V. Проблемы Юникода

1. Ошибка, касающаяся знаков U+0598 и U+05AE

В UnicodeData 3.0.1 [1] описания этих знаков таковы:

0598;HEBREW ACCENT ZARQA;Mn;230;NSM;;;;;N;;*;;; (230 = выше)
05AE;HEBREW ACCENT ZINOR;Mn;228;NSM;;;;;N;;;;; (228 = выше и слева)

Очевидно, что эти описания изменились со времени Unicode 1, однако в неверном направлении, что теперь сделало эти описания *еще более последовательно неверными.*

Все источники кроме таблиц кодов, т.е. различные книги по грамматике библейского иврита и книга Броера о знаках кантилляции (Знаки кантилляции Танаха в 21 книге и в книгах Ийов, Мишлей, Tehилим. Иерусалим, 5742 год (т.е. 1981 – 1982 год)),

מרדכי ברויאר .טעמי המקרא בכ"א ספרים ובספרי אמ"ת .ירושלים .תשמ"ב

мнение, выраженное в которой, для меня является решающим, согласны в следующем:

● **Зарка,** знак, который пишется после слова, – это разделительный символ кантилляции, используемый в 21 книге Библии. Такой же знак, также пишущийся после слова и также являющийся разделительным символом, в остальных 3 книгах называется **цинор.** Обычно такие пишущиеся после слова знаки кантилляции располагаются выше и слева последней буквы слова, но бывает, они располагаются выше последней буквы, и это не влияет на семантику.

● **Цинорит,** пишущийся на верхней части буквы в начале или в середине слова, относится к соединительным символам, используемым в трех книгах.

● Оба имени, цинор и цинорит, пишутся, начиная с буквы цади, в Юникоде TSADI (U+05E6), а не с зайин, в Юникоде ZAYIN (U+05D6). Следуя транскрипции, используемой в Юникоде для всех других названий на иврите, они, следовательно, должны были бы быть написаны, начиная с “TS-”, а не с “Z-”. Удвоенные согласные никогда не отражаются в транскрипции, используемой в названиях букв и знаков Юникода; поэтому мы последовательно пишем «цинор», «цинорит» вместо транскрипции «циннор», «циннорит», используемой в других главах этой статьи.

Таблица знаков кантилляции на странице 867 18-го тома Еврейской энциклопедии [издающейся с 1949 года на иврите в Израиле] содержит только знаки кантилляции из 21 книги и, следовательно, не содержит цинорит. Поэтому эта энциклопедия не может служить источником для того, чтобы преодолеть путаницу между цинорит и зарка тождественными цинору. То, что в энциклопедии написание синонима знака зарка приведено, как «цинори», а не как «цинор» немного добавляет путаницы, но не противоречит краткому описанию того, как обстоят дела, приведенному выше.

Несмотря на эти факты, Юникод (следуя в этом вопросе израильскому национальному стандарту SI 1311-2) считает различными **ZARQA** и **ZINOR** (используется именно такое написание!), где знак ZINOR, по-видимому, играет роль цинорит, и видимо, путаница вызвана тем, что имена «цинор» и «цинорит» гораздо ближе друг к другу, чем «зарка» и «цинор». Такая интерпретация (что вместо ZINOR следовало бы писать TSINORIT) поддерживается порядком, в котором приводятся знаки

