Proposal for a Hebrew Script Root Zone Label Generation Ruleset (LGR)

LGR Version: 3
Date: 2019-04-24
Document Version: 1.3
Authors: Hebrew Generation Panel (HGP)

1 General Information

The purpose of this document is to provide an overview of the proposed Hebrew LGR as provided in the XML format, and the rationale behind design decisions taken. It includes a discussion of relevant features of the script, languages and communities using it, the code points included, variant code points and information on the contributors. The formal specification of the LGR will be found in the accompanying XML document: `proposal-hebrew-lgr-24apr19-en.xml`. Labels for testing can be found in the accompanying text document: `hebrew-test-labels-24apr19-en.txt`.

2 The Script

ISO 15924 Code: Hebr
ISO 15924 Key: 125
ISO 15924 English Name: Hebrew
Latin Transliteration of Native Script Name: Ivrit
Native Script Name: עברית
Minimal Starting Repertoire (MSR) Version: MSR-4

3 Background About the Script and Languages Using it

3.1 The Hebrew Script

The Hebrew script, written from right to left, is one of the most ancient alphabetic scripts in the world. The first Hebrew inscriptions date back 3000 years, and are written in letters similar to those of the Phoenician script. This ancient Hebrew script was used mainly during the First Jewish Temple period (early 1st millennium BCE – 587 BCE). At first it was a purely consonantal script, but in the course of time, four of its 22 letters came to serve also as vowel letters, making the words easier to read.

From the Persian period onward, the Jews, like several other nations, adopted the Aramaic script, and gradually the Jewish script became what it is today: a 22-letter alphabet, with 5 of those letters also having a word-final form.

Towards the end of the first millennium C.E. new signs were invented by the Masoretes to mark vowels, stress and additional information the letters were not meant to convey. These signs are called nikkud (pointing) and te'amim (accents). The Masoretes' aim was to record their tradition for public reading of the Bible and transmit it to the reader preserving all details. Modern Hebrew is written without te'amim, but nikkud is still in use, mostly in children's books and poetry.

3.2 Languages Using the Hebrew Script

Today the Hebrew script is used primarily for modern Hebrew – the main language used in Israel, spoken by 8,330,000[1] people.

For centuries the Hebrew script was also used for Jewish languages – languages and dialects that developed in Jewish communities around the world (the largest being Judeo-Arabic, Yiddish and Ladino). Today, most of the Jewish languages are nearly extinct, and out of dozens of languages only Yiddish and Ladino are defined with EGIDS Scale 4. Since Ladino is written today mainly using the Latin script and is not the first language of its 137,000 users, we will not consider it here. Yiddish has 514,160 users worldwide, and is at status 4 for 55% of the users (Israel: 275k, Belarus: 7k); status 5 for 33% of the users (US: 156k, Canada: 13.6k, Moldova: 1.3k); status 7-9 for users in other countries. Being a language that uses primarily the Hebrew script, Yiddish has been considered when preparing the RZ-LGR and this document.

3.3 Spelling Variations

As described above, the Hebrew script is originally a consonantal script, but in modern Hebrew a few letters are heavily used as vowels. This orthography has developed over two millennia, often in an inconsistent way. While it is regulated today by the Academy of the Hebrew Language, the common spelling (regarding those letters) is still irregular. This phenomenon opens up possibilities for both confusion and the creation of misleading labels. Unfortunately, it cannot be resolved via LGR rules that would enforce a certain usage of those vowel-like letters (כתיב מלא vs. כתיב חסר).

[1] All the numerical data of this section is based on [ETHNOLOGUE]. Note that the numbers in [ETHNOLOGUE] are apparently based on older statistics and the real numbers may be higher.

4 Guiding Principles

Development of this LGR was done in accordance with [LGR-CONSIDERATIONS]. Constructing it, the panel focused on the principles of simplicity, security and robustness.

