No organization has endorsed
this document

Aram Airapetian
<aramhayr@hotmail.com>

RFC: EN-0001

Type: Draft

Date: 11/02/2000

Armenian Character Names, Descriptions, Classification, Collating, and Searching

This document defines the Basic Armenian Character Set for information exchange. It is based on [1, 2] and more than 10 years of development and usage of ArmSCII (Armenian Standard Code for Information Interchange) – an 8-bit Armenian character encoding - conforming software.

It is a recommendation for Armenian language related software developers.

The names (ISO-10646 [3]) of internationally accepted signs appear in the text in all capital letters. These signs usually have very close semantics in most European languages; for example, COLON, COMMA, etc.

The idea of terminology to define Armenian character set and to describe characters is taken from [4]. The terms defined in the following list are used in building those descriptions. They set up constraints for other, related to this one, documents and in describing the actions of an Armenian information processor:

may
Conforming documents and Armenian information processors are permitted to but need not behave as described.

must
Conforming documents and Armenian information processors are required to behave as described; otherwise they are in error.

error
A violation of the rules of this specification; results are undefined. Conforming software may detect and report an error and may recover from it.

at user option
Conforming software may or must (depending on the modal verb in the sentence) behave as described; if it does, it must provide users a means to enable or disable the behavior described.

The Armenian script is written in linear sequence from left to right. Spaces are used to separate words.

I. Classic Alphabet

The names of Armenian letters are transliterated according to Eastern Armenian pronunciation. Western Armenian pronunciation is given in parenthesis whenever it is different. Armenian Classic Alphabet consists of the following letters:

1. Capital Letter Ayb (Ayp)
2. Small Letter Ayb (Ayp)
3. Capital Letter Ben (Pen)
4. Small Letter Ben (Pen)
5. Capital Letter Gim (Kim)
6. Small Letter Gim (Kim)
7. Capital Letter Da (Ta)
8. Small Letter Da (Ta)
9. Capital Letter Yech
10. Small Letter Yech
11. Capital Letter Za
12. Small Letter Za
13. Capital Letter Eh
14. Small Letter Eh
15. Capital Letter Et
16. Small Letter Et
17. Capital Letter To
18. Small Letter To
19. Capital Letter Zheh
20. Small Letter Zheh
21. Capital Letter Ini
22. Small Letter Ini
23. Capital Letter Lyun
24. Small Letter Lyun
25. Capital Letter Xeh
26. Small Letter Xeh
27. Capital Letter Tsa (Dza)
28. Small Letter Tsa (Dza)
29. Capital Letter Ken (Gen)
30. Small Letter Ken (Gen)
31. Capital Letter Ho
32. Small Letter Ho
33. Capital Letter Dza (Tsa)
34. Small Letter Dza (Tsa)
35. Capital Letter Ghat
36. Small Letter Ghat
37. Capital Letter Tcheh (Jeh)
38. Small Letter Tcheh (Jeh)
39. Capital Letter Men
40. Small Letter Men

41. Capital Letter Yi
42. Small Letter Yi
43. Capital Letter Nu
44. Small Letter Nu
45. Capital Letter Sha
46. Small Letter Sha
47. Capital Letter Vo
48. Small Letter Vo
49. Capital Letter Cha
50. Small Letter Cha
51. Capital Letter Peh (Beh)
52. Small Letter Peh (Beh)
53. Capital Letter Jeh (Cheh)
54. Small Letter Jeh (Cheh)
55. Capital Letter Ra
56. Small Letter Ra
57. Capital Letter Seh
58. Small Letter Seh
59. Capital Letter Vev
60. Small Letter Vev
61. Capital Letter Tyun (Dyun)
62. Small Letter Tyun (Dyun)
63. Capital Letter Reh
64. Small Letter Reh
65. Capital Letter Co
66. Small Letter Co
67. Capital Letter Vyun (Hyun)
68. Small Letter Vyun (Hyun)
69. Capital Letter Pyur
70. Small Letter Pyur
71. Capital Letter Qeh (Keh)
72. Small Letter Qeh (Keh)
73. Capital Letter O
74. Small Letter O
75. Capital Letter Feh
76. Small Letter Feh

II Punctuation Marks

There are four subcategories of Armenian punctuation signs: a). disclosing, b). separating, c). uniting [joining], and d). accent signs. The first two categories are pretty similar: they separate units (words, clauses, sentences) of text. Also the separating sign SPACE is used in Armenian texts. The uniting signs serve for joining syllables of the word or for creating complex words. The accent signs have same semantics as ‘!’, ‘’’, or ‘?’, but in addition they allow for positioning the logical stress at the specific word in the sentence.