кантилляции: сначала все разделительные знаки в порядке убывания разделительной силы, затем соединительные знаки, и в каждом из этих классов сначала приводятся знаки, используемые в 21 книге (или во всех книгах), а затем – используемые в 3 книгах. Исходя из того, что порядок именно таков, видно, что код U+0598 предполагалось использовать для разделительного знака кантилляции, имеющего среднюю разделительную силу и используемого в 21 книге, каков и есть знак зарка. Можно спокойно заключить [2], что на самом деле ZARQA означает «зарка, тождественный цинору» [3], а ZINOR означает «цинорит». Однако [если принять такую интерпретацию, то возникает следующее затруднение:] таблица графических изображений символов показывает эти два знака переставленными один на место другого, и классы сочетаемости [1] (влияние которых на нормализацию [4] в данном конкретном случае минимально) согласуются с таблицей графических изображений символов и не согласуются с приведенной выше интерпретацией имен знаков. После версии Unicode 1.0, но до версии Unicode 3.0 примечание “=zinorit” [5] было добавлено к знаку ZARQA в соответствии с графическим изображением символа. Мнения разделились по вопросу, достаточно ли такого примечания, чтобы с уверенностью отвергнуть первоначальную интерпретацию, что ZARQA означает «зарка», а ZINOR означает «цинорит». В любом случае, обе интерпретации в недавних обсуждениях были приняты несколькими людьми как сами собой разумеющиеся, так что не является неправильным утверждать, что осталась неоднозначность и она является существенной.

Критерии кодирования

В этой статье проблемы были идентифицированы и возможность разных способов решения была оценена по следующим критериям:

1. Юникод (а также SI 1311-2) придерживаются стратегии, заключающейся в том, что графически одинаковые, но семантически разные символы должны считаться одними тем же символом, а не различными символами (Unicode 3.0, стр. 17); это правило имеет ряд исключений, которые здесь неприменимы. Юникод следует этой стратегии и в отношении других знаков кантилляции: типэха=тарха, мерха=йоред, метег=силук и т. д. Следовательно зарка и цинор должны считаться одним и тем же знаком. Должен ли цинорит считаться отличным от зарка=цинор, зависит от того, считают ли его иное положение по отношению к букве, на которой он ставится, свойством знака или деталью процесса визуализации. Поскольку теперь последним трем знакам присвоено два кода и существуют вариации графического изображения символа определенно неверные для цинорит и верные для зарка=цинор, то было бы шагом назад объединять коды всех трех знаков.

2. Если дана последовательность символов, стандарт должен точно указать, как ее закодировать.

   Однако возможна некоторая неизбежная неоднозначность, когда стандарт указывает, что символы различны, а они могут быть похожими друг на друга при *некоторых* способах визуализации. Могут быть приведены следующие примеры такой ситуации:

   ○ Не определено, требуется ли сохранять при кодировании текста, написанного латинским шрифтом в 18-ом веке, различие между долгим s (U+017F) и конечным s (U+0073) или оба эти символа могут быть представлены единственным кодом (U+0073), используемым сегодня.

   ○ Аналогично в случае, если встретится знак зарка, расположенный так же, как цинорит (т.е. над буквой), не определено может ли он кодироваться тем же кодом, что цинорит, хотя они являются различными знаками по критерию 1 (если применить принцип «кодируй то, что *написано,* а не то, что это *значит»* в духе объединения).

В таких случаях, когда абсолютная единственность кодирования не может быть достигнута, по крайней мере, требуется, чтобы стандарт стал однозначным после того, как его пользователи установят правила трактовки соответствующих символов. Другими словами, может быть, неизбежна неоднозначность того, как *применять* стандарт в данной ситуации, однако то, что стандарт *гласит,* не допускает неоднозначностей.

3. Несмотря на то, что графические изображения, приведенные в Юникоде, не являются обязательными для символов, применение графических изображений стандарта не должно препятствовать идентификации символов (Unicode 3, стр. 40, п. D2). В настоящее время этот принцип нарушен: если одно из двух графических изображений, которое используется для ZARQA и ZINOR, обозначает зарка, а другое нет, то совершенно однозначно, что это *не* графическое изображение ZARQA.

4. Имена символов Юникода больше не должны изменяться (правило, относящееся к вебсайту Юникода, есть ли оно также в документации стандарта? [6])

5. Изменение свойств символов Юникода [т.е. параметров его описания] должно быть сведено к минимуму (Unicode 3.0, стр. 73).

6. Желательно, но не обязательно, чтобы набор знаков кантилляции в Юникоде соответствовал израильскому стандарту SI 1311-2. Если это не так, то замечание на стр. 187 документа Unicode 3.0 должно быть изменено.