In accordance with these principles, the LGR is designed to be lean, observing that it would be very hard to remove a code point from it, once it made it to a root zone label.

5 Repertoire

As per the Procedure to Develop and Maintain the Label Generation Rules for the DNS Root Zone in Respect of IDNA Labels (hereinafter [Procedure]), only code points included in Maximal Starting Repertoire [MSR-4] will be considered.

The complete set of characters in the Hebrew script fall in the following Unicode ranges:

Hebrew: U+0591 – U+05F4 and U+FB1D – FB4F

5.1 Groups of Signs

The set of characters in the Hebrew script defined above can be divided into five types of signs:

5.1.1 Alphabet / Letters

The 22 alphabet letters (U+05D0 – U+05EA), 5 of which have a word-final form. All of these characters are included in the RZ-LGR repertoire.

5.1.2 Nikkud

A system of diacritical signs used to represent vowels or distinguish between alternative forms of pronunciation of letters of the Hebrew alphabet (U+05B0 – U+05BC, U+05C1, U+05C2, U+05C7).

Nikkud characters are zero-width. They are typographically positioned above, under or inside a letter. In some cases, multiple Nikkud signs may be positioned over a single letter.

Nikkud is helpful but is not obligatory in Hebrew.

According to the YIVO Institute for Jewish Research's Yiddish Alphabet listing [YIVO], some Nikkud signs are obligatory in Yiddish when combined with specific letters (combinations like U+FB1D –Yod with Hiriq ([YIVO]: Hirik Yod), U+FB2E –Alef with Patah ([YIVO]: Pasekh Alef), U+FB2F –Alef with Qamats ([YIVO]: Komets Alef), U+FB4E – Pe with Rafe ([YIVO]: Fey)). However, in common writing these combinations are often ignored or not used – see for example the Yiddish Wikipedia site [YIDDISH-WIKI].

All of the Nikkud code points are excluded from the RZ-LGR repertoire, for the following reasons:

- Non-unique appearance. There are multiple ways to generate a similarly looking letter with Nikkud (e.g. Shin + Holam versus Shin + Sin Dot; any consonant + Hataf Segol versus the same consonant + Sheva + Segol.
- Non-deterministic order. Two or more Nikkud symbols can be added to a letter in different orders, resulting in different code point streams generating the exact same typographical rendering.
- Simple to mislead and confuse: (a) Partial use of Nikkud can generate strings that will be construed by many users as the same (especially slight changes such as omitting Dagesh or omitting left/right dot on Shin); (b) Nikkud errors (e.g. replacing Qamats with Patah) are less likely to be detected than letters misspelling, since precise Nikkud rules are not common knowledge.
- Most users of the Hebrew language do not know how to generate Nikkud on their keyboards. Even those who do are not likely to be aware of the canonical order in which those symbols need to be typed, when two or more apply to a single letter.

5.1.3 Apostrophes

The apostrophe ׳ (Geresh, U+05F3) is used to signal an abbreviation and to expand the repertoire of written consonants in order to include consonants of borrowed foreign words. For example, ג=/g/ while ׳ג= /ʤ/.

The Hebrew apostrophe ״ (Gershaim, U+05F4) is mostly used to mark an acronym.

Both these apostrophes are quite common in Hebrew writing, but have been excluded from the RZ-LGR repertoire for the following reasons:

- They are excluded from MSR-4.
- Very few people know how to generate these symbols on the Hebrew keyboard. Inevitably, they will type their ASCII counterparts (Apostrophe ‘ and Quotation Mark “, also excluded from MSR-4), leading to resolution errors and / or deception.

5.1.4 Te’amim

Te’amim (U+0591 – U+05AF) are cantillation marks used for ritual reading from the Hebrew Bible. Te’amim also provide a reading structure to biblical sentences, much like modern punctuation marks. These characters are used exclusively in biblical text.

Te’amin are excluded from the RZ-LGR for the following reasons:

- They are excluded from MSR-4.
- They are used only in the context of biblical text.

