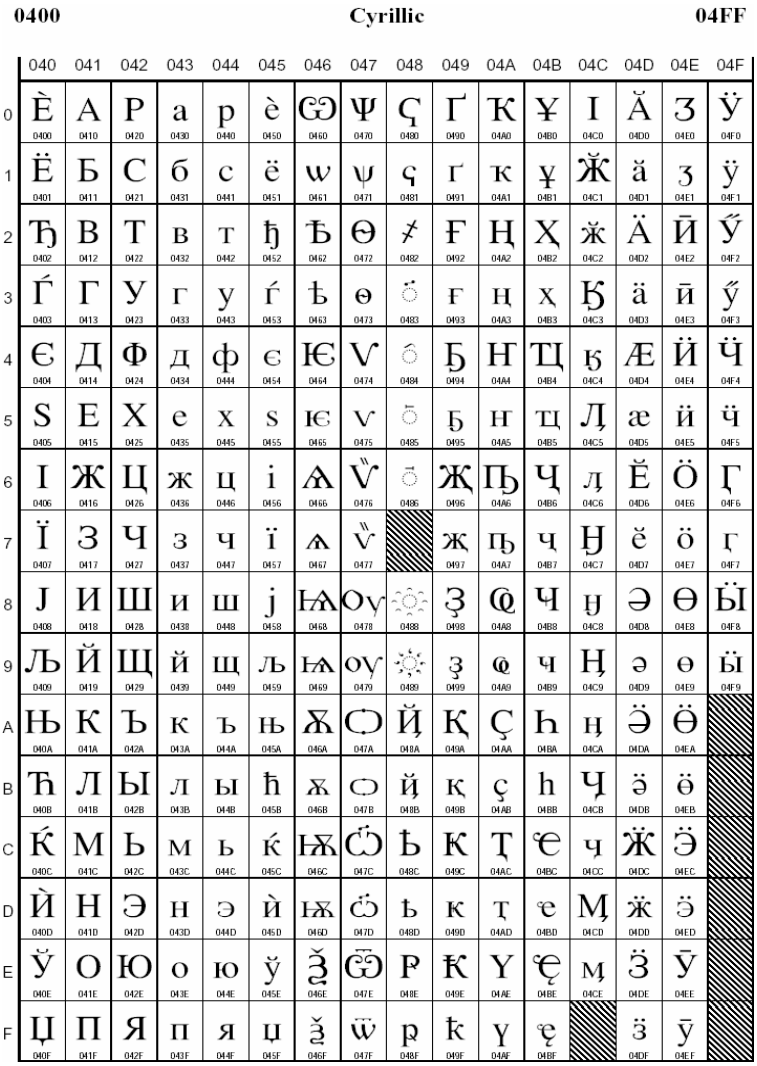


Рис. 2. Диапазон «Cyrillic» кодировки Unicode 4.1.0. Изображение взято с официального сайта The Unicode Consortium (http://www.unicode.org).

2. Проблема кодового пространства

Решения проблемы нехватки и распределения кодового пространства связаны с организацией кодировок, шрифтов и шрифтовым дизайном. Эти области используют похожие термины, но с разными значениями. Проблемы компьютерного шрифтового дизайна, графические типы шрифтов и стили не рассматриваются в данной работе. Вместо термина «письменность» в лингвистической литературе обычно используется термин «письмо».

Кодировкой в программировании называется сопоставление чисел (то есть номеров, кодов) набору символов. Создание кодировки называют также индексированием. Носителями кодировок являются шрифты. Программное обеспечение и специальные компоненты операционной системы содержат информацию о кодировках (таблицы для сопоставления кодов и символов, которые называются кодовыми таблицами).

По оценкам японских специалистов [4] для индексирования символов всех современных и устаревших (представляющих исторический интерес) языков необходимо 3 байта. Однако ещё до появления первоначально 2-байтной кодировки Unicode существовал проект 4-байтной кодировки UCS (Universal Coded Character Set или ISO DIS-10646.1:1990). Международный стандарт UCS важно не путать с альтернативным 8-битным стандартом кодировки для славянских символов UCS8 (Unified Church Slavonic Encoding). Совпадение названий было замечено, когда UCS8 была уже распространена.

Фигурное сочетание нескольких символов называется лигатурой. Существует ряд лигатур, необходимых по правилам языка и для корректного отображения графики письма. Эти лигатуры служат как новые буквы: для передачи определённых фонем или для отличия слов, обладающих разным значением, но одинаковым звучанием. Лигатуры присутствуют в древней и национальной кириллице, латинице, в греческой и в других алфавитных системах письма. Графически лигатура представляет собой особое начертание двух и более букв, в том числе – буквы и диакритического знака (надстрочного, подстрочного или другого символа). Отдельные диакритические знаки, а иногда и буквы с ними, называются диакритами. В подавляющем большинстве материалов сети Интернет буквы с диакритикой не считаются лигатурами, но в данной работе лигатура понимается в более широком смысле (к этому классу причисляются и диакритизированные буквы), так как другой, более подходящий, термин отсутствует. Далее будут рассмотрены методы представления лигатур в кодировках и шрифтах.

Полнолигатурный метод. Полнолигатурный метод – это представление лигатуры в качестве глифа, для которого выделена специальная ячейка кодовой таблицы. Глиф (glyphs) – графическое изображение символа, – обязательный компонент любого шрифта. Понятие глифа в компьютерной типографике соответствует понятию литеры из печатной типографики.

Если не хватает кодового пространства и в кодировке для лигатуры не выделено отдельной ячейки (то есть знакоместа), то лигатуру можно представить как композицию нескольких символов, используя для этого один из методов, рассмотренных ниже. В этом случае при наборе текста символы печатаются друг за другом, а их глифы должны быть автоматически совмещены или заменены на другой глиф. Для этого существуют метод накладной диакритики нулевой ширины, метод кернинга и метод лигатурной

подстановки. Эти методы можно назвать композиционными или составными. Далее приводятся описания композиционных методов и их недостатки по сравнению с полнолигатурным методом.

Под композицией или композиционным символом в данной работе понимается комбинация в тексте нескольких символов, помещённых друг за другом: например, чтобы печатать «Ять с тяжёлым ударением», сначала печатается «Ять», затем «Тяжёлое ударение», их коды помещаются друг за другом в текстовом файле, а их глифы автоматически замещаются на один специальный глиф или располагаются относительно друг друга таким образом, который предусмотрен в шрифте и доступен в текстовом процессоре. Похожий метод наложения используется и в шрифтовом дизайне при создании глифов для лигатур, поэтому во избежание путаницы в данной работе не используется графический термин «композит». Композиция есть форма представления лигатур.

Метод накладной диакритики. Как правило, вместо накладной диакритики говорят о накладных надстрочниках, так как описываемый метод обычно применяется к ним. Но он может быть применён к любым символам накладной диакритики (например, древнерусской цифровой диакритике высоких порядков).

Определённые индексы кодировки, то есть соответствующие им абстрактные ячейки символов, могут быть выделены под отдельные диакритические знаки (например, в Unicode для этих целей существует диапазон «Combining Diacritical Marks»). Метод накладной диакритики заключается в следующем. В шрифте для знаков диакритики ширина символа (величина площадки, которая выделяется для символа в тексте) задаётся равной нулю, но ширина изображения символа на глифе остаётся полной, за счёт этого диакритический символ накладывается на расположенную перед ним букву. На практике ширину символа устанавливают предельно близкой к нулю, так как многие программы и драйверы принтеров некорректно работают с символами, для формального расположения которых выделена площадь, равная нулю. Нулевая ширина диакритического знака необходима для того, чтобы следующая буква располагалась сразу после той, с которой совмещён диакритический знак. Изображение знака на глифе помещают со смещением в зависимости от того, где знак должен располагаться относительно буквы. Как правило, смещение производят в левую сторону, так как в подавляющем большинстве текстовых процессоров буква, с которой требуется совместить диакрит, печатается первой – левее диакрита. В качестве символов, обладающих нулевой шириной, могут быть использованы только накладные диакритические символы, так как столь сильное совмещение глифов неприемлемо для лигатур, состоящих из нескольких букв.

Метод кернинга. Кернинг – изменение расстояния между символами, входящими в определенные (кернинговые) пары. Важно не путать с тре-

кингом – изменением межсимвольных интервалов. Можно сказать, что кернинг является частным случаем трекинга: он используется лишь для пар букв и не регулируется пользователем. Кернинг задаётся в качестве свойств шрифта. Метод кернинга применим не только к диакритам, но и для некоторых композиций двух-, трёхбуквенных лигатур, имеющихся в системах письма, использующих славянскую графику. Для некоторых шрифтов кернинг делает недоступным трекинг.

Составной символ в композициях, в различных лигатурах, должен быть расположен на различном расстоянии относительно других символов по горизонтали и вертикали и также должен изменять форму. Например, буква ѣ («Ять») с надстрочником ̀ («тяжёлое ударение» – «варйя») должна выглядеть, как показано здесь: ѣ̀, но простое наложение не позволяет создать качественное изображение. Это требует отдельного знакоместа для каждой лигатуры. Кроме того, метод накладной диакритики и метод кернинга осложняют (не критически) программную обработку текста для некоторых языков, это проявляется в случае с двух- и трёхбуквенными лигатурами – например, в кавказских языках (см. следующий раздел).

Метод накладной диакритики и метод кернинга не позволяют качественно реализовывать лигатуры: такие методы снижают графическую аутентичность символов, читабельность и эстетику текста. Под графической аутентичностью символов в данной работе предлагается понимать соответствие изображения символа определённому символу письменной речи, то есть соответствие начертания символа одной из шрифтовых гарнитур. Читабельность – более лёгкий для произношения термин, вместо которого в шрифтовом дизайне обычно используется термин «удобочитаемость». Читабельность шрифта (и, следовательно, качество распознавания печатного текста) зависит от графических характеристик шрифта и от условий его применения (средств чтения и т.д.). Читабельность шрифта важно не путать с читабельностью текста, которая определяется по длине предложений и числу слогов в слове.

