ETSI TS 123 038 V18.0.0 (2024-05)

Digital cellular telecommunications system (Phase 2+) (GSM);
Universal Mobile Telecommunications System (UMTS);
LTE;
5G;
Alphabets and language-specific information
(3GPP TS 23.038 version 18.0.0 Release 18)

Reference
RTS/TSGC-0123038vi00

Keywords
5G,GSM,LTE,UMTS

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Important notice

The present document can be downloaded from:
https://www.etsi.org/standards-search

The present document may be made available in electronic versions and/or in print. The content of any electronic and/or print versions of the present document shall not be modified without the prior written authorization of ETSI. In case of any existing or perceived difference in contents between such versions and/or in print, the prevailing version of an ETSI deliverable is the one made publicly available in PDF format at www.etsi.org/deliver.

Users of the present document should be aware that the document may be subject to revision or change of status.
Information on the current status of this and other ETSI documents is available at
https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

If you find errors in the present document, please send your comment to one of the following services:
https://portal.etsi.org/People/CommiteeSupportStaff.aspx

If you find a security vulnerability in the present document, please report it through our
Coordinated Vulnerability Disclosure Program:
https://www.etsi.org/standards/coordinated-vulnerability-disclosure

Notice of disclaimer & limitation of liability

The information provided in the present deliverable is directed solely to professionals who have the appropriate degree of experience to understand and interpret its content in accordance with generally accepted engineering or other professional standard and applicable regulations.
No recommendation as to products and services or vendors is made or should be implied.
No representation or warranty is made that this deliverable is technically accurate or sufficient or conforms to any law and/or governmental rule and/or regulation and further, no representation or warranty is made of merchantability or fitness for any particular purpose or against infringement of intellectual property rights.
In no event shall ETSI be held liable for loss of profits or any other incidental or consequential damages.

Any software contained in this deliverable is provided "AS IS" with no warranties, express or implied, including but not limited to, the warranties of merchantability, fitness for a particular purpose and non-infringement of intellectual property rights and ETSI shall not be held liable in any event for any damages whatsoever (including, without limitation, damages for loss of profits, business interruption, loss of information, or any other pecuniary loss) arising out of or related to the use of or inability to use the software.

Copyright Notification

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™** and **LTE™** are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Legal Notice

This Technical Specification (TS) has been produced by ETSI 3rd Generation Partnership Project (3GPP).

The present document may refer to technical specifications or reports using their 3GPP identities. These shall be interpreted as being references to the corresponding ETSI deliverables.

The cross reference between 3GPP and ETSI identities can be found under https://webapp.etsi.org/key/queryform.asp.

Modal verbs terminology

In the present document "**shall**", "**shall not**", "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Contents

Foreword

This Technical Specification has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

x the first digit:

1 presented to TSG for information;

2 presented to TSG for approval;

3 Indicates TSG approved document under change control.

y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.

z the third digit is incremented when editorial only changes have been incorporated in the specification;

1 Scope

The present document defines the character sets, languages and message handling requirements for SMS, CBS and USSD and may additionally be used for Man Machine Interface (MMI) (3GPP TS 22.030 [2]).

The specification for the Data Circuit terminating Equipment/Data Terminal Equipment (DCE/DTE) interface (3GPP TS 27.005 [8]) will also use the codes specified herein for the transfer of SMS data to an external terminal.

2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

[1] void

[2] 3GPP TS 22.030: "Man-Machine Interface (MMI) of the User Equipment (UE)".

[3] 3GPP TS 23.090: "Unstructured Supplementary Service Data (USSD) - Stage 2".

[4] 3GPP TS 23.040: "Technical realization of the Short Message Service (SMS) ".

[5] 3GPP TS 23.041: "Technical realization of Cell Broadcast Service (CBS)".

[6] 3GPP TS 24.011: "Point-to-Point (PP) Short Message Service (SMS) support on mobile radio interface".

[7] Void.

[8] 3GPP TS 27.005: "Use of Data Terminal Equipment - Data Circuit terminating Equipment (DTE - DCE) interface for Short Message Service (SMS) and Cell Broadcast Service (CBS)".

[10] ISO/IEC 10646: "Information technology; Universal Multiple-Octet Coded Character Set (UCS)".

[11] 3GPP TS 24.090: "Unstructured Supplementary Service Data (USSD); Stage 3".

[12] ISO 639: "Code for the representation of names of languages".

[13] 3GPP TS 23.042: "Compression algorithm for text messaging services".

[14] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".

[15] "Wireless Datagram Protocol Specification", Wireless Application Protocol Forum Ltd.

