No organization has endorsed
this document

Aram Airapetian
<aramhayr@hotmail.com>

RFC: EN-0002

Type: Draft

Date: 11/02/2000

Armenian Character Set 8 and 16 bit encoding

The document maps Armenian characters defined in [1] into 8-bit region from xA0 to xFF and 16-bit region from 0530 to 058F. The mapping slightly differs from existing 8 [2,3] and 16 bit [4] Armenian Character Set encoding. The main reason for suggested modifications to existing encoding is reconciliation of ArmSCII [1] and Unicode [4]: one to one 8 to 16 bit character mapping. 8-bit encoding is made as close to ArmSCII or Armenian National Standard [2,3] as possible. 16-bit encoding is made as close to Unicode as possible. This document also defines what international characters may some Armenian characters merged with.

The following structure is selected for representing both code tables: code point (8 or 16 bit), character name [1], and character name from [2] or [4] in square brackets. The latter is provided only if the character name is different from [1].

This document is a recommendation for software developers who implement algorithms described in [1] as well as driver programmers, application developers, and font designers.

For naming and other conventions see [1].

I. 8-bit encoding

This encoding differs from [3] at the following code points:

- The Sign "Yev" is removed from the x26 code point to xA2 code point. It is done for the following reasons: 1). the Sign "Yev" has only slight semantic similarities with the AMPERSAND, 2). the glyph of the Sign "Yev" is completely different, 3). the category: the Sign "Yev" is a letter while the AMPERSAND is not.
- The 'Eternity Sign' (xA1) is deprecated because it is not important for Armenian information interchange.
- The signs Text's Right Parenthesis (xA4), Text's Left Parenthesis (xA5), and Text's Comma (xAB) are deprecated because their semantics and glyphs in Armenian are the same as RIGHT PARENTHESIS (x29), LEFT PARENTHESIS (x28), and COMMA (x2C).
- The sign Small Ampersand (xFF) is deprecated because it has no specific semantics or glyph and may cause confusion in text processing.

```
xA0 <Reserved>
xA1 <Deprecated> [Eternity Sign]
```

```
xA2 Sign 'Yev' [Ligature 'Yev', Ampersand '&', Logic 'AND']
xA3 Full Stop (Verjaket)
xA4 <Deprecated> [Text’s Right Parenthesis]
xA5 <Deprecated> [Text’s Left Parenthesis]
xA6 Right Quotation Mark (Aj Chakert)
xA7 Left Quotation Mark (Dzakh Chakert)
xA8 Joining Line (Miutyan Gtsik) [Joined Line]
xA9 Middle Dot (Mijaket)
xAA Separation Sign (Boot)
xAB <Deprecated> [Text’s Comma]
xAC Separating Line (Anjatman Gits) [Dash]
xAD Hyphen (Yentamna) [Hyphen Sign]
xAE Ellipsis (Kakhman Keter) [Ellipsis Points]
xAF Exclamation Mark (Batsakanchakan) [Exclamation Sign]
xB0 Emphasis Mark (Shesht) [Accent]
xB1 Question Mark (Paruyk)
xB2 Capital Letter Ayb
xB3 Small Letter Ayb
xB4 Capital Letter Ben
xB5 Small Letter Ben
xB6 Capital Letter Gim
xB7 Small Letter Gim
xB8 Capital Letter Da
xB9 Small Letter  Da
xBA Capital Letter Yech
xBB Small Letter Yech
xBC Capital Letter Za
xBD Small Letter Za
xBE Capital Letter Eh
xBF Small Letter Eh
xC0 Capital Letter Et [At]
xC1 Small Letter Et [At]
xC2 Capital Letter To
xC3 Small Letter To
xC4 Capital Letter Zheh
xC5 Small Letter Zheh
xC6 Capital Letter Ini
xC7 Small Letter Ini
xC8 Capital Letter Lyun
xC9 Small Letter Lyun
xCA Capital Letter Xeh
xCB Small Letter Xeh
xCC Capital Letter Tsa
xCD Small Letter Tsa
xCE Capital Letter Ken
xCF Small Letter Ken
xD0 Capital Letter Ho
```

```
xD1 Small Letter Ho
xD2 Capital Letter Dza
xD3 Small Letter Dza
xD4 Capital Letter Ghat
xD5 Small Letter Ghat
xD6 Capital Letter Tcheh
xD7 Small Letter Tcheh
xD8 Capital Letter Men
xD9 Small Letter Men
xDA Capital Letter Yi
xDB Small Letter Yi
xDC Capital Letter Nu
xDD Small Letter Nu
xDE Capital Letter Sha
xDF Small Letter Sha
xE0 Capital Letter Vo
xE1 Small Letter Vo
xE2 Capital Letter Cha
xE3 Small Letter Cha
xE4 Capital Letter Peh
xE5 Small Letter Peh
xE6 Capital Letter Jeh
xE7 Small Letter Jeh
xE8 Capital Letter Ra
xE9 Small Letter Ra
xEA Capital Letter Seh
xEB Small Letter Seh
xEC Capital Letter Vev
xED Small Letter Vev
xEE Capital Letter Tyun
xEF Small Letter Tyun
xF0 Capital Letter Reh
xF1 Small Letter Reh
xF2 Capital Letter Co
xF3 Small Letter Co
xF4 Capital Letter Vyun
xF5 Small Letter Vyun
xF6 Capital Letter Pyur
xF7 Small Letter Pyur
xF8 Capital Letter Qeh
xF9 Small Letter Qeh
xFA Capital Letter O
xFB Small Letter O
xFC Capital Letter Feh
xFD Small Letter Feh
xFE Apostrophe (Apatarts) [Capital Apostrophe]
xFF <Deprecated> [Small Apostrophe]
```