Метод лигатурной подстановки (Ligature substitution)**.** Можно задать соответствие между определёнными последовательностями символов и лигатурными глифами. Таблица соответствий определяется внутри конкретного шрифта. Таким образом будет осуществляться подстановка лигатур, не обладающих индексами в какой-либо кодировке (кроме внутренней своего рода «кодировки» шрифта). Это похоже на макросы автозамены в текстовых процессорах, но создавать такую таблицу соответствия в текстовом процессоре нерационально, так как для неё нет общего стандарта, она задаётся разработчиком шрифта. Метод лигатурной подстановки реализован в формате шрифтов OpenType [5]. Для полноценной работы с лигатурными подстановками, импортирования текстов в различные форматы и печати необходимо, чтобы редакторы, текстовые процессоры и браузеры

в достаточной мере поддерживали формат шрифтов OpenType. Но за 10 лет существования формата ни в одно известное средство поддержка не была внедрена в достаточной мере. MS Word 2003 поддержка осуществляется лишь для восточно-азиатских языков, а в Adobe InDesign ограничение на трекинг (+23) в рамках определения лигатур Microsoft: под ними подразумевается замена только рядом стоящих «слипающихся» символов на аккуратный конгломерат. В распространённых средствах обычно предоставлено мало возможностей для установки плагинов, надстроек третьих производителей, фирмы не заинтересованы создавать фичи для лигатурных подстановок.

Комбинированный метод. В связи с присущими для каждого метода недостатками большинство шрифтов используют одновременно несколько методов представления лигатур – в зависимости от того, какой метод подходит лучше для отображения того или иного символа в данной гарнитуре. Гарнитурой называется шрифтовое семейство, имеющее определённые стилевые особенности, иначе говоря – графический тип шрифта.

Методы представления лигатур, используемые в комбинированном методе (а также и в альтернативных кодировках), можно классифицировать следующим образом:

- полнолигатурный метод;
- композиционные методы:
    - метод накладной диакритики;
    - метод кернинга;
    - метод лигатурной подстановки (редко и экспериментально).

Дальнейшие выводы. Из описания представления лигатур следует, что графическим (и техническим) требованиям в полной мере удовлетворяет лишь полнолигатурный метод. Он подразумевает наличие большого числа знакомест. На практике размещение большого количества символов, отсутствующих в стандартных кодировках, осуществляется при помощи альтернативной кодировки:

- в рамках одной 8- или 16-битной кодовой таблицы;
- используя несколько 8-битных кодовых страниц.

Для создания альтернативной кодировки за основу берётся какая-либо стандартная (например, расширенная ANSII или Unicode). Не придерживаясь жёстких принципов, авторы стремятся разместить символы на те позиции, где в стандартной кодировке были расположены похожие на них символы (похожие лексически или только графически) – это упрощает набор текстов с клавиатуры. В качестве примера на рис. 3 приведена кодировка шрифта «Izhitsa.ttf», положившего основу одноимённой гарнитуре и ряду других художественных решений. В верхнем левом углу каждой ячейки указан её символ в исходной кодировке CP1251 – латинская область кодовой таблицы занята дополнительными древнерусскими буквами, размещёнными по принципу графической «схожести». Гарнитура «Ижица» об-

ладает шрифтами на основе кодировок CP1251 (Windows Cyrillic) и Unicode.

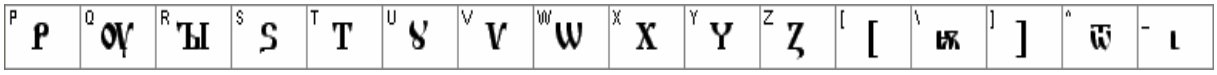


Рис. 3. Символы из диапазона 50_{16}–$5F_{16}$ шрифта «Ижица» (Izhitsa.ttf), (на рисунке в целях компактности глифы сжаты по ширине) – 6-я строка кодовой таблицы.

Альтернативная кодировка может использовать несколько кодовых страниц, требующих переключения раскладки: например, расположить кириллицу на страницах Windows 1251 (Cyrillic), 1252 (Western – Latin 1), 1253 (Greek) и т.д. с учётом пересечений в нижней странице, или воспользоваться семейством ISO 8859-X. Выбор этих стандартов и их страниц зависит от того, как с ними работает программное обеспечение, для которого создаётся альтернативная кодировка. Основные принципы для наиболее корректного проектирования альтернативных 8-битных кодировок изложены на сайте Министерства образования и науки РФ [6].

Ряд альтернативных кодировок (кодировочно-шрифтовые системы проектов «Ирмологий» [7] и «Манускрипт» [8]) применяют мультишрифтовой метод размещения большого числа символов. Мультишрифтовым методом называют метод альтернативного размещения символов с использованием нескольких шрифтов или нескольких (файлов) начертаний одного шрифта. Альтернативную кодировку, использующую несколько 8-битных кодовых страниц (например, страницы ISO 8859-X или Microsoft CP125X), можно было бы также назвать мультишрифтовой, так как для различных страниц данной кодировки могут использоваться различные шрифты. Но в этом случае может использоваться и один шрифт с заданными (прописанными в файл «Win.ini») шрифтовыми «сечениями» (разбиением на кодовые страницы), как это делается для приложений, не поддерживающих Unicode, но работающих с многостраничными Unicode-шрифтами.

Мультишрифтовой метод использует несколько шрифтов или несколько начертаний одного шрифта в качестве различных «страниц» кодировки. В обоих случаях работа в данной кодировке упрощается, если набор шрифтов снабжён специальным инструментарием. Один шрифт может обладать несколькими шрифтовыми файлами для нормального, полужирного, курсивного, полужирного курсивного и других начертаний. Например, в ОС Windows каждый шрифт формата TrueType может иметь до 4-х файлов с описанными выше начертаниями (остальные начертания система генерирует автоматически). Если необходимость в таком количестве начертаний отсутствует, излишние начертания (файлы и опции) можно использовать для размещения (и выбора пользователем) дополнительных символов. Тогда, в худшем случае, для смены раскладки клавиатуры пользователю придётся выбрать опцию с другим начертанием шрифта, а в лучшем

случае – воспользоваться комбинацией клавиш, предусмотренной установленным инструментарием (макросом или перехватчиком клавиатуры).

Создание альтернативных кодировок, как правило, начинается с создания одного шрифта (или, в физическом смысле, набора шрифтов для мультишрифтовой кодировки). Каждый новый шрифт может представлять и новую оригинальную кодировку, даже если впоследствии она не будет использована в каких-либо других шрифтах. А созданные на его основе шрифты могут иметь свои особенности расположения символов, представляя собой уже новые альтернативные кодировки.

В связи с понятием мультишрифтовых кодировок, а также в связи с потребностью некоторых шрифтовых систем (и использующих их программ) в дополнительном инструментарии и смене клавиатурных раскладок, в ряде публикаций используется термин «кодировочно-шрифтовая система» (КШС). Возможно, впервые данный термин был предложен в рамках проекта «Ирмологий» [7]. КШС понимается как совокупность какой-либо кодировки, набора шрифтов и инструментария. В данной работе понятие «КШС» не используется, дабы разделять, когда речь идёт о размещении символов в кодовых позициях (о кодировке) и когда приводится пример её реализации (даже если какой-либо шрифт использует оригинальную, присущую лишь ему, кодировку). То, какие файлы или сечения и управляющие элементы в совокупности составляют шрифт (как программный, а не только лексикографический продукт) или кодировку, является особенностями конкретного шрифта или кодировки.

Альтернативными кодировками пользуются многие шрифты и семейства шрифтов для древнерусского языка, например: «WP CirillicA», «Вертоград» (кодировка UCS8), «Евангелие» (UCS8), «Елизаветская» (ROOS), «Златоуст» (UCS8, ieUCS8 kUCS8), «Индиктион» (UCS8, ieUCS8), «Ирмологий» (Ирмологий, UCS8), «Ижица», «Киприан», «Ортодокс» (eROOS, UCS8), «Печерская» (UCS8), «Почаевск» (UCS8, ieUCS8, kUCS8), «Псалтирь» (UCS8, ieUCS8, kUCS8), «Путята», «Сатис», «Сергий», «Славянская библия», «Славянский» (UCS8, ieUCS8, kUCS8), «Старо-успенский» (UCS8, ieUCS8, kUCS8), «Триодин» (UCS8, ieUCS8, kUCS8), «Феофан» (UCS8). Существует большое количество шрифтов, построенных на альтернативных кодировках, содержащих кириллические символы языков неславянских народов (бывшего СССР). Из них наибольшее распространение получили 16-битные шрифты на базе Unicode: шрифт «SL_Times New Roman» и шрифты фирмы Paratype в альтернативной кодировке Cyrillic Asian.

Альтернативные кодировки удовлетворяют требованиям полиграфии. Однако такие кодировки и созданные для них шрифты не обладают совместимостью: выбор шрифта другой кодировки для набранного текста требует создания программного обеспечения для перекодировки. Отсутствие соответствия распространённым стандартам осложняет внедрение в

текстовые процессоры проверки правописания для текстов, набранных в альтернативных кодировках.

До распространения ОС Windows XP (например, в 2001 году) в статьях, посвящённых кодировкам, высказывались мнения, что с полным переходом на Unicode проблема кодового пространства, связанная с кириллическими алфавитами, исчезнет. К сожалению, эти предположения не оправдались. Проблема приводила даже к тому, что в ряде случаев создавались шрифты для отдельных приложений, использующих собственные альтернативные кодировки.

Методы представления символов с изменённым начертанием при помощи текстовой разметки представляют собой в определённом смысле тоже альтернативную кодировку, но только не кодировку «вместо», а кодировку «над» какой-либо терминальной кодировкой. Стандарт инвариантного представления (не зависящего от платформы: кодировки, шрифтовой системы, операционной системы) HIP [9] не является альтернативной кодировкой «вместо» какой-либо кодировки и не является разметкой «над» какой-либо конкретной кодировкой. Формат HIP представляют собой кодировку над множеством из минимум 136 символов (с возможностью использовать дополнительные), которые присутствуют почти во всех кодировках, имеющих кириллическую страницу. Формат HIP не чувствителен к тому, в какой кодировке представлены эти символы. Своеобразный «код» текста, хранящегося в формате HIP, представляет собой смесь кириллицы, латиницы, цифр, знаков препинания и некоторых других неалфавитных символов (= < > ~ _ \ | ‘ ` % & # ^ $ @ { } [ ] * +). Для читабельного отображения и упрощения редактирования текстов в формате HIP используются специальные конверторы и пакеты надстроек для браузеров и редакторов текста, например, для TeX/LaTeX, HTML и BibleQuote. HIP также используется в технологии смешанного набора [10].

Использование композиционных методов представления лигатур, методов разметки и HIP-формата является вынужденным и не предоставляет столь широких возможностей, которые могла бы предоставить полная индексация базиса необходимых символов в Unicode (в том числе с использованием полнолигатурного метода).

