3GPP TSG-T2 #5 / ETSI SMG4
Helsinki, FINLAND, 6-9 September 1999

TSGT2#5(99)

Agenda Item:

Source: Ericsson

Title: Proposal for additional information in GSM 03.38

Document for: Discussion in Helsinki meeting

The SMS default 7-bit alphabet is defined in GSM 03.38 by a coding table. Such tables may not in themselves provide unambiguous identification of all their characters, since in present-day communications and data processing technology several alphabets exist, with many different characters that have similar appearance in printing.

ISO/IEC in its new and revised standards therefore complements code tables with lists of unique character names taken from the multi-byte character coding standard ISO/IEC 10646-1. ETSI has also, in the latest edition of the ERMES standard, used this method.

This document briefly describes the subject, provides a proposed list identifying the characters of the default alphabet, and describes some additions to the 03.38 text that could be useful.

This page intentionally left blank

Proposal for additional information in GSM 03.38

1 Complementary character identification

The 7-bit coding of the default SMS alphabet is defined in section 6.2.1 of GSM 03.38 by a table of characters. This agrees with the traditional data processing view that it is not necessary to strictly specify what characters are intended in the different positions of code tables, the shapes of the characters being considered identification enough.

With the large number of different alphabets nowadays coexisting in computer and communications technology, the shapes may however no longer be sufficient as identification. In some fonts, characters like for instance the "Inverted exclamation mark" could be mistaken for the Turkish "Capital letter I with dot". Even if, when displayed on a terminal, these two characters may be represented by one and the same "glyph image" (i.e. the displayed shape of the character) it should still be completely clear from the specifying coding standard which one is intended.

This problem was recognized some years ago in ISO. As ISO IT standards are revised, unambiguous character names from the standard ISO/IEC 10646-1 are therefore introduced for identification of characters; sometimes also complemented by the characters' hexadecimal codings according to 10646.

It is proposed that the same principle is introduced in 03.38. In this connection it could be noticed that, in the latest revision of the ETSI ERMES standard, ETS 300 133-2, this method has been used.

On page 4 a proposed table of 10646 names and codings for the SMS default alphabet is given.

In the default table on page 5 there is one proposed modification as compared to the original table in 03.38 section 6.2.1: the character in position 0/09 is identified as "Small letter C with cedilla", not Capital. This is in line with the ERMES standard, which in its latest revision introduced the same change for its alphanumeric character set 0, (presumably mainly since there should be no large need for a "C with cedilla" at the beginning of any sentence).

2 Additional text for standard

If the code table is complemented as proposed in the previous section some text describing the relationship of the new names to ISO standards should be added. It seems suitable, also, to introduce some text describing the table.

The recently developed text for all ISO standards in the ISO/IEC 8859 series ("Latin-1", "Latin-2" ... "Latin/Cyrillic" etc) could be taken as a starting point for a possible 03.38 text extension. An extract from 8859 text describing its code tables is given on page 6.

