## ISO

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION ORGANISATION INTERNATIONALE DE NORMALISATION

## ISO/IEC JTC1/SC2/WG2 Universal Multiple-Octet Coded Character Set (UCS)

## ISO/IEC JTC1/SC2/WG2 N 2005

Date: 1999-05-29

## TITLE: ISO/IEC 10646-1 Second Edition text, Draft 2

SOURCE: Bruce Paterson, project editor
STATUS: Working paper of JTC1/SC2/WG2
ACTION: For review and comment by WG2
DISTRIBUTION: Members of JTC1/SC2/WG2

## 1. Scope

This paper provides a second draft of the text sections of the Second Edition of ISO/IEC 10646-1. It replaces the previous paper WG2 N 1796 (1998-06-01).
This draft text includes:

- Clauses 1 to 27 (replacing the previous clauses 1 to 26),
- Annexes A to R (replacing the previous Annexes A to T), and is attached here as "Draft 2 for ISO/IEC 10646-1 : 1999" (pages ii \& 1 to 77).
Published and Draft Amendments up to Amd. 31 (Tibetan extended), Technical Corrigenda nos. 1, 2, and 3, and editorial corrigenda approved by WG2 up to 1999-03-15, have been applied to the text.
The draft does not include:
- character glyph tables and name tables (these will be provided in a separate WG2 document from AFII),
- the alphabetically sorted list of character names in Annex E (now Annex G),
- markings to show the differences from the previous draft.

A separate WG2 paper will give the editorial corrigenda applied to this text since N 1796 . The editorial corrigenda are as agreed at WG2 meetings \#34 to \#36.
Editorial corrigenda applicable to the character glyph tables and name tables, as listed in N1796 pages 2 to 5 , have already been applied to the draft character tables prepared by AFII. in March 1999.

## 2. Electronic version of this text

The electronic version of this text is supplied as three separate files for convenience:

- WG2 cover sheet and Clauses 1 to 27 (pages i, ii, \& 1-21),
- Annexes A to Q (pages 22-69),
- Annex R (Informative annex on CJK unification, pages 70-77, file size 1.075 MB).


## Contents

Page
1 Scope ..... 1
2 Conformance ..... 1
3 Normative references ..... 2
4 Definitions ..... 2
5 General structure of the UCS ..... 4
6 Basic structure and nomenclature ..... 4
7 General requirements for the UCS ..... 8
8 The Basic Multilingual Plane ..... 8
9 Other planes ..... 8
10 Private use groups, planes, and zones ..... 8
11 Revision and updating of the UCS ..... 9
12 Subsets ..... 9
13 Coded representation forms of the UCS ..... 9
14 Implementation levels ..... 9
15 Use of control functions with the UCS ..... 10
16 Declaration of identification of features ..... 10
17 Structure of the code tables and lists ..... 11
18 Block names ..... 12
19 Characters in bi-directional context ..... 12
20 Special characters ..... 12
21 Presentation forms of characters ..... 13
22 Compatibility characters ..... 13
23 Order of characters ..... 13
24 Combining characters ..... 13
25 Special features of individual scripts ..... 14
26 Code tables and lists of character names ..... 15
27 CJK unified ideographs ..... 20
Annexes
A Collections of graphic characters for subsets ..... 22
B List of combining characters ..... 28
C Transformation format for 16 planes of Group 00 (UTF-16) ..... 33
D UCS Transformation Format 8 (UTF-8) ..... 36
E Mirrored characters in Arabic bi-directional context ..... 40
F Alternate format characters ..... 42
G Alphabetically sorted list of character names ..... 47
H The use of "signatures" to identify UCS ..... 48
$J$ Recommendation for combined receiving/originating devices with internal storage ..... 49
K Notations of octet value representations ..... 50
L Character naming guidelines ..... 51
M Sources of characters ..... 53
N External references to character repertoires ..... 55
P Additional information on characters ..... 57
Q Code mapping table for Hangul syllables ..... 60
R Procedure for the unification and arrangement of CJK Ideographs ..... 70

# Information technology — Universal Multiple-Octet Coded Character Set (UCS) 

## Part 1:

Architecture and Basic Multilingual Plane

## 1 Scope

ISO/IEC 10646 specifies the Universal MultipleOctet Coded Character Set (UCS). It is applicable to the representation, transmission, interchange, processing, storage, input, and presentation of the written form of the languages of the world as well as additional symbols.

This part of ISO/IEC 10646 specifies the overall architecture, and

- defines terms used in ISO/IEC 10646;
- describes the general structure of the coded character set;
- specifies the Basic Multilingual Plane (BMP) of the UCS, and defines a set of graphic characters used in scripts and the written form of languages on a world-wide scale;
- specifies the names for the graphic characters of the BMP, and the coded representations;
- specifies the four-octet (32-bit) canonical form of the UCS: UCS-4;
- specifies a two-octet (16-bit) BMP form of the UCS: UCS-2;
- specifies the coded representations for control functions;
- specifies the management of future additions to this coded character set

The UCS is a coding system different from that specified in ISO 2022. The method to designate UCS from ISO 2022 is specified in 16.2.

NOTE - It is intended that character code positions for additional scripts and symbols will be allocated in this Part 1 of this International Standard when sufficient input and review is provided by national standards organizations or other qualified experts.

## 2 Conformance

### 2.1 General

Whenever private use characters are used as specified in ISO/IEC 10646, the characters themselves shall not be covered by these conformance requirements.

### 2.2 Conformance of information interchange

A coded-character-data-element (CC-data-element) within coded information for interchange is in conformance with ISO/IEC 10646 if
a) all the coded representations of graphic characters within that CC-data-element conform to clauses 6 and 7, to an identified form chosen from clause 13 or Annex $C$ or Annex $D$, and to an identified implementation level chosen from clause 14;
b) all the graphic characters represented within that CC-data-element are taken from those within an identified subset (clause 12);
c) all the coded representations of control functions within that CC-data-element conform to clause 15.

A claim of conformance shall identify the adopted form, the adopted implementation level and the adopted subset by means of a list of collections and/or characters.

### 2.3 Conformance of devices

A device is in conformance with ISO/IEC 10646 if it conforms to the requirements of item a) below, and either or both of items b) and c).

NOTE - The term device is defined (in 4.18) as a component of information processing equipment which can transmit and/or receive coded information within CC-dataelements. A device may be a conventional input/output device, or a process such as an application program or gateway function.

A claim of conformance shall identify the document that contains the description specified in a) below, and shall identify the adopted form(s), the adopted implementation level, the adopted subset (by means
of a list of collections and/or characters), and the selection of control functions adopted in accordance with clause 15.
a) Device description: A device that conforms to ISO/IEC 10646 shall be the subject of a description that identifies the means by which the user may supply characters to the device and/or may recognize them when they are made available to the user, as specified respectively, in subclauses b), and c) below.
b) Originating device: An originating device shall allow its user to supply any characters from an adopted subset, and be capable of transmitting their coded representations within a CC-data-element in accordance with the adopted form and implementation level.
c) Receiving device: A receiving device shall be capable of receiving and interpreting any coded representation of characters that are within a CC-data-element in accordance with the adopted form and implementation level, and shall make any corresponding characters from the adopted subset available to the user in such a way that the user can identify them.

Any corresponding characters that are not within the adopted subset shall be indicated to the user. The way used for indicating them need not distinguish them from each other.

## NOTES

1 An indication to the user may consist of making available the same character to represent all characters not in the adopted subset, or providing a distinctive audible or visible signal when appropriate to the type of user.
2 See also annex $J$ for receiving devices with retransmission capability.

## 3 Normative references

The following standards contain provisions which, through reference in this text, constitute provisions of this part of ISO/IEC 10646. At the time of publication, the editions indicated were valid. All standards are subject to revision, and parties to agreements based on this part of ISO/IEC 10646 are encouraged to investigate the possibility of applying the most recent editions of the standards listed below. Members of IEC and ISO maintain registers of currently valid International Standards.
ISO/IEC 2022:1994 Information technology Character code structure and extension techniques.
ISO/IEC 6429:1992 Information technology Control functions for coded character sets.

## 4 Definitions

For the purposes of ISO/IEC 10646, the following definitions apply:

### 4.1 Basic Multilingual Plane (BMP): Plane 00 of Group 00.

4.2 block: A contiguous range of code positions to which a set of characters that share common characteristics, such as script, are allocated. A block cannot overlap another block. One or more of the code positions within a block may have no character allocated to it.
4.3 canonical form: The form with which characters of this coded character set are specified using four octets to represent each character.

### 4.4 CC-data-element (coded-character-data-

element): An element of interchanged information that is specified to consist of a sequence of coded representations of characters, in accordance with one or more identified standards for coded character sets.
4.5 cell: The place within a row at which an individual character may be allocated.
4.6 character: A member of a set of elements used for the organisation, control, or representation of data.
4.7 character boundary: Within a stream of octets the demarcation between the last octet of the coded representation of a character and the first octet of that of the next coded character.
4.8 coded character: A character together with its coded representation.
4.9 coded character set: A set of unambiguous rules that establishes a character set and the relationship between the characters of the set and their coded representation.
4.10 code table: A table showing the characters allocated to the octets in a code.
4.11 collection: A set of coded characters which is numbered and named and which consists of those coded characters whose code positions lie within one or more identified ranges.

NOTE - If any of the identified ranges include code positions to which no character is allocated, the repertoire of the collection will change if an additional character is assigned to any of those positions at a future amendment of this International Standard. However it is intended that the collection number and name will remain unchanged in future editions of this International Standard.
4.12 combining character: A member of an identified subset of the coded character set of ISO/IEC 10646 intended for combination with the preceding non-combining graphic character, or with
a sequence of combining characters preceded by a non-combining character (see also 4.14).

NOTE - This part of ISO/IEC 10646 specifies several subset collections which include combining characters.
4.13 compatibility character: A graphic character included as a coded character of ISO/IEC 10646 primarily for compatibility with existing coded character sets.
4.14 composite sequence: A sequence of graphic characters consisting of a non-combining character followed by one or more combining characters (see also 4.12).

## NOTES

1 A graphic symbol for a composite sequence generally consists of the combination of the graphic symbols of each character in the sequence.
2 A composite sequence is not a character and therefore is not a member of the repertoire of ISO/IEC 10646.
4.15 control function: An action that affects the recording, processing, transmission or interpretation of data, and that has a coded representation consisting of one or more octets.
4.16 default state: The state that is assumed when no state has been explicitly specified.
4.17 detailed code table: A code table showing the individual characters, and normally showing a partial row.
4.18 device: A component of information processing equipment which can transmit and/or receive coded information within CC-data-elements. (It may be an input/output device in the conventional sense, or a process such as an application program or gateway function.)
4.19 fixed collection: A collection in which every code position within the identified range(s) has a character allocated to it, and which is intended to remain unchanged in future editions of this International Standard.
4.20 graphic character: A character, other than a control function, that has a visual representation normally handwritten, printed, or displayed.
4.21 graphic symbol: The visual representation of a graphic character or of a composite sequence.
4.22 group: A subdivision of the coding space of this coded character set; of $256 \times 256 \times 256$ cells.
4.23 high-half zone: a set of cells reserved for use in UTF-16 (see Annex C); an RC-element corresponding to any of these cells may be used as the first of a pair of RC-elements which represents a character from a plane other than the BMP.
4.24 interchange: The transfer of character coded data from one user to another, using telecommunication means or interchangeable media.
4.25 interworking: The process of permitting two or more systems, each employing different coded character sets, meaningfully to interchange character coded data; conversion between the two codes may be involved.
4.26 low-half zone: a set of cells reserved for use in UTF-16 (see Annex C); an RC-element corresponding to any of these cells may be used as the second of a pair of RC-elements which represents a character from a plane other than the BMP.
4.27 octet: An ordered sequence of eight bits considered as a unit.
4.28 plane: A subdivision of a group; of $256 \times 256$ cells
4.29 presentation; to present: The process of writing, printing, or displaying a graphic symbol.
4.30 presentation form: In the presentation of some scripts, a form of a graphic symbol representing a character that depends on the position of the character relative to other characters.
4.31 private use plane: A plane within this coded character set the contents of which is not specified in ISO/IEC 10646 (see clause 10)
4.33 RC-element: a two-octet sequence comprising the R-octet and the C-octet (see 6.2) from the four octet sequence that corresponds to a cell in the coding space of this coded character set.
4.33 repertoire: A specified set of characters that are represented in a coded character set.
4.34 row: A subdivision of a plane; of 256 cells.
4.35 script: A set of graphic characters used for the written form of one or more languages.
4.36 supplementary plane: A plane that accommodates characters which have not been allocated to the Basic Multilingual Plane.
4.37 unpaired RC-element: An RC-element in a CC-data element that is either:

- an RC-element from the high-half zone that is not immediately followed by an RC-element from the low-half zone, or
- an RC-element from the low-half zone that is not immediately preceded by a high-half RC-element from the high-half zone.
4.38 user: A person or other entity that invokes the service provided by a device. (This entity may be a process such as an application program if the
"device" is a code converter or a gateway function, for example.)
4.39 zone: A sequence of cells of a code table, comprising one or more rows, either in whole or in part, containing characters of a particular class (see clause 8).


## 5 General structure of the UCS

The general structure of the Universal Multiple-Octet Coded Character Set (referred to hereafter as "this coded character set") is described in this explanatory clause, and is illustrated in figures 1 and 2. The normative specification of the structure is given in the following clauses.
The value of any octet is expressed in hexadecimal notation from 00 to FF in ISO/IEC 10646 (see annex K).

The canonical form of this coded character set the way in which it is to be conceived - uses a fourdimensional coding space, regarded as a single entity, consisting of 128 three-dimensional groups.

NOTE - Thus, bit 8 of the most significant octet in the canonical form of a coded character can be used for internal processing purposes within a device as long as it is set to zero within a conforming CC-data-element.
Each group consists of 256 two-dimensional planes. Each plane consists of 256 one-dimensional rows, each row containing 256 cells. A character is located and coded at a cell within this coding space or the cell is declared unused.

In the canonical form, four octets are used to represent each character, and they specify the group, plane, row and cell, respectively. The canonical form consists of four octets since two octets are not sufficient to cover all the characters in the world, and a 32-bit representation follows modern processor architectures.
The four-octet canonical form can be used as a fouroctet coded character set, in which case it is called UCS-4.

The first plane (Plane 00 of Group 00) is called the Basic Multilingual Plane. The Basic Multilingual Plane includes characters in general use in alphabetic, syllabic and ideographic scripts together with various symbols and digits.

The subsequent planes are regarded as supplementary or private use planes, which will accommodate additional graphic characters (see clause 9).
The planes that are reserved for private use are specified in clause 10. The contents of the cells in
private use zones are not specified in ISO/IEC 10646.

Each character is located within the coded character set in terms of its Group-octet, Plane-octet, Rowoctet, and Cell-octet.

In addition to the canonical form, a two-octet BMP form is specified. Thus, the Basic Multilingual Plane can be used as a two-octet coded character set identified as UCS-2.

Subsets of the coding space may be used in order to give a sub-repertoire of graphic characters.

A UCS Transformation Format (UTF-16) is specified in Annex C which can be used to represent characters from 16 planes of group 00, additional to the BMP, in a form that is compatible with the twooctet BMP form.

A UCS Transformation Format (UTF-8) is specified in Annex D which can be used to transmit text data through communication systems which are sensitive to octet values for control characters coded according to the 8-bit structure of ISO/IEC 2022, and to ISO/IEC 4873. UTF-8 also avoids the use of octet values according to ISO/IEC 4873 which have special significance during the parsing of file-name character strings in widely-used file-handling systems.

## 6 Basic structure and nomenclature

### 6.1 Structure

The Universal Multiple-Octet Coded Character Set as specified in ISO/IEC 10646 shall be regarded as a single entity.

This entire coded character set shall be conceived of as comprising 128 groups of 256 planes. Each plane shall be regarded as containing 256 rows of characters, each row containing 256 cells. In a code table representing the contents of a plane (such as in figure 2), the horizontal axis shall represent the least significant octet, with its smaller value to the left; and the vertical axis shall represent the more significant octet, with its smaller value at the top.
Each axis of the coding space shall be coded by one octet. Within each octet the most significant bit shall be bit 8 and the least significant bit shall be bit 1.

Accordingly, the weight allocated to each bit shall be

| bit 8 | bit 7 | bit 6 | bit 5 | bit 4 | bit 3 | bit 2 | bit 1 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 128 | 64 | 32 | 16 | 8 | 4 | 2 | 1 |



Figure 1 - Entire coding space of the Universal Multiple-Octet Coded Character Set


NOTE - Labels "S-zone" and "Private use zone" are specified in clause 8.
Figure 2 - Group 00 of the Universal Multiple-Octet Coded Character Set

### 6.2 Coding of characters

In the canonical form of the coded character set, each character within the entire coded character set shall be represented by a sequence of four octets. The most significant octet of this sequence shall be the group-octet. The least significant octet of this sequence shall be the cell-octet. Thus this sequence may be represented as
m.s.
l.s.

| Group-octet | Plane-octet | Row-octet | Cell-octet |
| :--- | :--- | :--- | :--- |

where m.s. means the most significant octet, and I.s. means the least significant octet.

For brevity, the octets may be termed
m.s. I.s.

| G-octet | P-octet | R-octet | C-octet |
| :--- | :--- | :--- | :--- |

Where appropriate, these may be further abbreviated to $\mathrm{G}, \mathrm{P}, \mathrm{R}$, and C .

The value of any octet shall be represented by two hexadecimal digits, for example: 31 or FE. When a single character is to be identified in terms of the values of its group, plane, row, and cell, this shall be represented such as:

00000030 for DIGIT ZERO
00000041 for LATIN CAPITAL LETTER A
When referring to characters within an identified plane, the leading four digits (for G-octet and Poctet) may be omitted. For example, within plane 00 , 0030 may be used to refer to DIGIT ZERO.

### 6.3 Octet order

The sequence of the octets that represent a character, and the most significant and least significant ends of it, shall be maintained as shown above. When serialized as octets, a more significant octet shall precede less significant octets. When not serialized as octets, the order of octets may be specified by agreement between sender and recipient (see 16.1 and annex H ).

### 6.4 Naming of characters

ISO/IEC 10646 assigns a unique name to each character. The name of a character either:
a. denotes the customary meaning of the character, or
b. describes the shape of the corresponding graphic symbol, or
c. follows the rule given in clause 27 for Chinese/Japanese/Korean (CJK) unified ideographs.

Guidelines to be used for constructing the names of characters in cases $a$. and $b$. are given in annex $L$.

### 6.5 Identifiers for characters

ISO/IEC 10646 defines a short identifier for each character. The short identifier for any character is distinct from the short identifier for any other character. These short identifiers are independent of the language in which this standard is written, and are thus retained in all translations of the text.

The following alternative forms of notation of a short identifier are defined here.
a. The eight-digit form of short identifier shall consist of the sequence of eight hexadecimal digits that represents the code position of the character (see 6.2).
b. The four-digit form of short identifier shall consist of the last four digits of the eight-digit form. It is not defined if the first four digits of the eight-digit form are not all zeroes; that is, for characters allocated outside the Basic Multilingual Plane.
c. The character "-" (HYPHEN-MINUS) may, as an option, precede the 8 -digit form of short identifier.
d. The character "+" (PLUS SIGN) may, as an option, precede the 4 -digit form of short identifier.
e. The prefix letter "U" (LATIN CAPITAL LETTER U) may, as an option, precede any of the four forms of short identifier defined in a. to d. above.

The CAPITAL letters $A$ to $F$, and $U$ that appear within identifiers may be replaced by the corresponding SMALL letters.
The full syntax of the notation of a short identifier, in Backus-Naur form, is:

## \{ U | U \} [ \{ +$\} x x x x \mid\{-\} x x x x x x x x]$

where "x" represents one hexadecimal digit (0 to 9, A to $F$, or a to $f$ ), for example:
-hhhhhhhh +kkkk
Uhhhhhhhh U+kkkk
where hhhhhhhh indicates the eight-digit form and kkkk indicates the four-digit form.

NOTES
1 As an example the identifier for LATIN SMALL LETTER LONG S (see tables for Row 01 in clause 26) may be notated in any of the following forms:

0000017F -0000017F U0000017F U-0000017F
$017 \mathrm{~F}+017 \mathrm{~F}$ U017F U+017F
Any of the capital letters may be replaced by the corresponding small letter.

2 Two special prefixed forms of notation have also been used, in which the letter T (LATIN CAPITAL LETTER T or LATIN SMALL LETTER T) replaces the letter $U$ in the corresponding prefixed forms. The
forms of notation that included the prefix letter $T$ indicated that the identifier refers to a character in ISO/IEC 10646-1 First Edition (before the application of any Amendments), whereas the forms of notation that include the prefix letter $U$ always indicate that the identifier refers to a character in ISO/IEC 10646 at the most recent state of amendment. Corresponding identifiers of the form T -xxxxxxxx and U -xxxxxxxx refer to the same character except when xxxxxxxx lies in the range 00003400 to 00004DFF inclusive. Forms of notation that include no prefix letter always indicate a reference to the most recent state of amendment of ISO/IEC 10646, unless otherwise qualified.

## 7 General requirements for the UCS

The following requirements apply to the entire coded character set.
a) The values of P -, and R-, and C-octets used for representing graphic characters shall be in the range 00 to FF . The values of G-octets used for representation of graphic characters shall be in the range 00 to 7 F . On any plane, code positions FFFE and FFFF shall not be used.
NOTE - Code position FFFE is reserved for "signature" (see annex H). Code position FFFF can be used for internal processing uses requiring a numeric value that is guaranteed not to be a coded character such as in terminating tables, or signaling end-of-text. Since it is the largest two-octet value, it may also be used as the final value in binary or sequential searching index.
b) Code positions to which a character is not allocated, except for the positions reserved for private use characters or for transformation formats, are reserved for future standardization and shall not be used for any other purpose. Future editions of ISO/IEC 10646 will not allocate any characters to code positions reserved for private use characters or for transformation formats.
c) The same graphic character shall not be allocated to more than one code position. There are graphic characters with similar shapes in the coded character set; they are used for different purposes and have different character names.