[16] ISO 1073-1 and ISO 1073-2 Alphanumeric character sets for optical recognition – Parts 1 and 2: Character sets OCR-A and OCR-B, respectively - Shapes and dimensions of the printed image.

[17] 3GPP TS 31.102: "Characteristics of the USIM application"

[18] 3GPP TS 51.011 Release 4 (version 4.x.x): “Specification of the Subscriber Identity Module - Mobile Equipment (SIM - ME) interface”

[19] 3GPP TS 24.294: "IMS Centralized Services (ICS) Protocol via I1 Interface".

[20] 3GPP TS 24.008: "Mobile radio interface Layer 3 specification; Core network protocols; Stage 3".

[21] 3GPP TS 24.301: " Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3".

3 Abbreviations and definitions

For the purposes of the present document, the following terms and definitions apply:

National Language Identifier: A code representing a specific language and thereby selecting a specific National Language Table.

National Language Locking Shift Table: A national language table which replaces the GSM 7 bit default alphabet table in the case where the locking shift mechanism as defined in subclause 6.2.1.2.3 is used.

National Language Single Shift Table: A national language table which replaces the GSM 7 bit default alphabet extension table in the case where the single shift mechanism as defined in subclause 6.2.1.2.2 is used.

National Language Table: A table containing the characters of a specific national language.

For the purposes of the present document, the abbreviations used in the present document are listed in 3GPP TR 21.905 [14].

4 SMS Data Coding Scheme

The TP-Data-Coding-Scheme field, defined in 3GPP TS 23.040 [4], indicates the data coding scheme of the TP-UD field, and may indicate a message class. Any reserved codings shall be assumed to be the GSM 7 bit default alphabet (the same as codepoint 00000000) by a receiving entity. The octet is used according to a coding group which is indicated in bits 7..4. The octet is then coded as follows:

Coding Group Bits 7..4	Use of bits 3..0
00xx	General Data Coding indication Bits 5..0 indicate the following: Bit 5, if set to 0, indicates the text is uncompressed Bit 5, if set to 1, indicates the text is compressed using the compression algorithm defined in 3GPP TS 23.042 [13] Bit 4, if set to 0, indicates that bits 1 to 0 are reserved and have no message class meaning Bit 4, if set to 1, indicates that bits 1 to 0 have a message class meaning:: Bit 1 Bit 0 Message Class 0 0 Class 0 0 1 Class 1 Default meaning: ME-specific. 1 0 Class 2 (U)SIM specific message 1 1 Class 3 Default meaning: TE specific (see 3GPP TS 27.005 [8]) Bits 3 and 2 indicate the character set being used, as follows : Bit 3 Bit2 Character set: 0 0 GSM 7 bit default alphabet 0 1 8 bit data 1 0 UCS2 (16bit) [10] 1 1 Reserved NOTE: The special case of bits 7..0 being 0000 0000 indicates the GSM 7 bit default alphabet with no message class
01xx	Message Marked for Automatic Deletion Group This group can be used by the SM originator to mark the message (stored in the ME or (U)SIM) for deletion after reading irrespective of the message class. The way the ME will process this deletion should be manufacturer specific but shall be done without the intervention of the End User or the targeted application. The mobile manufacturer may optionally provide a means for the user to prevent this automatic deletion. Bit 5..0 are coded exactly the same as Group 00xx
1000..1011	Reserved coding groups
1100	Message Waiting Indication Group: Discard Message The specification for this group is exactly the same as for Group 1101, except that: - after presenting an indication and storing the status, the ME may discard the contents of the message. The ME shall be able to receive, process and acknowledge messages in this group, irrespective of memory availability for other types of short message.