Table 1 – Character set, coded representation

Bit combi-nation	Hex	Identifier	Name
0/00	00	U+0040	COMMERCIAL AT
0/01	01	U+00A3	POUND SIGN
0/02	02	U+0024	DOLLAR SIGN
0/03	03	U+00A5	YEN SIGN
0/04	04	U+00E8	LATIN SMALL LETTER E WITH GRAVE
0/05	05	U+00E9	LATIN SMALL LETTER E WITH ACUTE
0/06	06	U+00F9	LATIN SMALL LETTER U WITH GRAVE
0/07	07	U+00EC	LATIN SMALL LETTER I WITH GRAVE
0/08	08	U+00F2	LATIN SMALL LETTER O WITH GRAVE
0/09	09	U+00E7	LATIN SMALL LETTER C WITH CEDILLA
0/10	0A		*Control character LINE FEED (see clause ...)*
0/11	0B	U+00D8	LATIN CAPITAL LETTER O WITH STROKE
0/12	0C	U+00F8	LATIN SMALL LETTER O WITH STROKE
0/13	0D		*Control character CARRIAGE RETURN (see clause ...)*
0/14	0E	U+00C5	LATIN CAPITAL LETTER A WITH RING ABOVE
0/15	0F	U+00E5	LATIN SMALL LETTER A WITH RING ABOVE
1/00	10	U+0394	GREEK CAPITAL LETTER DELTA
1/01	11	U+005F	LOW LINE
1/02	12	U+03A6	GREEK CAPITAL LETTER PHI
1/03	13	U+0393	GREEK CAPITAL LETTER GAMMA
1/04	14	U+039B	GREEK CAPITAL LETTER LAMDA
1/05	15	U+03A9	GREEK CAPITAL LETTER OMEGA
1/06	16	U+03A0	GREEK CAPITAL LETTER PI
1/07	17	U+03A8	GREEK CAPITAL LETTER PSI
1/08	18	U+03A3	GREEK CAPITAL LETTER SIGMA
1/09	19	U+0398	GREEK CAPITAL LETTER THETA
1/10	1A	U+039E	GREEK CAPITAL LETTER XI
1/11	1B		*Control character ESCAPE (see clause ...)*
1/12	1C	U+00C6	LATIN CAPITAL LETTER AE
1/13	1D	U+00E6	LATIN SMALL LETTER AE
1/14	1E	U+00DF	LATIN SMALL LETTER SHARP S (German)
1/15	1F	U+00C9	LATIN CAPITAL LETTER E WITH ACUTE
2/00	20	U+0020	SPACE
2/01	21	U+0021	EXCLAMATION MARK
2/02	22	U+0022	QUOTATION MARK
2/03	23	U+0023	NUMBER SIGN
2/04	24	U+00A4	CURRENCY SIGN
2/05	25	U+0025	PERCENT SIGN
2/06	26	U+0026	AMPERSAND
2/07	27	U+0027	APOSTROPHE
2/08	28	U+0028	LEFT PARENTHESIS
2/09	29	U+0029	RIGHT PARENTHESIS
2/10	2A	U+002A	ASTERISK
2/11	2B	U+002B	PLUS SIGN
2/12	2C	U+002C	COMMA
2/13	2D	U+002D	HYPHEN-MINUS
2/14	2E	U+002E	FULL STOP
2/15	2F	U+002F	SOLIDUS
3/00	30	U+0030	DIGIT ZERO
3/01	31	U+0031	DIGIT ONE
3/02	32	U+0032	DIGIT TWO
3/03	33	U+0033	DIGIT THREE
3/04	34	U+0034	DIGIT FOUR
3/05	35	U+0035	DIGIT FIVE
3/06	36	U+0036	DIGIT SIX
3/07	37	U+0037	DIGIT SEVEN
3/08	38	U+0038	DIGIT EIGHT
3/09	39	U+0039	DIGIT NINE
3/10	3A	U+003A	COLON
3/11	3B	U+003B	SEMICOLON
3/12	3C	U+003C	LESS-THAN SIGN
3/13	3D	U+003D	EQUALS SIGN
3/14	3E	U+003E	GREATER-THAN SIGN
3/15	3F	U+003F	QUESTION MARK