## 8 The Basic Multilingual Plane

Plane 00 of Group 00 shall be the Basic Multilingual Plane (BMP). The BMP can be used as a two-octet coded character set in which case it shall be called UCS-2 (see 13.1).
Code positions 00000000 to 0000 001F in the BMP are reserved for control characters, and code position 0000 007F is reserved for the character DELETE (see clause 15). Code positions 00000080 to 0000009 F are reserved for control characters.

Code positions 0000 D800 to 0000 DFFF are reserved for the use of UTF-16 (see Annex C). These positions are known as the S-zone.
Code positions 0000 E000 to 0000 F8FF are reserved for private use (see clause 10). These positions are known as the private use zone.
Code postions FFFE and FFFF are reserved.

## 9 Other planes

### 9.1 Planes reserved for future standardization

Planes 11 to DF in Group 00 and Planes 00 to FF in Groups 01 to 5F are reserved for future standardization, and thus those code positions shall not be used for any other purpose.

### 9.2 Planes accessible by UTF-16

Each code position in Planes 01 to 10 of Group 00 has a unique mapping to a four-octet sequence in accordance with the UTF-16 form of coded representation (see Annex C). This form is compatible with the two-octet BMP form of UCS-2 (see 13.1).

Code positions in Planes 11 to FF of Group 00, or in Planes 00 to FF of other groups, do not have a mapping to the UTF-16 form.

## 10 Private use groups, planes, and zones

### 10.1 Private use characters

Private use characters are not restrained in any way by ISO/IEC 10646. Private use characters can be used to provide user-defined characters. For example, this is a common requirement for users of ideographic scripts.

NOTE 1 - For meaningful interchange of private use characters, an agreement, independent of ISO/IEC 10646, is necessary between sender and recipient.

Private use characters can be used for dynamicallyredefinable character applications.

NOTE 2 - For meaningful interchange of dynamicallyredefinable characters, an agreement, independent of ISO/IEC 10646 is necessary between sender and recipient. ISO/IEC 10646 does not specify the techniques for defining or setting up dynamically-redefinable characters.

### 10.2 Code positions for private use characters

The code positions of the 32 groups from Group 60 to Group 7F shall be for private use.

The code positions of Plane 0F and Plane 10, and of the 32 planes from Plane E0 to Plane FF, of Group 00 shall be for private use.
The 6400 code positions E000 to F8FF of the Basic Multilingual Plane shall be for private use.

The contents of these code positions are not specified in ISO/IEC 10646 (see 10.1).

## 11 Revision and updating of the UCS

The revision and updating of this coded character set will be carried out by ISO/IEC JTC1/SC2.

NOTE - It is intended that in future editions of ISO/IEC 10646, the names and allocation of the characters in this edition will remain unchanged.

## 12 Subsets

ISO/IEC 10646 provides the specification of subsets of coded graphic characters for use in interchange, by originating devices, and by receiving devices.
There are two alternatives for the specification of subsets: limited subset and selected subset. An adopted subset may comprise either of them, or a combination of the two.

### 12.1 Limited subset

A limited subset consists of a list of graphic characters in the specified subset. This specification allows applications and devices that were developed using other codes to interwork with this coded character set.
A claim of conformance referring to a limited subset shall list the graphic characters in the subset by the names of graphic characters or code positions as defined in ISO/IEC 10646.

### 12.2 Selected subset

A selected subset consists of a list of collections of graphic characters as defined in ISO/IEC 10646. The collections from which the selection may be made are listed in annex A of each part of ISO/IEC 10646. A selected subset shall always automatically include the Cells 20 to 7E of Row 00 of Plane 00 of Group 00.

A claim of conformance referring to a selected subset shall list the collections chosen as defined in ISO/IEC 10646.

## 13 Coded representation forms of the UCS

ISO/IEC 10646 provides two alternative forms of coded representation of characters.

NOTE - The characters from the ISO/IEC 646 IRV repertoire are coded by simple zero extensions to their coded representations in ISO/IEC 646 IRV. Therefore, their coded representations have the same integer values when represented as 8 -bit, 16 -bit, or 32 -bit integers. For implementations sensitive to a zero-valued octet (e.g. for use as a string terminator), use of 8 -bit based array data type should be avoided as any zero-valued octet may be
interpreted incorrectly. Use of data types at least 16-bits wide is more suitable for UCS-2, and use of data types at least 32-bits wide is more suitable for UCS-4.

### 13.1 Two-octet BMP form

This coded representation form permits the use of characters from the Basic Multilingual Plane with each character represented by two octets.
Within a CC-data-element conforming to the twooctet BMP form, a character from the Basic Multilingual Plane shall be represented by two octets comprising the R -octet and the C -octet as specified in 6.2 (i.e. its RC-element).

NOTE - A coded graphic character using the two-octet BMP form may be implemented by a 16 -bit integer for processing.

### 13.2 Four-octet canonical form

The canonical form permits the use of all the characters of ISO/IEC 10646, with each character represented by four octets.

Within a CC-data-element conforming to the fouroctet canonical form, every character shall be represented by four octets comprising the G-octet, the P-octet, the R-octet, and the C-octet as specified in 6.2.

NOTE - A coded graphic character using the four-octet canonical form may be implemented by a 32-bit integer for processing.

## 14 Implementation levels

ISO/IEC 10646 specifies three levels of implementation. Combining characters are described in 24 and listed in annex B.

### 14.1 Implementation level 1

When implementation level 1 is used, a CC-dataelement shall not contain coded representations of combining characters (see clause B.1) nor of characters from HANGUL JAMO block (see clause 25). When implementation level 1 is used the unique-spelling rule shall apply (25.2).

### 14.2 Implementation level 2

When implementation level 2 is used, a CC-dataelement shall not contain coded representations of characters listed in clause B.2. When implementation level 2 is used the unique-spelling rule shall apply (25.2).

### 14.3 Implementation level 3

When implementation level 3 is used, a CC-dataelement may contain coded representations of any characters.

## 15 Use of control functions with the UCS

This coded character set provides for use of control functions encoded according to ISO/IEC 6429 or similarly structured standards for control functions, and standards derived from these. A set or subset of such coded control functions may be used in conjunction with this coded character set. These standards encode a control function as a sequence of one or more octets.

When a control character of ISO/IEC 6429 is used with this coded character set, its coded representation as specified in ISO/IEC 6429 shall be padded to correspond with the number of octets in the adopted form (see clause 13). Thus, the least significant octet shall be the bit combination specified in ISO/IEC 6429, and the more significant octet(s) shall be zeros.

For example, the control character FORM FEED is represented by "000C" in the two-octet form, and "0000 000C" in the four-octet form.
For escape sequences, control sequences, and control strings (see ISO/IEC 6429) consisting of a coded control character followed by additional bit combinations in the range 20 to 7 F , each bit combination shall be padded by octet(s) with value 00.

For example, the escape sequence "ESC 02/00 $04 / 00$ " is represented by "001B 0020 0040" in the two-octet form, and "0000 001B 000000200000 0040 " in the four-octet form.

NOTE - The term "character" appears in the definition of many of the control functions specified in ISO/IEC 6429, to identify the elements on which the control functions will act. When such control functions are applied to coded characters according to ISO/IEC 10646 the action of those control functions will depend on the type of element from ISO/IEC 10646 that has been chosen, by the application, to be the element (or character) on which the control functions act. These elements may be chosen to be characters (noncombining characters and/or combining characters) or may be chosen in other ways (such as composite sequences) when applicable.
Code extension control functions for the ISO/IEC 2022 code extension techniques (such as designation escape sequence, single shift, and locking shift) shall not be used with this coded character set.

## 16 Declaration of identification of features

### 16.1 Purpose and context of identification

CC-data-elements conforming to ISO/IEC 10646 are intended to form all or part of a composite unit of coded information that is interchanged between an originator and a recipient. The identification of

ISO/IEC 10646 (including the form), the implementation level, and any subset of the coding space that have been adopted by the originator must also be available to the recipient. The route by which such identification is communicated to the recipient is outside the scope of ISO/IEC 10646.

However, some standards for interchange of coded information may permit, or require, that the coded representation of the identification applicable to the CC-data-element forms a part of the interchanged information. This clause specifies a coded representation for the identification of UCS with an implementation level and a subset of ISO/IEC 10646, and also of a C0 and a C1 set of control functions from ISO/IEC 6429 for use in conjunction with ISO/IEC 10646. Such coded representations provide all or part of an identification data element, which may be included in information interchange in accordance with the relevant standard.

If two or more of the identifications are present, the order of those identifications shall follow the order as specified in this clause.

NOTE - An alternative method of identification is described in annex N .

### 16.2 Identification of UCS coded representation form with implementation level

When the escape sequences from ISO/IEC 2022 are used, the identification of a coded representation form of UCS (see clause 13) and an implementation level (see clause 14) specified by ISO/IEC 10646 shall be by a designation sequence chosen from the following list:

ESC 02/05 02/15 04/00
UCS-2 with implementation level 1
ESC 02/05 02/15 04/01
UCS-4 with implementation level 1
ESC 02/05 02/15 04/03
UCS-2 with implementation level 2
ESC 02/05 02/15 04/04
UCS-4 with implementation level 2
ESC 02/05 02/15 04/05
UCS-2 with implementation level 3
ESC 02/05 02/15 04/06
UCS-4 with implementation level 3
If such an escape sequence appears within a CC-data-element conforming to ISO/IEC 2022, it shall consist only of the sequences of bit combinations as shown above.

If such an escape sequence appears within a CC-data-element conforming to ISO/IEC 10646, it shall be padded in accordance with clause 15 .

### 16.3 Identification of subsets of graphic characters

When the control sequences of ISO/IEC 6429 are used, the identification of subsets (see clause 12) specified by ISO/IEC 10646 shall be by a control sequence IDENTIFY UNIVERSAL CHARACTER SUBSET (IUCS) as shown below.

CSI Ps... 02/00 06/13
Ps... means that there can be any number of selective parameters. The parameters are to be taken from the subset collection numbers as shown in annex A of each part of ISO/IEC 10646. When there is more than one parameter, each parameter value is separated by an octet with value 03/11.

Parameter values are represented by digits where octet values $03 / 00$ to 03/09 represent digits 0 to 9 .

If such an escape sequence appears within a CC-data-element conforming to ISO/IEC 2022, it shall consist only of the sequences of bit combinations as shown above.

If such a control sequence appears within a CC-data-element conforming to ISO/IEC 10646, it shall be padded in accordance with clause 15.

### 16.4 Identification of control function set

When the escape sequences from ISO/IEC 2022 are used, the identification of each set of control functions (see clause 15) of ISO/IEC 6429 to be used in conjunction with ISO/IEC 10646 shall be an identifier sequence of the type shown below.

ESC 02/01 04/00 identifies the full C0 set of ISO/IEC 6429

ESC 02/02 04/03 identifies the full C1 set of ISO/IEC 6429

For a subset of C 0 or C 1 sets, the final octet F shall be obtained from the International Register of Coded Character Sets. The identifier sequences for these sets shall be:

ESC 02/01 F identifies a C0 set
ESC 02/02 F identifies a C1 set
If such an escape sequence appears within a CC-data-element conforming to ISO/IEC 2022, it shall consist only of the sequences of bit combinations as shown above.

If such an escape sequence appears within a CC-data-element conforming to ISO/IEC 10646, it shall be padded in accordance with clause 15.

### 16.5 Identification of the coding system of ISO/IEC 2022

When the escape sequences from ISO/IEC 2022 are used, the identification of a return, or transfer, from

UCS to the coding system of ISO/IEC 2022 shall be by the escape sequence ESC 02/05 04/00. If such an escape sequence appears within a CC-dataelement conforming to ISO/IEC 10646, it shall be padded in accordance with clause 15.

If such an escape sequence appears within a CC-data-element conforming to ISO/IEC 2022, it shall consist only of the sequences of bit combinations as shown above.

> NOTE - Escape sequence ESC $02 / 0504 / 00$ is normally used for return to the restored state of ISO/IEC 2022 . The escape sequence ESC $02 / 0504 / 00$ specified here is sometimes not exactly as specified in ISO/IEC 2022 due to the presence of padding octets. For this reason the escape sequences in 16.2 for the identification of UCS include the octet $02 / 15$ to indicate that the return does not always conform to that standard.

## 17 Structure of the code tables and lists

The clauses 26 and 27 set out the detailed code tables and the lists of character names for the graphic characters. Together, these specify graphic characters, their coded representation, and the character name for each character.

The graphic symbols are to be regarded as typical visual representations of the characters. ISO/IEC 10646 does not attempt to prescribe the exact shape of each character. The shape is affected by the design of the font employed, which is outside the scope of ISO/IEC 10646.

Graphic characters specified in ISO/IEC 10646 are uniquely identified by their names. This does not imply that the graphic symbols by which they are commonly imaged are always different. Examples of graphic characters with similar graphic symbols are LATIN CAPITAL LETTER A, GREEK CAPITAL LETTER ALPHA, and CYRILLIC CAPITAL LETTER A.

The meaning attributed to any character is not specified by ISO/IEC 10646; it may differ from country to country, or from one application to another.

For the alphabetic scripts, the general principle has been to arrange the characters within any row in approximate alphabetic sequence; where the script has capital and small letters, these are arranged in pairs. However, this general principle has been overridden in some cases. For example, for those scripts for which a relevant standard exists, the characters are allocated according to that standard. This arrangement within the code tables will aid conversion between the existing standards and this coded character set. In general, however, it is anticipated that conversion between this coded
character set and any other coded character set will use a table lookup technique.
It is not intended, nor will it often be the case, that the characters needed by any one user will be found all grouped together in one part of the code table.
Furthermore, the user of any script will find that needed characters may have been coded elsewhere in this coded character set. This especially applies to the digits, to the symbols, and to the use of Latin letters in dual-script applications. Therefore, in using this coded character set, the reader is advised to refer first to the block names list in annex A. 2 or an overview of the BMP in figures 3 and 4 , and then to turn to the specific code table rows for the relevant script and for symbols and digits. In addition, annex $G$ contains an alphabetically sorted list of character names.

## 18 Block names

Named blocks of contiguous code positions are specified within a plane for the purpose of allocation of characters sharing some common characteristic, such as script. The blocks specified within the BMP are listed in A. 2 of Annex A, and are illustrated in Figures 3 and 4.

## 19 Characters in bi-directional context

A class of left/right handed pairs of characters have special significance in the context of bi-directional text. In this context the terms LEFT or RIGHT in the character name are also intended to imply "opening" or "closing" forms of character shape, rather than a strict left-hand or right-hand form. These characters are listed below.

| Code | Name |
| :---: | :---: |
| Position |  |
| 0028 | LEFT PARENTHESIS |
| 0029 | RIGHT PARENTHESIS |
| 005B | LEFT SQUARE BRACKET |
| 005D | RIGHT SQUARE BRACKET |
| 007B | LEFT CURLY BRACKET |
| 007D | RIGHT CURLY BRACKET |
| 2045 | LEFT SQUARE BRACKET WITH QUILL |
| 2046 | RIGHT SQUARE BRACKET WITH QUILL |
| 207D | SUPERSCRIPT LEFT PARENTHESIS |
| 207E | SUPERSCRIPT RIGHT PARENTHESIS |
| 208D | SUBSCRIPT LEFT PARENTHESIS |
| 208E | SUBSCRIPT RIGHT PARENTHESIS |
| 2329 | LEFT-POINTING ANGLE BRACKET |
| 232A | RIGHT-POINTING ANGLE BRACKET |
| 3008 | LEFT ANGLE BRACKET |
| 3009 | RIGHT ANGLE BRACKET |
| 300A | LEFT DOUBLE ANGLE BRACKET |
| 300B | RIGHT DOUBLE ANGLE BRACKET |
| 300C | LEFT CORNER BRACKET |
| 300D | RIGHT CORNER BRACKET |
| 300E | LEFT WHITE CORNER BRACKET |

3014
3015
3016
3017
3018
3019
301A
301B

## 300F RIGHT WHITE CORNER BRACKET <br> 3010 LEFT BLACK LENTICULAR BRACKET <br> 3011 RIGHT BLACK LENTICULAR BRACKET <br> LEFT TORTOISE SHELL BRACKET RIGHT TORTOISE SHELL BRACKET LEFT WHITE LENTICULAR BRACKET RIGHT WHITE LENTICULAR BRACKET LEFT WHITE TORTOISE SHELL BRACKET RIGHT WHITE TORTOISE SHELL BRACKET LEFT WHITE SQUARE BRACKET RIGHT WHITE SQUARE BRACKET

The interpretation and rendering of any of these characters depend on the state related to the symmetric swapping characters (see F.2.2) and on the direction of the character being rendered that are in effect at the point in the CC-data-element where the coded representation of the character appears.
For example, if the character ACTIVATE SYMMETRIC SWAPPING occurs and if the direction of the character is from right to left, the character shall be interpreted as if the term LEFT or RIGHT in its name had been replaced by the term RIGHT or LEFT, respectively.

NOTE - In the context of Arabic bi-directional text, certain mathematical symbols may also have special significance (see annex E).

## 20 Special characters

There are some characters that do not have printable graphic symbols. These characters include space characters. They are

| Code | Name |
| :---: | :---: |
| Position |  |
| 0020 | SPACE |
| 00A0 | NO-BREAK SPACE |
| 2000 | EN QUAD |
| 2001 | EM QUAD |
| 2002 | EN SPACE |
| 2003 | EM SPACE |
| 2004 | THREE-PER-EM SPACE |
| 2005 | FOUR-PER-EM SPACE |
| 2006 | SIX-PER-EM SPACE |
| 2007 | FIGURE SPACE |
| 2008 | PUNCTUATION SPACE |
| 2009 | THIN SPACE |
| 200A | HAIR SPACE |
| 3000 | IDEOGRAPHIC SPACE |

Currency symbols in ISO/IEC 10646 do not necessarily identify the currency of a country. For example, YEN SIGN can be used for Japanese yen and Chinese yuan. Also, DOLLAR SIGN is used in numerous countries including the United States of America.

There is a special class of characters called Alternate Format Characters which are included for compatibility with some industry practices. These are described in annex $F$.

## 21 Presentation forms of characters

Each presentation form of a character provides an alternative form, for use in a particular context, to the nominal form of the character or sequence of characters from the other zones of graphic characters. The transformation from the nominal form to the presentation forms may involve substitution, superimposition, or combination.
The rules for the superimposition, choice of differently shaped characters, or combination into ligatures, or conjuncts which are often of extreme complexity are not specified in ISO/IEC 10646.

In general, presentation forms are not intended to be used as a substitute for the nominal forms of the graphic characters specified elsewhere within this coded character set. However, specific applications may encode these presentation forms instead of the nominal forms for specific reasons among which is compatibility with existing devices. The rules for searching, sorting, and other processing operations on presentation forms are outside the scope of ISO/IEC 10646.

Within the BMP these characters are mostly allocated to positions in rows FB to FF.

## 22 Compatibility characters

Compatibility characters are included in ISO/IEC 10646 primarily for compatibility with existing coded character sets to allow two-way code conversion without loss of information.

Within the BMP many of these characters are allocated to positions within rows F9, FA, FE, and FF, and within rows 31 and 33 . Some compatibility characters are also allocated within other rows.

## 23 Order of characters

Usually, coded characters appear in a CC-dataelement in logical order (logical or backing store order corresponds approximately to the order in which characters are entered from the keyboard, after corrections such as insertions, deletions, and overtyping have taken place). This applies even when characters of different dominant direction are mixed: left-to-right (Greek, Latin, Thai) with right-toleft (Arabic, Hebrew), or with vertical (Mongolian) script.
Some characters may not appear linearly in final rendered text. For example, the medial form of the short i in Devanagari is displayed before the character that it logically follows in the CC-dataelement.

## 24 Combining characters

This clause specifies the use of combining characters. A list of combining characters is shown in clause B.1. A list of combining characters not allowed in implementation level 2 is shown in clause B.2.

NOTE - The names of many script-independent combining characters contain the word "COMBINING".

### 24.1 Order of combining characters

Coded representations of combining characters shall follow that of the graphic character with which they are associated (for example, coded representations of LATIN SMALL LETTER A followed by COMBINING TILDE represent a composite sequence for Latin "ã").
If a combining character is to be regarded as a composite sequence in its own right, it shall be coded as a composite sequence by association with the character SPACE. For example, grave accent can be composed as SPACE followed by COMBINING GRAVE ACCENT.


#### Abstract

NOTE - Indic matras form a special category of combining characters, since the presentation can depend on more than one of the surrounding characters. Thus it might not be desirable to associate Indic matra with the character SPACE.


### 24.2 Appearance in code tables

Combining characters intended to be positioned relative to the associated character are depicted within the character code tables above, below, to the right of, to the left of, in, around, or through a dotted circle. In presentation, these characters are intended to be positioned relative to the preceding base character in some manner, and not to stand alone or function as base characters. This is the motivation for the term "combining". Diacritics are the principal class of combining characters used in European alphabets.

In the code tables for some scripts, such as Hebrew, Arabic, and the scripts of India and South East Asia, combining characters are indicated in relation to dotted circles to show their position relative to the base character. Many of these combining characters encode vowel letters; as such they are not generally referred to as "diacritical marks".