7. Желательно, но не обязательно, чтобы знаки кантилляции в коде шли в том порядке, который представляется логичным с точки зрения семантики знаков.

Проблема и возможные решения

Выше, когда мы начинаем с критерия 1, мы находим, что два различных имени символов Юникода, ZARQA и ZINOR, обозначают один и тот же символ, а имя символа, который не имеет аналога и должен быть закодирован, TSINORIT, пропущено и не существует как имя символа Юникода. Из-за критерия 4 эта проблема не имеет решения. Аналогично критерий 3 не может быть полностью выполненным, поскольку невозможно по-разному идентифицировать символы, названные двумя именами, и поставить этим именам в соответствие разные графические изображения символов, если на деле это имена одного и того же символа. Решение будет состоять в том, чтобы предоставить пользователю достаточно информативный комментарий, сообщающий, что несмотря на неточность имен символов, которой уже невозможно избежать, критерий 2 выполнен, и определить способы идентификации символов так, чтобы критерий 3 был не так уж грубо нарушен.

Если принять всерьез заявление, что Юникод определяет символы, а не их графические изображения, и в особенности формулировку принципа D2, выражающего эту идею, тогда необходимо изменить графическое изображение таким образом, чтобы оно соответствовало символу, вместо того, чтобы оставить и графическое изображение, и имя символа такими, какие они есть: если графическое изображение, соответствующее имени "LATIN CAPITAL LETTER A", показывает “B”, то изображение неверно, а "LATIN CAPITAL LETTER A" не является несколько странным именем “B”. Исходя из подобных рассуждений, эта глава первоначально содержала просьбу изменить одно имя (ZINOR на TSINORIT), а затем, чтобы было соблюдено правило не изменять имена, – просьбу, чтобы графические изображения были восстановлены в надлежащем порядке несмотря на то, что имена не могут быть исправлены. То, что показанные графические изображения уже реализуются в шрифтах (так что они де-факто трактуются как стандартные) и что по меньшей мере одно имя невозможно

полностью исправить, может служить аргументом за то, чтобы сохранить оба символа переставленными, такими, какими они сейчас включены в таблицу графических изображений символов.

Далее, существует четыре способа сделать это:

● Сделать акцент на идентификации символов, интерпретировать имена символов самым прямым способом (так, как это первоначально подразумевалось делать), прекратить использовать текущие значения параметров, присвоенные этим символам в системе Юникод, и соответствующие изображения в таблицах графических изображений символов: другими словами, сделать так, чтобы были выполнены критерии 3 и 7 за счет критерия 5. Такой подход применяется ниже, в Решении 1.

● Сделать акцент на стабильности значений параметров, присвоенных символам в системе Юникод, в особенности классов сочетаемости [1], а также таблиц графических изображений символов; другими словами, сделать так, чтобы был выполнен критерий 5 за счет критериев 3 и 7. Так как положение знаков относительно буквы, над которой они ставятся, не *всегда* одинаково, определение символов может быть сформулировано так, чтобы критерий 3 был *по меньшей мере иногда* выполнен для каждого из символов. Такой подход применяется ниже, в Решении 2.

● Решение 3, приведенное ниже, немногим отличается от решения 2, а именно тем, что существующая *иногда* неопределенность положения знаков не эксплуатируется для того, чтобы представить неточные имена символов более достоверными. Вместо этого одно из имен символов без обиняков охарактеризовано как неверное.

● Наиболее кардинальный способ обеспечить выполнение критерия 3 – это исключить из игры одно из неверных имен символов, объявив использование символа нежелательным, и ввести пропущенное имя как новый символ. Это может быть сделано многими способами. Один из них представлен ниже как решение 4.

