Network Working Group
Request for Comments: DRAFT
<draft-melikyan-armenian-charsets-00.txt>
H.Melikyan
Armenian Computer Center
17 May 1998

Armenian Character Sets: Implementation Guide

Status of this Memo

Abstract

The document presents the set of Armenian characters that is used in the information systems in accordance to AST 34.001 and AST 34.002 standards of the State Standards Commission of the Republic of Armenia, as well as provides classification and sorting thereof and recommendations for implementation of basic algorithms of text processing.

Table of Contents

1. Introduction

The publication of comments in reference to the standards is due to the following considerations:

(1) The Armenian character sets have been used in different computer systems approx. since 1982, whereas the state standard was established only in 1997. This time lag resulted in emergence of incompatible coding systems. The existing discrepancies are also due to the existence of two different grammars of the Armenian language.

(2) The emergence of internationalized operating systems and an important number of multi-lingual applications result in situations when the national language support is implemented by programmers that are not familiar with the given language.

The present memo is a recommendation rather than a binding standard.

The recommendations set forth herein are elaborated on the basis of the state standards AST 34.001 (reg.no. 166-97) and AST 34.002 (reg.no. 167-97), as well as ArmSCII standard.

2. Basic Character Set

2.1. Naming

The Basic Armenian character set presented below follows the standard AST 34.001. The first column contains full naming of the characters, and the second column provides abbreviations thereof that can be used in the systems confined to the Latin character set. The detailed classification of the characters follows in the points below.

In spite of the fact that the space, numbers and Latin letters are also part of the Armenian character set, these were not included in the AST 34.001 standard since these are present in all systems.

Table 1. Basic Character Set

Armenian Eternity Sign	armeternity
Armenian Ligature "ew"	armew
Armenian Section Sign	armsection
Armenian Full Stop (Verjaket)	armfullstop
Armenian Right Parenthesis	armparenright
Armenian Left Parenthesis	armparenleft
Armenian Right Quotation Mark	armquotright
Armenian Left Quotation Mark	armquotleft
Armenian EM Dash	armemdash
Armenian Dot (Mijaket)	armdot
Armenian Separation Mark (But)	armsep
Armenian Comma	armcomma
Armenian EN Dash	armendash
Armenian Hyphen (Yentamna)	armyentamna
Armenian Ellipsis	armellipsis
Armenian Apostrophe	armapostrophe
Armenian Exclamation Mark (Amanak)	armexclam
Armenian Accent (Shesht)	armaccent
Armenian Question Mark (Paruyk)	armquestion
Armenian Capital Letter [ayb]	Armayb
Armenian Small Letter [ayb]	armayb
Armenian Capital Letter [ben]	Armben
Armenian Small Letter [ben]	armben
Armenian Capital Letter [gim]	Armgim
Armenian Small Letter [gim]	armgim
Armenian Capital Letter [da]	Armda
Armenian Small Letter [da]	armda
Armenian Capital Letter [yech]	Armyech
Armenian Small Letter [yech]	armyech
Armenian Capital Letter [za]	Armza
Armenian Small Letter [za]	armza
Armenian Capital Letter [e]	Arme
Armenian Small Letter [e]	arme
Armenian Capital Letter [at]	Armat
Armenian Small Letter [at]	armat
Armenian Capital Letter [to]	Armto
Armenian Small Letter [to]	armto
Armenian Capital Letter [zhe]	Armzhe
Armenian Small Letter [zhe]	armzhe
Armenian Capital Letter [ini]	Armini
Armenian Small Letter [ini]	armini
Armenian Capital Letter [lyun]	Armlyun
Armenian Small Letter [lyun]	armlyun
Armenian Capital Letter [khe]	Armkhe
Armenian Small Letter [khe]	armkhe
Armenian Capital Letter [tsa]	Armtsa
Armenian Small Letter [tsa]	armtsa
Armenian Capital Letter [ken]	Armken