Coding Group Bits 7..4	Use of bits 3..0
1101	Message Waiting Indication Group: Store Message This Group defines an indication to be provided to the user about the status of types of message waiting on systems connected to the GSM/UMTS PLMN. The ME should present this indication as an icon on the screen, or other MMI indication. The ME shall update the contents of the Message Waiting Indication Status on the SIM (see 3GPP TS 51.011 [18]) or USIM (see 3GPP TS 31.102 [17]) when present or otherwise should store the status in the ME. In case there are multiple records of EF_{MWIS} this information shall be stored within the first record. The contents of the Message Waiting Indication Status should control the ME indicator. For each indication supported, the mobile may provide storage for the Origination Address. The ME may take note of the Origination Address for messages in this group and group 1100. Text included in the user data is coded in the GSM 7 bit default alphabet. Where a message is received with bits 7..4 set to 1101, the mobile shall store the text of the SMS message in addition to setting the indication. The indication setting should take place irrespective of memory availability to store the short message. Bits 3 indicates Indication Sense: Bit 3 0 Set Indication Inactive 1 Set Indication Active Bit 2 is reserved, and set to 0 Bit 1 Bit 0 Indication Type: 0 0 Voicemail Message Waiting 0 1 Fax Message Waiting 1 0 Electronic Mail Message Waiting 1 1 Other Message Waiting* * Mobile manufacturers may implement the "Other Message Waiting" indication as an additional indication without specifying the meaning.
1110	Message Waiting Indication Group: Store Message The coding of bits 3..0 and functionality of this feature are the same as for the Message Waiting Indication Group above, (bits 7..4 set to 1101) with the exception that the text included in the user data is coded in the uncompressed UCS2 character set.
1111	Data coding/message class Bit 3 is reserved, set to 0. Bit 2 Message coding: 0 GSM 7 bit default alphabet 1 8-bit data Bit 1 Bit 0 Message Class: 0 0 Class 0 0 1 Class 1 default meaning: ME-specific. 1 0 Class 2 (U)SIM-specific message. 1 1 Class 3 default meaning: TE specific (see 3GPP TS 27.005 [8])

GSM 7 bit default alphabet indicates that the TP-UD is coded from the GSM 7 bit default alphabet given in clause 6.2.1. When this character set is used, the characters of the message are packed in octets as shown in clause 6.1.2.1.1, and the message can consist of up to 160 characters. The GSM 7 bit default alphabet shall be supported by all MSs and SCs offering the service. If the GSM 7 bit default alphabet extension mechanism is used then the number of displayable characters will reduce by one for every instance where the GSM 7 bit default alphabet extension table is used. 8-bit data indicates that the TP-UD has user-defined coding, and the message can consist of up to 140 octets.

UCS2 character set indicates that the TP-UD has a UCS2 [10] coded message, and the message can consist of up to 140 octets, i.e. up to 70 UCS2 characters. The General notes specified in clause 6.1.1 override any contrary

specification in UCS2, so for example even in UCS2 a <CR> character will cause the MS to return to the beginning of the current line and overwrite any existing text with the characters which follow the <CR>.

When a message is compressed, the TP-UD consists of the GSM 7 bit default alphabet or UCS2 character set compressed message, and the compressed message itself can consist of up to 140 octets in total.

When a mobile terminated message is class 0 and the MS has the capability of displaying short messages, the MS shall display the message immediately and send an acknowledgement to the SC when the message has successfully reached the MS irrespective of whether there is memory available in the (U)SIM or ME. The message shall not be automatically stored in the (U)SIM or ME.

The ME may make provision through MMI for the user to selectively prevent the message from being displayed immediately.

If the ME is incapable of displaying short messages or if the immediate display of the message has been disabled through MMI then the ME shall treat the short message as though there was no message class, i.e. it will ignore bits 0 and 1 in the TP-DCS and normal rules for memory capacity exceeded shall apply.

When a mobile terminated message is Class 1, the MS shall send an acknowledgement to the SC when the message has successfully reached the MS and can be stored. The MS shall normally store the message in the ME by default, if that is possible, but otherwise the message may be stored elsewhere, e.g. in the (U)SIM. The user may be able to override the default meaning and select their own routing.

When a mobile terminated message is Class 2 ((U)SIM-specific), an MS shall ensure that the message has been transferred to the SMS data field in the (U)SIM before sending an acknowledgement to the SC. The MS shall return a "protocol error, unspecified" error message (see 3GPP TS 24.011 [6]) if the short message cannot be stored in the (U)SIM and there is other short message storage available at the MS. If all the short message storage at the MS is already in use, the MS shall return "memory capacity exceeded". This behaviour applies in all cases except for an MS supporting (U)SIM Application Toolkit when the Protocol Identifier (TP-PID) of the mobile terminated message is set to "(U)SIM Data download" (see 3GPP TS 23.040 [4]).

When a mobile terminated message is Class 3, the MS shall send an acknowledgement to the SC when the message has successfully reached the MS and can be stored, irrespectively of whether the MS supports an SMS interface to a TE, and without waiting for the message to be transferred to the TE. Thus the acknowledgement to the SC of a TE-specific message does not imply that the message has reached the TE. Class 3 messages shall normally be transferred to the TE when the TE requests "TE-specific" messages (see 3GPP TS 27.005 [8]). The user may be able to override the default meaning and select their own routing.