Первоначально решение 1 предлагалось в качестве единственного решения. Теперь же, поскольку *точное* совпадение между именами символов и тем, как они идентифицируются, все равно не может быть достигнуто, решение 2 могло бы быть удовлетворительным компромиссом, обеспечивающим баланс между стабильностью стандарта и верностью содержащихся в нем определений. Я оставляю за различными организациями, в чьем ведении находятся стандарты, право самим найти решение, которое они считают лучше всего соответствующим процедурам, принятым для этих стандартов. Лично я предпочел бы решение 1 (или даже более последовательное решение 4), но гораздо важнее, чтобы стандарт как можно скорее стал однозначным, причем также для его пользователей (людей, производящих кодирование текста и обработку закодированного текста), а не только для людей, занимающихся разработкой шрифтов.

Решение 1

● Заменить соответствующий текст на стр.387 стандарта следующим текстом:

0598 HEBREW ACCENT ZARQA
= tsinor, zinor
05AE HEBREW ACCENT ZINOR

= tsinorit, zinorit
* этот знак кантилляции – не Tsinor (=Zinor), а Tsinorit (=Zinorit), что не соответствует имени знака
-> 0598 zarqa

● Изменить значения классов сочетаемости [1] следующим образом:

0598;HEBREW ACCENT ZARQA;Mn;228;NSM;;;;;N;;*;;; (228=выше и слева)
05AE;HEBREW ACCENT ZINOR;Mn;230;NSM;;;;;N;;;;; (230=выше)

Модификация не должна повлиять на результат нормализации [4], т.к. над одной буквой никогда не ставится более одного знака кантилляции этих классов.

● Изменить положения знаков в таблицах графических изображений символов, чтобы они отражали исправление значений классов сочетаемости. Иными словами, просто переставить два изображения.

Решение 2

● Заменить соответствующий текст на стр.387 стандарта следующим текстом:

0598 HEBREW ACCENT ZARQA
= tsinor, zinor, tsinorit, zinorit
* должен использоваться для знака кантилляции Tsinorit (=Zinorit)
* может быть использован для знаков кантилляции Zarqa и Tsinor (=Zinor), чтобы недвусмысленно указать, что знак кантилляции должен ставится над буквой подобно Tsinorit
-> 05AE zinor
05AE HEBREW ACCENT ZINOR
= zarqa, tsinor
* должен регулярно использоваться для знаков кантилляции Zarqa и Tsinor (=Zinor), кроме случая, когда нужно недвусмысленно указать, что знак кантилляции должен ставиться над буквой
-> 0598 zarqa [7]

Решение 3

● Заменить соответствующий текст на стр.387 стандарта следующим текстом:

0598 HEBREW ACCENT ZARQA
= tsinorit, zinorit
* этот знак кантилляции – не Zarqa, а Tsinorit (=Zinorit), что не соответствует имени знака
-> 05AE zinor
05AE HEBREW ACCENT ZINOR
= zarqa, tsinor
-> 0598 zarqa

Решение 4

● Заменить соответствующий текст на стр. 387 стандарта следующим текстом:

0598 HEBREW ACCENT ZARQA
= tsinor, zinor

05A2 HEBREW ACCENT TSINORIT
= zinorit
05AE HEBREW ACCENT ZINOR
* не рекомендуется для использования
-> 0598 zarqa
-> 05A2 tsinorit

● Изменить значения классов сочетаемости [1] следующим образом:

0598;HEBREW ACCENT ZARQA;Mn;228;NSM;;;;;N;;*;;; (228 = выше и слева)
05A2;HEBREW ACCENT TSINORIT;Mn;230;NSM;;;;;N;;;;; (230=выше)

Модификация (только знака ZARQA) не должна повлиять на результат нормализации [4], т.к. над одной буквой никогда не ставится более одного знака кантилляции этих классов.

● Изменить положения знаков в таблицах графических изображений символов, чтобы они отражали исправление значений классов сочетаемости.

2. Порядок символов холам [8] **и вав** [9].