II. 16-bit encoding

This encoding differs from [4] at the following code points:
- The ‘Modifier Letter Left Half Ring’ (0559) and ‘Abbreviation Mark’ (055F) are deprecated because they are not important for Armenian information interchange (these signs are very rare and not in use today).

The table below shows Armenian characters in the Unicode’s Armenian block: 0530 –058F [4]. In addition it provides with the mapping of Armenian signs with international characters (this information is missing in [4]). Armenian names of these characters are presented. See [4] for international names.

00AB	Left Quotation Mark (Dzakh Chakert)
00BB	Right Quotation Mark (Aj Chakert)
0530	<Reserved>
0531	Capital Letter Ayb
0532	Capital Letter Ben
0533	Capital Letter Gim
0534	Capital Letter Da
0535	Capital Letter Yech [Ech]
0536	Capital Letter Za
0537	Capital Letter Eh
0538	Capital Letter Et
0539	Capital Letter To
053A	Capital Letter Zheh [Zhe]
053B	Capital Letter Ini
053C	Capital Letter Lyun [Liwn]
053D	Capital Letter Xeh
053E	Capital Letter Tsa [Ca]
053F	Capital Letter Ken
0540	Capital Letter Ho
0541	Capital Letter Dza [Ja]
0542	Capital Letter Ghat [Ghad]
0543	Capital Letter Tcheh [Cheh]
0544	Capital Letter Men
0545	Capital Letter Yi
0546	Capital Letter Nu [Now]
0547	Capital Letter Sha
0548	Capital Letter Vo
0549	Capital Letter Cha
054A	Capital Letter Peh
054B	Capital Letter Jeh [Jheh]
054C	Capital Letter Ra
054D	Capital Letter Seh
054E	Capital Letter Vev [Vew]

```
054F Capital Letter Tyun [Tiwn]
0550 Capital Letter Reh
0551 Capital Letter Co
0552 Capital Letter Vyun [Yiwn]
0553 Capital Letter Pyur [Piwr]
0554 Capital Letter Qeh [Keh]
0555 Capital Letter O [Oh]
0556 Capital Letter Feh
0557 <Reserved>
0558 <Reserved>
0559 <Deprecated> [Modifier Letter Left Half Ring]
055A Apostrophe (Apatarts)
055B Emphasis Mark (Shesht)
055C Exclamation Mark (Batsakanchakan) [(Batsaganchakan nshan)]
055D Separation Sign (Boot) [Comma (bowt)]
055E Question Mark (Paruyk) [(Hartsakan nshan)]
055F <Deprecated> [Abbreviation Mark (patiw)]
0560 <Reserved>
0561 Small Letter Ayb
0562 Small Letter Ben
0563 Small Letter Gim
0564 Small Letter Da
0565 Small Letter Yech [Ech]
0566 Small Letter Za
0567 Small Letter Eh
0568 Small Letter Et
0569 Small Letter To
056A Small Letter Zheh [Zhe]
056B Small Letter Ini
056C Small Letter Lyun [Liwn]
056D Small Letter Xeh
056E Small Letter Tsa [Ca]
056F Small Letter Ken
0570 Small Letter Ho
0571 Small Letter Dza [Ja]
0572 Small Letter Ghat [Ghad]
0573 Small Letter Tcheh [Cheh]
0574 Small Letter Men
0575 Small Letter Yi
0576 Small Letter Nu [Now]
0577 Small Letter Sha
0578 Small Letter Vo
0579 Small Letter Cha
057A Small Letter Peh
057B Small Letter Jeh [Jheh]
057C Small Letter Ra
057D Small Letter Seh
```

```
057E Small Letter Vev [Vew]
057F Small Letter Tyun [Tiwn]
0580 Small Letter Reh
0581 Small Letter Co
0582 Small Letter Vyun [Yiwn]
0583 Small Letter Pyur [Piwr]
0584 Small Letter Qeh [Keh]
0585 Small Letter O [Oh]
0586 Small Letter Feh
0587 Sign 'Yev' [Small ligature EchYiwn]
0588 <Reserved>
0589 Full Stop (Verjaket) [(Vertsaket)]
058A Hyphen (Yentamna)
058B <Reserved>
058C <Reserved>
058D <Reserved>
058E <Reserved>
058F <Reserved>
2011 Joining Line (Miatsman Gits)
2014 Separating Line (Anjatman Gits)
2024 Middle Dot (Mijaket)
2026 Ellipsis (Kakhman Keter)
```