Armenian Small Letter [ken]	armken
Armenian Capital Letter [ho]	Armho
Armenian Small Letter [ho]	armho
Armenian Capital Letter [dza]	Armdza
Armenian Small Letter [dza]	armdza
Armenian Capital Letter [ghat]	Armghat
Armenian Small Letter [ghat]	armghat
Armenian Capital Letter [tche]	Armtche
Armenian Small Letter [tche]	armtche
Armenian Capital Letter [men]	Armmen
Armenian Small Letter [men]	armmen
Armenian Capital Letter [hi]	Armhi
Armenian Small Letter [hi]	armhi
Armenian Capital Letter [nu]	Armnu
Armenian Small Letter [nu]	armnu
Armenian Capital Letter [sha]	Armsha
Armenian Small Letter [sha]	armsha
Armenian Capital Letter [vo]	Armvo
Armenian Small Letter [vo]	armvo
Armenian Capital Letter [cha]	Armcha
Armenian Small Letter [cha]	armcha
Armenian Capital Letter [pe]	Armpe
Armenian Small Letter [pe]	armpe
Armenian Capital Letter [je]	Armje
Armenian Small Letter [je]	armje
Armenian Capital Letter [ra]	Armra
Armenian Small Letter [ra]	armra
Armenian Capital Letter [se]	Armse
Armenian Small Letter [se]	armse
Armenian Capital Letter [vev]	Armvev
Armenian Small Letter [vev]	armvev
Armenian Capital Letter [tyun]	Armtyun
Armenian Small Letter [tyun]	armtyun
Armenian Capital Letter [re]	Armre
Armenian Small Letter [re]	armre
Armenian Capital Letter [tso]	Armtso
Armenian Small Letter [tso]	armtso
Armenian Capital Letter [vyun]	Armvyun
Armenian Small Letter [vyun]	armvyun
Armenian Capital Letter [pyur]	Armpyur
Armenian Small Letter [pyur]	armpyur
Armenian Capital Letter [ke]	Armke
Armenian Small Letter [ke]	armke
Armenian Capital Letter [o]	Armo
Armenian Small Letter [o]	armo
Armenian Capital Letter [fe]	Armfe
Armenian Small Letter [fe]	armfe

End of Table 1.

The naming of characters are hereinafter referred to in abbreviated forms contained in the second column.

2.2. Classification and Sorting

unclassified-symbols ::= {armeternity, armew, armsection}

punctuation-signs ::= {armfullstop, armparenright, armparenleft, armquotright, armquotleft, armemdash, armdot, armsep, armcomma, armendash}

pseudo-letters ::= {armyentamna, armellipsis, armapostrophe}

diacritic-signs ::= {armexclam, armaccent, armquestion}

letters ::= {capital-letters, small-letters}

capital-letters ::= {Armayb, Armben, Armgim, Armda, Armyech, Armza, Arme, Armat, Armto, Armzhe, Armini, Armlyun, Armkhe, Armtsa, Armken, Armho, Armdza, Armghat, Armtche, Armmen, Armhi, Armnu, Armsha, Armvo, Armcha, Armpe, Armje, Armra, Armse, Armvev, Armtyun, Armre, Armtso, Armvyun, Armpyur, Armke, Armo, Armfe}

small-letters ::= {armayb, armben, armgim, armda, armyech, armza, arme, armat, armto, armzhe, armini, armlyun, armkhe, armtsa, armken, armho, armdza, armghat, armtche, armmen, armhi, armnu, armsha, armvo, armcha, armpe, armje, armra, armse, armvev, armtyun, armre, armtso, armvyun, armpyur, armke, armo, armfe}

The sorting order is important for letter characters only and is made in the order presented in the Table 1.

The case shift applies for letter characters only. The shift from the upper case to the lower case replaces the capital letter character with the subsequent character as per the Table 1. Accordingly, the shift from lower case to the upper case replaces the small letter character with the preceding character as per the Table 1.

The text search and dictionary applications should take into account the following factors: (1) in the Armenian language, a word is a sequence of letter characters, diacritic-signs, and pseudo-letters; (2) in comparison of words in the text or dictionary, the diacritic-signs and pseudo-letters may be ignored.

In reference to the diacritic-signs, the following factors are

important: (1) the diacritic-sign refers to the preceding letter (only vowel in Armenian), (2) a letter can be followed by more than one diacritic sign.

2.3. Ligatures

Ligature is a traditional or convenience graphical presentation of a sequence of letters, e.g. the Latin ligature "ft", the German ligature "ss", the Armenian ligature "armmen, armnu", etc. The ligatures can be officially registered and codified (like in UNICODE standard), but the systems supporting ligatures substitute them automatically only on the screen, printer, or other graphical devices.

The Armenian ligature armew that is a combination of armyech and armvyun was included in the AST 34.001 standard in view of the following considerations: (1) armew is a "ligature symbol" rather than a ligature, and (2) armew carries an "and" denotation similar to the "&" character.

3. Encoding

3.1. Basic Principles

The Coded Character Set is a mapping of a set of characters into a set of integer numbers, e.g. ArmSCII-7, ArmSCII-8 and ArmSCII-8A tables.

The term "unification" is used in the following denotation: as a rule, the mapping of an Armenian character set takes place in operating environments where other character sets are already available; thus, certain characters, in particular punctuation marks, may have identical graphical mapping and similar functions. In such cases, some characters of the Armenian character set may be mapped into already existing codified characters. The details of unification of Armenian punctuation marks are reviewed below.