Во избежание вышеописанных проблем страницы Unicode содержат лигатуры, необходимые для максимально полной поддержки языков, использующих латинские и греческие символы. Например, для латинских двухбуквенных лигатур выделено 29 знакомест (не считая диапазона IPA – International Phonetic Alphabet): Æ, æ, Ǣ, ǣ, Ǽ, ǽ, DZ, Dz, dz, DŽ, Dž, dž, ff, fi, fl, ffi, ffl, ſt, IJ, ij, LJ, Lj, lj, NJ, Nj, nj, Œ, œ, st, для кириллических – только 12: Ӕ, ӕ, Љ, љ, Ҥ, ҥ, Њ, њ, Оу, оу, Ҷ, ҷ. Латинские диакритизированные варианты буквы «A» занимают 58 знакомест: À, à, Á, á, Â, â, Ã, ã, Ä, ä, Å, å, Ā, ā, Ă, ă, Ą, ą, Ǎ,

ă, A̋, a̋, Ā, ā, Ǻ, ǻ, À, à, Â, â, Ą, ą, Ạ, ạ, Ả, ả, Ấ, ấ, Ầ, ầ, Ẩ, ẩ, Ẫ, ẫ, Ậ, ậ, Ắ, ắ, Ằ, ằ, Ẳ, ẳ, Ẵ, ẵ, Ặ, ặ, ⒜, Ⓐ. Кроме них в ячейках Unicode представлены и другие разновидности начертаний латинских букв, например, формы с различной шириной, несмотря на то что ширину можно задавать при помощи масштабирования символов любого шрифта. Греческие диакритизированные варианты буквы «А» занимают 45 знакомест: Ἀ, Ἁ, ἀ, Ἄ, ἁ, Ἂ, ἂ, Ἃ, ἃ, Ἅ, ἅ, Ἆ, ἆ, Ἇ, ἇ, Ᾰ, ᾰ, Ᾱ, ᾱ, Ὰ, Ά, Α͂, ᾶ, ᾳ, ᾴ, ᾲ, ᾷ, Ὰι, Άι, ᾼ, ᾀ, ᾁ, ᾂ, ᾃ, ᾄ, ᾅ, ᾆ, ᾇ, ᾈ, ᾉ, ᾊ, ᾋ, ᾌ, ᾍ, ᾎ, ᾏ, ᾲ, ᾳ, ᾴ, ᾶ, ᾷ, Ὰι, Άι, ᾼ, ᾀ, ᾁ, ᾂ, ᾃ, ᾄ, ᾅ, ᾆ, ᾇ, ᾈ, ᾉ, ᾊ, ᾋ, ᾌ, ᾍ, ᾎ, ᾏ, кириллические – всего четыре знакоместа: Ӑ, ӑ, Ӓ, ӓ. Несмотря на то что все необходимые диакритические знаки имеются в диапазоне составных надстрочников Unicode, для данных символов используются отдельные ячейки кодовой таблицы. Диакритика современных европейских языков реализована в Unicode полнолигатурным методом. В версии Unicode 5.0 ожидается ещё два блока расширений латиницы.

С учётом полного лигатурирования для поздней кириллицы (XVIII-XIX век) требуется максимум 500 знакомест, а для всей древней кириллицы по некоторым оценкам [7] – до 10000 знакомест.

3. Лексикографическая сортировка символов славянского происхождения в кодировке Unicode

В данной работе речь идёт только о лексикографической, алфавитно-числовой сортировке текстов. Под текстами понимаются любые последовательности символов: строки, абзацы, словарные статьи, содержание ячеек таблиц.

Для создания словарей и различных баз данных программные средства осуществляют лексикографическую сортировку для одного языка из предлагаемого ими списка. Однако существуют задачи, в которых это не устраивает лексикографов [1-2]. Они заключаются в составлении смешанных пользовательских словарей, словарей на языках, сортировка для которых не предусмотрена. Если такие средства, как MS Office, предоставляют более-менее широкий список языков для сортировки текстов, то менее распространённые программы вызывают затруднения у пользователей, интересующихся языком, который полностью не поддерживается. Кроме сортировки для конкретного языка лексикографов интересует такая сортировка всех славянских символов, в которой, например, за буквой «К» следуют лексически и фонетически близкие к ней символы: «Ҟ», «Ҡ», «Ӄ», «Ҝ», «Ҟ» (по названиям в Unicode: «KA WITH DESCENDER», «KA WITH HOOK», «KA WITH STROKE», «KA WITH VERTICAL STROKE», «BASHKIR KA») и т.д. Расположение символов в таблице Unicode не является полностью алфавитным и не обеспечивает сортировку. Например, в кириллическом диапазоне в соответствии

с алфавитным порядком размещены лишь буквы современного русского языка (кроме факультативной буквы «Ё»).

Для решения описанных выше задач предлагается проект общего алфавита славянских символов различных народов и времён – своего рода гиперазбука (азбукой называют славянский алфавит), представленная в приложении. Подобный лексикографический алфавит, содержащий все славянские символы Unicode, не может претендовать на соответствие всем национальным алфавитам. Создание гиперазбуки преследует собой следующие цели:

- обеспечить лексикографически корректную сортировку «по умолчанию» (если язык не выбран или тексты составлены на различных языках);
- облегчить программную реализацию лексикографической сортировки для большинства распространённых языков с письмом на основе славянской графики для электронных словарей и других баз данных различных разработчиков;
- предложить наглядный и удобный порядок расположения в кодировке символов, которыми необходимо дополнить Unicode;
- предложить стандарт размещения славянских символов для проектируемых кодировок в целях упрощения сортировки в любом программном обеспечении, поддерживающем данные кодировки.

Гиперазбука основана на базе славянских символов Unicode и может быть в дальнейшем расширена. Гиперазбука способна облегчить программную сортировку для текстов на старорусском, древнерусском и церковнославянском языках, на глаголических азбуках, на словио и на 27 современных распространённых языках с кириллической письменностью: абазинском, адыгском, балкарском, белорусском, болгарском, бурятском, гагаузском, даргинском, ингушском, карачаевском, киргизском, языках коми (считающихся здесь за один), марийском, молдавском, монгольском, ногайском, осетинском, русском, татарском, тувинском, удмуртском, узбекском, украинском, цыганском, чеченском, чувашском и чукотском. Упомянутый здесь сло́вио – искусственный язык с логичной, безысключительной грамматикой, созданный с целью быть понятным друг другу для говорящих на языках славянской группы без их дополнительного изучения. Автор словио – лингвист Марк Хучко (info@slovio.com).

Например, в MS Office поддерживается сортировка для восьми из перечисленных языков: белорусского, болгарского, киргизского, молдавского, русского, татарского, узбекского и украинского. Однако если бы в MS Office хотя бы при сортировке для русского языка использовалась разработанная гиперазбука, данная сортировка оказалась бы корректной для всех языков, которым соответствует гиперазбука.

Из-за иного порядка символов проектируемый алфавит определённо не подходит (по имеющимся материалам) для 13 распространённых кириллических алфавитов: аварского, азербайджанского, алтайского, баш-

кирского, казахского, калмыцкого, курдского, македонского, румынского, саамского, сербского, туркменского, черногорского. Порядок некоторых символов из последних перечисленных алфавитов также учитывается в общем алфавите (т.е. гиперазбуке) – в тех случаях, где это не противоречит доминирующей группе алфавитов. Под символами в приложении указаны языки (не только из доминирующей группы), повлиявшие на размещение данных символов в гиперазбуке.

Список языков с алфавитами на основе кириллицы можно найти по ссылке [11]. Стоит напомнить, что ряд славянских языков, таких, как польский, чешский, словацкий, обладают латинской письменностью, поэтому в данном списке отсутствуют. Международный язык эсперанто не имеет устоявшегося кириллического варианта алфавита.

Если какой-либо язык использует два алфавита одновременно (официально или неофициально), в нашем случае рассматривается кирилличекий вариант алфавита. Даже если язык официально переведён на латиницу, на нём сохраняется масса текстов в кириллическом представлении. В соответствии с ныне действующим российским законодательством субъекты РФ не имеют права самостоятельно менять графическую основу национальных алфавитов. Это означает, что, скорее всего, кириллические алфавиты неславянских народов РФ не будут официально переведены на основу латинской или какой-либо другой графики. Например, если русская кириллица не передаёт таких татарских букв как твёрдое и мягкое «К», то сегодня такая возможность представлена в кириллице Unicode и, возможно, было бы проще расширить татарский кириллический алфавит, чем переводить его на латиницу. При этом не возникала бы проблема с чтением СМИ у пожилых представителей татарской нации и «переобучения» этнических татар, живущих за пределами Татарстана, например, в Башкирии [12]. Расширенные возможности Unicode сегодня позволяют использовать кириллическую татарскую письменность в информационных технологиях, включая Интернет.

Оценка и критика существующих кодировок (см. предыдущий раздел) могла бы позволить предположить появление в будущем новых трёх- или четырёхбайтных кодировок или, что почти то же самое, изменение структуры диапазонов Unicode, но это вызывает слишком большие технические сложности. Более вероятно усовершенствование Unicode в плане некоторого расширения, без изменения позиций уже индексированных символов. Однако хотелось бы отметить, что проектируемый алфавит может оказаться полезным в случае создания новой кодировки, и предлагает определённый стандарт размещения славянских символов в кодовой таблице. Поэтому, если какой-либо распространённый символ встречается во многих языках, алфавиты которых не соответствуют гиперазбуке, то символ размещается в гиперазбуке соответственно алфавитам данных языков, даже если по этой причине из списка языков гиперазбуки придётся исключить какой-

либо малораспространённый язык. Так, например, без ущерба языкам гиперазбуки символ «J» (U+0408, U+0458) можно было расположить соответственно алтайскому алфавиту, но он сохраняет своё «логичное» лексикографическое место расположения – после символа «И». В приложении алфавиты, не соответствующие гиперазбуке, указаны под символами, на порядок расположения которых они влияют, в квадратных скобках.

Из множества Unicode-символов, имеющих отношение к славянскому письму, можно выделить четыре пересекающихся подмножества (на основе материалов энциклопедии «Википедия» [11]):

1. Древние символы, бывшие в употреблении у славян (включая, например, русскую, сербскую, хорватскую, моравийскую глаголицу и символы, используемые в церковнославянском языке);

2. Символы, входящие в состав алфавитов славянских народов, не перешедших на латиницу, либо использующих два письма (официально или неофициально);

3. Символы, входящие в состав алфавитов неславянских народов бывшего СССР, графика которых была либо построена на основе кириллицы, либо переведена на неё (например, с арабского письма);

4. Символы, входящие в состав алфавитов плановых языков (Словио, Лингуа Франка Нова в кириллическом представлении) – подмножества букв современной русской кириллицы.