5.1.5 Special Hebrew Code Points

The Unicode Standard contains a few other, less common code points.

- U+05F0 – HEBREW LIGATURE YIDDISH DOUBLE VAV. Intended for use in Yiddish texts, this code point provides a special combined ligature for two consecutive HEBREW LETTER VAV.
- U+05F1 – HEBREW LIGATURE YIDDISH VAV YOD. Intended for use in Yiddish texts, this code point provides a special combined ligature for HEBREW LETTER VAV followed by HEBREW LETTER YOD.
- U+05F2 – HEBREW LIGATURE YIDDISH DOUBLE YOD. Intended for use in Yiddish texts, this code point provides a special combined ligature for two consecutive HEBREW LETTER YOD.
- U+05BF – HEBREW POINT RAFE. This code point is a diacritic symbol, a subtle horizontal overbar placed above certain letters to indicate that they are to be pronounced as fricatives. NSM (NonSpacing Mark).
- U+FB1E – HEBREW POINT JUDEO-SPANISH VARIKA. This code point is used in rare Judeo-Spanish (Ladino) texts. NSM (NonSpacing Mark). See also 5.2 below.

All five of these code points are excluded from the RZ-LGR.

The first three might be confused with their respective combinations of two single letters. In addition, they can be adequately replaced by their respective combination of two consecutive single letters – DOUBLE YOD by two consecutive YOD, etc. Another advantage of these equivalent replacements is that they can be typed using a standard Hebrew-mapped keyboard.

The latter two bear the same problems as described in 5.1.2, and share their reasons for exclusion – augmented by their specific and rather uncommon use cases.

5.2 Hebrew-Related Code Points that are not PVALID

Code points that are not IDNA2008 Protocol Valid (PVALID) are excluded from the RZ-LGR.

Among these are Hebrew-related code points in the Unicode Alphabetic Presentation Forms block (U+FB1D through U+FB4F), and code points in the Unicode Letterlike Symbols block (U+2135 through U+2138).

5.3 Hebrew section of Maximal Starting Repertoire (MSR) Version 4

Code points shown below are from [MSR-4]. Code points with white background are not IDNA2008 PVALID. Pink background denotes IDNA2008 PVALID code points that are excluded from [MSR-4]. The repertoire of the RZ-LGR is denoted by the orange line.

5.4 Included Code Point Table

No.	Unicode Code Point	Glyph	Character Name	Refs
1	05D0	א	HEBREW LETTER ALEF	[OMNI]
2	05D1	ב	HEBREW LETTER BET	[OMNI]
3	05D2	ג	HEBREW LETTER GIMEL	[OMNI]
4	05D3	ד	HEBREW LETTER DALET	[OMNI]
5	05D4	ה	HEBREW LETTER HE	[OMNI]
6	05D5	ו	HEBREW LETTER VAV	[OMNI]
7	05D6	ז	HEBREW LETTER ZAYIN	[OMNI]
8	05D7	ח	HEBREW LETTER HET	[OMNI]
9	05D8	ט	HEBREW LETTER TET	[OMNI]
10	05D9	י	HEBREW LETTER YOD	[OMNI]
11	05DA	ך	HEBREW LETTER FINAL KAF	[OMNI]
12	05DB	כ	HEBREW LETTER KAF	[OMNI]
13	05DC	ל	HEBREW LETTER LAMED	[OMNI]
14	05DD	ם	HEBREW LETTER FINAL MEM	[OMNI]
15	05DE	מ	HEBREW LETTER MEM	[OMNI]
16	05DF	ן	HEBREW LETTER FINAL NUN	[OMNI]
17	05E0	נ	HEBREW LETTER NUN	[OMNI]
18	05E1	ס	HEBREW LETTER SAMEKH	[OMNI]
19	05E2	ע	HEBREW LETTER AYIN	[OMNI]
20	05E3	ף	HEBREW LETTER FINAL PE	[OMNI]
21	05E4	פ	HEBREW LETTER PE	[OMNI]
22	05E5	ץ	HEBREW LETTER FINAL TSADI	[OMNI]
23	05E6	צ	HEBREW LETTER TSADI	[OMNI]
24	05E7	ק	HEBREW LETTER QOF	[OMNI]
25	05E8	ר	HEBREW LETTER RESH	[OMNI]
26	05E9	ש	HEBREW LETTER SHIN	[OMNI]
27	05EA	ת	HEBREW LETTER TAV	[OMNI]