The mapping of characters in coding tables has several aspects (in order of priority): (1) scope of the character mapping, (2) sequence of mapping, (3) character unification requirements, (4) general requirements of a given operating environment.

The encoding in every new operating environment should, to the extent possible, use the already existing coding tables (see the next section). Should this be impossible, the newly created coding tables should follow as much as possible the following general principles:

(1) The Armenian character set should be comprehensive (with due regard to the unification)

(2) The Armenian character set should be mapped into a continual sequence of codes in the order these are presented in the Table 1. The unified character codes should be left absolute, i.e. should not be used for other purposes. The most important is the letter sequence.

(3) The unification implies both graphical and functional identity of characters. For example, mapping of the parenthesis (armparenleft and armparenright) into the parenthesis existing in the ASCII is not an error. On the other hand, the similarity of the Armenian full stop (armfullstop) and the colon is purely graphical. The armdot and armsep bear functions different from the Latin dot and the grave accent character accordingly. Another important factor of character unification is the use of the Latin alphabet and punctuation marks in formal languages. It should be born in mind, for example, that a comma is often used as a separator in lists (e.g. in a keyword list in HTML document header), and in order to avoid confusion, the armcomma character may be mapped into a Latin comma.

(4) It may often happen that the requirements of a given operating environment may contradict the above principles. For example, the pseudo-graphical characters in DOS that were supported by video-adapters ("ninth pixel" factor), resulted in the creation of an alternative 8-bit coding table ArmSCII-8A. Another example is Macintosh OS where codes like ellipsis, nbsp and soft hyphen are recognized and interpreted in a special by numerous applications, which rendered the meaningful use the ArmSCII standard in this system impossible (the ArmSCII-8A table is used in OS Macintosh).

ArmSCII does not fully correspond to the above principles, and the Armenian block in the current version of UNICODE (2.1) corresponds to neither (1), (2), nor (3).

3.2. Cross Reference of Coding Tables

Table 2. Cross reference of coding tables

1 - Short name
2 - ArmSCII-7
3 - ArmSCII-8 (AST 34.002, Basic coding table)
4 - ArmSCII-8A (AST 34.002, Alternative coding table)
5 - ArmSCII-16U
6 - UNICODE Version 2.1

1	2	3	4	5	6
armeternity	21	A1	DC	0521	-
armew	-	26	26	0512	0587
armsection	22	A2	1A	0522	00A7
armfullstop	23	A3	3A	0523	0589
armparenright	24	A4	29	0524	0029
armparenleft	25	A5	28	0525	002A
armquotright	26	A6	AF	0526	00BB
armquotleft	27	A7	AE	0527	00AB
armemdash	28	A8	2D	0528	2014
armdot	29	A9	2E	0529	002E
armsep	2A	AA	60	052A	055D
armcomma	2B	AB	2C	052B	002C
armendash	2C	AC	5F	052C	2013
armyentamna	2D	AD	DD	052D	058A
armellipsis	2E	AE	DE	052E	2026
armapostrophe	7E	FE	FE	057E	02BC
armexclam	2F	AF	7E	052F	055C
armaccent	30	B0	27	0530	055B
armquestion	31	B1	DF	0531	055E
Armayb	32	B2	80	0532	0531
armayb	33	B3	81	0533	0561
Armben	34	B4	82	0534	0532
armben	35	B5	83	0535	0562
Armgim	36	B6	84	0536	0533
armgim	37	B7	85	0537	0563
Armda	38	B8	86	0538	0534
armda	39	B9	87	0539	0564
Armyech	3A	BA	88	053A	0535
armyech	3B	BB	89	053B	0565
Armza	3C	BC	8A	053C	0536
armza	3D	BD	8B	053D	0566
Arme	3E	BE	8C	053E	0537
arme	3F	BF	8D	053F	0567
Armat	40	C0	8E	0540	0538
armat	41	C1	8F	0541	0568
Armto	42	C2	90	0542	0539
armto	43	C3	91	0543	0569
Armzhe	44	C4	92	0544	053A
armzhe	45	C5	93	0545	056A
Armini	46	C6	94	0546	053B
armini	47	C7	95	0547	056B
Armlyun	48	C8	96	0548	053C
armlyun	49	C9	97	0549	056C
Armkhe	4A	CA	98	054A	053D
armkhe	4B	CB	99	054B	056D