Упомянутый здесь Лингуа Франка Нова (Lingua Franca Nova) — международный искусственный язык с упрощённой грамматикой, созданный Джорджем Буре (George Boeree, http://lingua-franca-nova.net) на основе лексики романских языков: французского, итальянского, испанского, португальского и каталонского. Алфавиты славянских народов и младописьменные алфавиты народов бывшего СССР по расположению общих букв, как правило, соответствуют русскому алфавиту. К языкам народов бывшего СССР, обладающим младописьменными алфавитами, причисляют абазинский, аварский, адыгейский, ингушский, алтайский, корякский, хантыйский, хакасский, чукотский и т.д. (около 50 языков). Некоторые языки приняли русский алфавит в неизмененном виде, но в большинстве случаев алфавит был расширен. Дополнительные буквы национальных алфавитов представляют собой диакритизированные и видоизменённые буквы современного русского языка и заимствованные латинские буквы. Большинство их включено в кириллический диапазон Unicode. Стандарт Unicode 4.1.0 поддерживает в смысле наличия символов большинство языков народов России с дополнительными кириллическими буквами: алтайский, башкирский, бурятский, долганский, калмыцкий, коми, корякский, марийский, нанайский, ненецкий, осетинский, саамский (без указания долготы гласных), татарский, тувинский, удмуртский, хакасский, хантыйский, чувашский, эвенкийский, эвенский, якутский, кавказские языки с буквой «палочка» (абазинский, адыгейский, аварский, чеченский, даргинский, ингушский,

кабардинский, лакский, лезгинский, табасаранский,…) и другие. Буквы, построенные на русской графической основе, в алфавитах обычно размещены следом за буквами, традиционно входящими в состав русской кириллицы или за фонетически близкими к ним буквами. Буквы, заимствованные из латинского алфавита, также часто расположены следом за лексически и фонетически близкими буквами. Во избежание недоразумений необходимо напомнить, что ранее порядок букв многих национальных алфавитов был иным: все дополнительные символы располагались в конце. Некоторое время назад в ряде языков была проведена реформа размещения дополнительных символов в алфавитах. Она являлась необходимой мерой для создания удобных словарей и прочих баз данных. Так, например, порядковый номер буквы «Ә» («CYRILLIC SCHWA») в татарском алфавите изменился с 34-го на второй: она стала располагаться следом за созвучной ей буквой «А». Как правило, для каждой буквы русского языка найдётся не более одной фонетически близкой дополнительной буквы в каждом национальном алфавите, поэтому часто порядок расположения дополнительных символов национальных алфавитов непротиворечив. Это делает возможной и целесообразной разработку единого алфавита сортировки символов, обладающих графикой славянского происхождения.

Следует отметить, что пока речь идёт лишь об общем алфавите, а не об алгоритме сортировки. Алгоритм лексикографического сравнения Unicode-текстов обладает пятью основными уровнями, на трёх из которых будет использоваться составленный алфавит. Для языков алфавита порядок и содержание данных уровней могут быть различными.

Существует группа символов, которые присутствуют одновременно в нескольких национальных алфавитах. Если они размещены в разных алфавитах в одинаковой последовательности, для них также возможно создать общий алфавит. В противных случаях при составлении гиперазбуки необходимо жертвовать её соответствием тому алфавиту, чей порядок символов находится в «меньшинстве», то есть встречается редко и в малораспространённых алфавитах. Алфавит с таким расположением символов не удастся включить в гиперазбуку. Для такого языка придётся создать свой алгоритм сортировки.

Существует группа таких дополнительных символов, что каждый из них используется лишь в одном из национальных алфавитов. Для данных символов не имеет значения, в какой последовательности они должны быть расположены относительно друг друга. Поэтому данные символы можно располагать относительно друг друга по определённым особенностям начертания, порядок которых определяется частотой их встречаемости в языках.

Сопоставляя кириллические алфавиты, можно выделить несколько групп языков, для которых возможно составить общие алфавиты. Для ги-

неразбуки выбрана наибольшая из данных групп как по количеству входящих в неё языков, так и по распространённости данных языков.

Большинство народов за относительно малый исторический промежуток обладало алфавитами в разных системах письма (до четырёх систем письма для одного языка), а в настоящее время использует по 2 алфавита (как правило, неофициально). В случае интересующих нас языков (обладающих кириллической письменностью) основная масса текстов составлена на кириллице. Кириллические алфавиты также претерпевали изменения, в том числе реформу размещения дополнительных символов. Разрозненные и неполные материалы не всегда позволяют выяснить, была ли проведена реформа того или иного алфавита и как она была проведена.

Современный русский и древний кириллический алфавиты непротиворечивы, так как общие для них буквы расположены в одинаковой последовательности. Поэтому гиперазбуку требуется лишь дополнить старославянскими буквами таким образом, как они размещены в древней азбуке. Порядок символов другой славянской азбуки – глаголицы полностью соответствует кириллической азбуке в пересечении множеств их символов. Глаголица представляет собой ещё одно расширение базы славянских символов.

Интегрированная лексикографическая сортировка текстов на кириллице и глаголице облегчит работу лингвиста с базой данных. Допустим, пользователь не знает, на какой из двух славянских систем письма набрано слово, искомое в словаре. Так как обычно пользователю приходится работать с кириллицей, то, наблюдая определённый диапазон, он упустит слово, набранное на глаголице, если сортировка не была интегрирована для данных азбук. В иных случаях интегрированная сортировка способна принести неудобства, поэтому она должна быть включена опционально (факультативно). Специалисты по истории языка и языкознанию не имеют устоявшегося представления о том, каким был порядок букв в глаголической азбуке и в скором будущем вряд ли придут к общему мнению. Однако, например, в случае необходимости сортировки слов для создания словаря, требуется придерживаться какого-либо алфавита, в качестве которого можно временно выбрать глаголическую азбуку, чаще всего встречающуюся в энциклопедиях [13-15]. Эта азбука совпадает по порядку размещения общих по значению символов с кириллицей. Несмотря на довольно малое количество текстов на глаголице, она представляет определённый интерес для лингвистов. Это доказывает наличие в Интернете большого числа глаголических не-Unicode шрифтов, а также программного обеспечения, для печати на глаголице. Начиная с версии 4.1.0, Unicode поддерживает глаголицу. В ближайшее время ожидается появление Unicode-шрифтов, содержащих глаголицу [16-17].

Как было сказано выше, известные программные средства не производят лексикографически правильной сортировки всех славянских симво-

лов Unicode и текстов из данных символов. Например, пакеты MS Office и KOffice (пакет «офисных» программ для графической оболочки KDE над Linux) лексикографически упорядочивают в основном лишь символы современного русского языка и некоторые национальные символы. Прочие символы, в том числе древнекириллические, группируются в начале или в конце: иногда – по Unicode-индексам, иногда – по не вполне понятному принципу. Некоторые лексически не связанные символы сгруппированы вместе лишь из-за того, что они похожи графически. Корректная сортировка глаголических текстов в кодировке Unicode не поддерживается сегодня ни одной известной программой. Пакеты MS Office, KOffice и прочие упорядочивают тексты на глаголице по возрастанию Unicode-индексов. Но расположение символов на странице Unicode «Glagolitic» также полностью не соответствует ни одной из глаголических азбук и не может дать правильной сортировки хотя бы потому, что представляет собой два отдельных блока для разных регистров. Программное обеспечение третьих фирм, не использующее для сортировки текстов таблицы файла локали ОС Windows mswdat10.dll, осуществляет лексикографическую сортировку для ещё меньшего количества языков. Файлами локали называются файлы комплекта языковых настроек.

Основные принципы организации гиперазбуки следующие:

1. Символы русского и древнерусского языков располагаются соответственно русской кириллической азбуке.

2. Символы письма других славянских народов и письма народов бывшего СССР, если это не противоречит их алфавитам, располагаются непосредственно за символом русского языка, к которому они близки по лексическому значению (то есть по звучанию и по правилам использования в устной и письменной речи). Символы с изменённым начертанием и диакритикой обладают рядом лексикографических особенностей, которые можно классифицировать (пункты 2.1, 2.2). Данные лексикографические характеристики упорядочены, в том числе, по степени их распространённости в языках и по степени распространённости языков, в которых они встречаются:

2.1. Символы, обладающие диакритикой, упорядочиваются преимущественно по следующим диакритическим особенностям (в терминологии Unicode [18]): «DIAERESIS», «BREVE», «MACRON», «GRAVE», «DOUBLE ACUTE».

2.2. Символы с изменённым начертанием упорядочиваются преимущественно по следующим особенностям графем: «UPTURN», «DESCENDER», «TAIL», «HOOK», «STROKE», «VERTICAL STROKE», «TICK», «KOMI».

2.3. Символы с изменённым начертанием и диакритикой упорядочиваются аналогично пунктам 2.1 и 2.2.

2.4. Лигатуры: если первая буква в диграфах (слитых парах) одинакова, диграфы упорядочиваются между собой по второй букве.

2.5. Далее располагаются символы, заимствованные кириллическими алфавитами из латинской письменности.

2.6. Далее располагаются символы, заимствованные кириллическими алфавитами из латинской письменности и видоизменённые.

3. Символ глаголицы располагается ниже соответствующего ему символа кириллицы после символов, описанных в пункте 2 (порядок глаголической азбуки сохраняется).

4. Строчный символ располагается непосредственно после заглавного, если таковые имеются (то есть если символ представлении в двух регистрах).

5. Древнеславянские цифровые неалфавитные символы (в том числе составные диакритические) размещаются выше всех других символов в соответствии с правилами алфавитно-числовой сортировки по возрастанию.

6. Нецифровые составные диакритические символы размещаются следом за цифровыми, если они не имеют самостоятельного расположения в национальных алфавитах.

В спорных случаях символы могут быть упорядочены по их числовому значению, если таковое имеется в древних алфавитах (большинство древних символов кириллицы и глаголицы обозначало и букву, и цифру).

Таким образом, за каждым русским или древнерусским символом образуется группа родственных ему символов.

Если расположение каких-либо символов в гиперазбуке не соответствует порядку, описанному в пункте 2, колонки алфавита в приложении позволяют легко выяснить, для какого языка определено данное исключение. Для этого достаточно найти алфавит, названный под теми символами, взаимное расположение которых представляет интерес.