### 24.3 Multiple combining characters

There are instances where more than one combining character is applied to a single graphic character. ISO/IEC 10646 does not restrict the number of combining characters that can follow a base character. The following rules shall apply:
a) If the combining characters can interact in presentation (for example, COMBINING MACRON and COMBINING DIAERESIS), then the position of
the combining characters in the resulting graphic display is determined by the order of the coded representation of the combining characters. The presentations of combining characters are to be positioned from the base character outward. For example, combining characters placed above a base character are stacked vertically, starting with the first encountered in the sequence of coded representations and continuing for as many marks above as are required by the coded combining characters following the coded base character. For combining characters placed below a base character, the situation is inverted, with the combining characters starting from the base character and stacking downward.
An example of multiple combining characters above the base character is found in Thai, where a consonant letter can have above it one of the vowels 00000 E 34 to 00000 E 37 and, above that, one of four tone marks 00000 E 48 to 0000 0E4B. The order of the coded representation is: base consonant, followed by a vowel, followed by a tone mark.
b) Some specific combining characters override the default stacking behaviour by being positioned horizontally rather than stacking, or by forming a ligature with an adjacent combining character. When positioned horizontally, the order of coded representations is reflected by positioning in the dominant order of the script with which they are used. For example, horizontal accents in a left-toright script are coded left-to-right. Prominent characters that show such override behaviour are associated with specific scripts or alphabets. For example, the COMBINING GREEK KORONIS (0000 0343) requires that, together with a following acute or grave accent, they be rendered side-by-side above a letter, rather than the accent marks being stacked above the COMBINING GREEK KORONIS. The order of the coded representations is: the letter itself, followed by that of the breathing mark, followed by that of the accent marks. Two Vietnamese tone marks which have the same graphic appearance as the Latin acute and grave accent marks do not stack above the three Vietnamese vowel letters which already contain the circumflex diacritic (â, ê, ô). Instead, they form ligatures with the circumflex component of the vowel letters.
c) If the combining characters do not interact in presentation (for example, when one combining character is above a graphic character and another is below), the resultant graphic symbol from the base character and combining characters in different orders may appear the same. For example, the coded representations of LATIN SMALL LETTER A, followed by COMBINING CARON, followed by

COMBINING OGONEK may result in the same graphic symbol as the coded representations of LATIN SMALL LETTER A, followed by COMBINING OGONEK, followed by COMBINING CARON.

Combining characters in Hebrew or Arabic scripts do not normally interact. Therefore, the sequence of their coded representations in a composite sequence does not affect its graphic symbol. The rules for forming the combined graphic symbol are beyond the scope of ISO/IEC 10646.


#### Abstract

NOTE - Where combining characters are used for the generation of composite sequences in implementation level 3 , this facility may be used to provide an alternative coded representation of text. For example, in implementation level 3 the French word "là" may be represented by the characters LATIN SMALL LETTER L followed by LATIN SMALL LETTER A WITH GRAVE, or may be represented by the characters LATIN SMALL LETTER L followed by LATIN SMALL LETTER A followed by COMBINING GRAVE ACCENT.


### 24.4 Collections containing combining characters

In some collections of characters listed in annex A, such as collections 14 (BASIC ARABIC) or 25 (THAI), both combining characters and noncombining characters are included.
When implementation level 1 or 2 is adopted, a CC-data-element shall not contain the coded representations of combining characters listed in annex B, even though the adopted subset may include them.
Other collections of characters listed in annex A comprise only combining characters, for example collection 7 (COMBINING DIACRITICAL MARKS). Such a collection shall not be included in the adopted subset when implementation level 1 is adopted.

## 25 Special features of individual scripts

### 25.1 Hangul syllable composition method

In rendering, a sequence of Hangul Jamo (from HANGUL JAMO block: 1100 to 11FF) are displayed as a series of syllable blocks. Jamo can be classified into three classes: Choseong (syllable-initial character), Jungseong (syllable-peak character), and Jongseong (syllable-final character). A complete syllable block is composed of a Choseong and a Jungseong, and optionally a Jongseong.
An incomplete syllable is a string of one or more characters which does not constitute a complete syllable (for example, a Choseong alone, a Jungseong alone, a Jongseong alone, or a Jungseong followed by a Jongseong). An incomplete syllable which starts with a Jungseong or a Jongseong must be preceded by a CHOSEONG

FILLER (0000 115F). An incomplete syllable composed of a Choseong alone must be followed by a JUNGSEONG FILLER (0000 1160).

The implementation level 3 shall be used for the Hangul syllable composition method.

## NOTES

1 Hangul Jamo are not combining characters.
2 When a combining character such as HANGUL SINGLE DOT TONE MARK (0000 302E) is intended to apply to a sequence of Hangul Jamo it should be placed at the end of the sequence, after the Hangul Jamo character which completes the syllable block.

### 25.2 Features of Indic alphabetic scripts

In the tables for Rows 09 to 0D and 0F, and for the MYANMAR block in Row 10, of the BMP (see 26) the graphic symbols shown for some characters appear to be formed as compounds of the graphic symbols for two other characters in the same table.

## Examples:

Row OB Tamil. The graphic symbol for ob94 TAMIL LETTER AU appears is if it is constructed from the graphic symbols for:

0B93 TAMIL LETTER OO and OBD7 TAMIL AU LENGTH MARK
Row OD Malayalam. The graphic symbol for OD4A MALAYALAM VOWEL SIGN o appears as if it is constructed from the graphic symbols for:
OD46 MALAYALAM VOWEL SIGN E and OD3E MALAYALAM VOWEL SIGN AA

In such cases a single coded character may appear to the user to be equivalent to the sequence of two coded characters whose graphic symbols, when combined, are visually similar to the graphic symbol of that single character, as in a composite sequence (4.14).

In Levels 1 and 2 a "unique-spelling" rule shall apply. When this rule applies, no coded character from a table for Rows 09 to OD or OF, or for the MYANMAR block in Row 10, shall be regarded as equivalent to a sequence of two or more other coded characters taken from the same table.

NOTE - In Levels 1 and 2, if such a sequence occurs in a CC-data-element it is always made available to the user as two distinct characters in accordance with their respective character names.

## 26 Code tables and lists of character names

### 26.1 General

An overview of the Basic Multilingual Plane is shown in figure 3. Detailed code tables and lists of character names for the Basic Multilingual Plane are
shown on the following pages and in applicable Amendments.

Guidelines to be used for constructing names of characters are given in annex $L$ for information. In some cases, a name of a character is followed by additional explanatory statements not part of the name. These statements are in parentheses and not in capital letters except for the initials of the word, where required.

### 26.2 Character names and annotations for Hangul syllables

Names for the Hangul syllable characters in code positions (hex) 0000 AC00-0000 D7A3 are derived from their code position numbers by the numerical procedure described below. Lists of names for these characters are not provided.

1. Obtain the code position number of the Hangul syllable character. It is of the form $0000 h_{1} h_{2} h_{3} h_{4}$ where $h_{1}, h_{2}, h_{3}$, and $h_{4}$ are hexadecimal digits; $h_{1} h_{2}$ is the Row number within the BMP and $h_{3} h_{4}$ is the cell number within the row. The number $h_{1} h_{2} h_{3} h_{4}$ lies within the range AC00 to D7A3.
2. Derive the decimal numbers $d_{1}, d_{2}, d_{3}, d_{4}$ that are numerically equal to the hexadecimal digits $h_{1}, h_{2}, h_{3}$, $h_{4}$ respectively.
3. Calculate the character index $C$ from the formula:

$$
\begin{aligned}
C= & 4096 \times\left(d_{1}-10\right)+256 \times\left(d_{2}-12\right) \\
& +16 \times d_{3}+d_{4}
\end{aligned}
$$

Note: If $C<0$ or $>11,171$ then the character is not in the HANGUL SYLLABLES block.
4. Calculate the syllable component indices $I, P, F$ from the following formulae:

$$
\begin{array}{ll}
I=C / 588 & (\text { Note: } 0 \leq I \leq 18) \\
P=(C \% 588) / 28 & (\text { Note: } 0 \leq P \leq 20) \\
F=C \% 28 & (\text { Note: } 0 \leq F \leq 27)
\end{array}
$$

where "/" indicates integer division (i.e. $x / y$ is the integer quotient of the division), and "\%" indicates the modulo operation (i.e. $x \% y$ is the remainder after the integer division $x / y$ ).
5. Obtain the Latin character strings that correspond to the three indices $I, P, F$ from columns 2, 3, and 4 respectively of Table 1 below (for $I=11$ and for $F=0$ the corresponding strings are null). Concatenate these three strings in left-to-right order to make a single string, the syllable-name.
6. The character name for the character at position $0000 h_{1} h_{2} h_{3} h_{4}$ is then:
HANGUL SYLLABLE s-n
where "s-n" indicates the syllable-name string derived in step 5.
Example.
For the character in code position D4DE:
$d_{1}=13, d_{2}=4, d_{3}=13, d_{4}=14$.
$C=10462$
$I=17, \quad P=16, \quad F=18$.
The corresponding Latin character strings are:
P , WI, BS.
The syllable-name is PWIBS, and the character name is:

HANGUL SYLLABLE PWIBS
Annotations for the Hangul syllable characters in code positions (hex) 0000 AC00 - 0000 D7A3 are also derived from their code position numbers by a similar numerical procedure described below.
8. Obtain the Latin character strings that correspond to the three indices $I, P, F$ from columns 5, 6, and 7 respectively of Table 1 below (for $I=11$ and for $F=0$ the corresponding strings are null). Concatenate these three strings in left-to-right order to make a single string, and enclose it within parentheses to form the annotation.
Example.
For the character in code position D4DE:
$d_{1}=13, d_{2}=4, d_{3}=13, d_{4}=14$.
$C=10462$
$I=17, \quad P=16, \quad F=18$.
The corresponding Latin character strings are: ph, wi, ps,
and the annotation is (phwips).
7. Carry out steps 1 to 4 as described above.

Table 1: Elements of Hangul syllable names and annotations

| Index number | Syllable name elements |  |  | Annotation elements |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\begin{aligned} & \hline I \\ & \text { string } \\ & \hline \end{aligned}$ | $\begin{aligned} & \hline P \\ & \text { string } \end{aligned}$ | F <br> string | I string | P string | F string |
| 0 | G | A |  | k | a |  |
| 1 | GG | AE | G | kk | ae | k |
| 2 | N | YA | GG | n | ya | kk |
| 3 | D | YAE | GS | t | yae | ks |
| 4 | DD | EO | N | tt | eo | n |
| 5 | R | E | NJ | r | e | nc |
| 6 | M | YEO | NH | m | yeo | nh |
| 7 | B | YE | D | p | ye | , |
| 8 | BB | O | L | pp | 0 | I |
| 9 | S | WA | LG | s | wa | Ik |
| 10 | SS | WAE | LM | SS | wae | Im |
| 11 |  | OE | LB |  | Oe | Ip |
| 12 | J | YO | LS | C | yo | Is |
| 13 | JJ | U | LT | CC | u | Ith |
| 14 | C | WEO | LP | ch | weo | Iph |
| 15 | K | WE | LH | kh | we | Ih |
| 16 | T | WI | M | th | wi | m |
| 17 | P | YU | B | ph | yu | p |
| 18 | H | EU | BS | h | eu | ps |
| 19 |  | YI | S |  | yi | S |
| 20 |  | I | SS |  | i | SS |
| 21 |  |  | NG |  |  | ng |
| 22 |  |  | J |  |  | c |
| 23 |  |  | C |  |  | ch |
| 24 |  |  | K |  |  | kh |
| 25 |  |  | T |  |  | th |
| 26 |  |  | P |  |  | ph |
| 27 |  |  | H |  |  | h |


| Rows 00 to 33 (see Figure 4) |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
| CJK Unified Ideographs Extension A |  |  |  |  |
| CJK Unified Ideographs |  |  |  |  |
| Yi Syllables |  |  |  |  |
| Yi Radicals |  |  |  |  |
|  |  |  |  |  |
| Hangul Syllables |  |  |  |  |
| S-zone (for use in UTF-16 only) |  |  |  |  |
| Private Use Zone |  |  |  |  |
| CJK Compatibility Ideographs |  |  |  |  |
| Alphabetic Presentation Forms |  |  |  |  |
| Arabic Presentation Forms-A |  |  |  |  |
| Comb. Half M'ks | CJK Compat. F'ms | Small Form Vars. | Arabic Prese |  |
| Halfwidth And Fullwidth Forms |  |  |  | Special |
| = not graphic characters |  |  | = reserved for future standardization |  |

Figure 3-Overview of the Basic Multilingual Plane


Figure 4 - Overview of Rows 00 to 33 of the Basic Multilingual Plane
[This page is left intentionally blank]

Tables of character graphic symbols and character names for Rows 00 to 33, A0 to D7, and F9 to FF will appear on the following pages in the Final Text. (estimate: 280 pages)

## 27 CJK unified ideographs

Detailed code tables for:

- CJK (Chinese / Japanese / Korean) Unified Ideographs Extension A (starting at code position 3400), and
- CJK Unified Ideographs (starting at code position 4E00),
are shown on the following pages.
Entries in the code tables for both CJK Unified Ideographs and its Extension A are arranged as follows.

| Row/Cell Hex code |  | -T | $\begin{gathered} \mathrm{J} \\ \text { Kanji } \end{gathered}$ | $\begin{gathered} \text { K } \\ \text { Hanja } \end{gathered}$ |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 078/000 |  |  |  |  |  |
| 4E00 | 0-523B | 1-4421 | 0-306C | 0-6C69 | 1-2121 |
|  | 0-5027 | 1-3601 | 0-1676 | 0-7673 | 1-0101 |

NOTE - Under each ideograph the two lines of numbers indicate the source code positions; the first line shows hexdecimal values, the second line shows decimal values.
The leftmost column of an entry shows the code position in ISO/IEC 10646, giving the code representation both in decimal and in hexadecimal notation.

Each of the other columns shows the graphic symbol for the character, and its coded representation, as specified in a source standard for character sets that is also identified in the table entry. Each of these source standards is assigned to one of five groups indicated by $\mathrm{G}, \mathrm{T}, \mathrm{J}, \mathrm{K}$, or V as shown in the lists below. In each table entry, a separate column is assigned for the corresponding character (if any) from each of those groups of source standards.
An entry in any of the G, T, J, K, or V columns includes a sample graphic symbol from the source character set standard, together with its coded representation in that standard. The first line below the graphic symbol shows the coded representation in hexadecimal notation. The second line shows the coded representation in decimal notation which comprises two digits for section number followed by two digits for position number. Each of the coded representations is prefixed by a one-character source code identification followed by a hyphen. This source code character identifies the coded character set standard from which the character is taken as shown in the lists below.

Hanzi G sources are
G0 GB2312-80
G1 GB12345-90 with 58 Hong Kong and 92 Korean "Idu" characters
G3 GB7589-87 unsimplified forms
G5 GB7590-87 unsimplified forms
G7 General Purpose Hanzi List for Modern Chinese Language, and General List of Simplified Hanzi
GS Singapore Characters
G8 GB8565-88
GE GB16500-95
Hanzi T sources are
T1 TCA-CNS 11643-1992 1st plane
T2 TCA-CNS 11643-1992 2nd plane
T3 TCA-CNS 11643-1992 3rd plane with some additional characters
T4 TCA-CNS 11643-1992 4th plane
T5 TCA-CNS 11643-1992 5th plane
T6 TCA-CNS 11643-1992 6th plane
T7 TCA-CNS 11643-1992 7th plane
TF TCA-CNS 11643-1992 15th plane
Kanji J sources are
JO JIS X 0208-1990
J1 JIS X 0212-1990
JA Unified Japanese IT Vendors Contemporary Ideographs, 1993
Hanja K sources are
K0 KS C 5601-1987
K1 KS C 5657-1991
K2 PKS C 5700-1 1994
K3 PKS C 5700-2 1994
ChuNom V sources are
Vo TCVN 5773:1993
V1 TCVN 6056:1995
For CJK (Chinese/Japanese/Korean) Ideographs in the BMP, the names shall be algorithmically constructed by appending their two-octet coded representation in hexadecimal notation to "CJK UNIFIED IDEOGRAPH-". For example, the first CJK ideograph character in the BMP has the name "CJK UNIFIED IDEOGRAPH-3400".

Tables of character graphic symbols for Rows 34 to 9F will appear on this and following pages in the Final Text. (estimate 574 pages)

## Collections of graphic characters for subsets

## A. 1 Collections of coded graphic characters

The collections listed below are ordered by collection number. An * in the "positions" column indicates that the collection is a fixed collection.

See Note 2 for an alphabetically-ordered index of the principal terms used in the names of these collections.

NOTE 1 - Use of implementation levels 1 and 2 restricts the repertoire of some character collections (see 24.4). Collections which include combining characters are 7, 10, 13 to $26,35,49,50,63,65,72,84,85,86,87,88,89,90$, and 91.

| Collection number and name | Positions |
| :---: | :---: |
| 1 BASIC LATIN | 0020-007E * |
| $2{ }_{*}^{\text {LATIN-1 }}$ SUPPLEMENT | 00A0-00FF |
| LATIN EXTENDED-A | 0100-017F * |
| LATIN EXTENDED-B | 0180-024F |
| 5 IPA EXTENSIONS | 0250-02AF |
| 6 SPACING MODIFIER LETTERS | 02B0-02FF |
| 7 COMBINING DIACRITICAL MAR | $\begin{aligned} & \text { KKS } \\ & 0300-036 \mathrm{~F} \end{aligned}$ |
| 8 BASIC GREEK | 0370-03CF |
| 9 GREEK SYMBOLS AND COPTIC |  |
| 10 CYRILLIC | 0400-04FF |
| 11 ARMENIAN | 0530-058F |
| 12 BASIC HEBREW | 05D0-05EA |
| 13 HEBREW EXTENDED | $\begin{aligned} & 0590-05 \mathrm{CF} \\ & 05 \mathrm{~EB}-05 \mathrm{FF} \end{aligned}$ |
| 14 BASIC ARABIC | 0600-065F |
| 15 ARABIC EXTENDED | 0660-06FF |
| 16 DEVANAGARI | $\begin{aligned} & 0900-097 \mathrm{~F} \\ & 200 \mathrm{C}, 200 \mathrm{D} \end{aligned}$ |
| 17 BENGALI | $\begin{aligned} & \text { 0980-09FF } \\ & \text { 200C, 200D } \end{aligned}$ |
| 18 GURMUKHI | OAOO- OA7F <br> 200C, 200D |


| 19 | GUJARATI | $\begin{aligned} & \text { 0A80-0AFF } \\ & 200 \mathrm{C}, 200 \mathrm{D} \end{aligned}$ |
| :---: | :---: | :---: |
| 20 | ORIYA | $\begin{aligned} & \text { 0B00 - OB7F } \\ & \text { 200C, 200D } \end{aligned}$ |
| 21 | TAMIL | $\begin{aligned} & \text { 0B80 - OBFF } \\ & 200 \mathrm{C}, 200 \mathrm{D} \end{aligned}$ |
| 22 | TELUGU | $\begin{aligned} & 0 C 00-0 C 7 F \\ & 200 \mathrm{C}, 200 \mathrm{D} \end{aligned}$ |
| 23 | KANNADA | $\begin{aligned} & \text { 0C80-0CFF } \\ & 200 \mathrm{C}, 200 \mathrm{D} \end{aligned}$ |
| 24 | MALAYALAM | $\begin{aligned} & \text { 0D00 - OD7F } \\ & \text { 200C, 200D } \end{aligned}$ |
| 25 | THAI | 0E00-0E7F |
| 26 | LAO | 0E80-0EFF |
| 27 | BASIC GEORGIAN | 10D0-10FF |
| 28 | GEORGIAN EXTENDED | 10A0-10CF |
| 29 | HANGUL JAMO | 1100-11FF |
| 30 | LATIN EXTENDED ADDITIONAL | 1E00-1EFF |
| 31 | GREEK EXTENDED | 1F00-1FFF |
| 32 | GENERAL PUNCTUATION | 2000-206F |
| 33 SUPERSCRIPTS AND SUBSCRIPTS |  |  |
| ${ }_{3} 4$ | CURRENCY SYMBOLS | 20A0-20CF |
| 35 | COMBINING DIACRITICAL MARK SYMBOLS | $\begin{aligned} & \text { KS FOR } \\ & 20 D 0-20 F F \end{aligned}$ |
| 36 | LETTERLIKE SYMBOLS | 2100-214F |
| 37 | NUMBER FORMS | 2150-218F |
| 38 | ARROWS | 2190-21FF |
| 39 | MATHEMATICAL OPERATORS | 2200-22FF |
| 40 | MISCELLANEOUS TECHNICAL | 2300-23FF |
| 41 | CONTROL PICTURES | 2400-243F |
| 42 | OPTICAL CHARACTER RECOG | $\begin{aligned} & \text { NITION } \\ & 2440-245 F \end{aligned}$ |
| 43 | ENCLOSED ALPHANUMERICS | 2460-24FF |
| 44 | BOX DRAWING | 2500-257F |
| 45 | BLOCK ELEMENTS | 2580-259F |
| 46 | GEOMETRIC SHAPES | 25A0-25FF |


| 47 | MISCELLANEOUS SYMBOLS | 2600-26FF |
| :---: | :---: | :---: |
| 48 | DINGBATS | 2700-27BF |
| 49 | CJK SYMBOLS AND PUNCTUAT | $\begin{aligned} & \text { ION } \\ & 3000-303 F \end{aligned}$ |
| 50 | HIRAGANA | 3040-309F |
| 51 | KATAKANA | 30A0-30FF |
| 52 | BOPOMOFO | $\begin{aligned} & 3100-312 F \\ & 31 \mathrm{AO}-31 \mathrm{BF} \end{aligned}$ |
| 53 | HANGUL COMPATIBILITY JAMO | 3130 |
| 54 | CJK MISCELLANEOUS | 3190-319F |
| 55 | ENCLOSED CJK LETTERS AND | MONTHS $3200-32 F F$ |
| 56 | CJK COMPATIBILITY | 3300-33FF |
| 57 | [deleted at Amd.5] |  |
| 58 | [deleted at Amd.5] |  |
| 58 | [deleted at Amd.5] |  |
| 60 | CJK UNIFIED IDEOGRAPHS | 4E00-9FFF |
| 61 | PRIVATE USE AREA | E000-F8FF |
| 62 | CJK COMPATIBILITY IDEOGRAP | $\begin{aligned} & \text { FHS } \\ & \text { F900 - FAFF } \end{aligned}$ |
| 63 | ALPHABETIC PRESENTATION F | ORMS FB00 - FB4F |
| 64 | ARABIC PRESENTATION FORM | S-A FB50 - FDFF |
| 65 | COMBINING HALF MARKS | FE20-FE2F |
| 66 | CJK COMPATIBILITY FORMS | FE30-FE4F |
| 67 | SMALL FORM VARIANTS | FE50-FE6F |
| 68 | ARABIC PRESENTATION FORM | S-B FE70 - FEFE |
| 69 | HALFWIDTH AND FULLWIDTH | FORMS FF00-FFEF |
| 70 | SPECIALS | FFF0-FFFD |
| 71 | HANGUL SYLLABLES | AC00-D7A3 |
| 72 | BASIC TIBETAN | OFOO- OFBF |
| 73 | ETHIOPIC | 1200-137F |
| 74 | UNIFIED CANADIAN ABORIGINA SYLLABICS | $\begin{aligned} & \text { AL } \\ & 1400-167 F \end{aligned}$ |
| 75 | CHEROKEE | 13A0-13FF |
| 76 | YI SYLLABLES | A000-A48F |
| 77 | YI RADICALS | A490-A4CF |
| 78 | KANGXI RADICALS | 2F00-2FDF |
| 79 | CJK RADICALS SUPPLEMENT | 2E80-2EFF |


| COMBINING CHARACTERS B-2 characters specified in annex B. 2 |  |  |  |
| :---: | :---: | :---: | :---: |
| [299 | BMP FIRST EDITION] | see A. 3 | ] |
| 300 | BMP | $\begin{aligned} & 0000-D \\ & \text { E000 - } \end{aligned}$ |  |
| 301 | BMP-AMD. 7 | see A. 3 |  |
| 302 | BMP SECOND EDITION | see A. 3 |  |