III Conversion

The above described encoding allows for 8-bit to 16-bit round trip conversion without loss of information. Conforming to this document ArmSCII to Unicode converter software must convert code points from x00 to x7F to respective values from 0000 to 007F. It may convert code points from x80 to x9F to respective values from 0080 to 009F or to any other valid Unicode values except listed in Section II. The region from x80 to x9F is not in the scope of the document. The code points from xA0 to xFF must be converted according the following C style array:

```
const wchar_t armsciiToUnicode[] = {
0xFFFD, 0xFFFD, 0x0587, 0x0589, 0xFFFD, 0xFFFD, 0x00BB, 0x00AB,
0x2011, 0x2024, 0x055D, 0xFFFD, 0x2014, 0x058A, 0x2026, 0x055C,
0x055B, 0x055E, 0x0531, 0x0561, 0x0532, 0x0562, 0x0533, 0x0563,
0x0534, 0x0564, 0x0535, 0x0565, 0x0536, 0x0566, 0x0537, 0x0567,
0x0538, 0x0568, 0x0539, 0x0569, 0x053A, 0x056A, 0x053B, 0x056B,
0x053C, 0x056C, 0x053D, 0x056D, 0x053E, 0x056E, 0x053F, 0x056F,
0x0540, 0x0570, 0x0541, 0x0571, 0x0542, 0x0572, 0x0543, 0x0573,
0x0544, 0x0574, 0x0545, 0x0575, 0x0546, 0x0576, 0x0547, 0x0577,
0x0548, 0x0578, 0x0549, 0x0579, 0x054A, 0x057A, 0x054B, 0x057B,
0x054C, 0x057C, 0x054D, 0x057D, 0x054E, 0x057E, 0x054F, 0x057F,
0x0550, 0x0580, 0x0551, 0x0581, 0x0552, 0x0582, 0x0553, 0x0583,
0x0554, 0x0584, 0x0555, 0x0585, 0x0556, 0x0586, 0x055A, 0xFFFD
};
```

The items (elements) in the array are separated by comma. The 1-st (0 in C/C++/Java) element corresponds to xA0, the 2-nd - to xA1, and so on. The last corresponds to xFF. There are 96 elements in the array.

The ArmSCII to Unicode converter software may convert code points x26, xA1, xA4, xA5, and xAB to 0587, 2043, 0029, 0028, and 002C respectively to process current Armenian National Standard compliant data. This mode of conversion is not recommended to be default. It may be set at user option via environment, configuration, command line variable, or other user interface parameter. Note, that this mode is applicable only from ArmSCII to Unicode conversion and is not recommended for opposite conversion.

Unicode to ArmSCII converter software must convert code points from 0000 to 007F to x00 to x7F respectively. Values from 0080 to 009F or any other 32 valid Unicode values not listed in Section II may be converted to code points from x80 to x9F. All values from armsciiToUnicode array (except FFFD) must be converted according to their position: index + xA0. The 'index' starts from 0 and goes to 95. Any Unicode character value other than above mentioned is error. Note that all <Reserved> or <Deprecated> values from Armenian block (0530-058F) are also errors. Conforming software may convert erred Unicode characters to x3F (QUESTION MARK).

VI Acknowledgements

The author greatly appreciates R.Youatt's input during discussions of the topic and recognizes his effort for corresponding software evaluation. His analysis [5] of encoding reconciliation has had significant influence on this effort.

V. References

1. RFC EN-001. Armenian Character Names, Descriptions, Classification, Collating, and Searching.
2. Character Sets. Standard of Republic of Armenia (HST 165-97)
3. 8 Bit Encoding Character Sets. Standard of Republic of Armenia (HST 166-97)
4. The Unicode Standard Version 3.0. Addison-Wesley. Reading, MA. Jan 2000.
5. R. Youatt, "Unicode and ARMSCII", Journal of the Society of Armenian Studies 9, 1999, pp. 87-97.