Цифры (знаки для составного обозначения чисел) в славянских азбуках обладали порядками и выше 1-го. В качестве цифровых символов относительно низких порядков (до третьего включительно) славяне использовали буквы алфавита, как правило (но не всегда) помеченные диакритикой (титлом или одной-двумя точками слева и справа от буквы). Не каждая буква в алфавитах обладала цифровым значением, но порядок возрастания цифр почти полностью совпадал с алфавитным порядком соответствующих им букв в кириллической азбуке, что удобно для сортировки. Для глаголической азбуки порядок цифровых и алфавитных значений символов совпадал полностью [19]. Размещение составной диакритики (в том числе цифровой) во главе алфавита также позволяет упростить алгоритм сортировки. Это связано с тем, что составные символы в текстовых форматах располагаются после буквы, к которой принадлежат (для которой служат, например, в качестве надстрочника). Составные символы диакритики Unicode из интересующих нас языков используются главным образом в церковнославянском. Наличие составной диакритики позволяет осуществить

поддержку церковнославянских символов, но осложняет шрифтовые решения. Составные символы диакритики не позволяют качественно печатать более одной комбинации символов. Поэтому, если проект получит развитие, гиперазбука должна будет пополниться за счёт церковнославянских лигатур с диакритикой, используемых вместо составных символов.

В приложении под каждым символом гиперазбуки указан его Unicode-индекс в шестнадцатеричном и десятеричном форматах для облегчения поиска и программирования. Под каждым символом, дополняющим современную русскую кириллицу, указаны языки, которые влияли на расположение данного символа. Это не означает, что среди списков указаны все соответствующие гиперазбуке языки: например, болгарский и словио не указаны ни под одним символом, так как не имеют расширений. В целях сохранения «читабельности» таблицы ряд пояснений под символами не приводится. Например, если в каком-либо алфавите для определённого символа используется лишь его строчный вариант, об этом не упоминается. Глаголические символы подписаны без пояснений, к какому типу глаголицы они относятся. В приложении под сербским алфавитом понимается сербохорватский кириллический алфавит, используемый в Сербии, Черногории и некоторых частях Боснии, под саамским – алфавит кильдинского диалекта. Названия символов в терминологии Unicode можно найти на официальном сайте Unicode [18] по их индексам. Почти полностью данный набор символов представлен в шрифте «Arial Unicode MS».

Все приведённые в приложении символы можно найти в диапазонах 0400-052F (Cyrillic 0400-04FF и расширение Cyrillic Supplement, содержащее в Unicode 4.1.0 лишь один раздел – Komi letters 0500-052F), Combining Diacritical Marks 0300-036F, General Punctuation 0200-026F, Spacing Modifier Letters 02B0-02FF и Glagolitic 2C00-2C5F. Во избежание недоразумений стоит отметить, что над индексами U+048C и U+048D изображён полумягкий знак, похожий на букву «Ять», а ячейка U+04C0 содержит символ «Палочка» для письма кавказских языков, графически похожий на «И десятеричное заглавное». Под символами апострофов и символом «Палочка» указаны и те алфавиты, в которых они не присутствуют как отдельные символы – это сделано для использования при сортировке составных лигатур. Соответственно ряду кавказских алфавитов над индексами U+2019 и U+02BC размещены варианты знака «Одинарный апостроф». Графически похожий на них символ U+0027 из раздела «Controls and Basic Latin» является «Апострофом-кавычкой», он не включён в алфавит. Знак «Апостроф-кавычка» может быть по ошибке введён пользователем при наборе текста вместо знака «Одинарный апостроф», но включать по этой причине знак «Апостроф-кавычка» в гиперазбуку нельзя, так как соответственно 4-му уровню стандарта сортировки Unicode [18] для знака «Апостроф-кавычка» применяются другие правила сортировки, чем для знака «Одинарный апостроф».

Наиболее похожие лексически символы в таблице Unicode совмещены. Это относится ко всем буквам древнеславянской письменности, сохранившимся в современных письменностях. Например, гаммаобразный вариант буквы «Ук» и «А йотированная» совмещены соответственно в обоих регистрах с буквами «У» и «Я», несмотря на существенное различие в начертаниях, а древнеславянская буква «Есть» не совмещена с «Е», так как отличается от неё в современном украинском алфавите.

Несмотря на наличие в Unicode некоторых древнерусских диакритических цифровых символов, на сегодня нет известных Unicode-шрифтов, в которых данные символы были корректно реализованы, то есть с использованием хотя бы одного из методов представления лигатур (например, при помощи нулевой ширины диакрита или по методу кернинга).

Принципы построения гиперазбуки можно рассмотреть на примере упорядочивания букв, фонетически близких к букве «И». Порядок букв древнерусской азбуки в различных источниках указан несколько различный. Например, «И десятеричное» встречается и до, и после «И». В подобных случаях символы гиперазбуки предлагается упорядочивать по их древнему числовому значению, а также соответственно национальным алфавитам (в нашем случае «И» перед «И десятеричным» – по числовому значению в древнерусской азбуке и соответственно алфавитам белорусского, украинского, коми и других языков). Непосредственно после строчной буквы «И» следуют буквы национальных алфавитов с диакритикой, образованные из «И». Исключением является «И-краткое», которое в русском алфавите следует после буквы «И», поэтому должно быть расположено ниже всех исторических и национальных символов, лексически близких к «И», что также соответствует алфавитам младописьменных языков. К тому же, как самостоятельная буква, «И-краткое» обладает своей диакритической вариацией. Для буквы «Ё», введённой факультативно, такое правило не действует (пример тому – чувашский алфавит). Если ввести данное правило для факультативной буквы, то порядок сортировки будет сильно различаться при использовании и не использовании факультативной буквы. Далее следует «И десятеричное». За ним следует похожий символ с двумя точками над «І» – буква «Yi». Далее следуют глаголические синонимы буквы «И». Далее следует «И-краткое», затем – его видоизменённые аналоги и его глаголический эквивалент, затем – заимствованная национальными алфавитами из латыни буква «Йот».

Кириллические буквы «Н» и «К» обладают наибольшим количеством вариаций с изменённым начертанием без диакритических знаков – шестью и пятью соответственно (считая диграфы). Так как интерес лексикографов чаще обращён к букве «К», на её примере будет рассмотрен принцип расположения вариантов букв с изменённым начертанием. Среди рассмотренных алфавитов лишь абхазский обладает сразу двумя изменёнными вариантами буквы «К». Он определяет порядок их взаимного расположения:

«Қ», «Ҟ» («KA WITH DESCENDER», «KA WITH STROKE»). «Қ» («KA WITH DESCENDER») является наиболее распространённой вариацией буквы «К» и особенность «DESCENDER» является наиболее распространённой среди символов с изменённым начертанием – она применяется для десяти букв, в одиннадцати языках (суммировались лишь те языки, чьи алфавиты повлияли на порядок расположения символов в гиперазбуке). Поэтому буквы «WITH DESCENDER» размещаются первыми среди производных букв. Графическая особенность, названная в Unicode как «HOOK», похожа на «DESCENDER», но встречается реже (она имеется у четырёх букв, используется в трёх языках). Буквы «WITH HOOK» размещаются следом за буквами «WITH DESCENDER». Это соответствует, в частности, алфавиту абхазского языка. Далее должны следовать буквы с особенностью «STROKE» (она свойственна для четырёх букв, используется в семи языках). Символически похожая на неё особенность «VERTICAL STROKE» свойственна двум буквам азербайджанского языка, они должны размещаться ниже. Далее размещается единственная в своём роде буква «BASHKIR KA» башкирского языка. В данном примере описано взаимное расположение символов с особенностями, характерными для вариаций буквы «К», в других случаях символы размещаются по тем же принципам.

Среди особенностей начертания букв и диакритических знаков в рамках славянских символов Unicode наиболее распространённой особенностью является наличие у буквы надстрочника, представляющего собой две точки, расположенные над символом («DIAERESIS»). Кроме базовых символов современной русской кириллицы наиболее часто в алфавитах встречается символ «Ү» («STRAIGTH U»), за ним следует символ «Ө» («BARRED O»).

Существует 5 основных уровней сравнения для сортировки Unicode-текста [18]: уровень базовых символов (L1); уровень акцентов (L2); уровень регистров (L3); уровень пунктуации (L4); уровень разделителей (Ln). Гиперазбука может использоваться в уровнях L1, L2 и L3. Для уровней L4 и Ln может использоваться стандартный алгоритм сравнения.

Специальные диакритические символы расположены в гиперазбуке таким образом, который позволяет производить правильную сортировку без дополнительного кода в алгоритме. Однако гиперазбука разработана в рамках возможностей Unicode, в котором ряд лигатур можно представить лишь в качестве составных символов. Следовательно, в ряде случаев (например, для двух кавказских языков: абазинского и адыгского) алгоритм сортировки должен пользоваться не только гиперазбукой, но и словарём. Словарь необходим для определения принадлежности составных символов к тем или иным лигатурам, то есть для контекстуальной чувствительности. Это можно назвать нулевым уровнем сравнения. Речь идёт о довольно редких лигатурах (например, «Джь» и «Джв»), поэтому пользователь может быть вполне удовлетворён даже алгоритмом сортировки, основанным на одной гиперазбуке.

Возможное расширение Unicode и кириллицы в Unicode влечёт развитие проекта гиперазбуки. Ячейки, которые ещё могут быть предоставлены для кириллицы в двух первых байтах Unicode, скорее всего, будут отданы под национальные символы. Ближайшее расширение Unicode-кириллицы, ориентированной на древнюю письменность, возможно за счёт размещения символов церковно-славянского языка, так как имеются заинтересованные в этом специалисты. Если отложить задачу дальнейшего полного лигатурирования древней кириллицы, то свободные ячейки основной кириллической страницы Unicode можно было бы заполнить четырьмя нелигатурными древними символами (использующимися в церковно-славянском языке) и символами, по какой-то причине не размещёнными среди древних цифровых символов высоких порядков.

В случае если представится возможность не только пополнения, но и расширения кириллического диапазона Unicode, размещение большинства символов желательно произвести по возможности единовременно, на основании разработанных принципов и определённого на данный момент базиса уникальных символов. При индексировании желательно не только учесть порядок, представленный, например, в гиперазбуке, но и разместить символы по возможности таким образом, чтобы младший байт совпадал у символов с одинаковым признаком. Это технически облегчит, например, поиск тяжёлых ударений в тексте.