Bit combi-nation	Hex	Identifier	Name
4/00	40	U+00A1	INVERTED EXCLAMATION MARK
4/01	41	U+0041	LATIN CAPITAL LETTER A
4/02	42	U+0042	LATIN CAPITAL LETTER B
4/03	43	U+0043	LATIN CAPITAL LETTER C
4/04	44	U+0044	LATIN CAPITAL LETTER D
4/05	45	U+0045	LATIN CAPITAL LETTER E
4/06	46	U+0046	LATIN CAPITAL LETTER F
4/07	47	U+0047	LATIN CAPITAL LETTER G
4/08	48	U+0048	LATIN CAPITAL LETTER H
4/09	49	U+0049	LATIN CAPITAL LETTER I
4/10	4A	U+004A	LATIN CAPITAL LETTER J
4/11	4B	U+004B	LATIN CAPITAL LETTER K
4/12	4C	U+004C	LATIN CAPITAL LETTER L
4/13	4D	U+004D	LATIN CAPITAL LETTER M
4/14	4E	U+004E	LATIN CAPITAL LETTER N
4/15	4F	U+004F	LATIN CAPITAL LETTER O
5/00	50	U+0050	LATIN CAPITAL LETTER P
5/01	51	U+0051	LATIN CAPITAL LETTER Q
5/02	52	U+0052	LATIN CAPITAL LETTER R
5/03	53	U+0053	LATIN CAPITAL LETTER S
5/04	54	U+0054	LATIN CAPITAL LETTER T
5/05	55	U+0055	LATIN CAPITAL LETTER U
5/06	56	U+0056	LATIN CAPITAL LETTER V
5/07	57	U+0057	LATIN CAPITAL LETTER W
5/08	58	U+0058	LATIN CAPITAL LETTER X
5/09	59	U+0059	LATIN CAPITAL LETTER Y
5/10	5A	U+005A	LATIN CAPITAL LETTER Z
5/11	5B	U+00C4	LATIN CAPITAL LETTER A WITH DIAERESIS
5/12	5C	U+00D6	LATIN CAPITAL LETTER O WITH DIAERESIS
5/13	5D	U+00D1	LATIN CAPITAL LETTER N WITH TILDE
5/14	5E	U+00DC	LATIN CAPITAL LETTER U WITH DIAERESIS
5/15	5F	U+00A7	SECTION SIGN
6/00	60	U+00BF	INVERTED QUESTION MARK
6/01	61	U+0061	LATIN SMALL LETTER A
6/02	62	U+0062	LATIN SMALL LETTER B
6/03	63	U+0063	LATIN SMALL LETTER C
6/04	64	U+0064	LATIN SMALL LETTER D
6/05	65	U+0065	LATIN SMALL LETTER E
6/06	66	U+0066	LATIN SMALL LETTER F
6/07	67	U+0067	LATIN SMALL LETTER G
6/08	68	U+0068	LATIN SMALL LETTER H
6/09	69	U+0069	LATIN SMALL LETTER I
6/10	6A	U+006A	LATIN SMALL LETTER J
6/11	6B	U+006B	LATIN SMALL LETTER K
6/12	6C	U+006C	LATIN SMALL LETTER L
6/13	6D	U+006D	LATIN SMALL LETTER M
6/14	6E	U+006E	LATIN SMALL LETTER N
6/15	6F	U+006F	LATIN SMALL LETTER O
7/00	70	U+0070	LATIN SMALL LETTER P
7/01	71	U+0071	LATIN SMALL LETTER Q
7/02	72	U+0072	LATIN SMALL LETTER R
7/03	73	U+0073	LATIN SMALL LETTER S
7/04	74	U+0074	LATIN SMALL LETTER T
7/05	75	U+0075	LATIN SMALL LETTER U
7/06	76	U+0076	LATIN SMALL LETTER V
7/07	77	U+0077	LATIN SMALL LETTER W
7/08	78	U+0078	LATIN SMALL LETTER X
7/09	79	U+0079	LATIN SMALL LETTER Y
7/10	7A	U+007A	LATIN SMALL LETTER Z
7/11	7B	U+00E4	LATIN SMALL LETTER A WITH DIAERESIS
7/12	7C	U+00F6	LATIN SMALL LETTER O WITH DIAERESIS
7/13	7D	U+00F1	LATIN SMALL LETTER N WITH TILDE
7/14	7E	U+00FC	LATIN SMALL LETTER U WITH DIAERESIS
7/15	7F	U+00E0	LATIN SMALL LETTER A WITH GRAVE

For each character in the set the code table (table 2) shows a graphic symbol at the position in the code table corresponding to the bit combination specified in table 1.

The shaded positions in the code table correspond to bit combinations that represent control characters. Their use is specified in clause ....