The following collections are outside the Basic Multilingual Plane.

```
4 0 0 ~ P R I V A T E ~ U S E ~ P L A N E S ~ G = 0 0 , ~ P = 0 F , 1 0 , ~ \& ~
    E0 - FF
500 PRIVATE USE GROUPS G=60-7F
```

NOTE 2 - The principal terms (keywords) used in the collection names shown above are listed below in alphabetical order. The entry for a term shows the collection number of every collection whose name includes the term. These terms do not provide a complete cross-reference to all the collections where characters sharing a particular attribute, such as script name, may be found. Although most of the terms identify an attribute of the characters within the collection, some characters that possess that attribute may be present in other collections whose numbers do not appear in the entry for that term.

| Alphabetic | 63 |
| :--- | :--- |
| Alphanumeric | 43 |
| Arabic | 14156468 |
| Armenian | 11 |
| Arrows | 38 |
| Bengali | 17 |
| Bi-directional | 202203 |
| Block elements | 45 |
| BMP | 300301302 (299) |
| Box drawing | 44 |
| Bopomofo | 52 |
| Braille patterns | 80 |
| Canadian Aboriginal | 74 |
| Cherokee | 75 |
| CJK | 495455566062667881 |
| Combining | 73565270271 |
| Compatibility | 53566266 |
| Control pictures | 41 |
| Coptic | 9 |
| Currency | 34 |
| Cyrillic | 10 |
| Devanagari | 16 |
| Diacritical marks | 735 |
| Dingbats | 48 |
| Enclosed | 4355 |
| Ethiopic | 73 |
| Format | 201202203250251 |
| Fullwidth | 69 |
| Geometric shapes | 46 |
| Georgian | 2728 |
| Greek | 8931 |
| Gujarati | 19 |
| Gurmukhi | 18 |
| Half (marks, width) | 6569 |
| Hangul | 295371204 |
|  |  |
|  |  |
| Bal |  |


| Hebrew | 1213 |
| :--- | :--- |
| Hiragana | 50 |
| ldeographs | 606281207 |
| IPA extensions | 5 |
| Jamo | 2953 |
| Kangxi | 78 |
| Kannada | 23 |
| Katakana | 51 |
| Khmer | 88 |
| Lao | 26 |
| Latin | 123430 |
| Letter | 3655 |
| Malayalam | 24 |
| Mathematical operators | 39 |
| Mongolian | 89 |
| Months | 55 |
| Myanmar | 8790 |
| Number | 37 |
| Ogham | 82 |
| Optical character recognition 42 |  |
| Oriya | 20 |
| Presentation forms | 636468 |
| Private use | 61400500 |
| Punctuation | 3249 |
| Radicals | 777879 |
| Runic | 83 |
| Shape, shaping | 205206 |
| Sinhala | 84 |
| Small form | 67 |
| Spacing modifier | 6 |
| Specials | 70 |
| Subscripts, superscripts | 33 |
| Syllables, syllabics | 717476 |
| Symbols | 93435364749 |
| Syriac | 85 |
| Tamil | 21 |
| Technical | 40 |
| Telugu | 22 |
| Thaana | 86 |
| Thai | 25 |
| Tibetan | 7291 |
| Yi | 7677 |
| Zero-width | 200 |
|  |  |
|  |  |

## A. 2 Blocks in the BMP

The following blocks are specified in the Basic Multilingual Plane. They are ordered by code position.

| Block name | from to |
| :--- | ---: |
| BASIC LATIN | $0020-007 E$ |
| LATIN-1 SUPPLEMENT | $00 A 0-00 F F$ |
| LATIN EXTENDED-A | $0100-017 \mathrm{~F}$ |
| LATIN EXTENDED-B | $0180-024 \mathrm{~F}$ |
| IPA (INTERNATIONAL PHONETIC ALPHABET) |  |
| EXTENSIONS | $0250-02 A F$ |
| SPACING MODIFIER LETTERS | $02 B 0-02 F F$ |
| COMBINING DIACRITICAL MARKS | $0300-036 F$ |
| GREEK AND COPTIC | $0370-03 F F$ |
| CYRILLIC | $0400-04 F F$ |
| ARMENIAN | $0530-058 F$ |
| HEBREW | $0590-05 F F$ |
| ARABIC | $0600-06 F F$ |
| SYRIAC | $0700-074 F$ |
| THAANA | $0780-07 B F$ |


| DEVANAGARI | 0900-097F |
| :---: | :---: |
| BENGALI | 0980-09FF |
| GURMUKHI | 0A00-0A7F |
| GUJARATI | 0A80-0AFF |
| ORIYA | 0B00-0B7F |
| TAMIL | 0B80-0BFF |
| TELUGU | 0C00-0C7F |
| KANNADA | 0C80-0CFF |
| MALAYALAM | 0D00-0D7F |
| SINHALA | 0D80-0DFF |
| THAI | 0E00-0E7F |
| LAO | 0E80-0EFF |
| TIBETAN | 0F00-0FFF |
| MYANMAR | 1000-109F |
| GEORGIAN | 10A0-10FF |
| HANGUL JAMO | 1100-11FF |
| ETHIOPIC | 1200-137F |
| CHEROKEE | 13A0-13FF |
| UNIFIED CANADIAN ABORIGINAL SYLLABICS |  |
|  | 1400-167F |
| OGHAM | 1680-169F |
| RUNIC | 16A0-16FF |
| KHMER | 1780-17FF |
| MONGOLIAN | 1800-18AF |
| LATIN EXTENDED ADDITIONAL | 1E00-1EFF |
| GREEK EXTENDED | 1F00-1FFF |
| GENERAL PUNCTUATION | 2000-206F |
| SUPERSCRIPTS AND SUBSCRIPTS | 2070-209F |
| CURRENCY SYMBOLS | 20A0-20CF |
| COMBINING DIACRITICAL MARKS FOR SYMBOLS$20 \mathrm{DO}-20 \mathrm{FF}$ |  |
|  |  |
| LETTERLIKE SYMBOLS | 2100-214F |
| NUMBER FORMS | 2150-218F |
| ARROWS | 2190-21FF |
| MATHEMATICAL OPERATORS | 2200-22FF |
| MISCELLANEOUS TECHNICAL | 2300-23FF |
| CONTROL PICTURES | 2400-243F |
| OPTICAL CHARACTER RECOGNITION | 2440-245F |
| ENCLOSED ALPHANUMERICS | 2460-24FF |
| BOX DRAWING | 2500-257F |
| BLOCK ELEMENTS | 2580-259F |
| GEOMETRIC SHAPES | 25A0-25FF |
| MISCELLANEOUS SYMBOLS | 2600-26FF |
| DINGBATS | 2700-27BF |
| BRAILLE PATTERNS | 2800-28FF |
| CJK RADICALS SUPPLEMENT | 2E80-2EFF |
| KANGXI RADICALS | 2F00-2FDF |
| IDEOGRAPHIC DESCRIPTION CHARACTERS |  |
|  | 2FF0-2FFF |
| CJK SYMBOLS AND PUNCTUATION | 3000-303F |
| HIRAGANA | 3040-309F |
| KATAKANA | 30A0-30FF |
| BOPOMOFO | 3100-312F |
| HANGUL COMPATIBILITY JAMO | 3130-318F |
| KANBUN (CJK miscellaneous) | 3190-319F |
| BOPOMOFO EXTENDED | 31A0-31BF |
| ENCLOSED CJK LETTERS AND MONTHS |  |
|  | 3200-32FF |
| CJK COMPATIBILITY | 3300-33FF |
| CJK UNIFIED IDEOGRAPHS EXTENSION A |  |
|  | $3400-4 D B F$ |
| CJK UNIFIED IDEOGRAPHS | 4E00-9FFF |
| YI SYLLABLES | A000-A48F |
| YI RADICALS | A490-A4CF |
| HANGUL SYLLABLES | AC00-D7A3 |
| PRIVATE USE AREA | E000-F8FF |


| CJK COMPATIBILITY IDEOGRAPHS | F900 - FAFF |
| :--- | :--- |
| ALPHABETIC PRESENTATION FORMS | FB00 - FB4F |
| ARABIC PRESENTATION FORMS-A | FB50 - FDFF |
| COMBINING HALF MARKS | FE20 - FE2F |
| CJK COMPATIBILITY FORMS | FE30 - FE4F |
| SMALL FORM VARIANTS | FE50 - FE6F |
| ARABIC PRESENTATION FORMS-B | FE70 - FEFE |
| HALFWIDTH AND FULLWIDTH FORMS | FF00 - FFEF |
| SPECIALS | FFF0 - FFFD |

## A. 3 Fixed collections of the whole BMP

## A.3.1 301 BMP-AMD. 7

The collection 301 BMP-AMD. 7 is specified below as a fixed collection (4.19). It comprises only those coded characters that were in the BMP after amendments up to, but not after, AMD. 7 were appplied to this International Standard. Accordingly the repertoire of this collection is not subject to change if new characters are added to the BMP by any subsequent amendments.

NOTE - The repertoire of the collection 300 BMP is subject to change if new characters are added to the BMP by an amendment to this International Standard.

301 BMP-AMD. 7 is specified by the following ranges of code positions as indicated for each row or contiguous series of rows.

| Rows | Positions (cells) |
| :---: | :---: |
| 00 | 20-7E AO-FF |
| 01 | 00-F5 FA-FF |
| 02 | 00-17 50-A8 B0-DE E0-E9 |
| 03 | $00-45$ 60-61 74-75 7A 7E 84-8A 8C 8E-A1 A3-CE DO-D6 DA DC DE EO E2-F3 |
| 04 | $01-0 \mathrm{C} 0 \mathrm{E}-4 \mathrm{~F} 51-5 \mathrm{C} 5 \mathrm{E}-86$ 90-C4 C7-C8 CBCC DO-EB EE-F5 F8-F9 |
| 05 | $31-56$ D0-EA F0-F4 61-87 89 $81-A 1$ A3-B9 BB-C4 |
| 06 | OC 1B 1F 21-3A 40-52 60-6D 70-B7 BA-BE CO-CE DO-ED FO-F9 |
| 09 | 01-03 05-39 3C-4D 50-54 58-70 81-83 85-8C |
|  | 8F-90 93-A8 AA-B0 B2 B6-B9 BC BE-C4 C7-C8 CB-CD D7 DC-DD DF-E3 E6-FA |
| OA | 02 05-0A OF-10 13-28 2A-30 32-33 |
|  | 38-39 3C 3E-42 47-48 4B-4D 59-5C |
|  | 74 81-83 85-8B 8D 8F-91 93-A8 AA-B0 |
|  | B3 EF B5-B9 BC-C5 C7-C9 CB-CD DO E0 E6 |
| OB | 01-03 05-0C 0F-10 13-28 2A-30 32-33 36-39 |
|  | 3C-43 47-48 4B-4D 56-57 5C-5D 5F-61 |
|  | 70 82-83 85-8A 8E-90 92-25 99-9A 9C |
|  | 9F A3-A4 A8-AA AE-B5 B7-B9 BE-C2 C6-C CA-CD D7 E7-F2 |
| OC | 01-03 05-0C 0E-10 12-28 2A-33 35-39 3E-44 |
|  | 46-48 4A-4D 55-56 60-61 66-6F 82-83 85-8C |
|  | 8E-90 92-A8 AA-B3 B5-B9 BE-C4 C6-C8 CA |
|  | CD D5-D6 DE E0-E1 E6-EF |
| OD | 02-03 05-0C 0E-10 12-28 2A-39 3E-43 |
|  | 4A-4D 57 60-61 66-6F |
| OE | 01-3A 3F-5B 81-82 84 87-88 8A 8D 94-97 |
|  | $99-9 \mathrm{~F}$ A1-A3 A5 A7 AA-AB AD-B9 BB-B |
|  | C0-C4 C6 C8-CD D0-D9 DC-DD |


| OF | 00-47 49-69 71-8B 90-95 97 99-AD B1-B7 |
| :---: | :---: |
|  | B9 |
| 10 | A0-C5 D0-F6 FB |
| 11 | 00-59 5F-A2 A8-F9 |
| 1E | 00-9B A0-F9 |
| 1F | 00-15 18-1D 20-45 48-4D 50-57 59 5B 5D |
|  | 5F-7D 80-B4 B6-C4 C6-D3 D6-DB DD-EF |
|  | F2-F4 F6-FE |
| 20 | 00-2E 30-46 6A-70 74-8E A0-AB D0-E1 |
| 21 | 00-38 53-82 90-EA |
| 22 | 00-F1 |
| 23 | 00 02-7A |
| 24 | 00-24 40-4A 60-EA |
| 25 | 00-95 A0-EF |
| 26 | 00-13 1A-6F |
| 27 | 01-04 06-09 0C-27 29-4B 4D 4F-52 56 58- |
|  | 5E 61-67 76-94 98-AF B1-BE |
| 30 | 00-37 3F 41-94 99-9E A1-FE |
| 31 | 05-2C 31-8E 90-9F |
| 32 | 00-1C 20-43 60-7B 7F-B0 C0-CB D0-FE |
| 33 | 00-76 7B-DD E0-FE |
| 4E-9F | 4E00-9FA5 |
| AC-D7 | AC00-D7A3 |
| E0-F8 | E000-F8FF |
| F9-FA | F900-FA2D |
| FB | 00-06 13-17 $1 \mathrm{E}-36 \quad 38-3 \mathrm{C}$ 3E $\quad 40-41 \quad 43-44$ |
|  | 46-B1 D3-FF |
| FC | 00-FF |
| FD | 00-3F 50-8F 92-C7 F0-FB |
| FE | 20-23 30-44 49-52 54-66 68-6B 70-72 74 76- |
|  | FC FF |
| FF | 01-5E 61-BE C2-C7 CA-CF D2-D7 DA-DC |
|  | E0-E6 E8-EE FD |

## A.3.2 299 BMP FIRST EDITION

The collection number and collection name:

## 299 BMP FIRST EDITION

have been reserved to identify the fixed collection comprising all of the coded characters that were in the BMP in the First Edition of this International Standard. This collection is not now in conformity with this International Standard.

NOTE - The specification of collection 299 BMP FIRST EDITION consisted of the specification of collection 301 BMP-AMD. 7 except for the replacement of the corresponding entries in the list above with the entries shown below:

| rows | positions |
| :---: | :---: |
| 05 | 31-56 59-5F 61-87 89 B0-B9 BB-C3 |
|  | D0-EA F0-F4 |
| OF | [no positions] |
| 1E | 00-9A A0-F9 |
| 20 | 00-2E 30-46 6A-70 74-8E A0-AA D0-E1 |
| AC-D7 | [no positions] |
| and by including an additional entry: |  |
| 34-4D | 3400-4DFF |
| for the code position ranges of three collections (57,58,59) |  |
| of coded characters which have been deleted from this |  |
| internatior | al Standard since the First Edition. |

## A.3.3 302 BMP SECOND EDITION

The fixed collection 302 BMP SECOND EDITION comprises only those coded characters that are in the BMP in this Second Edition of ISO/IEC 10646-1. The repertoire of this collection is not subject to change if new characters are added to the BMP by any subsequent amendments.

302 BMP SECOND EDITION is specified by the following ranges of code positions as indicated for each row or contiguous series of rows.

| Rows | Positions (cells) |
| :---: | :---: |
| 00 | 20-7E A0-FF |
| 01 | 00-FF |
| 02 | 00-33 50-AD B0-EE |
| 03 | 00-4E 60-62 74-75 7A 7E 84-8A 8C 8E-A1 A3-CE D0-D7 DA-F3 |
| 04 | 00-86 88-89 8C-CE D0-F5 F8-F9 |
| 05 | 31-56 59-5F 61-87 89-8A 91-A1 A3-B9 BBC4 D0-EA F0-F4 |
| 06 | 0C 1B 1F 21-3A 40-55 60-6D 70-ED..F0-FE |
| 07 | 00-0D 0F-2C 30-4A 80-BF |
| 09 | 01-03 05-39 3C-4D 50-54 58-70 81-83 85-8C |
|  | 8F-90 93-A8 AA-B0 B2 B6-B9 BC BE-C4 |
|  | C7-C8 CB-CD D7 DC-DD DF-E3 E6-FA |
| OA | 02 05-0A OF-10 13-28 2A-30 32-33 35-36 |
|  | 38-39 3C 3E-42 47-48 4B-4D 59-5C 5E 66- |
|  | 74 81-83 85-8B 8D 8F-91 93-A8 AA-B0 B2- |
|  | B3 B5-B9 BC-C5 C7-C9 CB-CD D0 E0 E6- |
|  | EF |
| OB | 01-03 05-0C 0F-10 13-28 2A-30 32-33 36-39 |
|  | 3C-43 47-48 4B-4D 56-57 5C-5D 5F-61 66- |
|  | 70 82-83 85-8A 8E-90 92-25 99-9A 9C 9E- |
|  | 9F A3-A4 A8-AA AE-B5 B7-B9 BE-C2 C6-C8 |
|  | CA-CD D7 E7-F2 |
| OC | 01-03 05-0C 0E-10 12-28 2A-33 35-39 3E-44 |
|  | 46-48 4A-4D 55-56 60-61 66-6F 82-83 85-8C |
|  | 8E-90 92-A8 AA-B3 B5-B9 BE-C4 C6-C8 CA- |
|  | CD D5-D6 DE E0-E1 E6-EF |
| OD | 02-03 05-0C 0E-10 12-28 2A-39 3E-43 46-48 |
|  | 4A-4D 57 60-61 66-6F 82-83 $85-96$ 9A-B1 |
|  | B3-BB BD C0-C6 CA CF-D4 D6 D8-DF F2- |
|  | F4 |
| OE | 01-3A 3F-5B 81-82 84 87-88 8A 8D 94-97 |
|  | 99-9F A1-A3 A5 A7 AA-AB AD-B9 BB-BD |
|  | C0-C4 C6 C8-CD D0-D9 DC-DD |
| 0F | 00-47 49-6A 71-8B 90-97 99-BC BE-CC CF |
| 10 | 00-21 23-27 29-2A 2C-32 36-39 40-59 A0-C5 |
|  | D0-F6 FB |
| 11 | 00-59 5F-A2 A8-F9 |
| 12 | 20-26 28-46 48 4A-4D 50-56 58 5A-5D 60- |
|  | 8688 8A-8D 90-AE B0 B2-B5 B8-BE C0 |
|  | C2-C5 C8-CE D0-D6 D8-EE F0-FF |
| 13 | 00-0E 10 12-15 18-1E 20-46 48-5A $61-7 \mathrm{C}$ |
|  | A0-F4 |
| 14-15 | 1401-15FF |
| 16 | 00-76 80-9C A0-F0 |
| 17 | 80-DC E0-E9 |
| 18 | 00-0E 10-19 20-77 80-A9 |
| 1E | 00-9B A0-F9 |

4E-9F 4E00
A0-A3 A000-A3FF

A4 00-8C 90-A1 A4-B3 B5-C0 C2-C4 C6
AC-D7 AC00-D7A3
E0-F8 E000-F8FF
F9-FA F900-FA2D
$\begin{array}{llllllll}\text { FB } & 00-06 & 13-17 & 1 D-36 & 38-3 C & 3 E & 40-41 & 43-44\end{array}$ 46-B1 D3-FF
FC 00-FF
FD 00-3F 50-8F 92-C7 F0-FB
FE 20-23 30-44 49-52 54-66 68-6B 70-72 74 76FC FF
FF 01-5E 61-BE C2-C7 CA-CF D2-D7 DA-DC E0-E6 E8-EE F9-FD
[Editor's note: The details of the above entries will be adjusted as necessary when the exact character repertoire of ISO/IEC 10646-1 Second Edition is finalised.]