В настоящий момент гиперазбука используется в алгоритме сортировки в инсталлируемом пакете макросов для MS Word «Генератор словарей». Проект гиперазбуки предусматривает возможность обсуждения и внесения поправок.

Литература

1. Сайт поддержки электронного словаря ABBYY Lingvo – http://www.lingvo.ru.
2. Ассоциация лексикографов Lingvo – http://www.lingvoda.ru.
3. Сообщество славянской типографики – http://cslav.org.
4. Steven J. Searle. Brief History of Character Codes in North America, Europe, and East Asia – http://tronweb.super-nova.co.jp/characcodehist.html.
5. Техническая спецификация формата шрифтов OpenType // Microsoft Corporation – http://www.microsoft.com/OpenType/OTSpec.
6. Интернет и алфавиты языков России // Российский общеобразовательный портал, Министерство образования и науки РФ – http://www.peoples.org.ru/alfavit.html.
7. Проект по разработке шрифтов «Ирмологий» – **http://irmologion.ru**.
8. Информационно-поисковая система «Манускрипт» // Лаборатория по автоматизации филологических работ УдГУ – http://manuscripts.ru.
9. Описание формата HIP / Проект по сотрудничеству в области разработки методов электронного представления текстов «Печатный двор» – http://www.pechatnyj-dvor.narod.ru/docs.html.
10. Технология смешанного набора – http://www.znamen.ru/tsn.htm.
11. Энциклопедия «Википедия»: Список языков с алфавитами на основе кириллицы – http://ru.wikipedia.org/wiki/Languages_using_Cyrillic.

12. Интернет-версия газеты «Известия науки» – http://www.inauka.ru

13. Русский язык: Энциклопедия. – М.: Научное издательство «Большая Российская энциклопедия», 2003.

14. Лингвистический энциклопедический словарь. – М.: Советская энциклопедия, 1990.

15. Большой энциклопедический словарь: Языкознание. – М.: Научное издательство «Большая Российская энциклопедия», 1998.

16. Проект по разработке шрифтов «Alphabetum Unicode font» – http://guindo.cnice.mecd.es/~jmag0042/alphaeng.html.

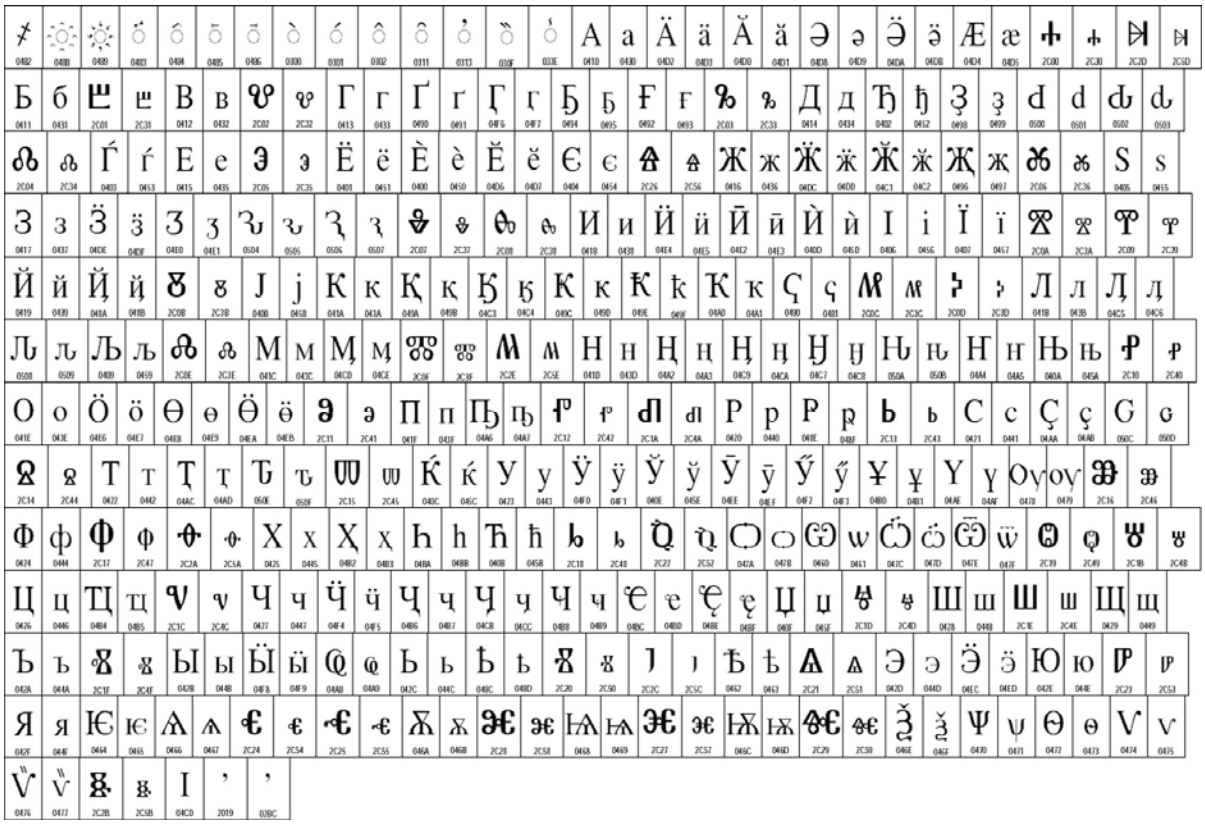
17. MUFI (Medieval Unicode Font Initiative) – http://gandalf.aksis.uib.no/mufi.

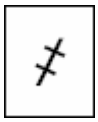
18. The Unicode Consortium – http://www.unicode.org.

19. Проект по разработке шрифтов «Русское шрифтовое зало» – http://rp.spb.su/zalo.

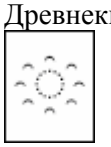
Приложение. Гиперазбука (версия Гиперазбука Unicode 4.1.0 Linguistic)

Гиперазбука (гипералфавит символов славянского происхождения различных народов и эпох), созданная в рамках множества символов стандарта Unicode 4.1.0.

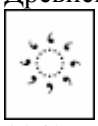




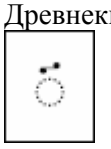
1154, U+0482
Древнекирилл.



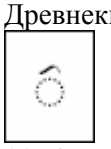
1160, U+0488
Древнекирилл.



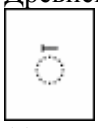
1161, U+0489
Древнекирилл.



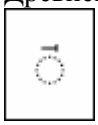
1155, U+0483
Древнекирилл.



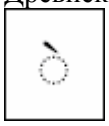
1156, U+0484
Древнекирилл.



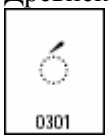
1157, U+0485
Древнекирилл.



1158, U+0486
Древнекирилл.

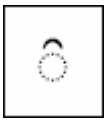


768, U+0300
Древнекирилл.

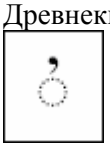


769, U+0301
Древнекирилл.

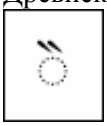
770, U+0302
Древнекирилл.



785, U+0311
Древнекирилл.

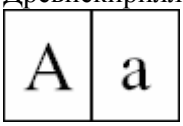


787, U+0313
Древнекирилл.

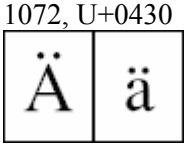


783, U+030F
Древнекирилл.

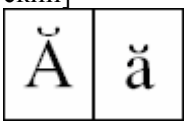
830, U+033E
Древнекирилл.



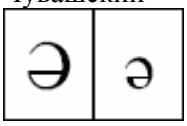
1040, U+0410
1072, U+0430



1234, U+04D2
1235, U+04D3
Марийский, гагаузский, [саамский]

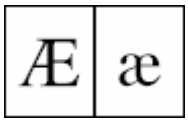


1232, U+04D0
1233, U+04D1
Чувашский

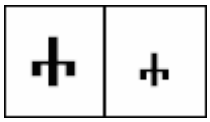


1240, U+04D8
1241, U+04D9
[Казахский], [калмыцкий], татарский

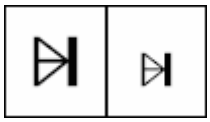
1242, U+04DA
1243, U+04DB



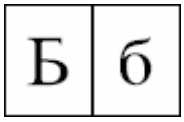
1236, U+04D4
1237, U+04D5
Осетинский



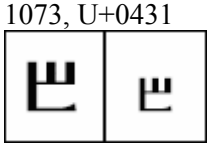
11264, U+2C00
11312, U+2C30
Глаголический



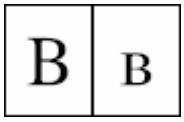
11309, U+2C2D
11357, U+2C5D
Глаголический



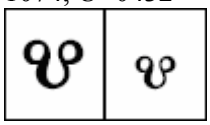
1041, U+0411
1073, U+0431



11265, U+2C01
11313, U+2C31
Глаголический

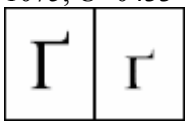


1042, U+0412
1074, U+0432



11266, U+2C02
11314, U+2C32
Глаголический

1043, U+0413
1075, U+0433

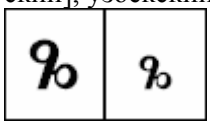


1168, U+0490
1169, U+0491
Украинский, цыганский

1270, U+04F6
1271, U+04F7
[Абхазский]

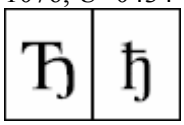
1172, U+0494
1173, U+0495
[Абхазский]

1170, U+0492
1171, U+0493
[Азербайджанский], [башкирский], [казахский], [таджикский], узбекский

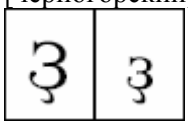


11267, U+2C03
11315, U+2C33
Глаголический

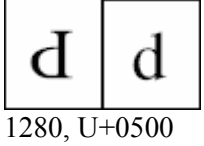
1044, U+0414
1076, U+0434



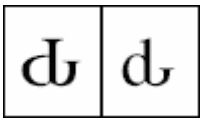
1026, U+0402
1106, U+0452
[Сербский], [черногорский]



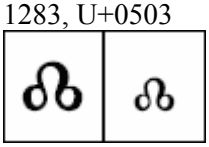
1176, U+0498
1177, U+0499
[Башкирский]



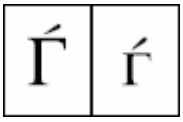
1280, U+0500
1281, U+0501



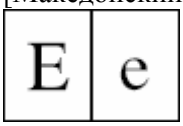
1282, U+0502
1283, U+0503



11268, U+2C04
11316, U+2C34
Глаголический



1027, U+0403
1107, U+0453
[Македонский]