В случае, когда в огласованном тексте гласная представлена *и* знаком огласовки, *и* буквой, которая может использоваться как согласная ([такое использование согласной для обозначения гласного звука называется] мать чтения, по латыни *mater lectionis),* в прежних версиях Unicode не достает инструкций, определяющих, в каком порядке должны следовать эти знаки. Почти во всех случаях этот порядок очевиден, и о нем можно судить по виду типографских знаков, а они выглядят так же, как если бы одна из согласной писалась без знака огласовки. Однако в случае, когда друг за другом следуют знаки холам и вав ([эта последовательность называется] *холам мале,* [т. е. «полный холам»]), существует необходимость определить, в каком порядке предлагается их использовать. Вне зависимости, какой порядок избрать, это влияет на определение символа VAV WITH HOLAM (U+FB4B). Например:

Должно ли слово «шалом» писаться так:

SHIN + SHIN DOT + QAMATS
LAMED + HOLAM
VAV
FINAL MEM

или так:

SHIN + SHIN DOT + QAMATS
LAMED
VAV + HOLAM
FINAL MEM

и является ли последовательность

SHIN WITH SHIN DOT + QAMATS
LAMED
VAV WITH HOLAM
FINAL MEM

эквивалентной?

Подходящим местом, куда можно было бы вставить такое дополнительное объяснение в стандарт Юникода, является абзац в конце стр. 186, начинающийся со слова «Гласные». Приведенный там текст может быть улучшен добавлением объяснения, предлагаемого в следующем абзаце. Его формулировка следует тому же правилу, что используется в главах 7 – 11 стандарта в отношении других шрифтов: принципы, которым следуют шрифты, объясняются несколько подробней, чем это нужно для читателей уже знакомых со шрифтом:

Следующие знаки огласовки используются в литургических текстах, включая Библию, в поэтических произведениях, в словарях и всегда, когда огласовка должна быть однозначно указана. В большинстве других текстов они опускаются. Независимо от присутствия знаков огласовки гласные часто представляются буквами U+05D0 HEBREW LETTER ALEF, U+05D5 HEBREW LETTER VAV, U+05D9 HEBREW LETTER YOD, а также исключительно в конце слова – буквой U+05D4 HEBREW LETTER HE. Когда присутствуют знаки огласовки, они не только обозначают гласные, не представленные одной из этих букв, но они также определяют, в каких случаях эти буквы используются вместо гласных, и если это происходит, то для каких гласных. Таким образом, гласная может быть представлена и буквой, и огласовкой. В случае гласной *шурук,* т.е. гласной /u/ [т. е. «у»] в открытом слоге и в последнем слоге слова, знак огласовки U+05BC HEBREW POINT DAGESH OR MAPIQ используется вместе с буквой *вав,* играющей роль гласной. Во всех других случаях знак огласовки используется вместе с предшествующей согласной, а буква, представляющая гласную, остается без знака огласовки.

Если такая интерпретация нежелательна, этот текст может быть изменен следующим образом:

Следующие знаки огласовки […] и буквой, и огласовкой. Если это буква *вав,* то знак огласовки (либо U+05B9 HEBREW POINT HOLAM, либо U+05BC HEBREW POINT DAGESH OR MAPIQ) используется вместе с буквой *вав.* Во всех других случаях знак огласовки используется вместе с предшествующей согласной, а буква, представляющая гласную, остается без знака огласовки.

Первая из этих альтернатив требует изменения VAV WITH HOLAM (U+FB4B) на HOLAM + VAV; в противном случае этот символ не имеет применения.

С другой стороны, вследствие второй из этих альтернатив HOLAM может использоваться вместе с последней буквой слова, так что теоретически он может быть типографически несовместимым со знаком, который ставится выше и слева слова, как зарка (вне зависимости от того, какое имя Юникод будет использовать для этого знака кантилляции). На практике эти знаки не мешают друг другу, ибо такой холам пишется над вав, и таким образом справа от знака кантилляции, как и предписывает значение параметра класс сочетаемости [1].

3. Неясные графические изображения символов

Изображения следующих знаков кантилляции в таблице знаков неясны. Вот как они должны выглядеть:

● *Знак кантилляции* SEGOL (U+0592) состоит из трех точек, которые близки друг к другу, а не разбросаны в стороны вдоль верхней части буквы, над которой ставится этот знак. Нужно взять

изображение *знака огласовки* SEGOL (U+05B6), повернуть все изображение на 180° и сместить знак кантилляции в положение вверху и слева подобно U+0599.