Table 2 – Code table of default alphabet

b_7	0	0	0	0	1	1	1	1	
b_6	0	0	1	1	0	0	1	1	
b_5	0	1	0	1	0	1	0	1	
b_4 b_3 b_2 b_1	0	1	2	3	4	5	6	7	
0 0 0 0	@	Δ	SP	0	¡	P	¿	p	0
0 0 0 1	£	_	!	1	A	Q	a	q	1
0 0 1 0	$	Φ	"	2	B	R	b	r	2
0 0 1 1	¥	Γ	#	3	C	S	c	s	3
0 1 0 0	è	Λ	¤	4	D	T	d	t	4
0 1 0 1	é	Ω	%	5	E	U	e	u	5
0 1 1 0	ù	Π	&	6	F	V	f	v	6
0 1 1 1	ì	Ψ	'	7	G	W	g	w	7
1 0 0 0	ò	Σ	(	8	H	X	h	x	8
1 0 0 1	ç	Θ	)	9	I	Y	i	y	9
1 0 1 0	LF	Ξ	*	:	J	Z	j	z	A
1 0 1 1	Ø	ESC	+	;	K	Ä	k	ä	B
1 1 0 0	ø	Æ	,	<	L	Ö	l	ö	C
1 1 0 1	CR	æ	-	=	M	Ñ	m	ñ	D
1 1 1 0	Å	ß	.	>	N	Ü	n	ü	E
1 1 1 1	å	É	/	?	O	§	o	à	F
	0	1	2	3	4	5	6	7	hex

Text extract from ISO/IEC 8859-1 *(Document copyright ISO/IEC)*

5 Notation, code table and names

5.1 Notation

The bits of the bit combinations of the 8-bit code are identified by b_8, b_7, b_6, b_5, b_4, b_3, b_2, and b_1, where b_8 is the highest-order, or most-significant bit and b_1 is the lowest-order, or least-significant bit.

The bit combinations may be interpreted to represent numbers in binary notation by attributing the following weights to the individual bits:

Bit	b_8	b_7	b_6	b_5	b_4	b_3	b_2	b_1
Weight	128	64	32	16	8	4	2	1

Using these weights, the bit combinations are identified by notations of the form xx/yy, where xx and yy are numbers in the range 00 to 15. The correspondence between the notations of the form xx/yy and the bit combinations consisting of the bits b_8 to b_1 is as follows:

– xx is the number represented by b_8, b_7, b_6 and b_5 where these bits are given the weights 8, 4, 2, and 1 respectively.

– yy is the number represented by b_4, b_3, b_2 and b_1 where these bits are given the weights 8, 4, 2, and 1 respectively.

The bit combinations are also identified by notations of the form hk, where h and k are numbers in the range 0 to F in hexadecimal notation. The number h is the same as the number xx described above, and the number k the same as the number yy described above.

5.2 Layout of the code table

An 8-bit code table consists of 256 positions arranged in 16 columns and 16 rows. The columns and the rows are numbered 00 to 15. In hexa-decimal notation the columns and the rows are numbered 0 to F.

The code table positions are identified by notations of the form xx/yy, where xx is the column number and yy is the row number. The column and row numbers are shown at the top and left edges of the table respectively. The code table positions are also identified by notations of the form hk, where h is the column number and k is the row number in hexadecimal notation. The column and row numbers are shown at the bottom and right edges of the table respectively.

The positions of the code table are in one-to-one correspondence with the bit combinations of the code. The notation of a code table position, of the form xx/yy, or of the form hk, is the same as that of the corresponding bit combination.

5.3 Names and meanings

This part of ISO/IEC 8859 assigns a unique name and a unique identifier to each graphic character. These names and identifiers have been taken from ISO/IEC 10646-1 (E). This part of ISO/IEC 8859 also specifies an acronym for each of the characters SPACE, NO-BREAK SPACE and SOFT HYPHEN. For acronyms only Latin capital letters A to Z are used. It is intended that the acronyms be retained in all translations of the text.

Except for SPACE (SP), NO-BREAK SPACE (NBSP) and SOFT HYPHEN (SHY), this part of ISO/IEC 8859 does not define and does not restrict the meanings of graphic characters.

This part of ISO/IEC 8859 specifies a graphic symbol for each graphic character. This symbol is shown in the corresponding position of the code table. However, this part, or any other part, of ISO/IEC 8859 does not specify a particular style or font design for imaging graphic characters. Annex B of ISO/IEC 10367 gives further information on this subject.

5.3.1 SPACE (SP)

A graphic character the visual representation of which consists of the absence of a graphic symbol.

5.3.2 NO-BREAK SPACE (NBSP)

.....