1045, U+0415
1077, U+0435

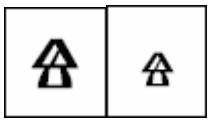
11269, U+2C05
11317, U+2C35
Глаголический

1025, U+0401
1105, U+0451

1024, U+0400
1104, U+0450

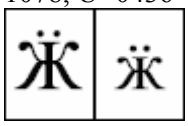
1238, U+04D6
1239, U+04D7
Чувашский

1028, U+0404
1108, U+0454
Древнекирилл., украинский

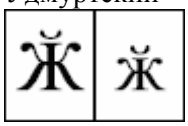


11302, U+2C26
11350, U+2C56
Глаголический

1046, U+0416
1078, U+0436

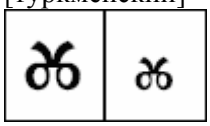


1244, U+04DC
1245, U+04DD
Удмуртский

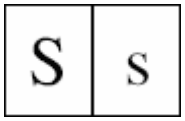


1217, U+04C1
1218, U+04C2
Молдавский

1174, U+0496
1175, U+0497
[Калмыцкий], татарский, [туркменский]

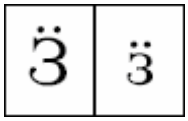


11270, U+2C06
11318, U+2C36
Глаголический



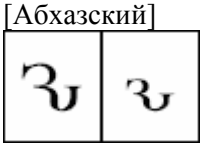
1029, U+0405
1109, U+0455
Древнекирилл.

1047, U+0417
1079. U+0437

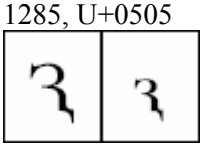


1246, U+04DE
1247, U+04DF
Удмуртский

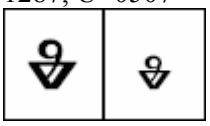
1248, U+04E0
1249, U+04E1
[Абхазский]



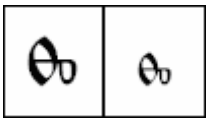
1284, U+0504
1285, U+0505



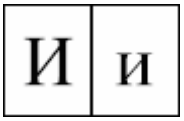
1286, U+0506
1287, U+0507



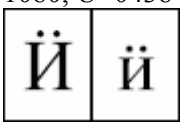
11271, U+2C07
11319, U+2C37
Глаголический



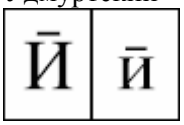
11272, U+2C08
11320, U+2C38
Глаголический



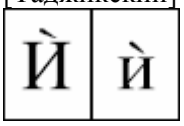
1048, U+0418
1080, U+0438



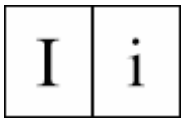
1252, U+04E4
1253, U+04E5
Удмуртский



1250, U+04E2
1251, U+04E3
[Таджикский]



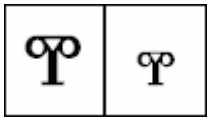
1037, U+040D
1117, U+045D



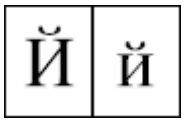
1030, U+0406
1110, U+0456
Древнекирилл., белорусский, коми, ногайский, украинский

1031, U+0407
1111, U+0457
Древнекирилл., украинский

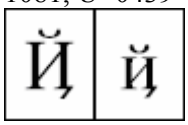
11274, U+2C0A
11322, U+2C3A
Глаголический



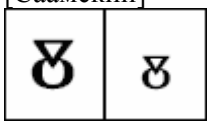
11273, U+2C09
11321, U+2C39
Глаголический



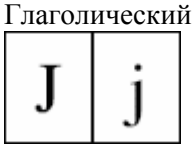
1049, U+0419
1081, U+0439



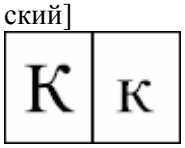
1162, U+048A
1163, U+048B
[Саамский]



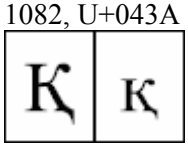
11275, U+2C0B
11323, U+2C3B
Глаголический



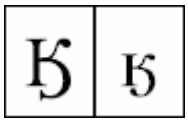
1032, U+0408
1112, U+0458
[Азербайджанский], [македонский], [саамский], [сербский], [черногорский]



1050, U+041A
1082, U+043A

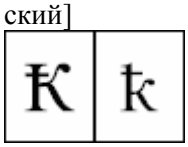


1178, U+049A
1179, U+049B
[Абхазский], [казахский], [таджикский], узбекский

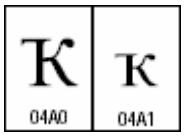


1219, U+04C3
1220, U+04C4
Чукотский

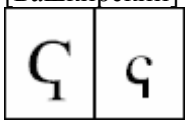
1180, U+049C
1181, U+049D
[Азербайджанский]



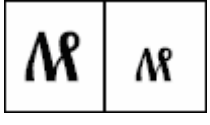
1182, U+049E
1183, U+049F
Абхазский



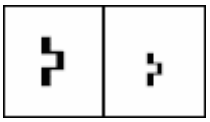
1184, U+04A0
1185, U+04A1
[Башкирский]



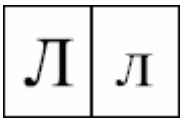
1152, U+0480
1153, U+0481
Древнекирилл.



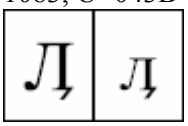
11276, U+2C0C
11324, U+2C3C
Глаголический



11277, U+2C0D
11325, U+2C3D
Глаголический

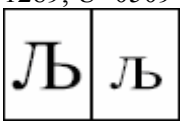


1051, U+041B
1083, U+043B

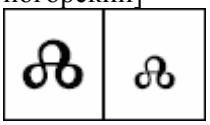


1221, U+04C5
1222, U+04C6
[Саамский]

1288, U+0508
1289, U+0509



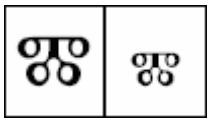
1033, U+0409
1113, U+0459
[Македонский], [сербский], [черногорский]



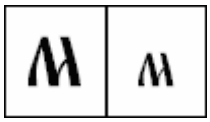
11278, U+2C0E
11326, U+2C3E
Глаголический

1052, U+041C
1084, U+043C

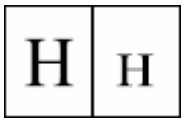
1229, U+04CD
1230, U+04CE
[Саамский]



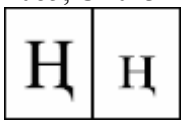
11279, U+2C0F
11327, U+2C3F
Глаголический



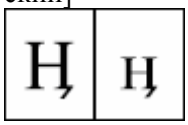
11310, U+2C2E
11358, U+2C5E
Глаголический



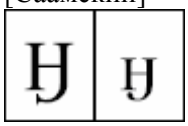
1053, U+041D
1085, U+043D



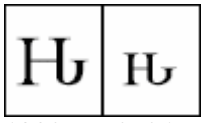
1186, U+04A2
1187, U+04A3
[Башкирский], [казахский], [калмыцкий], киргизский, татарский, тувинский, [туркменский]



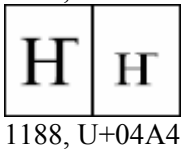
1225, U+04C9
1226, U+04CA
[Саамский]



1223, U+04C7
1224, U+04C8
[Саамский], чукотский



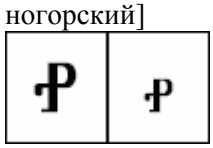
1290, U+050A
1291, U+050B



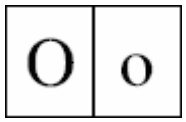
1188, U+04A4
1189, U+04A5

[Алтайский], марийский, узбекский

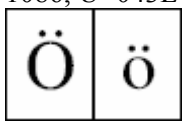
1034, U+040A
1114, U+045A
[Македонский], [сербский], [черногорский]



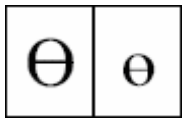
11280, U+2C10
11328, U+2C40
Глаголический



1054, U+041E
1086, U+043E

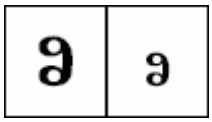


1254, U+04E6
1255, U+04E7
[Алтайский], марийский, гагаузский, коми, ногайский, удмуртский

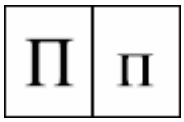


1256, U+04E8
1257, U+04E9
[Азербайджанский], [башкирский], бурятский, [казахский], [калмыцкий], киргизский, монгольский, татарский, тувинский, [туркменский]

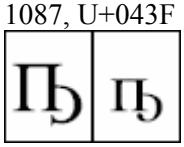
1258, U+04EA
1259, U+04EB



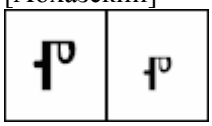
11281, U+2C11
11329, U+2C41
Глаголический



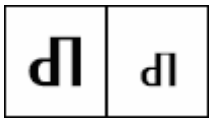
1055, U+041F
1087, U+043F



1190, U+04A6
1191, U+04A7
[Абхазский]

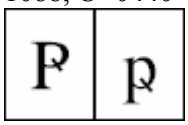


11282, U+2C12
11330, U+2C42
Глаголический



11290, U+2C1A
11338, U+2C4A
Глаголический

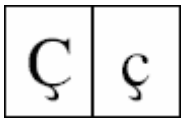
1056, U+0420
1088, U+0440



1166, U+048E
1167, U+048F
[Саамский]

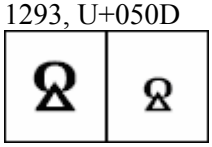
11283, U+2C13
11331, U+2C43
Глаголический

1057, U+0421
1089, U+0441

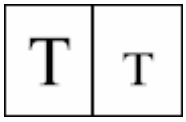


1194, U+04AA
1195, U+04AB
[Башкирский], чувашский

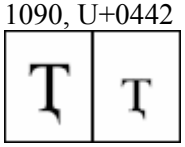
1292, U+050C
1293, U+050D



11284, U+2C14
11332, U+2C44
Глаголический

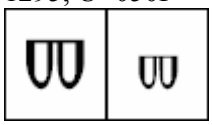


1058, U+0422
1090, U+0442



1196, U+04AC
1197, U+04AD
[Абхазский]

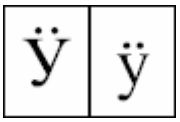
1294, U+050E
1295, U+050F



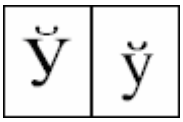
11285, U+2C15
11333, U+2C45
Глаголический

1036, U+040C
1116, U+045C
[Македонский]

1059, U+0423
1091, U+0443

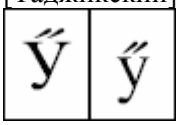


1264, U+04F0
1265, U+04F1
[Алтайский], гагаузский, марийский, ногайский

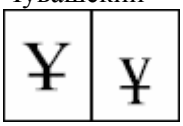


1038, U+040E
1118, U+045E
Белорусский, узбекский

1262, U+04EE
1263, U+04EF
[Таджикский]

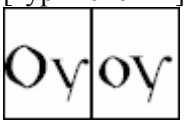


1266, U+04F2
1267, U+04F3
Чувашский



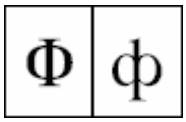
1200, U+04B0
1201, U+04B1
[Казахский]

1198, U+04AE
1199, U+04AF
[Азербайджанский], балкарский, [башкирский], бурятский, [казахский], [калмыцкий], карачаевский, киргизский, монгольский, татарский, тувинский, [туркменский]

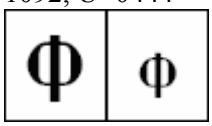


1144, U+0478
1145, U+0479
Древнекирилл.