## Annex B <br> (normative)

## List of combining characters

## B. 1 List of all combining characters

The characters in the subset collections COMBINING DIACRITICAL MARKS (0300 to 036F), COMBINING DIACRITICAL MARKS FOR SYMBOLS (20D0 to 20FF), and COMBINING HALF MARKS (FE20 to FE2F) are combining characters. In addition, the following characters are combining characters.

```
\begin{tabular}{ll} 
0483 & COMBINING CYRILLIC TITLO \\
0484 & COMBINING CYRILLIC PALATALIZATION \\
0485 & COMBINING CYRILLC DASIA PNEUMATA \\
0486 & COMBINING CYRILLIC PSILI PNEUMATA \\
0488 & COMBINING CYRILLIC HUNDRED THOUSANDS \\
& SIGN \\
0489 & COMBINING CYRILLIC MILLIONS SIGN \\
0591 & HEBREW ACCENT ETNAHTA \\
0592 & HEBREW ACCENT SEGOL \\
0593 & HEBREW ACCENT SHALSHELET \\
0594 & HEBREW ACCENT ZAQEF QATAN \\
0595 & HEBREW ACCENT ZAQEF GADOL \\
0596 & HEBREW ACCENT TIPEHA \\
0597 & HEBREW ACCENT REVIA \\
0598 & HEBREW ACCENT ZARQA \\
0599 & HEBREW ACCENT PASHTA \\
059A & HEBREW ACCENTYETIV \\
059B & HEBREW ACCENT TEVIR \\
059C & HEBREW ACCENT GERESH \\
059D & HEBREW ACCENT GERESH MUQDAM \\
059E & HEBREW ACCENT GERSHAYIM \\
059F & HEBREW ACCENT QARNEY PARA \\
05A0 & HEBREW ACCENT TELISHA GEDOLA \\
05A1 & HEBREW ACCENT PAZER \\
05A3 & HEBREW ACCENT MUNAH \\
05A4 & HEBREW ACCENT MA HAPAKH \\
05A5 & HEBREW ACCENT MERKHA \\
05A6 & HEBREW ACCENT MERKHA KEFULA \\
05A7 & HEBREW ACCENT DARGA \\
05A8 & HEBREW ACCENT QADMA \\
05A9 & HEBREW ACCENT TELISHA QETANA \\
05AA & HEBREW ACCENT YERAH BEN YOMO \\
05AB & HEBREW ACCENT OLE \\
05AC & HEBREW ACCENT ILUY \\
05AD & HEBREW ACCENT DEHI \\
05AE & HEBREW ACCENT ZINOR \\
05AF & HEBREW MARK MASORA CIRCLE \\
05B0 & HEBREW POINT SHEVA \\
05B1 & HEBREW POINT HATAF SEGOL \\
05B2 & HEBREW POINT HATAF PATAH \\
05B3 & HEBREW POINT HATAF QAMATS \\
05B4 & HEBREW POINT HIRIQ \\
05B5 & HEBREW POINT TSERE \\
05B6 & HEBREW POINT SEGOL \\
05B7 & HEBREW POINT PATAH \\
&
\end{tabular}
```

05B8
05B9
05BB
05BC
05BD
05BF
05C1
05C2
05C4
064B
064C
064D
064E
064F
0650
0651
0652
0653
0654
0655
0670
06 D 7
06D8
06D9
06DA
06DB
06DC
06DD
06DE
06DF
06E0
06E1
06E2
06E3
06E4
06E7
06E8
06EA
06EB
06EC
06ED
0711
0730
0731
0732
0733
0734
0735
0736
0737
0738
0739

HEBREW POINT QAMATS HEBREW POINT HOLAM
HEBREW POINT QUBUTS
HEBREW POINT DAGESH OR MAPIQ
HEBREW POINT METEG
HEBREW POINT RAFE
HEBREW POINT SHIN DOT
HEBREW POINT SIN DOT
HEBREW MARK UPPER DOT
ARABIC FATHATAN
ARABIC DAMMATAN
ARABIC KASRATAN
ARABIC FATHA
ARABIC DAMMA
ARABIC KASRA
ARABIC SHADDA
ARABIC SUKUN
ARABIC MADDAH ABOVE
ARABIC HAMZA ABOVE
ARABIC HAMZA BELOW
ARABIC LETTER SUPERSCRIPT ALEF
ARABIC SMALL HIGH LIGATURE QAF WITH LAM WITH ALEF MAKSURA
ARABIC SMALL HIGH MEEM INITIAL FORM
ARABIC SMALL HIGH LAM ALEF
ARABIC SMALL HIGH JEEM
ARABIC SMALL HIGH THREE DOTS
ARABIC SMALL HIGH SEEN
ARABIC END OF AYAH
ARABIC START OF RUB EL HIZB
ARABIC SMALL HIGH ROUNDED ZERO
ARABIC SMALL HIGH UPRIGHT RECTANGULAR ZERO
ARABIC SMALL HIGH DOTLESS HEAD OF KHAH
ARABIC SMALL HIGH MEEM ISOLATED FORM
ARABIC SMALL LOW SEEN
ARABIC SMALL HIGH MADDA
ARABIC SMALL HIGH YEH
ARABIC SMALL HIGH NOON
ARABIC EMPTY CENTRE LOW STOP
ARABIC EMPTY CENTRE HIGH STOP ARABIC ROUNDED HIGH STOP WITH FILLED CENTRE
ARABIC SMALL LOW MEEM
SYRIAC LETTER SUPERSCRIPT ALAPH
SYRIAC PTHAHA ABOVE
SYRIAC PTHAHA BELOW
SYRIAC PTHAHA DOTTED
SYRIAC ZQAPHA ABOVE
SYRIAC ZQAPHA BELOW
SYRIAC ZQAPHA DOTTED
SYRIAC RBASA ABOVE
SYRIAC RBASA BELOW
SYRIAC DOTTED ZLAMA HORIZONTAL
SYRIAC DOTTED ZLAMA ANGULAR

| 073A | SYRIAC HBASA ABOVE |
| :---: | :---: |
| 073B | SYRIAC HBASA BELOW |
| 073C | SYRIAC HBASA-ESASA DOTTED |
| 073D | SYRIAC ESASA ABOVE |
| 073E | SYRIAC ESASA BELOW |
| 073F | SYRIAC RWAHA |
| 0740 | SYRIAC FEMININE DOT |
| 0741 | SYRIAC QUSHSHAYA |
| 0742 | SYRIAC RUKKAKHA |
| 0743 | SYRIAC TWO VERTICAL DOTS ABOVE |
| 0744 | SYRIAC TWO VERTICAL DOTS BELOW |
| 0745 | SYRIAC THREE DOTS ABOVE |
| 0746 | SYRIAC THREE DOTS BELOW |
| 0747 | SYRIAC OBLIQUE LINE ABOVE |
| 0748 | SYRIAC OBLIQUE LINE BELOW |
| 0749 | SYRIAC MUSIC |
| 074A | SYRIAC BARREKH |
| 07A6 | THAANA ABAFILI |
| 07A7 | THAANA AABAAFILI |
| 07A8 | THAANA IBIFILI |
| 07A9 | THAANA EEBEEFILI |
| 07AA | THAANA UBUFILI |
| 07AB | THAANA OOBOOFILI |
| 07AC | THAANA EBEFILI |
| 07AD | THAANA EYBEYFILI |
| 07AE | THAANA OBOFILI |
| 07AF | THAANA OABOAFILI |
| 07B0 | THAANA SUKUN |
| 0901 | DEVANAGARI SIGN CANDRABINDU |
| 0902 | DEVANAGARI SIGN ANUSVARA |
| 0903 | DEVANAGARI SIGN VISARGA |
| 093C | DEVANAGARI SIGN NUKTA |
| 093E | DEVANAGARI VOWEL SIGN AA |
| 093F | DEVANAGARI VOWEL SIGN I |
| 0940 | DEVANAGARI VOWEL SIGN II |
| 0941 | DEVANAGARI VOWEL SIGN U |
| 0942 | DEVANAGARI VOWEL SIGN UU |
| 0943 | DEVANAGARI VOWEL SIGN VOCALIC R |
| 0944 | DEVANAGARI VOWEL SIGN VOCALIC RR |
| 0945 | DEVANAGARI VOWEL SIGN CANDRA E |
| 0946 | DEVANAGARI VOWEL SIGN SHORT E |
| 0947 | DEVANAGARI VOWEL SIGN E |
| 0948 | DEVANAGARI VOWEL SIGN AI |
| 0949 | DEVANAGARI VOWEL SIGN CANDRA O |
| 094A | DEVANAGARI VOWEL SIGN SHORT O |
| 094B | DEVANAGARI VOWEL SIGN O |
| 094C | DEVANAGARI VOWEL SIGN AU |
| 094D | DEVANAGARI SIG N VIRAMA |
| 0951 | DEVANAGARI STRESS SIGN UDATTA |
| 0952 | DEVANAGARI STRESS SIGN ANUDATTA |
| 0953 | DEVANAGARI GRAVE ACCENT |
| 0954 | DEVANAGARI ACUTE ACCENT |
| 0962 | DEVANAGARI VOWEL SIGN VOCALIC L |
| 0963 | DEVANAGARI VOWEL SIGN VOCALIC LL |
| 0981 | BENGALI SIGN CANDRABINDU |
| 0982 | BENGALISIGN ANUSVARA |
| 0983 | BENGALI SIGN VISARGA |
| 09BC | BENGALI SIGN NUKTA |
| 09BE | BENGALI VOWEL SIGN AA |
| 09BF | BENGALI VOWEL SIGN I |
| 09C0 | BENGALI VOWEL SIGN II |
| 09 C 1 | BENGALI VOWEL SIGN U |
| 09C2 | BENGALI VOWEL SIGN UU |
| 09 C 3 | BENGALI VOWEL SIGN VOCALIC R |
| 09 C 4 | BENGALI VOWEL SIGN VOCALIC RR |
| $09 \mathrm{C7}$ | BENGALI VOWEL SIGN E |

09C8
09CB
09CC
09CD
$09 \mathrm{D7}$
09E2
09E3
0A02
0A3C
0A3E
0A3F
0A40
0A41
0A42
0A47
0A48
$0 A 4 B$
0A4C
0A4D
0A70
0A71
0A81
0A82
0A83
OABC
OABE
OABF
OACO
0AC1
0AC2
0AC3
OAC4
0AC5
0AC7
0AC8
0AC9
OACB
OACC
OACD
0B01
0B02
0B03
0B3C
0B3E
0B3F
0B40
0B41
0B42
0B43
0B47
0B48
0B4B
0B4C
0B4D
0B56
0B57
0B82
0B83
OBBE
OBBF
OBCO
0BC1
0BC2
0BC6
0BC7
0BC8

BENGALI VOWEL SIGN AI
BENGALI VOWEL SIGN O
BENGALI VOWEL SIGN AU
BENGALI SIGN VIRAMA
BENGALI AU LENGTH MARK
BENGALI VOWEL SIGN VOCALIC L
BENGALI VOWEL SIGN VOCALIC LL
GURMUKHI SIG N BINDI
GURMUKHI SIGN NUKTA
GURMUKHI VOWEL SIGN AA
GURMUKHI VOWEL SIGN I
GURMUKHI VOWEL SIGN II
GURMUKHI VOWEL SIGN U
GURMUKHI VOWEL SIGN UU
GURMUKHI VOWEL SIGN EE
GURMUKHI VOWEL SIGN AI
GURMUKHI VOWEL SIGN OO
GURMUKHI VOWEL SIGN AU
GURMUKHI SIGN VIRAMA
GURMUKHI TIPPI
GURMUKHI ADDAK
GUJARATI SIGN CANDRABINDU
GUJARATI SIGN ANUSVARA
GUJARATI SIGN VISARGA
GUJARATI SIGN NUKTA
GUJARATI VOWEL SIGN AA
GUJARATI VOWEL SIGN I
GUJARATI VOWEL SIGN II
GUJARATI VOWEL SIGN U
GUJARATI VOWEL SIGN UU
GUJARATI VOWEL SIGN VOCALIC R
GUJARATI VOWEL SIGN VOCALIC RR
GUJARATI VOWEL SIGN CANDRA E
GUJARATI VOWEL SIGN E
GUJARATI VOWEL SIGN AI
GUJARATI VOWEL SIGN CANDRA O GUJARATI VOWEL SIGN O
GUJARATI VOWEL SIGN AU
GUJARATI SIGN VIRAMA
ORIYA SIGN CANDRABINDU
ORIYA SIGN ANUSVARA
ORIYA SIGN VISARGA
ORIYA SIGN NUKTA
ORIYA VOWEL SIGN AA
ORIYA VOWEL SIGN I
ORIYA VOWEL SIGN II
ORIYA VOWEL SIGN U
ORIYA VOWEL SIGN UU
ORIYA VOWEL SIGN VOCALIC R
ORIYA VOWEL SIGN E
ORIYA VOWEL SIGN AI
ORIYA VOWEL SIGN O
ORIYA VOWEL SIGN AU
ORIYA SIGN VIRAMA
ORIYA AI LENGTH MARK
ORIYA AU LENGTH MARK
TAMIL SIGN ANUSVARA
TAMIL SIGN VISARGA
TAMIL VOWEL SIGN AA
TAMIL VOWEL SIGN I
TAMIL VOWEL SIGN II
TAMIL VOWEL SIGN U
TAMIL VOWEL SIGN UU
TAMIL VOWEL SIGN E
TAMIL VOWEL SIGN EE
TAMIL VOWB SIGN AI

| OBCA | TAMIL VOWEL SIGN O |
| :---: | :---: |
| OBCB | TAMIL VOWEL SIGN OO |
| OBCC | TAMIL VOWEL SIGN AU |
| OBCD | TAMIL SIGN VIRAMA |
| OBD7 | TAMIL AU LENGTH MARK |
| 0C01 | TELUGU SIGN CANDRABINDU |
| 0C02 | TELUGU SIGN ANUSVARA |
| 0 C 03 | TELUGU SIGN VISARGA |
| 0C3E | TELUGU VOWEL SIGN AA |
| 0C3F | TELUGU VOWEL SIGN I |
| 0C40 | TELUGU VOWEL SIGN II |
| 0C41 | TELUGU VOWEL SIGN U |
| 0C42 | TELUGU VOWEL SIGN UU |
| 0 C 43 | TELUGU VOWEL SIGN VOCALIC R |
| 0C44 | TELUGU VOWEL SIGN VOCALIC RR |
| 0C46 | TELUGU VOWEL SIGN E |
| 0 C 47 | TELUGU VOWEL SIGN EE |
| 0 C 48 | TELUGU VOWEL SIGN AI |
| 0C4A | TELUGU VOWEL SIGN O |
| 0C4B | TELUGU VOWEL SIGN OO |
| 0C4C | TELUGU VOWEL SIGN AU |
| 0C4D | TELUGU SIGN VIRAMA |
| 0 C 55 | TELUGU LENGTH MARK |
| 0 C 56 | TELUGU AI LENGTH MARK |
| 0 C 82 | KANNADA SIGN ANUSVARA |
| 0 C 83 | KANNADA SIGN VISARGA |
| OCBE | KANNADA VOWEL SIGN AA |
| OCBF | KANNADA VOWEL SIGN I |
| OCCO | KANNADA VOWEL SIGN II |
| 0CC1 | KANNADA VOWEL SIGN U |
| 0CC2 | KANNADA VOWEL SIGN UU |
| 0CC3 | KANNADA VOWEL SIGN VOCALIC R |
| 0CC4 | KANNADA VOWEL SIGN VOCALIC RR |
| 0CC6 | KANNADA VOWEL SIGN E |
| 0CC7 | KANNADA VOWEL SIGN EE |
| 0CC8 | KANNADA VOWEL SIGN AI |
| OCCA | KANNADA VOWEL SIGN O |
| OCCB | KANNADA VOWEL SIGN OO |
| OCCC | KANNADA VOWEL SIGN AU |
| OCCD | KANNADA SIGN VIRAMA |
| OCD5 | KANNADA LENGTH MARK |
| 0CD6 | KANNADA AI LENGTH MARK |
| 0D02 | MALAYALAM SIGN ANUSVARA |
| OD03 | MALAYALAM SIGN VISARGA |
| OD3E | MALAYALAM VOWEL SIGN AA |
| 0D3F | MALAYALAM VOWEL SIGN I |
| 0D40 | MALAYALAM VOWEL SIGN II |
| OD41 | MALAYALAM VOWEL SIGN U |
| OD42 | MALAYALAM VOWEL SIGN UU |
| 0D43 | MALAYALAM VOWEL SIGN VOCALIC R |
| OD46 | MALAYALAM VOWEL SIGN E |
| 0D47 | MALAYALAM VOWEL SIGN EE |
| OD48 | MALAYALAM VOWEL SIGN AI |
| 0D4A | MALAYALAM VOWEL SIGN O |
| 0D4B | MALAYALAM VOWEL SIGN OO |
| OD4C | MALAYALAM VOWEL SIGN AU |
| 0D4D | MALAYALAM SIGN VIRAMA |
| 0D57 | MALAYALAM AU LENGTH MARK |
| 0D82 | SINHALA SIGN ANUSVARAYA |
| 0D83 | SINHALA SIGN VISARGAYA |
| ODCA | SINHALA SIGN AL-LAKUNA |
| ODCF | SINHALA VOWEL SIGN AELA-PILLA |
| ODD0 | SINHALA VOWEL SIGN KETTI AEDA-PILLA |
| 0DD1 | SINHALA VOWEL SIGN DIGA AEDA-PILLA |
| 0DD2 | SINHALA VOWEL SIGN KETTI IS-PILLA |
| ODD3 | SINHALA VOWEL SIGN DIGA IS-PILLA |

0DD4
0DD6
0DD8
0DD9
ODDA
ODDB
ODDC
ODDD
ODDE
ODDF
0DF2
0DF3
0E31
0E34
0E35
0E36
0E37
0E38
0E39
0E3A
0E47
0E48
0E49
0E4A
0E4B
0E4C
0E4D
0E4E
0EB1
0EB4
0EB5
0EB6
0EB7
0EB8
0EB9
0EBB
OEBC
0EC8
0EC9
OECA
OECB
OECC
OECD
0F18
0F19
0F35
0F37
0F39
0F3E
0F3F
0F71
0F72
0F73
0F74
0F75
0F76
0F77
0F78
0F79
0F7A
0F7B
0F7C
0F7D

SINHALA VOWEL SIGN KETTI PAA-PILLA SINHALA VOWEL SIGN DIGA PAA-PILLA SINHALA VOWEL SIGN GAETTA-PILLA SINHALA VOWEL SIGN KOMBUVA SINHALA VOWEL SIGN DIGA KOMBUVA SINHALA VOWEL SIGN KOMBU DEKA SINHALA VOWEL SIGN KOMBUVA HAA AELAPILLA
SINHALA VOWEL SIGN KOMBUVA HAA DIGA AELA-PILLA
SINHALA VOWEL SIGN KOMBUVA HAA GAYANUKITTA SINHALA VOWEL SIGN GAYANUKITTA SINHALA VOWEL SIGN DIGA GAETTA-PILLA SINHALA VOWEL SIGN DIGA GAYANUKITTA THAI CHARACTER MAI HAN-AKAT
THAI CHARACTER SARA I
THAI CHARACTER SARA II
THAI CHARACTER SARA UE
THAI CHARACTER SARA UEE
THAI CHARACTER SARA U
THAI CHARACTER SARA UU
THAI CHARACTER PHINTHU
THAI CHARACTER MAITAIKHU
THAI CHARACTER MAI EK
THAI CHARACTER MAI THO
THAI CHARACTER MAI TRI
THAI CHARACTER MAI CHATTAWA
THAI CHARACTER THANTHAKHAT
THAI CHARACTER NIKHAHIT THAI CHARACTER YAMAKKAN
LAO VOWEL SIGN MAI KAN
LAO VOWEL SIGN I
LAO VOWEL SIGN II
LAO VOWEL SIGN Y
LAO VOWEL SIGN YY
LAO VOWEL SIGN U
LAO VOWEL SIGN UU
LAO VOWEL SIGN MAI KON
LAO SEMIVOWEL SIGN LO
LAO TONE MAI EK
LAO TONE MAI THO
LAO TONE MAI TI
LAO TONE MAI CATAWA
LAO CANCELLATION MARK
LAO NIGGAHITA
TIBETAN ASTROLOGICAL SIGN -KHYUD PA TIBETAN ASTROLOGICAL SIGN SDONG TSHUGS
TIBETAN MARK NGAS BZUNG NYI ZLA
TIBETAN MARK NGAS BZUNG SGOR RTAGS
TIBETAN MARK TSA -PHRU
TIBETAN SIGN YAR TSHES
TIBETAN SIGN MAR TSHES
TIBETAN VOWEL SIGN AA
TIBETAN VOWEL SIGN I
TIBETAN VOWEL SIGN II
TIBETAN VOWEL SIGN U
TIBETAN VOWEL SIGN UU
TIBETAN VOWEL SIGN VOCA LIC R TIBETAN VOWEL SIGN VOCALIC RR TIBETAN VOWEL SIGN VOCALIC L
TIBETAN VOWEL SIGN VOCALIC LL
TIBETAN VOWEL SIGN E
TIBETAN VOWEL SIGN EE
TIBETAN VOWEL SIGN O
TIBETAN VOWEL SIGN OO