The message class codes may also be used for mobile originated messages, to provide an indication to the destination SME of how the message was handled at the MS.

The MS will not interpret reserved or unsupported values but shall store them as received. The SC may reject messages with a Data Coding Scheme containing a reserved value or one which is not supported.

5 CBS Data Coding Scheme

The CBS Data Coding Scheme indicates the intended handling of the message at the MS, the character set/coding, and the language (when applicable). Any reserved codings shall be assumed to be the GSM 7 bit default alphabet (the same as codepoint 00001111) by a receiving entity. The octet is used according to a coding group which is indicated in bits 7..4. The octet is then coded as follows:

Coding Group Bits 7..4	Use of bits 3..0
0000	Language using the GSM 7 bit default alphabet Bits 3..0 indicate the language: 0000 German 0001 English 0010 Italian 0011 French 0100 Spanish 0101 Dutch 0110 Swedish 0111 Danish 1000 Portuguese 1001 Finnish 1010 Norwegian 1011 Greek 1100 Turkish 1101 Hungarian 1110 Polish 1111 Language unspecified
0001	0000 GSM 7 bit default alphabet; message preceded by language indication. The first 3 characters of the message are a two-character representation of the language encoded according to ISO 639 [12], followed by a CR character. The CR character is then followed by 90 characters of text (NOTE 1). 0001 UCS2; message preceded by language indication The message starts with a two GSM 7-bit default alphabet character representation of the language encoded according to ISO 639 [12]. This is padded to the octet boundary with two bits set to 0 and then followed by 40 characters of UCS2-encoded message (NOTE 1). An MS not supporting UCS2 coding will present the two character language identifier followed by improperly interpreted user data. 0010..1111 Reserved
0010..	0000 Czech 0001 Hebrew (NOTE 2) 0010 Arabic (NOTE 2) 0011 Russian (NOTE 2) 0100 Icelandic 0101..1111 Reserved for other languages using the GSM 7 bit default alphabet, with unspecified handling at the MS
0011	0000..1111 Reserved for other languages using the GSM 7 bit default alphabet, with unspecified handling at the MS

Coding Group Bits 7..4	Use of bits 3..0
01xx	General Data Coding indication Bits 5..0 indicate the following: Bit 5, if set to 0, indicates the text is uncompressed Bit 5, if set to 1, indicates the text is compressed using the compression algorithm defined in 3GPP TS 23.042 [13] Bit 4, if set to 0, indicates that bits 1 to 0 are reserved and have no message class meaning Bit 4, if set to 1, indicates that bits 1 to 0 have a message class meaning: Bit 1 Bit 0 Message Class: 0 0 Class 0 0 1 Class 1 Default meaning: ME-specific. 1 0 Class 2 (U)SIM specific message. 1 1 Class 3 Default meaning: TE-specific (see 3GPP TS 27.005 [8]) Bits 3 and 2 indicate the character set being used, as follows: Bit 3 Bit 2 Character set: 0 0 GSM 7 bit default alphabet 0 1 8 bit data 1 0 UCS2 (16 bit) [10] 1 1 Reserved
1000	Reserved coding groups
1001	Message with User Data Header (UDH) structure: Bit 1 Bit 0 Message Class: 0 0 Class 0 0 1 Class 1 Default meaning: ME-specific. 1 0 Class 2 (U)SIM specific message. 1 1 Class 3 Default meaning: TE-specific (see 3GPP TS 27.005 [8]) Bits 3 and 2 indicate the alphabet being used, as follows: Bit 3 Bit 2 Alphabet: 0 0 GSM 7 bit default alphabet 0 1 8 bit data 1 0 USC2 (16 bit) [10] 1 1 Reserved
1010..1100	Reserved coding groups
1101	I1 protocol message defined in 3GPP TS 24.294 [19]
1110	Defined by the WAP Forum [15]
1111	Data coding / message handling Bit 3 is reserved, set to 0. Bit 2 Message coding: 0 GSM 7 bit default alphabet 1 8 bit data Bit 1 Bit 0 Message Class: 0 0 No message class. 0 1 Class 1 user defined. 1 0 Class 2 user defined. 1 1 Class 3 default meaning: TE specific (see 3GPP TS 27.005 [8])
NOTE 1: The language indication shall appear at the start of each Message Information Page (see 3GPP TS 23.041 [5]) and the language indication on each Message Information Page shall be for the same language. NOTE 2: Message text in Hebrew, Arabic and Russian cannot be encoded in the GSM 7-bit default alphabet. For these languages UCS2 encoding shall be used.	

These codings may also be used for USSD and MMI/display purposes.