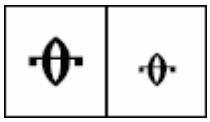
11286, U+2C16
11334, U+2C46
Глаголический



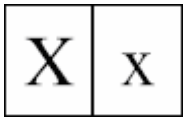
1060, U+0424
1092, U+0444



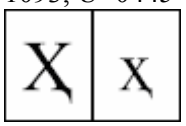
11287, U+2C17
11335, U+2C47
Глаголический



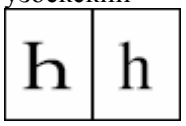
11306, U+2C2A
11354, U+2C5A
Глаголический



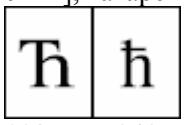
1061, U+0425
1093, U+0445



1202, U+04B2
1203, U+04B3
[Абхазский], [таджикский], узбекский

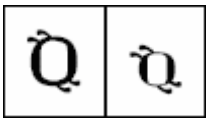


1210, U+04BA
1211, U+04BB
[Азербайджанский], [башкирский], [казахский], татарский

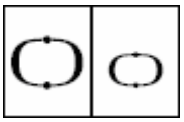


1035, U+040B
1115, U+045B
Бурятский

11288, U+2C18
11336, U+2C48
Глаголический

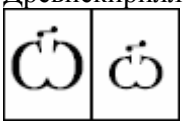


11298, U+2C22
11346, U+2C52
Глаголический

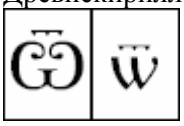


1146, U+047A
1147, U+047B
Древнекирилл.

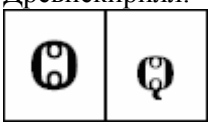
1120, U+0460
1121, U+0461
Древнекирилл.



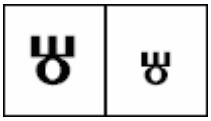
1148, U+047C
1149, U+047D
Древнекирилл.



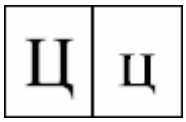
1150, U+047E
1151, U+047F
Древнекирилл.



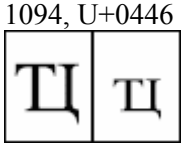
11289, U+2C19
11337, U+2C49
Глаголический



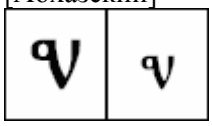
11291, U+2C1B
11339, U+2C4B
Глаголический



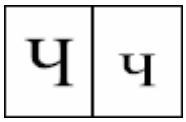
1062, U+0426
1094, U+0446



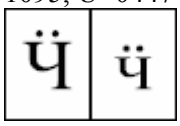
1204, U+04B4
1205, U+04B5
[Абхазский]



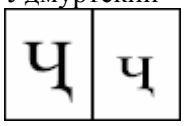
11292, U+2C1C
11340, U+2C4C
Глаголический



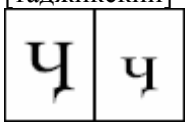
1063, U+0427
1095, U+0447



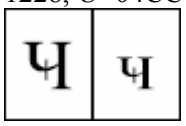
1268, U+04F4
1269, U+04F5
Удмуртский



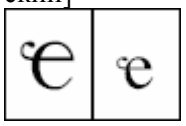
1206, U+04B6
1207, U+04B7
[Абхазский], [таджикский]



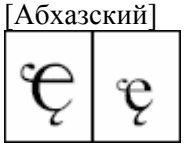
1227, U+04CB
1228, U+04CC



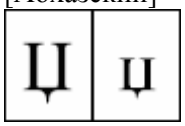
1208, U+04B8
1209, U+04B9
[Азербайджанский]



1212, U+04BC
1213, U+04BD
[Абхазский]

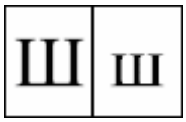


1214, U+04BE
1215, U+04BF
[Абхазский]

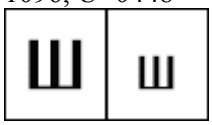


1039, U+040F
1119, U+045F
[Македонский], [сербский], [черногорский]

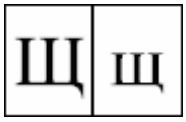
11293, U+2C1D
11341, U+2C4D
Глаголический



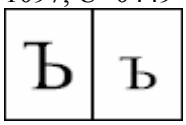
1064, U+0428
1096, U+0448



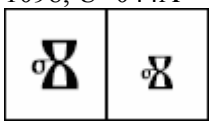
11294, U+2C1E
11342, U+2C4E
Глаголический



1065, U+0429
1097, U+0449

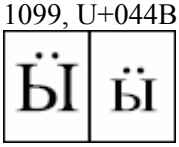


1066, U+042A
1098, U+044A

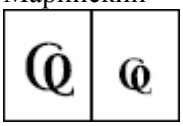


11295, U+2C1F
11343, U+2C4F
Глаголический

1067, U+042B
1099, U+044B

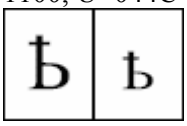


1272, U+04F8
1273, U+04F9
Марийский

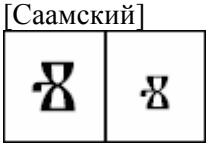


1192, U+04A8
1193, U+04A9
[Абхазский]

1068, U+042C
1100, U+044C



1164, U+048C
1165, U+048D
[Саамский]



11296, U+2C20
11344, U+2C50
Глаголический

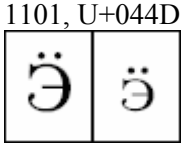
11308, U+2C2C
11356, U+2C5C
Глаголический

1122, U+0462
1123, U+0463
Древнекирилл.

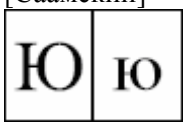
11297, U+2C21
11345, U+2C51
Глаголический

1069, U+042D

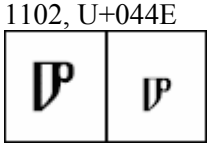
1101, U+044D



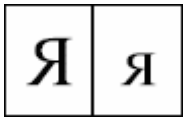
1260, U+04EC
1261, U+04ED
[Саамский]



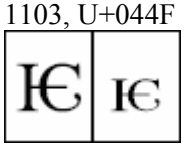
1070, U+042E
1102, U+044E



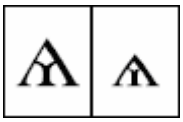
11299, U+2C23
11347, U+2C53
Глаголический



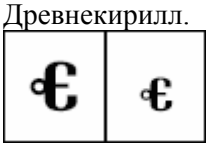
1071, U+042F
1103, U+044F



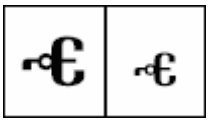
1124, U+0464
1125, U+0465
Древнекирилл.



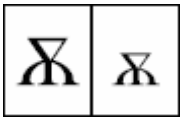
1126, U+0466
1127, U+0467
Древнекирилл.



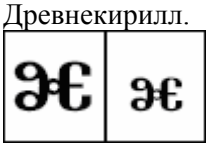
11300, U+2C24
11348, U+2C54
Глаголический



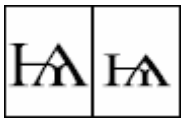
11301, U+2C25
11349, U+2C55
Глаголический



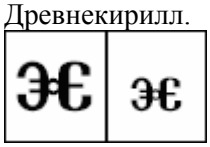
1130, U+046A
1131, U+046B
Древнекирилл.



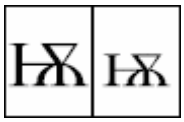
11304, U+2C28
11352, U+2C58
Глаголический



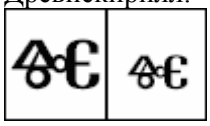
1128, U+0468
1129, U+0469
Древнекирилл.



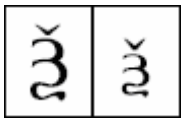
11303, U+2C27
11351, U+2C57
Глаголический



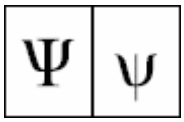
1132, U+046C
1133, U+046D
Древнекирилл.



11305, U+2C29
11353, U+2C59
Глаголический

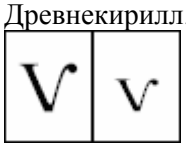


1134, U+046E
1135, U+046F
Древнекирилл.

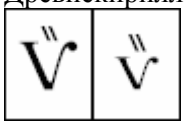


1136, U+0470
1137, U+0471
Древнекирилл.

1138, U+0472
1139, U+0473
Древнекирилл.



1140, U+0474
1141, U+0475
Древнекирилл.

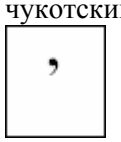


1142, U+0476
1143, U+0477
Древнекирилл.

11307, U+2C2B
11355, U+2C5B
Глаголический

1216, U+04C0
Абазинский, [аварский], адыгский, даргинский, ингушский, [кабардинский], [чеченский]

8217, U+2019
[Азербайджанский], [саамский], узбекский, чукотский



700, U+02BC
[Азербайджанский], [саамский], узбекский, чукотский