| 0F7E | TIBETAN SIGN RJES SU NGA RO |
| :---: | :---: |
| 0F7F | TIBETAN SIGN RNAM BCAD |
| 0F80 | TIBETAN VOWEL SIGN REVERSED I |
| 0 F81 | TIBETAN VOWEL SIGN REVERSED II |
| 0F82 | TIBETAN SIGN NYI ZLA NAA DA |
| 0 F83 | TIBETAN SIGN SNA LDAN |
| 0F84 | TIBETAN MARK HALANTA |
| 0F86 | TIBETAN MARK LCI RTAGS |
| 0 F87 | TIBETAN MARK YANG RTAGS |
| 0F90 | TIBETAN SUBJOINED LETTER KA |
| 0F91 | TIBETAN SUBJOINED LETTER KHA |
| 0F92 | TIBETAN SUBJOINED LETTER GA |
| 0 F93 | TIBETAN SUBJOINED LETTER GHA |
| 0F94 | TIBETAN SUBJOINED LETTER NGA |
| 0F95 | TIBETAN SUBJOINED LETTER CA |
| 0F96 | TIBETAN SUBJOINED LETTER CHA |
| 0 F97 | TIBETAN SUBJOINED LETTER JA |
| 0 F99 | TIBETAN SUBJOINED LETTER NYA |
| 0F9A | TIBETAN SUBJOINED LETTER TTA |
| 0F9B | TIBETAN SUBJOINED LETTER TTHA |
| 0F9C | TIBETAN SUBJOINED LETTER DDA |
| 0F9D | TIBETAN SUBJOINED LETTER DDHA |
| 0F9E | TIBETAN SUBJOINED LETTER NNA |
| 0F9F | TIBETAN SUBJOINED LETTER TA |
| OFA0 | TIBETAN SUBJOINED LETTER THA |
| 0FA1 | TIBETAN SUBJOINED LETTER DA |
| 0FA2 | TIBETAN SUBJOINED LETTER DHA |
| 0FA3 | TIBETAN SUBJOINED LETTER NA |
| 0FA4 | TIBETAN SUBJOINED LETTER PA |
| 0FA5 | TIBETAN SUBJOINED LETTER PHA |
| 0FA6 | TIBETAN SUBJOINED LETTER BA |
| 0FA7 | TIBETAN SUBJOINED LETTER BHA |
| 0FA8 | TIBETAN SUBJOINED LETTER MA |
| 0FA9 | TIBETAN SUBJOINED LETTER TSA |
| OFAA | TIBETAN SUBJOINED LETTER TSHA |
| OFAB | TIBETAN SUBJOINED LETTER DZA |
| OFAC | TIBETAN SUBJOINED LETTER DZHA |
| OFAD | TIBETAN SUBJOINED LETTER WA |
| OFAE | TIBETAN SUBJOINED LETTER ZHA |
| OFAF | TIBETAN SUBJOINED LETTER ZA |
| OFBO | TIBETAN SUBJOINED LETTER -A |
| 0FB1 | TIBETAN SUBJOINED LETTER YA |
| 0FB2 | TIBETAN SUBJOINED LETTER RA |
| 0FB3 | TIBETAN SUBJOINED LETTER LA |
| OFB4 | TIBETAN SUBJOINED LETTER SHA |
| 0FB5 | TIBETAN SUBJOINED LETTER SSA |
| 0FB6 | TIBETAN SUBJOINED LETTER SA |
| 0FB7 | TIBETAN SUBJOINED LETTER HA |
| 0FB8 | TIBETAN SUBJOINED LETTER A |
| 0FB9 | TIBETAN SUBJOINED LETTER KSSA |
| OFBA | TIBETAN SUBJOINED LETTER FIXED-FORM WA |
| OFBB | TIBETAN SUBJOINED LETTER FIXED-FORM YA |
| OFBC | TIBETAN SUBJOINED LETTER FIXED-FORM RA |
| OFC6 | TIBETAN SYMBOL PADMA GDAN |
| 102C | MYANMAR VOWEL SIGN AA |
| 102D | MYANMAR VOWEL SIGN I |
| 102E | MYANMAR VOWEL SIGN II |
| 102F | MYANMAR VOWEL SIGN U |
| 1030 | MYANMAR VOWEL SIGN UU |
| 1031 | MYANMAR VOWEL SIGN E |
| 1032 | MYANMAR VOWEL SIGN AI |
| 1036 | MYANMAR SIGN ANUSVARA |
| 1037 | MYANMAR SIGN DOT BELOW |
| 1038 | MYANMAR SIGN VISARGA |
| 1039 | MYANMAR SIGN VIRAMA |
| 1056 | MYANMAR VOWEL SIGN VOCALIC R |


| 1057 | MYANMAR VOWEL SIGN VOCALIC RR |
| :--- | :--- |
| 1058 | MYANMAR VOWEL SIGN VOCALIC L |
| 1059 | MYANMAR VOWEL SIGN VOCALIC LL |
| 17B4 | KHMER VOWEL INHERENT AQ |
| 17B5 | KHMER VOWEL INHERENT AA |
| 17B6 | KHMER VOWEL SIGN AA |
| 17B7 | KHMER VOWEL SIGN I |
| 17B8 | KHMER VOWEL SIGN II |
| 17B9 | KHMER VOWEL SIGN Y |
| 17BA | KHMER VOWEL SIGN YY |
| 17BB | KHMER VOWEL SIGN U |
| 17BC | KHMER VOWEL SIGN UU |
| 17BD | KHMER VOWEL SIGN UA |
| 17BE | KHMER VOWEL SIGN OE |
| 17BF | KHMER VOWEL SIGN YA |
| 17C0 | KHMER VOWEL SIGN IE |
| 17C1 | KHMER VOWEL SIGN E |
| 17C2 | KHMER VOWEL SIGN AE |
| 17C3 | KHMER VOWEL SIGN AI |
| 17C4 | KHMER VOWEL SIGN OO |
| 17C5 | KHMER VOWEL SIGN AU |
| 17C6 | KHMER SIGN NIKAHIT |
| 17C7 | KHMER SIGN REAHMUK |
| 17C8 | KHMER SIGN YUUKALEAPINTU |
| 17C9 | KHMER SIGN MUUSIKATOAN |
| 17CA | KHMER SIGN TRIISAP |
| 17CB | KHMER SIGN BANTOC |
| 17CC | KHMER SIGN ROBAT |
| 17CD | KHMER SIGN TOANDAKHIAT |
| 17CE | KHMER SIGN KAKABAT |
| 17CF | KHMER SIGN AHSDA |
| 17DO | KHMER SIGN SAMYOK SANNYA |
| 17D1 | KHMER SIGN VIRIAM |
| 17D2 | KHMER SIGN COENG |
| 17D3 | KHMER SIGN BATHAMASAT |
| 18A9 | MONGOLIAN LETTER AG DAGALGA |
| 302A | IDEOGRAPHIC LEVEL TONE MARK |
| 302B | IDEOGRAPHIC RISING TONE MARK |
| 302C | IDEOGRAPHIC DEPARTING TONE MARK |
| 302D | IDEOGRAPHIC ENTERING TONE MARK |
| 302E | HANGUL SINGLE DOT TONE MARK |
| 302F | HANGUL DOUBLE DOT TONE MARK |
| 3099 | COMBINING KATAKANA-HIRAGANA VOICED |
|  | SOUND MARK |
| 309A | COMBINING KATAKANA-HIRAGANA |
| FB1E | SEMI-VOICED SOUND MARK |
| HEBREW POINT JUDEO-SPANISH VARIKA |  |
|  |  |

## B. 2 List of characters not allowed in implementation level 2

The characters in the subset collections COMBINING DIACRITICAL MARKS (0300 to 036F), COMBINING DIACRITICAL MARKS FOR SYMBOLS (20D0 to 20FF), HANGUL JAMO (1100 to 11 FF ) and COMBINING HALF MARKS (FE20 to FE2F) are not allowed in implementation level 2. In addition, the following individual characters are also not allowed.

NOTE - This list is a subset of the list in clause B. 1 except for HANGUL JAMO (see 25.1).

```
0483
0484
0485
0486 COMBINING CYRILLIC PSILI PNEUMATA
```

| 0591 | HEBREW ACCENT ETNAHTA | $05 C 4$ |
| :--- | :--- | :--- |
| 0592 | HEBREW ACCENT SEGOL | 093 C |
| 0593 | HEBREW ACCENT SHALSHELET | 0953 |
| 0594 | HEBREW ACCENT ZAQEF QATAN | 0954 |
| 0595 | HEBREW ACCENT ZAQEF GADOL | $09 B C$ |
| 0596 | HEBREW ACCENT TIPEHA | $09 D 7$ |
| 0597 | HEBREW ACCENT REVIA | $0 A 3 C$ |
| 0598 | HEBREW ACCENT ZARQA | $0 A 70$ |
| 0599 | HEBREW ACCENT PASHTA | $0 A 71$ |
| $059 A$ | HEBREW ACCENT YETIV | $0 A B C$ |
| 059B | HEBREW ACCENT TEVIR | $0 B 3 C$ |
| 059C | HEBREW ACCENT GERESH | $0 B 56$ |
| 059D | HEBREW ACCENT GERESH MUQDAM | $0 B 57$ |
| 059E | HEBREW ACCENT GERSHAYIM | $0 B D 7$ |
| 059F | HEBREW ACCENT QARNEY PARA | $0 C 55$ |
| 05A0 | HEBREW ACCENT TELISHA GEDOLA | $0 C 56$ |
| 05A1 | HEBREW ACCENT PAZER | $0 C D 5$ |
| 05A3 | HEBREW ACCENT MUNAH | $0 C D 6$ |
| 05A4 | HEBREW ACCENT MAHAPAKH | $0 D 57$ |
| 05A5 | HEBREW ACCENT MERKHA | $0 F 39$ |
| 05A6 | HEBREW ACCENT MERKHA KEFULA | $302 A$ |
| 05A7 | HEBREW ACCENT DARGA | $302 B$ |
| 05A8 | HEBREW ACCENT QADMA | $302 C$ |
| 05A9 | HEBREW ACCENT TELISHA QETANA | $302 D$ |
| 05AA | HEBREW ACCENT YERAH BEN YOMO | $302 E$ |
| 05AB | HEBREW ACCENT OLE | $302 F$ |
| 05AC | HEBREW ACCENT ILUY | 3099 |
| 05AD | HEBREW ACCENT DEHI |  |
| 05AE | HEBREW ACCENT ZINOR | $309 A$ |
| 05AF | HEBREW MARK MASORA CIRCLE |  |

HEBREW MARK UPPER DOT
DEVANAGARI SIGN NUKTA
DEVANAGARI GRAVE ACCENT
DEVANAGARI ACUTE ACCENT
BENGALI SIGN NUKTA
BENGALI AU LENGTH MARK
GURMUKHI SIGN NUKTA
GURMUKHI TIPPI
GURMUKHI ADDAK
GUJARATI SIGN NUKTA
ORIYA SIGN NUKTA
ORIYA AI LENGTH MARK
ORIYA AU LENGTH MARK
TAMIL AU LENGTH MARK
TELUGU LENGTH MARK
TELUGU AI LENGTH MARK
KANNADA LENGTH MARK
KANNADA AI LENGTH MARK
MALAYALAM AU LENGTH MARK TIBETAN MARK TSA -PHRU IDEOGRAPHIC LEVEL TONE MARK IDEOGRAPHIC RISING TONE MARK IDEOGRAPHIC DEPARTING TONE MARK IDEOGRAPHIC ENTERING TONE MARK HANGUL SINGLE DOT TONE MARK HANGUL DOUBLE DOT TONE MARK COMBINING KATAKANA-HIRAGANA VOICED SOUND MARK COMBINING KATAKANA-HIRAGANA SEMI-VOICED SOUND MARK

## Transformation format for 16 planes of Group 00 (UTF-16)

UTF-16 provides a coded representation of over a million graphic characters of UCS-4 in a form that is compatible with the two-octet BMP form of UCS-2 (13.1). This permits the coexistence of those characters from UCS-4 within coded character data that is in accordance with UCS-2.

In UTF-16 each graphic character from the UCS-2 repertoire retains its UCS-2 coded representation. In addition, the coded representation of any character from a single contiguous block of 16 Planes in Group 00 ( $1,048,576$ code positions) consists of a pair of RC-elements (4.33), where each such RCelement corresponds to a cell in a single contiguous block of 8 Rows in the BMP ( 2,048 code positions). These code positions are reserved for the use of this coded representation form, and shall not be allocated for any other purpose.

## C. 1 Specification of UTF-16

The specification of UTF-16 is as follows:

1. The high-half zone shall be the 4 rows D 8 to DB of the BMP, i.e., the 1,024 cells in the S-zone whose code positions are from D800 through DBFF.
2. The low-half zone shall be the 4 rows $D C$ to DF of the BMP, i.e., the 1,024 cells in the S-zone whose code positions are from DC00 through DFFF.
3. All cells in the high-half zone and the low-half zone shall be permanently reserved for the use of the UTF-16 coded representation form.
4. In UTF-16, any UCS character from the BMP shall be represented by its UCS-2 coded representation as specified by the body of this international standard.
5. In UTF-16, any UCS character whose UCS-4 coded representation is in the range 00010000 to 0010 FFFF shall be represented by a sequence of two RC-elements from the S-zone, of which the first is an RC-element from the high-half zone, and the second is an RCelement from the low-half zone.

The mapping between UCS-4 and UTF-16 for these characters shall be as shown in C.3; the reverse mapping is shown in C.4.

## C. 2 Notation

1. All numbers are in hexadecimal notation.
2. Double-octet boundaries in the notations for UTF-16 are indicated with semicolons.
3. The symbol "\%" indicates the modulo operation, e.g.: x \% y = x modulo y .
4. The symbol "/" indicates the integer division operation, e.g.: $7 / 3=2$.
5. Precedence is integer-division $>$ modulooperation > integer-multiplication > integeraddition.

## C. 3 Mapping from UCS-4 form to UTF-16 form

| UCS-4 (4-octet) | UTF-16, 2-octet elements |  |
| :--- | :--- | :--- |
| $x=$ | $00000000 .$. | $x \% 00010000 ;$ |
|  | 0000 FFFF (see Note 1) |  |
| $x=$ | 00010000. |  |
|  | 0010 FFFF |  |

where $\mathrm{y}=((\mathrm{x}-00010000) / 400)+$ D800
$z=((x-00010000) \% 400)+$ DC00
$x \quad 00110000$.. (no mapping
7FFF FFFF (is defined
NOTE 1 - Code positions from 0000 D800 to 0000 DFFF are reserved for the UTF-16 form and do not occur in UCS-4. The values 0000 FFFE and 0000 FFFF also do not occur (see clause 8). The mapping of these code positions in UTF-16 is undefined.

## Example:

The UCS-4 sequence [0000 0048] [0000 0069]
[0001 0000] [0000 0021] [0000 0021]
represents "Hi<0001 0000>!!".
It is mapped to UTF-16 as:
[0048] [0069] [D800] [DC00] [0021] [0021]
If interpreted as UCS-2 this sequence will be
"Hi<RC-element from high-half zone> <RC-element from low-half zone>!!"

## C. 4 Mapping from UTF-16 form to UCS-4 form

UTF-16, 2-octet elements UCS-4 (4-octet)
$x=0000 ;$.. D7FF; $\quad x$
$x=$ E000; .. FFFF; $\quad x$
pair ( $x, y$ ) such that

$$
\begin{array}{lll}
x= & \text { D800; .. DBFF; } & ((x-D 800) * 400 \\
y= & \text { DC00; .. DFFF; } & +(y-D C 00)) \\
& & +00010000
\end{array}
$$

## Example:

The UTF-16 sequence
[0048] [0069] [D800] [DC00] [0021] [0021]
is mapped to UCS-4 as
[0000 0048] [0000 0069] [0001 0000]
[0000 0021] [0000 0021]
and represents "Hi<0001 0000>!!".

## C. 5 Identification of UTF-16

When the escape sequences from ISO/IEC 2022 are used, the identification of UTF-16 and an implementation level (see clause 14) shall be by a designation sequence chosen from the following list:
ESC 02/05 02/15 04/10
UTF-16 with implementation level 1
ESC 02/05 02/15 04/11
UTF-16 with implementation level 2
ESC 02/05 02/15 04/12
UTF-16 with implementation level 3
If such an escape sequence appears within a CC-data-element conforming to ISO/IEC 2022, it shall consist only of the sequences of bit combinations as shown above.

If such an escape sequence appears within a CC-data-element conforming to ISO/IEC 10646, it shall be padded in accordance with clause 15.

When the escape sequences from ISO 2022 are used, the identification of a return, or transfer, from UTF-16 to the coding system of ISO 2022 shall be as specified in 16.5 for a return or transfer from UCS.

## C. 6 Unpaired RC-elements: Interpretation by receiving devices

According to C. 1 an unpaired RC-element (4.33) is not in conformance with the requirements of UTF-16.

If a receiving device that has adopted the UTF-16 form receives an unpaired RC-element because of error conditions either:

- in an originating device, or
- in the interchange between an originating and the receiving device, or
- in the receiving device itself,
then it shall interpret that unpaired RC-element in the same way that it interprets a character that is outside the adopted subset that has been identified for the device (see 2.3c).

NOTE 2 - Since a high-half RC-element followed by a lowhalf RC-element is a sequence that is in accordance with UTF-16, the only possible type of syntactically malformed sequence is an unpaired RC-element.

## Example:

A receiving/originating device which only handles the Latin- 1 repertoire, and uses boxes to display missing glyphs would display:
"The Greek letter <alpha> corresponds to<hieroglyphicHigh>."
as:
"The Greek letter <box> corresponds to<box>."
Accordingly a similar device that can also interpret a UTF-16 data stream should display an unpaired RCelement as a <box> also.

## C. 7 Receiving devices, advisory notes

When a receiving device interprets a CC-dataelement that is in accordance with UTF-16 the following advisory notes apply.

1. UTF-16 is designed to be compatible with the UCS-2 two-octet BMP Form (13.1). The highhalf and low-half zones are assigned to separate ranges of code positions, to which characters can never be assigned. Thus the function of every RC-element (two-octet unit) within a UTF-16 data stream is always immediately identifiable from its value, without regard to context.
For example, the valid UTF-16 sequence [0048] [0069] [D800] [DC00] [0021] [0021] may also be interpreted, by a receiving device, that has adopted only UCS-2, as the coded representation of
"Hi<unrecognized><unrecognized>!!"
This form of compatibility is possible because RC-elements from the S-zone are interpreted according to UTF-16 by receiving devices that have adopted UTF-16, and as unrecognized characters by receiving devices that have only adopted UCS-2. Consequently an originating device may transmit UTF-16 data even if the receiving device can only interpret that data as UCS-2 characters.
2. Designers of devices may choose to use UTF16 as an internal representation for processing or other purposes. There are two primary issues for such devices:

- Does the device interpret (i.e., process according to the assigned semantics) some subset of the pairs (high-half + low-half) of RCelements, e.g., render the pair as the intended single character?
- Does the device guarantee the integrity of every pair (high-half + low-half) of RC-elements, e.g., never separate such pairs in operations such as string truncation, insertion, or other modifications of the coded character sequence?
The decisions on these issues give rise to four possible combinations of capability in a device:
(U) UCS-2 implementations:
- Interpret no pairs.
- Do not guarantee integrity of pairs.
(W) Weak UTF-16 implementations:
- Interpret a non-null subset of pairs.
- Do not guarantee integrity of pairs.
(A) Aware UTF-16 implementations:
- Interpret no pairs.
- Guarantee integrity of pairs.
(S) Strong UTF-16 implementations:
- Interpret a non-null subset of pairs.
- Guarantee integrity of pairs.


## Example:

The following sentence could be displayed in three different ways, assuming that both the weak and strong implementations have Phoenician fonts but no hieroglyphics:
"The Greek letter <alpha> corresponds to<hieroglyphicHigh><hieroglyphicLow> and to <phoenicianHigh><phoenicianLow>."
U: "The Greek letter <alpha> corresponds to<box><box> and to <box><box>."

W: "The Greek letter <alpha> corresponds to <box><box> and to<Phoenician>."
A: "The Greek letter <alpha> corresponds to <box> and to<box>."

S: "The Greek letter <alpha> corresponds to <box> and to<Phoenician>."

# Annex D <br> (normative) 

## UCS Transformation Format 8 (UTF-8)

UTF-8 is an alternative coded representation form for all of the characters of the UCS. It can be used to transmit text data through communication systems which assume that individual octets in the range 00 to 7 F have a definition according to ISO/IEC 4873, including a CO set of control functions according to the 8-bit structure of ISO/IEC 2022. UTF-8 also avoids the use of octet values in this range which have special significance during the parsing of filename character strings in widely-used file-handling systems.

The number of octets in the UTF-8 coded representation of the characters of the UCS ranges from one to six; the value of the first octet indicates the number of octets in that coded representation.

## D. 1 Features of UTF-8

- UCS characters from the BASIC LATIN collection are represented in UTF-8 in accordance with ISO/IEC 4873, i.e. single octets with values ranging from 20 to 7 E .
- Control functions in positions 00000000 to 0000 001F, and the DELETE character in position 0000 007F, are represented without the padding octets specified in clause 15 , i.e. as single octets with values ranging from 00 to 1 F , and 7 F respectively in accordance with ISO/IEC 4873 and with the 8-bit structure of ISO/IEC 2022.
- Octet values 00 to 7 F do not otherwise occur in the UTF-8 coded representation of any character. This provides compatibility with existing filehandling systems and communications subsystems which parse CC-data-elements for these octet values.
- The first octet in the UTF-8 coded representation of any character can be directly identified when a CC-data-element is examined, one octet at a time, starting from an arbitrary location. It indicates the number of continuing octets (if any) in the multi-octet sequence that constitutes the coded representation of that character.


## D. 2 Specification of UTF-8

In the UTF-8 coded representation form each character from this International Standard shall have a coded representation that comprises a sequence of octets of length 1, 2, 3, 4, 5, or 6 octets.

For all sequences of one octet the most significant bit shall be a ZERO bit.

For all sequences of more than one octet, the number of ONE bits in the first octet, starting from the most significant bit position, shall indicate the number of octets in the sequence. The next most significant bit shall be a ZERO bit.

NOTE 1 - For example, the first octet of a 2 -octet sequence has bits 110 in the most significant positions, and the first octet of a 6 -octet sequence has bits 1111110 in the most significant positions.

All of the octets, other than the first in a sequence, are known as continuing octets. The two most significant bits of a continuing octet shall be a ONE bit followed by a ZERO bit.

The remaining bit positions in the octets of the sequence shall be "free bit positions" that are used to distinguish between the characters of this International Standard. These free bit positions shall be used, in order of increasing significance, for the bits of the UCS-4 coded representation of the character, starting from its least significant bit. Some of the high-order ZERO bits of the UCS-4 representation shall be omitted, as specified below.

Table D. 1 below shows the format of the octets of a coded character according to UTF-8. Each free bit position available for distinguishing between the characters is indicated by an $x$. Each entry in the column "Maximum UCS-4 value" indicates the upper end of the range of coded representations from UCS-4 that may be represented in a UTF-8 sequence having the length indicated in the "Octet usage" column.

| Table D.1 | Format of octets in a UTF-8 sequence |
| :--- | :---: | :---: | :---: |

Table D. 1 shows that, in a CC-data-element conforming to UTF-8, the range of values for each octet indicates its usage as follows:

00 to 7 F first and only octet of a sequence;
80 to BF continuing octet of a multi-octet sequence;

C0 to FD first octet of a multi-octet sequence;
FE or FF not used.
The mapping between UCS-4 and UTF-8 shall be as shown in D.4; the reverse mapping is shown in D.5.