The following punctuation marks and signs are used in classic and contemporary Armenian texts (Armenian names are in parenthesis):

77. Full Stop (Verjaket). Semantics: equal to FULL STOP (end of sentence). In multi-language environments, where sentence recognition as a unit is not important, may be merged with COLON because of similar glyph. Subcategory: Separating sign.
78. Right Parenthesis (Aj Pakagits). Must be merged with RIGHT PARENTHESIS. Subcategory: Disclosing sign.
79. Left Parenthesis (Dzakh Pakagits). Must be merged with LEFT PARENTHESIS. Subcategory: Disclosing sign.
80. Right Square Bracket (Aj Ughigh Pakagits). Must be merged with RIGHT SQUARE BRACKET.
81. Left Square Bracket (Dzakh Ughigh Pakagits). Must be merged with LEFT SQUARE BRACKET.
82. Right Curly Bracket (Aj Dzevavor Pakagits). Must be merged with RIGHT CURLY BRACKET.
83. Left Curly Bracket (Dzakh Dzevavor Pakagits). Must be merged with LEFT CURLY BRACKET.
84. Right Quotation Mark (Aj Chakert). May be merged with RIGHT-POINTING DOUBLE ANGLE QUOTATION MARK. Subcategory: Disclosing sign.
85. Left Quotation Mark (Dzakh Chakert). May be merged with LEFT-POINTING DOUBLE ANGLE QUOTATION MARK. Subcategory: Disclosing sign.
86. Joining Line (Miutyan Gtsik). May be merged either with DASH or EN DASH. Subcategory: Uniting sign.
87. Middle Dot (Mijaket). Semantics: close to SEMICOLON. In multi-language environments, where sentence recognition as a unit is not important, may be merged with FULL STOP. Also may be merged with ONE DOT LEADER. Subcategory: Separating sign.
88. Separation Mark (Boot). Semantics: close to COLON. Subcategory: Separating sign.
89. Comma (Storaket). Must be merged with COMMA.
90. Separating Line (Anjatman Gits). May be merged with EM DASH. Subcategory: Disclosing sign.
91. Ellipsis (Kakhman Keter). Semantics: used at the end of incomplete clause or sentence, hence may function as FULL STOP. May be merged with HORIZONTAL ELLIPSIS or represented by tree consecutive dots (Middle Dots). Subcategory: Separating sign.
92. Suspension points (Bazmaket). Semantics: used to denote missing words or lines. May be represented by more than tree consecutive dots (Middle Dots) or two Ellipses. Subcategory: Disclosing sign.
93. Exclamation Mark (Batsakanchakan). Semantics: Exclamation mark. Goes next to the last vowel in the word. Subcategory: Accent sign.
94. Emphasis Mark (Shesht). Semantics: Emphasis (Stress) mark. Goes next to the last vowel in the word. Subcategory: Accent sign.
95. Question Mark (Paruyk). Semantics: Question mark. Goes next to the last vowel in the word. Subcategory: Accent sign.
96. Hyphen (Yentamna). May be merged either with DASH or HYPHEN. Semantics: HYPHEN. Subcategory: Uniting sign.

97. Apostrophe (Apatarts). Semantics: used [Western Armenian orthography only] for hidden Letter Et. May be merged with the APOSTROPHE. Subcategory: [Conditionally] uniting sign.