Armtsa	4C	CC	9A	054C	053E
armtsa	4D	CD	9B	054D	056E
Armken	4E	CE	9C	054E	053F
armken	4F	CF	9D	054F	056F
Armho	50	D0	9E	0550	0540
armho	51	D1	9F	0551	0570
Armdza	52	D2	A0	0552	0541
armdza	53	D3	A1	0553	0571
Armghat	54	D4	A2	0554	0542
armghat	55	D5	A3	0555	0572
Armtche	56	D6	A4	0556	0543
armtche	57	D7	A5	0557	0573
Armmen	58	D8	A6	0558	0544
armmen	59	D9	A7	0559	0574
Armhi	5A	DA	A8	055A	0545
armhi	5B	DB	A9	055B	0575
Armnu	5C	DC	AA	055C	0546
armnu	5D	DD	AB	055D	0576
Armsha	5E	DE	AC	055E	0547
armsha	5F	DF	AD	055F	0577
Armvo	60	E0	E0	0560	0548
armvo	61	E1	E1	0561	0578
Armcha	62	E2	E2	0562	0549
armcha	63	E3	E3	0563	0579
Armpe	64	E4	E4	0564	054A
armpe	65	E5	E5	0565	057A
Armje	66	E6	E6	0566	054B
armje	67	E7	E7	0567	057B
Armra	68	E8	E8	0568	054C
armra	69	E9	E9	0569	057C
Armse	6A	EA	EA	056A	054D
armse	6B	EB	EB	056B	057D
Armvev	6C	EC	EC	056C	054E
armvev	6D	ED	ED	056D	057E
Armtyun	6E	EE	EE	056E	054F
armtyun	6F	EF	EF	056F	057F
Armre	70	F0	F0	0570	0550
armre	71	F1	F1	0571	0580
Armtso	72	F2	F2	0572	0551
armtso	73	F3	F3	0573	0581
Armvyun	74	F4	F4	0574	0552
armvyun	75	F5	F5	0575	0582
Armpyur	76	F6	F6	0576	0553
armpyur	77	F7	F7	0577	0583
Armke	78	F8	F8	0578	0554
armke	79	F9	F9	0579	0584
Armo	7A	FA	FA	057A	0555
armo	7B	FB	FB	057B	0585

Armfe	7C	FC	FC	057C	0556
armfe	7D	FD	FD	057D	0586

End of Table 2.

4. Character Set and Language Tags

4.1. Coded Character Set Tags

In the systems and protocols using mnemonic tags for coded character sets, the following tags should be used (name, official source, optional alias):

Name: armscii-8
Source: Armenian State Standard AST 34.002 Basic 8-bit coded character set
Alias: AST_34.002

Name: armscii-8a
Source: Armenian State Standard AST 34.002 Alternative 8-bit coded character set
Alias: AST_34.002-A

4.2. Language Tags

Dictionaries, spelling checkers and other linguistic systems, as well as operating environments distinguishing human languages and locale identification should take into consideration the existence of 4 mutually incomprehensible forms (dialects) of the Armenian language: Eastern, Western, Grabar and Middle. Table 3 presents two forms of suggested mnemonic tags: MIME-style (RFC-1766) and Windows-style 3-letter abbreviations.

Table 3. Language tags

MIME-style name	3-letter code	Full name
hy-eastern	AME	Armenian Eastern
hy-western	AMW	Armenian Western
hy-grabar	AMG	Armenian Grabar
hy-middle	AMM	Armenian Middle

5. Acknowledgements

This document is the result of long and intensive consultations and cooperation with the staff of the Standards Working Group of the Armenian Computer Center. Special thanks for most valuable inputs and comments go to (in alphabetical order):

Hovhannes Gizoghian
Tigran Haroutunian
Aram Hayrapetian
Ivan Lulukian
Vahram Mekhitarian
Rouben Taroumian-Hakobian
Hovhannes Zakaryan

6. Author's Address

Hovik Melikyan
Armenian Computer Center
email: hovik@moon.yerphi.am

7. References

[AST 34.001]
Information Technologies -- Character Set And Information Encoding: Character Set -- State Standardization Committee of the Republic of Armenia, July 1997

[AST 34.002]
Information Technologies -- Character Set And Information Encoding: 8-bit Coded Character Sets -- State Standardization Committee of the Republic of Armenia, July 1997

[ArmSCII]
Armenian Standard Code for Information Interchange -- Center of Humane Technologies "Armenian Computer", June 1991

[RFC-1766]
Alvestrand, H., "Tags for the Identification of Languages", RFC 1766, March 1995.

[UNICODE]
The Unicode Consortium, "The Unicode Standard -- Version 2.0", Addison-Wesley, 1996.

[MIME]

N. Freed, N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies", RFC 2045. N. Freed, N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types", RFC 2046. K. Moore, "MIME (Multipurpose Internet Mail Extensions) Part Three: Message Header Extensions for Non-ASCII Text", RFC 2047. N. Freed, J. Klensin, J. Postel, "Multipurpose Internet Mail Extensions (MIME) Part Four: Registration Procedures", RFC 2048. N. Freed, N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part Five: Conformance Criteria and Examples", RFC 2049. All November 1996.