NOTE 2 - Examples of UCS-4 coded representations and the corresponding UTF-8 coded representations are shown in Tables D. 2 and D. 3.

Table D. 2 shows the UCS-4 and the UTF-8 coded representations, in binary notation, for a selection of code positions from the UCS.
Table D. 3 shows the UCS-4 and the UTF-8 coded representations, in hexadecimal notation, for the same selection of code positions from the UCS.


Table D. 2 - Examples in binary notation

## Four-octet form - UCS-4

00000000000000000000000000000001 ;
000000000000000000000000 01111111;
000000000000000000000000 10000000;
000000000000000000000111 11111111; 00000000000000000000100000000000 ; 000000000000000011111111 11111111; 00000000000000010000000000000000 ; 000000000001111111111111 11111111; 00000000001000000000000000000000 ; 000000111111111111111111 11111111; 00000100000000000000000000000000 ; 011111111111111111111111 11111111;

## UTF-8 form

```
00000001;
```

01111111;
11000010; 10000000;
11011111; 10111111;
11100000; 10100000; 10000000;
11101111; 10111111; 10111111;
11110000; 10010000; 10000000;10000000; 11110111; 10111111; 10111111;10111111; 11111000; 10001000; 10000000;10000000; 10000000; 11111011; 10111111; 10111111;10111111; 10111111; 11111100; 10000100; 10000000;10000000; 10000000; 10000000; 11111101; 10111111; 10111111;10111111; 10111111; 10111111;

## D. 3 Notation

1. All numbers are in hexadecimal notation, except for the decimal numbers used in the power-of operation (see 5 below).
2. Boundaries of code elements are indicated with semicolons; these are single-octet boundaries within UTF-8 coded representations, and fouroctet boundaries within UCS-4 coded representations.
3. The symbol "\%" indicates the modulo operation, e.g.: $x \% y=x$ modulo $y$
4. The symbol "/" indicates the integer division operation, e.g.: $7 / 3=2$
5. Superscripting indicates the power-of operation, e.g.: $2^{3}=8$
6. Precedence is: power-of operation > integer division > modulo operation > integer multiplication > integer addition.
e.g.: $x / y^{z} \% \mathrm{w}=\left(\left(\mathrm{x} /\left(\mathrm{y}^{\mathrm{z}}\right)\right) \% \mathrm{w}\right)$

## D. 4 Mapping from UCS-4 form to UTF-8 form

Table D. 4 defines in mathematical notation the mapping from the UCS-4 coded representation form to the UTF-8 coded representation form.

In the left column (UCS-4) the notation $x$ indicates the four-octet coded representation of a single character of the UCS. In the right column (UTF-8) x indicates the corresponding integer value.

NOTE 3 - Values of $x$ in the range 0000 D800 .. 0000 DFFF are reserved for the UTF-16 form and do not occur in UCS4. The values 0000 FFFE and 0000 FFFF also do not occur (see clause 8). The mappings of these code positions in UTF-8 are undefined.

NOTE 4 - The algorithm for converting from UCS-4 to UTF-8 can be summarised as follows.
For each coded character in UCS-4 the length of octet sequence in UTF-8 is determined by the entry in the right column of Table D.1. The bits in the UCS-4 coded representation, starting from the least significant bit, are then distributed across the free bit positions in order of increasing significance until no more free bit positions are available.
Table D. 4 - Mapping from UCS-4 to UTF-8

| Range of values <br> in UCS-4 | Sequence of <br> octets in UTF-8 |
| :--- | :--- |

```
x = 0000 0000 .. 0000 007F; x;
x = 0000 0080 .. 0000 07FF; C0 +x / 26;
    80+x %26;
x = 0000 0800 .. 0000 FFFF; EO + x/2 12;
    (see Note 3) 80+x/26% % 6;
    80+x%26;
x=00010000 .. 001F FFFF; F0 + x/2 18;
80+x/2 12%/2}\mp@subsup{}{}{6}
80+x/26%%26;
80+x%2}\mp@subsup{}{}{6
x = 0020 0000 .. 03FF FFFF;
F8 + x/2 24;
80+x/218%%2
80+x/2 12%%26;
80+x/26% % }\mp@subsup{}{}{6
80 + x%26;
x = 0400 0000 .. 7FFF FFFF;
```


## D. 5 Mapping from UTF-8 form to UCS-4 form

Table D. 5 defines in mathematical notation the mapping from the UTF-8 coded representation form to the UCS-4 coded representation form.
In the left column (UTF-8) the following notations apply:
$z$ is the first octet of a sequence. Its value determines the number of continuing octets in the sequence.
$y$ is the 2nd octet in the sequence.
$x$ is the 3rd octet in the sequence.
w is the 4th octet in the sequence.
$v$ is the 5th octet in the sequence.
u is the 6th octet in the sequence.
The ranges of values applicable to these octets are shown in D. 2 above, following Table D.1.

NOTE 5- The algorithm for converting from UTF-8 to UCS-4 can be summarised as follows.
For each coded character in UTF-8 the bits in the free bit positions are concatenated as a bit-string. The bits from this string, in increasing order of significance, are then distributed across the bit positions of a four-octet sequence, starting from the least significant bit position. The remaining bit positions of that sequence are filled with ZERO bits.

Table D. 5 - Mapping from UTF-8 to UCS-4

| Sequence of <br> octets in UTF-8 | Four-octet <br> sequences in UCS-4 |
| :--- | :--- |

$z=00$.. $7 F ; \quad z$
$z=C 0 . . D F ; y ;$
$(z-C 0) * 2^{6}+(y-80)$;
$z=E 0$.. EF; $y ; x ;$
$(z-E 0))^{12}+(y-80) 2^{6}+(x-80)$;
$z=F 0$.. F7; $y ; x ; w ;$
$(z-F 0) * 2^{18}+(y-80) * 2^{12}+(x-80) * 2^{6}+(w-80) ;$
$z=F 8$..FB; $y ; x ; w ;$
$(z-F 8) * 2^{24}+(y-80) * 2^{18}+(x-80) * 2^{12}+(w-80) 2^{6}+(v-80) ;$
$z=$ FC, FD; $y ; x ; w ; v ; u ; \quad(z-F C) * 2^{30}+(y-80) 2^{24}+(x-80) 2^{18}+(w-80) * 2^{12}+(v-80) 2^{6}+(u-80) ;$

## D. 6 Identification of UTF-8

When the escape sequences from ISO/IEC 2022 are used, the identification of UTF-8 and an implementation level (see clause 14) shall be by a designation sequence chosen from the following list:
ESC 02/05 02/15 04/07
UTF-8 with implementation level 1
ESC 02/05 02/15 04/08
UTF-8 with implementation level 2
ESC 02/05 02/15 04/09
UTF-8 with implementation level 3
If such an escape sequence appears within a CC-data-element conforming to ISO/IEC 2022, it shall consist only of the sequences of bit combinations as shown above.

If such an escape sequence appears within a CC-data-element conforming to ISO/IEC 10646, it shall be padded in accordance with clause 15.
When the escape sequences from ISO/IEC 2022 are used, the identification of a return, or transfer, from UTF-8 to the coding system of ISO/IEC 2022 shall be as specified in 16.5 for a return or transfer from UCS.

NOTE 6 - The following escape sequence may also be used: ESC 02/05 04/07 UTF-8.
The implementation level is not defined. The escape sequence used for a return to the coding system of ISO/IEC 2022 is not padded as specified in 16.5.

## D. 7 Incorrect sequences of octets: Interpretation by receiving devices

According to D. 2 an octet in the range 00 .. 7F or C0 .. FB is the first octet of a UTF-8 sequence, and is followed by the appropriate number (from 0 to 5) of continuing octets in the range 80 .. BF. Furthermore, octets whose value is FE or FF are not used; thus they are invalid in UTF-8.
If a CC-data-element includes either:

- a first octet that is not immediately followed by the correct number of continuing octets, or
- one or more continuing octets that are not required to complete a sequence of first and continuing octets, or
- an invalid octet,
then according to D .2 such a sequence of octets is not in conformance with the requirements of UTF-8. It is known as a malformed sequence.
If a receiving device that has adopted the UTF-8 form receives a malformed sequence, because of error conditions either:
- in an originating device, or
- in the interchange between an originating and a receiving device, or
- in the receiving device itself,
then it shall interpret that malformed sequence in the same way that it interprets a character that is outside the adopted subset that has been identified for the device (see 2.3c).


## Annex E <br> (informative)

## Mirrored characters in Arabic bi-directional context

In the context of Arabic right-to-left (bi-directional) text, the following characters have semantic meaning. To preserve the meaning in right-to-left text, the graphic symbol representing the character may be rendered as the mirror image of the associated graphical symbol from the left-to-right context. These characters include mathematical symbols and paired characters such as the SQUARE BRACKETS. For example, in a right-to-left text segment, the GREATER-THAN SIGN (rendered as ">" in left-to-right text) may be rendered as the "<" graphic symbol.

| 0028 | LEFT PARENTHESIS |
| :---: | :---: |
| 0029 | RIGHT PARENTHESIS |
| 003C | LESS-THAN SIGN |
| 003E | GREATER-THAN SIGN |
| 005B | LEFT SQUARE BRACKET |
| 005D | RIGHT SQUARE BRACKET |
| 007B | LEFT CURLY BRACKET |
| 007D | RIGHT CURLY BRACKET |
| 00AB | LEFT-POINTING DOUBLE ANGLE QUOTATION MARK |
| 00BB | RIGHT-POINTING DOUBLE ANGLE QUOTATION |
|  | MARK |
| 2039 | SINGLE LEFT-POINTING ANGLE QUOTATION MARK |
| 203A | SINGLE RIGHT-POINTING ANGLE QUOTATION MARK |
| 2045 | LEFT SQUARE BRACKET WITH QUILL |
| 2046 | RIGHT SQUARE BRACKET WITH QUILL |
| 207D | SUPERSCRIPT LEFT PARENTHESIS |
| 207E | SUPERSCRIPT RIGHT PARENTHESIS |
| 208D | SUBSCRIPT LEFT PARENTHESIS |
| 208E | SUBSCRIPT RIGHT PARENTHESIS |
| 2201 | COMPLEMENT |
| 2202 | PARTIAL DIFFERENTIAL |
| 2203 | THERE EXISTS |
| 2204 | THERE DOES NOT EXIST |
| 2208 | ELEMENT OF |
| 2209 | NOT AN ELEMENT OF |
| 220A | SMALL ELEMENT OF |
| 220B | CONTAINS AS MEMBER |
| 220C | DOES NOT CONTAIN AS MEMBER |
| 220D | SMALL CONTAINS AS MEMBER |
| 2211 | N-ARY SUMMATION |
| 2215 | DIVISION SLASH |
| 2216 | SET MINUS |
| 221A | SQUARE ROOT |
| 221B | CUBE ROOT |
| 221 C | FOURTH ROOT |
| 221D | PROPORTIONAL TO |
| 221F | RIGHT ANGLE |
| 2220 | ANGLE |

2221
2222
2224
2226
222B
222C
222D
222E
222F
2230
2231
2232
2233
2239
223B
223C
223D
223E
223F
2240
2241
2242
2243
2244
2245
2246

```
MEASURED ANGLE SPHERICAL ANGLE
DOES NOT DIVIDE
NOT PARALLEL TO
INTEGRAL
DOUBLE INTEGRAL
TRIPLE INTEGRAL
CONTOUR INTEGRAL
SURFACE INTEGRAL
VOLUME INTEGRAL
CLOCKWISE INTEGRAL
CLOCKWISE CONTOUR INTEGRAL
ANTICLOCKWISE CONTOUR INTEGRAL
EXCESS
HOMOTHETIC
TILDE OPERATOR
REVERSED TILDE
INVERTED LAZY S
SINE WAVE
WREATH PRODUCT
NOT TILDE
MINUS TILDE
ASYMPTOTICALLY EQUAL TO
NOT ASYMPTOTICALLY EQUAL TO
APPROXIMATELY EQUAL TO
APPROXIMATELY BUT NOT ACTUALLY EQUAL TO
NEITHER APPROXIMATELY NOR ACTUALLY
EQUAL TO
ALMOST EQUAL TO
NOT ALMOST EQUAL TO
ALMOST EQUAL OR EQUAL TO
TRIPLE TILDE
ALL EQUAL TO
APPROXIMATELY EQUAL TO OR THE IMAGE OF
IMAGE OF OR APPROXIMATELY EQUAL TO
COLON EQUALS
EQUALS COLON
QUESTIONED EQUAL TO
NOT EQUAL TO
NOT IDENTICAL TO
LESS-THAN OR EQUAL TO
GREATER-THAN OR EQUAL TO
LESS-THAN OVER EQUAL TO
GREATER-THAN OVER EQUAL TO
LESS-THAN BUT NOT EQUAL TO
GREATER-THAN BUT NOT EQUAL TO
MUCH LESS-THAN
MUCH GREATER-THAN
NOT LESS-THAN
NOT GREATER-THAN
NEITHER LESS-THAN NOR EQUAL TO
NEITHER GREATER-THAN NOR EQUAL TO
LESS-THAN OR EQUIVALENT TO
GREATER-THAN OR EQUIVALENT TO
NEITHER LESS-THAN NOR EQUIVALENT TO
NEITHER GREATER-THAN NOR EQUIVALENT TO
```

| 2276 | LESS-THAN OR GREATER-THAN | 22CC |
| :---: | :---: | :---: |
| 2277 | GREATER-THAN OR LESS-THAN | 22CD |
| 2278 | NEITHER LESS-THAN NOR GREATER-THAN | 22D0 |
| 2279 | NEITHER GREATER-THAN NOR LESS-THAN | 22D1 |
| 227A | PRECEDES | 22D6 |
| 227B | SUCCEEDS | 22 D 7 |
| 227C | PRECEDES OR EQUAL TO | 22D8 |
| 227D | SUCCEEDS OR EQUAL TO | $22 \mathrm{D9}$ |
| 227E | PRECEDES OR EQUIVALENT TO | 22DA |
| 227F | SUCCEEDS OR EQUIVALENT TO | 22DB |
| 2280 | DOES NOT PRECEDE | 22DC |
| 2281 | DOES NOT SUCCEED | 22 DD |
| 2282 | SUBSET OF | 22DE |
| 2283 | SUPERSET OF | 22DF |
| 2284 | NOT A SUBSET OF | 22E0 |
| 2285 | NOT A SUPERSET OF | 22E1 |
| 2286 | SUBSET OF OR EQUAL TO | 22E2 |
| 2287 | SUPERSET OF OR EQUAL TO | 22E3 |
| 2288 | NEITHER A SUBSET OF NOR EQUAL TO | 22E4 |
| 2289 | NEITHER A SUPERSET OF NOR EQUAL TO | 22E5 |
| 228A | SUBSET OF WITH NOT EQUAL TO | 22E6 |
| 228B | SUPERSET OF WITH NOT EQUAL TO | 22E7 |
| 228 C | MULTISET | 22E8 |
| 228F | SQUARE IMAGE OF | 22E9 |
| 2290 | SQUARE ORIGINAL OF | 22EA |
| 2291 | SQUARE IMAGE OF OR EQUAL TO | 22EB |
| 2292 | SQUARE ORIGINAL OF OR EQUAL TO | 22EC |
| 2298 | CIRCLED DIVISION SLASH | 22ED |
| 22A2 | RIGHT TACK |  |
| 22A3 | LEFT TACK | 22F0 |
| 22A6 | ASSERTION | 22F1 |
| 22A7 | MODELS | 2308 |
| 22A8 | TRUE | 2309 |
| 22A9 | FORCES | 230A |
| 22AA | TRIPLE VERTICAL BAR TURNSTILE | 230B |
| 22AB | DOUBLE VERTICAL BAR DOUBLE RIGHT | 2320 |
|  | TURNSTILE | 2321 |
| 22AC | DOES NOT PROVE | 2329 |
| 22AD | NOT TRUE | 232A |
| 22AE | DOES NOT FORCE | 3008 |
| 22AF | NEGATED DOUBLE VERTICAL BAR DOUBLE | 3009 |
|  | RIGHT TURNSTILE | 300A |
| 22B0 | PRECEDES UNDER RELATION | 300B |
| 22B1 | SUCCEEDS UNDER RELATION | 300 C |
| 22B2 | NORMAL SUBGROUP OF | 300D |
| 22B3 | CONTAINS AS NORMAL SUBGROUP | 300E |
| 22B4 | NORMAL SUBGROUP OF OR EQUAL TO | 300F |
| 22B5 | CONTAINS AS NORMAL SUBGROUP OR EQUAL | 3010 |
|  | TO | 3011 |
| 22B6 | ORIGINAL OF | 3014 |
| 22 B 7 | IMAGE OF | 3015 |
| 22B8 | MULTIMAP | 3016 |
| 22BE | RIGHT ANGLE WITH ARC | 3017 |
| 22BF | RIGHT TRIANGLE | 3018 |
| 22C9 | LEFT NORMAL FACTOR SEMIDIRECT PRODUCT | 3019 |
| 22CA | RIGHT NORMAL FACTOR SEMIDIRECT PRODUCT | 301A |
| 22CB | LEFT SEMIDIRECT PRODUCT | 301B |

RIGHT SEMIDIRECT PRODUCT
REVERSE TILDE EQUALS
DOUBLE SUBSET
DOUBLE SUPERSET
LESS-THAN WITH DOT
GREATER-THAN WITH DOT
VERY MUCH LESS-THAN
VERY MUCH GREATER-THAN
LESS-THAN EQUAL TO OR GREATER-THAN
GREATER-THAN EQUAL TO OR LESS-THAN
EQUAL TO OR LESS-THAN
EQUAL TO OR GREATER-THAN
EQUAL TO OR PRECEDES
EQUAL TO OR SUCCEEDS
DOES NOT PRECEDE OR EQUAL
DOES NOT SUCCEED OR EQUAL
NOT SQUARE IMAGE OF OR EQUAL TO
NOT SQUARE ORIGINAL OF OR EQUAL TO
SQUARE IMAGE OF OR NOT EQUAL TO
SQUARE ORIGINAL OF OR NOT EQUAL TO
LESS-THAN BUT NOT EQUIVALENT TO
GREATER-THAN BUT NOT EQUIVALENT TO
PRECEDES BUT NOT EQUIVALENT TO
SUCCEEDS BUT NOT EQUIVALENT TO
NOT NORMAL SUBGROUP OF
DOES NOT CONTAIN AS NORMAL SUBGROUP
NOT NORMAL SUBGROUP OF OR EQUAL TO
DOES NOT CONTAIN AS NORMAL SUBGROUP OR EQUAL
UP RIGHT DIAGONAL ELLIPSIS
DOWN RIGHT DIAGONAL ELLIPSIS
LEFT CEILING
RIGHT CEILING
LEFT FLOOR
RIGHT FLOOR
TOP HALF INTEGRAL
BOTTOM HALF INTEGRAL
LEFT-POINTING ANGLE BRACKET
RIGHT-POINTING ANGLE BRACKET
LEFT ANGLE BRACKET
RIGHT ANGLE BRACKET
LEFT DOUBLE ANGLE BRACKET
RIGHT DOUBLE ANGLE BRACKET
LEFT CORNER BRACKET
RIGHT CORNER BRACKET
LEFT WHITE CORNER BRACKET
RIGHT WHITE CORNER BRACKET
LEFT BLAOK LENTICULAR BRACKET
RIGHT BLACK LENTICULAR BRACKET
LEFT TORTOISE SHELL BRACKET
RIGHT TORTOISE SHELL BRACKET
LEFT WHITE LENTICULAR BRACKET
RIGHT WHITE LENTICULAR BRACKET
LEFT WHITE TORTOISE SHELL BRACKET
RIGHT WHITE TORTOISE SHELL BRACKET
LEFT WHITE SQUARE BRACKET
RIGHT WHITE SQUARE BRACKET

## Annex F <br> (informative)

## Alternate format characters

There is a special class of characters called Alternate Format Characters which are included for compatibility with some industry practices. These characters do not have printable graphic symbols, and are thus represented in the character code tables by dotted boxes.
The function of most of these characters is to indicate the correct presentation of a sequence of characters. For any text processing other than presentation (such as sorting and searching), the alternate format characters, except for ZWJ and ZWNJ described in F.1.1, can be ignored by filtering them out. The alternate format characters are not intended to be used in conjunction with bi-directional control functions from ISO/IEC 6429.
There are collections of graphic characters for selected subsets which consist of Alternate Format Characters (see annex A).

## F. 1 General format characters

## F.1.1 Zero-width boundary indicators

The following characters are used to indicate whether or not the adjacent characters are separated by a word boundary. Each of these zero-width boundary indicators has no width in its own presentation.
ZERO WIDTH SPACE (200B): This character behaves like a SPACE in that it indicates a word boundary, but unlike SPACE it has no presentational width. For example, this character could be used to indicate word boundaries in Thai, which does not use visible gaps to separate words.
ZERO WIDTH NO-BREAK SPACE (FEFF): This character behaves like a NO-BREAK SPACE in that it indicates the absence of word boundaries, but unlike NO-BREAK SPACE it has no presentational width. For example, this character could be inserted after the fourth character in the text "base+delta" to indicate that there is to be no word break between the "e" and the " + ".

NOTE - For additional usages of this character for "signature", see annex H.
The following characters are used to indicate whether or not the adjacent characters are joined together in rendering (cursive joiners).

ZERO WIDTH NON-JOINER (200C): This character indicates that the adjacent characters are not joined together in cursive connection even when they would normally join together as cursive letter forms. For example, ZERO WIDTH NON-JOINER between ARABIC LETTER NOON and ARABIC LETTER MEEM indicates that the characters are not rendered with the normal cursive connection.