6 Variants

6.1 In-Script Variants

The RZ-LGR defines five variant pairs, one for each of the letters that have a word-final form. The final and non-final form are defined as variants of each other, resulting in the desired state where a label can be created with code points of either form, regardless of their position in the label (final or not), but then a label with a similar sequence of letters differing only by the final/non-final letter form of the code points will be blocked.

Set	Unicode Code Point	Glyph	Character Name	Refs
1	05DA 05DB	ך כ	HEBREW LETTER FINAL KAF HEBREW LETTER KAF	[OMNI]
2	05DD 05DE	ם מ	HEBREW LETTER FINAL MEM HEBREW LETTER MEM	[OMNI]
3	05DF 05E0	ן נ	HEBREW LETTER FINAL NUN HEBREW LETTER NUN	[OMNI]
4	05E3 05E4	ף פ	HEBREW LETTER FINAL PE HEBREW LETTER PE	[OMNI]
5	05E5 05E6	ץ צ	HEBREW LETTER FINAL TSADI HEBREW LETTER TSADI	[OMNI]

7 Whole Label Evaluation Rules

This LGR does not require WLE rules.

8 Contributors

- Mr. Doron Shikmoni, Founder, Israel Internet Association (ISOC-IL) and Forescout Technologies – Chair
- Ms. Dorit Lerer, Deputy CEO, The Academy for the Hebrew Language – member
- Mr. Matitiahu Allouche, Private Expert (Linguistics and computers) – member
- Mr. Meir Keraushar, DNS expert, Israel Internet Association (ISOC-IL) – member
- Mr. Yoram Hacohen, CEO, Israel Internet Association (ISOC-IL) – member

9 References

[MSR-4] ICANN Maximum Starting Repertoire 4 (MSR-4)
https://www.icann.org/sites/default/files/packages/lgr/msr/msr-4-wle-rules-09nov18-en.html

[UNICODE-630] The Unicode Standard 6.3.0
http://unicode.org/versions/Unicode6.3.0/

[HEB-IDN-TAB] IL IDN Table
https://www.iana.org/domains/idn-tables/tables/il_he_1.0.html

[IL-REG-RULES] IL Registry Rules
https://www.isoc.org.il/files/docs/ISOC-IL_Registration_Rules_v1.6_ENGLISH_-_18.12.2017.pdf

[2016-2LD-LGR] ICANN Second Level LGR
https://www.icann.org/sites/default/files/packages/lgr/lgr-second-level-hebrew-30aug16-en.xml , https://www.icann.org/sites/default/files/packages/lgr/lgr-second-level-hebrew-30aug16-en.html

[OMNI] Omniglot Hebrew
http://www.omniglot.com/writing/hebrew.htm

[ETHNOLOGUE] Ethnologue – Languages of the World
https://www.ethnologue.com/

[LGR-CONSIDERATIONS] Considerations for Designing a Label Generation Ruleset for the Root Zone
https://community.icann.org/download/attachments/43989034/Considerations-for-LGR-2017-09-15.pdf

[YIVO] YIVO Institute for Jewish Research, Yiddish Alef-Beys (Alphabet)
https://www.yivo.org/Yiddish-Alphabet

[YIDDISH-WIKI] Yiddish Wikipedia (וויקיפעדיע)
https://yi.wikipedia.org/