III. Additional Characters

The following additional letters are used in contemporary Armenian texts:

98. Sign 'Yev'. Category: Small Letter. Doesn't have Capital Letter. In Western Armenian orthography it is a shorthand for 'Yech' + 'Vyun' sequence. During capitalization gets transformed to the sequence of Capital 'Yech' + Small or Capital 'Vyun'. Sequence of Capital 'Yech' + Small or Capital 'Vyun' during de-capitalization always gets transformed to Sign 'Yev'. In Eastern Armenian orthography it is a shorthand for 'Yech' + 'Vev' sequence. During capitalization gets transformed to the sequence of Capital 'Yech' + Small or Capital 'Vev'. Sequence of Capital 'Yech' + Small or Capital 'Vev' during de-capitalization not always gets transformed to Sign 'Yev'. There are cases when the sequence of 'Yech' + 'Vev' is not a sign 'Yev'. Accent Signs go next to Sign "Yev". After capitalization they go next to the Letter 'Yech'.
99. Eastern Armenian Letter 'U'/Digraph 'U'. The sequence of the letters 'Vo' and 'Vyun' always represents vowel 'U' and, other way around, vowel 'U' is always represented by the sequence of the letters 'Vo' and 'Vyun'. In Eastern Armenian alphabet the 'Vo' + 'Vyun' digraph is considered a single letter 'U' and positioned at the place of the letter 'Vyun'. The Small Letter 'U' has the image of Small 'Vo' + Small 'Vyun' digraph and the Capital Letter 'U' has the image of Capital 'Vo' + Capital 'Vyun' digraph. Special collating must take place while processing an Eastern Armenian text. The sequence of the letters 'Vo' and 'Vyun' gets the weight more than the Letter 'Co' and less than Letter 'Pyur'. Software applications (for example, Armenian language support driver) may interpret the sequence of the letters 'Vo' and 'Vyun' as one letter.

IV Collating

In practice, the most important unit of processing (collating, sorting, searching, highlighting, etc.) is a word. Armenian word is a sequence of letters (1 - 76), inner signs (93-96), Apostrophe (97), and Sign 'Yev' (98).

Due to different interpretations of the Sign 'Yev' and digraph 'U' by Eastern and Western Armenian orthography two different sets of rules for collating are presented.

Let's introduce a set of inner signs. It consists of Accent Signs and Hyphen. The set of inner signs should be treated specially during search and collating.

Rules for Western Armenian orthography.

1. Remove all inner signs from each collating unit and append them to each other in the order of their appearance in the corresponding units.
2. For each collating unit transform Sign ‘Yev’ to the sequence of Small Letter ‘Yech’ and Small Letter ‘Vyun’.
3. Compare transformed units using the bullet-numbers of the characters in this document as their weights.
4. If the transformed units are not equal, then the result is achieved. Otherwise, compare the sequences of corresponding inner signs. The result of this comparison is the final result.

Rules for Eastern Armenian orthography.

The only difference for the Eastern Armenian orthography is the step 2:

2. For each collating unit transform Sign ‘Yev’ to the sequence of Small Letter ‘Yech’ and Small Letter ‘Vev’, each sequence of the Capital Letter ‘Vo’ and any (Capital or Small) Letter ‘Vyun’ to Capital Letter ‘Vyun’, and each sequence of the Small Letter ‘Vo’ and Small Letter ‘Vyun’ to Small Letter ‘Vyun’.

Note 1. These rules do not guarantee start order independent result for sorting (to guarantee that it is necessary to take into consideration also the location of inner characters in collating units). However, practically, even less complicated collation is possible. It is enough just to discard inner signs (not to perform step 4 for tie-break).

Note 2. For Classic Armenian texts step 2 is meaningless because these texts can not have Sign ‘Yev’. For Classic Armenian texts step 2 may be skipped.

V. Search

There are two parameters for search: a). case sensitivity and b). inner sign sensitivity. Each of them may be ‘true’ or ‘false’. Hence, there are four different modes of search. If case sensitivity is ‘false’, then the units of comparison are transformed to lower case before collation. If inner sign sensitivity is ‘false’, then inner signs are removed from the units of comparison before collation.

VI Acknowledgements

The author greatly appreciates B.Agopian’s, H.Melikyan’s, and R.Youatt’s input during countless discussions of the topic.

VII References

1. Jahukian. The Theoretical Basis of Modern Armenian Language. The Academy of Sciences of Armenian SSR, Yerevan, 1974.
2. On the Comprehensive Character Set of Eastern Armenian Alphabet and Signs. Order #518 of the State Language Inspection, Feb 12, 1999, Yerevan.
3. The Unicode Standard Version 3.0. Addison-Wesley. Reading, MA. Jan 2000.
4. XML Version 1.0. http://www.w3.org/TR/PR-xml.html