● SHALSHELET (U+0593) состоит из трех полных знаков < , касающихся друг друга, а не из двух с половиной.

● DARGA (U+05A7) имеет острые углы; он выглядит, как зеркальное отражение Z, а не как S.

Примечания переводчика

---

[1] UnicodeData – это файл с описаниями всех символов, поддерживаемых Юникодом. Каждая строка соответствует символу. Значения параметров, используемых для описания символа, разделены точками с запятой. Для этой статьи важны только три параметра: первый – значение кода, присвоенного символу в Юникоде; второй – имя символа и четвертый – канонический класс сочетаемости. Сочетаемость означает, что есть символы, используемые сами по себе, для которых значение этого параметра – ноль, и есть символы, имеющие смысл только в сочетании с одним из предшествующих символов. Это относится к знакам кантилляции, которые ставятся на буквах или около них и не имеют без них смысла (и эти буквы при кодировании текста Юникодом предшествуют символам, имеющим смысл только в сочетании с ними). Для них значение этого параметра не равно нулю. Объясним также слово «канонический». В Юникоде некоторые последовательности кодов эквивалентны друг другу, например, потому, что символы повторяются. Это связано с тем, что Юникод предоставляет полный набор символов каждого языка, а символы повторяются от языка к языку. Это приводит к тому, что один и тот же текст может быть закодирован разными способами. В Юникоде определено два вида эквивалентности: 1) последовательности символов, которые выглядят одинаково и имеют то же значение, называются канонически эквивалентными; 2) последовательности, которые имеют то же значение, но могут выглядеть по-разному, называются совместимыми. Выражение «канонический класс сочетаемости» означает, что этот класс используется для выявления канонической эквивалентности последовательностей символов. Отметим, что автор ссылается файл UnicodeData, созданный в 2001 году. Однако в текущем файле, созданном в 2013 году, значения этих трех параметров для двух упомянутых в тексте знаков кантилляции остались прежними.

[2] Можно заключить – руководствуясь порядком и не руководствуясь параметром «канонический класс сочетаемости» [1], а также, не руководствуясь таблицами, в которых изображены символы. Заметим, что в своих таблицах [см. глава IV] автор как раз поступает наоборот.

[3] Чтобы это заключить, нужно сказать, что U+0598 предполагалось использовать не для знака, используемого в 21 книге, а для знака, используемого во всех книгах.

[4] Для того, чтобы выявить эквивалентные последовательности символов [1], их приводят в одну из нормальных форм (существует четыре типа нормальных форм – по две для каждого вида эквивалентности). Этот процесс называется нормализацией. В результате нормализации эквивалентных последовательностей символов получается один и тот же результат, а в результате нормализации неэквивалентных последовательностей – различные результаты.

[5] Старые версии таблиц графических изображений символов не доступны в Интернете, а в текущей версии примечание выглядит “= tsinorit, zinorit; tsinor, zinor. Этот знак должен использоваться, когда Zarqa или Tsinor ставятся вверху, а также для Tsinorit”. К знаку ZINOR добавлено примечание: “= tsinor, zarqa. Этот знак должен использоваться, когда Zarqa или Tsinor ставятся вверху слева”. Т.е. новые примечания (введенные в Юникоде после опубликования этой статьи) соответствуют написанному выше, в этой главе.

[6] Этот линк ведет к главной странице правил Юникода, там нет этой цитаты, но на нужную страницу ведет ссылка, и там в том числе содержится указание, что это относится к стандарту Юникода.

[7] В Юникод были внесены изменения, основанные на решении 2. См. [5].

[8] Холам – это огласовка, обозначающая звук «о».
[9] Вав – это буква, которая может использоваться для обозначения согласного звука «в» или гласных звуков «о» или «у». Здесь рассматривается случай, когда буква вав вместе с огласовкой холам обозначает гласный звук «о».