ZERO WIDTH JOINER (200D): This character indicates that the adjacent characters are represented with joining forms in cursive connection even when they would not normally join together as cursive letter forms. For example, in the sequence SPACE followed by ARABIC LETTER BEH followed by SPACE, ZERO WIDTH JOINER can be inserted between the first two characters to display the final form of the ARABIC LETTER BEH.

## F.1.2 Format separators

The following characters are used to indicate formatting boundaries between lines or paragraphs.
LINE SEPARATOR (2028): This character indicates where a new line starts; although the text continues to the next line, it does not start a new paragraph; e.g. no inter-paragraph indentation might be applied.

PARAGRAPH SEPARATOR (2029): This character indicates where a new paragraph starts; e.g. the text continues on the next line and inter-paragraph line spacing or paragraph indentation might be applied.

## F.1.3 Bi-directional text formatting

The following characters are used in formatting bidirectional text. If the specification of a subset includes these characters, then text containing right-to-left characters are to be rendered with an implicit bi-directional algorithm.
An implicit algorithm uses the directional character properties to determine the correct display order of characters on a horizontal line of text.

The following characters are format characters that act exactly like right-to-left or left-to-right characters in terms of affecting ordering (Bi-directional format marks). They have no visible graphic symbols, and they do not have any other semantic effect.

Their use can be more convenient than the explicit embeddings or overrides, since their scope is more local.

LEFT-TO-RIGHT MARK (200E): In bi-directional formatting, this character acts like a left-to-right character (such as LATIN SMALL LETTER A). RIGHT-TO-LEFT MARK (200F): In bi-directional formatting, this character acts like a right-to-left character (such as ARABIC LETTER NOON).

The following format characters indicate that a piece of text is to be treated as embedded, and is to have a particular ordering attached to it (Bi-directional format embeddings). For example, an English quotation in the middle of an Arabic sentence can be marked as being an embedded left-to-right string. These format characters nest in blocks, with the embedding and override characters initiating (pushing) a block, and the pop character terminating (popping) a block.

The function of the embedding and override characters are very similar; the main difference is that the embedding characters specify the implicit direction of the text, while the override characters specify the explicit direction of the text. When text has an explicit direction, the normal directional character properties are ignored, and all of the text is assumed to have the ordering direction determined by the override character.

LEFT-TO-RIGHT EMBEDDING (202A): This character is used to indicate the start of a left-to-right implicit embedding.
RIGHT-TO-LEFT EMBEDDING (202B): This character is used to indicate the start of a right-to-left implicit embedding.
LEFT-TO-RIGHT OVERRIDE (202D): This character is used to indicate the start of a left-to-right explicit embedding.
RIGHT-TO-LEFT OVERRIDE (202E): This character is used to indicate the start of a right-to-left explicit embedding.

POP DIRECTIONAL FORMATTING (202C): This character is used to indicate the termination of an implicit or explicit directional embedding initiated by the above characters.

## F.1.4 Other boundary indicators

NARROW NO-BREAK SPACE (202F): This character is a non-breaking space. It is similar to OOAO NO-BREAK SPACE, except that it is rendered with a narrower width. When used with the Mongolian script this character is usually rendered at one-third of the width of a normal space, and it separates a suffix from the Mongolian word-stem. This allows for the normal rules of Mongolian
character shaping to apply, while indicating that there is no word boundary at that position.

## F. 2 Script-specific format characters

F.2.1 Hangul fill characters

The following format characters have a special usage for Hangul characters.

HANGUL FILLER (3164): This character represents the fill value used with the standard spacing Jamos.
HALFWIDTH HANGUL FILLER (FFAO): As with the other halfwidth characters, this character is included for compatibility with certain systems that provide halfwidth forms of characters.

## F.2.2 Symmetric swapping format characters

The following characters are used in conjunction with the class of left/right handed pairs of characters listed in clause 19. The following format characters indicate whether the interpretation of the term LEFT or RIGHT in the character names is OPENING or CLOSING respectively. The following characters do not nest.

The default state of interpretation may be set by a higer level protocol or standard, such as ISO/IEC 6429. In the absence of such a protocol, the default state is as established by ACTIVATE SYMMETRIC SWAPPING.
INHIBIT SYMMETRIC SWAPPING (206A): Between this character and the following ACTIVATE SYMMETRIC SWAPPING format character (if any), the stored characters listed in clause 19 are interpreted and rendered as LEFT and RIGHT, and the processing specified in that clause is not performed.
ACTIVATE SYMMETRIC SWAPPING (206B):
Between this character and the following INHIBIT SYMMETRIC SWAPPING format character (if any), the stored characters listed in clause 19 are interpreted and rendered as OPENING and CLOSING characters as specified in that clause.

## F.2.3 Character shaping selectors

The following characters are used in conjunction with Arabic presentation forms. During the presentation process, certain characters may be joined together in cursive connection or ligatures. The following characters indicate that the character shape determination process used to achieve this presentation effect is either activated or inhibited. The following characters do not nest.
INHIBIT ARABIC FORM SHAPING (206C): Between this character and the following ACTIVATE ARABIC FORM SHAPING format character (if any), the character shaping determination process is inhibited. The stored Arabic presentation forms are
presented without shape modification. This is the default state.

ACTIVATE ARABIC FORM SHAPING (206D): Between this character and the following INHIBIT ARABIC FORM SHAPING format character (if any), the stored Arabic presentation forms are presented with shape modification by means of the character shaping determination process.

NOTE - These characters have no effect on characters that are not presentation forms: in particular, Arabic nominal characters as from 0600 to 06FF are always subject to character shaping, and are unaffected by these formatting characters.

## F.2.4 Numeric shape selectors

The following characters allow the selection of the shapes in which the digits from 0030 to 0039 are rendered. The following characters do not nest. NATIONAL DIGIT SHAPES (206E): Between this character and the following NOMINAL DIGIT SHAPES format character (if any), digits from 0030 to 0039 are rendered with the appropriate national digit shapes as specified by means of appropriate agreements. For example, they could be displayed with shapes such as the ARABIC-INDIC digits from 0660 to 0669.

NOMINAL DIGIT SHAPES (206F): Between this character and the following NATIONAL DIGIT SHAPES format character (if any), the digits from 0030 to 0039 are rendered with the shapes as those shown in the code tables for those digits. This is the default state.

## F.2.5 Mongolian shaping selectors

The following characters are used in conjunction with the letters in the Mongolian script.

## mongolian free variation selector one (180B):

MONGOLIAN FREE VARIATION SELECTOR TWO (180C):
MONGOLIAN FREE VARIATION SELECTOR THREE (180D):
A Mongolian Free Variation Selector character may immediately follow another character from the Mongolian collection to indicate a specific variant form of graphic symbol for that character, when the appropriate variant cannot be determined from the context. For each Mongolian character the number of variant forms that it can take is predetermined within each context. This number does not exceed three for any character.

MONGOLIAN VOWEL SEPARATOR (180E): This character may be used between the Mongolian letter A or the Mongolian letter E and the preceding consonant letter. It indicates a special form of the graphic symbol for the letter A or E and the
preceding consonant. When rendered in visible form it is generally shown as a narrow space between the letters, but it may sometimes be shown as a distinct graphic symbol to assist the user.

## F. 3 Ideographic description characters

An Ideographic Description Character (IDC) is a graphic character, which is used with a sequence of other graphic characters to form an Ideographic Description Sequence (IDS). Such a sequence may be used to describe an ideographic character which is not specified within this International Standard.
The IDS describes the ideograph in the abstract form. It is not interpreted as a composed character and does not imply any specific form of rendering.

## F.3.1 Syntax of an ideographic description sequence

An IDS consists of an IDC followed by a fixed number of Description Components (DC). A DC may be any one of the following :

- a coded ideograph
- a coded radical
- another IDS

NOTE - The above description implies that any IDS may be nested within another IDS.

Each IDC has four properties as summarized in Table-F. 1 below.

- the number of DCs used in the IDS that commences with that IDC,
- the definition of its acronym,
- the syntax of the corresponding IDS,
- the relative positions of the DCs in the visual representation of the ideograph that is being described in its abstract form.

The syntax of the IDS introduced by each IDC is indicated in the "IDS Acronym and Syntax" column of the table by the abbreviated name of the IDC (e.g. IDC-LTR) followed by the corresponding number of DCs, i.e. ( $D_{1} D_{2}$ ) or ( $D_{1} D_{2} D_{3}$ ).

## F.3.2 Individual definitions of the ideographic description characters <br> IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO RIGHT (2FFO):

The IDS introduced by this character describes the abstract form of the ideograph with $D_{1}$ on the left and $\mathrm{D}_{2}$ on the right.

## IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO BELOW (2FF1):

The IDS introduced by this character describes the abstract form of the ideograph with $D_{1}$ above $D_{2}$.

## IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO MIDDLE AND RIGHT (2FF2):

The IDS introduced by this character describes the abstract form of the ideograph with $D_{1}$ on the left of $D_{2}$, and $D_{2}$ on the left of $D_{3}$.

## IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO MIDDLE AND BELOW (2FF3):

The IDS introduced by this character describes the abstract form of the ideograph with $D_{1}$ above $D_{2}$, and $D_{2}$ above $D_{3}$.

## IDEOGRAPHIC DESCRIPTION CHARACTER FULL SURROUND (2FF4):

The IDS introduced by this character describes the abstract form of the ideograph with D surrounding $\mathrm{D}_{2}$.

## IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM ABOVE (2FF5):

The IDS introduced by this character describes the abstract form of the ideograph with $D_{1}$ above $D_{2}$, and surrounding $D_{2}$ on both sides.

## IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM BELOW (2FF6):

The IDS introduced by this character describes the abstract form of the ideograph with $D_{1}$ below $D_{2}$, and surrounding $D_{2}$ on both sides

## IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LEFT (2FF7):

The IDS introduced by this character describes the abstract form of the ideograph with $D_{1}$ on the left of $D_{2}$, and surrounding $D_{2}$ above and below.

## IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER LEFT (2FF8):

The IDS introduced by this character describes the abstract form of the ideograph with $D_{1}$ at the top left corner of $D_{2}$, and partly surrounding $D_{2}$ above and to the left.

## IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER RIGHT (2FF9):

The IDS introduced by this character describes the abstract form of the ideograph with $D_{1}$ at the top
right corner of $D_{2}$, and partly surrounding $D_{2}$ above and to the right.

## IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LOWER LEFT (2FFA):

The IDS introduced by this character describes the abstract form of the ideograph with $D_{1}$ at the bottom left corner of $D_{2}$, and partly surrounding $D_{2}$ below and to the left.

## IDEOGRAPHIC DESCRIPTION CHARACTER OVERLAID (2FFB):

The IDS introduced by this character describes the abstract form of the ideograph with $D_{1}$ and $D_{2}$ overlaying each other.

## F. 4 Interlinear annotation characters

The following characters are used to indicate that an identified character string (the annotation string) is regarded as providing an annotation for another identified character string (the base string).

## INTERLINEAR ANNOTATION ANCHOR (FFF9):

This character indicates the beginning of the base string.

## INTERLINEAR ANNOTATION SEPARATOR

(FFFA): This character indicates the end of the base string and the beginning of the annotation string.

INTERLINEAR ANNOTATION TERMINATOR
(FFFB): This character indicates the end of the annotation string.

The relationship between the annotation string and the base string is defined by agreement between the user of the originating device and the user of the receiving device. For example, if the base string is rendered in a visible form the annotation string may be rendered on a different line from the base string, in a position close to the base string.

If the interlinear annotation characters are filtered out during processing, then all characters between the Interlinear Annotation Separator and the Interlinear Annotation Terminator should also be filtered out.

Table F．1：Properties of ideographic description characters

| Character Name： IDEOGRAPHIC DESCRIPTION CHARACTER ．．． | $\begin{gathered} \text { no. of } \\ \text { DCs } \end{gathered}$ | IDS Acronym and Syntax | Relative positions of DCs | Example of IDS | IDS example represents： |
| :---: | :---: | :---: | :---: | :---: | :---: |
| LEFT TO RIGHT | 2 | IDC－LTR $\mathrm{D}_{1} \mathrm{D}_{2}$ |  | ¢ 产 | 付 |
| ABOVE TO BELOW | 2 | IDC－ATB $\mathrm{D}_{1} \mathrm{D}_{2}$ |  |  | $\xrightarrow{x}$ |
| LEFT TO MIDDLE AND RIGHT | 3 | IDC－LMR $\mathrm{D}_{1} \mathrm{D}_{2} \mathrm{D}_{3}$ |  | 隹 言 च | 信 |
| ABOVE TO MIDDLE AND BELOW | 3 | IDC－AMB $\mathrm{D}_{1} \mathrm{D}_{2} \mathrm{D}_{3}$ | $:$ | —上 | 替 |
| FULL SURROUND | 2 | IDC－FSD $\mathrm{D}_{1} \mathrm{D}_{2}$ |  |  | 巷 |
| SURROUND FROM ABOVE | 2 | IDC－SAV $\mathrm{D}_{1} \mathrm{D}_{2}$ |  |  | 旦包 |
| SURROUND FROM BELOW | 2 | IDC－SBL $\mathrm{D}_{1} \mathrm{D}_{2}$ | $: D_{2}$ |  | $1$ |
| SURROUND FROM LEFT | 2 | IDC－SLT $\mathrm{D}_{1} \mathrm{D}_{2}$ |  |  | $\sqrt{\text { 虎 }}$ |
| SURROUND FROM UPPER LEFT | 2 | IDC－SUL $\mathrm{D}_{1} \mathrm{D}_{2}$ | D | 疾 | 舞 |
| SURROUND FROM UPPER RIGHT | 2 | IDC－SUR $\mathrm{D}_{1} \mathrm{D}_{2}$ |  | 左 | 寺 |
| SURROUND FROM LOWER LEFT | 2 | IDC－SLL $\mathrm{D}_{1} \mathrm{D}_{2}$ | ： |  |  |
| OVERLAID | 2 | IDC－OVL $\mathrm{D}_{1} \mathrm{D}_{2}$ |  |  | 人人 |

＊NOTE－$D_{1}$ and $D_{2}$ overlap each other．This diagram does not imply that $D_{1}$ is on the top left corner and $D_{2}$ is on the bottom right corner．

## Annex G <br> (informative)

## Alphabetically sorted list of character names

This annex lists all the character names from this part of ISO/IEC 10646 except Hangul syllables and CJK-ideographs (these are characters from blocks HANGUL SYLLABLES, CJK UNIFIED IDEOGRAPHS, CJK UNIFIED IDEOGRAPHS EXTENSION A and CJK COMPATIBILITY IDEOGRAPHS). They are shown with their code positions in the two-octet form.

2100 ACCOUNT OF<br>206D ACTIVATE ARABIC FORM SHAPING<br>206B ACTIVATE SYMMETRIC SWAPPING<br>00B4 ACUTE ACCENT<br>2101 ADDRESSED TO THE SUBJECT<br>262C ADI SHAKTI<br>2708 AIRPLANE<br>2135 ALEF SYMBOL<br>232E ALL AROUND-PROFILE<br>224C ALL EQUAL TO<br>224A ALMOST EQUAL OR EQUAL TO<br>2248 ALMOST EQUAL TO<br>2387 ALTERNATIVE KEY SYMBOL<br>0026 AMPERSAND<br>2220 ANGLE<br>$212 B$ ANGSTROM SIGN<br>2625 ANKH

Editor's note: The complete list of character names will be provided in the Final Text of the Second Edition. Estimate: 55 pages.

## Annex H <br> (informative)

## The use of "signatures" to identify UCS

This annex describes a convention for the identification of features of the UCS, by the use of "signatures" within data streams of coded characters. The convention makes use of the character ZERO WIDTH NO-BREAK SPACE, and is applied by a certain class of applications.

When this convention is used, a signature at the beginning of a stream of coded characters indicates that the characters following are encoded in the UCS-2 or UCS-4 coded representation, and indicates the ordering of the octets within the coded representation of each character (see 6.3). It is typical of the class of applications mentioned above, that some make use of the signatures when receiving data, while others do not. The signatures are therefore designed in a way that makes it easy to ignore them. In this convention, the ZERO WIDTH NO-BREAK SPACE character has the following significance when it is present at the beginningof a stream of coded characters:

UCS-2 signature: FEFF
UCS-4 signature: 0000 FEFF
UTF-8 signature: EF BB BF
UTF-16 signature: FEFF
An application receiving data may either use these signatures to identify the coded representation form, or may ignore them and treat FEFF as the ZERO WIDTH NO-BREAK SPACE character.

If an application which uses one of these signatures recognises its coded representation in reverse sequence (e.g. hexadecimal FFFE), the application can identify that the coded representations of the following characters use the opposite octet sequence to the sequence expected, and may take the necessary action to recognise the characters correctly.

NOTE - The hexadecimal value FFFE does not correspond to any coded character within ISO/IEC 10646.

## Annex J <br> (informative)

## Recommendation for combined receiving/originating devices with internal storage

This annex is applicable to a widely-used class of devices that can store received CC-data elements for subsequent retransmission.
This recommendation is intended to ensure that loss of information is minimised between the receipt of a CC-data-element and its retransmission.
A device of this class includes a receiving device component and an originating device component as in 2.3, and can also store received CC-dataelements for retransmission, with or without modification by the actions of the user on the corresponding characters represented within it. Within this class of device, two distinct types are identified here, as follows.

1. Receiving device with full retransmission capability
The originating device component will retransmit the coded representations of any received characters, including those that are outside the identified subset of the receiving device component, without change to their coded representation, unless modified by the user.
2. Receiving device with subset retransmission capability
The originating device component can retransmit only the coded representations of the characters of the subset adopted by the receiving device component.

## Annex K <br> (informative)

## Notations of octet value representations

Representation of octet values in ISO/IEC 10646 except in clause 16 is different from other character coding standards such as ISO/IEC 2022, ISO/IEC 6429 and ISO 8859. This annex clarifies the relationship between the two notations.

- In ISO/IEC 10646, the notation used to express an octet value is $z$, where $z$ is a hexadecimal number in the range 00 to FF .

For example, the character ESCAPE (ESC) of ISO/IEC 2022 is represented by 1B.

- In other character coding standards, the notation used to express an octet value is $\mathrm{x} / \mathrm{y}$, where x and $y$ are two numbers in the range 00 to 15 . The correspondence between the notations of the form $\mathrm{x} / \mathrm{y}$ and the octet value is as follows.
$x$ is the number represented by bit 8 , bit 7 , bit 6 and bit 5 where these bits are given the weight 8, 4, 2 and 1 respectively;
$y$ is the number represented by bit 4 , bit 3 , bit 2 and bit 1 where these bits are given the weight $8,4,2$ and 1 respectively.

For example, the character ESC of ISO/IEC 2022 is represented by 01/11.
Thus ISO/IEC 2022 (and other character coding standards) octet value notation can be converted to ISO/IEC 10646 octet value notation by converting the value of $x$ and $y$ to hexadecimal notation. For example; $04 / 15$ is equivalent to 4 F .

## Annex L <br> (informative)

## Character naming guidelines

Guidelines for generating and presenting unique names of characters in ISO/IEC JTC1/SC2 standards are listed in this annex for reference. These guidelines are used in information technology coded character set standards such as ISO/IEC 646, ISO/IEC 6937, ISO/IEC 8859, ISO/IEC 10367 as well as in ISO/IEC 10646.

These Guidelines specify rules for generating and presenting unique names of characters in those versions of the standards that are in the English language.

NOTE. In a version of such a standard in another language:
a) these rules may be amended to permit names of characters to be generated using words and syntax that are considered appropriate within that language;
b) the names of the characters from this version of the standard may be replaced by equivalent unique names constructed according to the rules amended as in a) above.

Rules 1 to 3 are implemented without exceptions. However it must be accepted that in some cases (e.g. historical or traditional usage, unforeseen special cases, difficulties inherent to the nature of the character considered), exceptions to some of the other rules will have to be tolerated. Nonetheless, these rules are applied wherever possible.

## Rule 1

By convention, only Latin capital letters A to Z, space, and hyphen are used for writing the names of characters.

> NOTE - Names of characters may also include digits 0 to 9 (provided that a digit is not the first character in a word) if inclusion of the name of the corresponding digit(s) would be inappropriate. As an example the name of the character at position 201 is SolNGLE LOWW-9 QUOTATION MARK; the symbol for the digit 9 is included in this name to illustrate the shape of the character, and has no numerical significance.

## Rule 2

The names of control functions are coupled with an acronym consisting of Latin capital letters A to Z and, where required, digits. Once the name has been specified for the first time, the acronym may be used in the remainder of the text where required for simplification and clarity of the text. Exceptionally, acronyms may be used for graphic characters where usage already exists and clarity requires it, in particular in code tables.

Examples:<br>Name: LOCKING-SHIFT TWO RIGHT<br>Acronym: LS2R<br>Name: SOFT-HYPHEN<br>Acronym: SHY

NOTE - In ISO/IEC 6429, also the names of the modes have been presented in the same way as control functions.

## Rule 3

In some cases, the name of a character can be followed by an additional explanatory statement not part of the name. These statements are in parentheses and not in capital Latin letters except the initials of the word where required. See examples in rule 12.
The name of a character may also be followed by a single * symbol. This indicates that additional information on the character appears in Annex P. Any * symbols are omitted from the character names listed in Annex G.

## Rule 4

The name of a character wherever possible denotes its customary meaning, for examples PLUS SIGN. Where this is not possible, names describe shapes, not usage; for example: UPWARDS ARROW.
The name of a character is not intended to identify its properties or attributes, or to provide information on its linguistic characteristics, except as defined in Rule 6 below.

## Rule 5

Only one name is given to each character.

## Rule 6

The names are constructed from an appropriate set of the applicable terms of the following grid and ordered in the sequence of this grid. Exceptions are specified in Rule 11. The words WITH and AND may be included for additional clarity when needed.

| 1 | Script |
| :--- | :--- |
| 2 | Case |
| 3 | Type |
| 4 | Language |
| 5 | Attribute |
| 6 | Designation |
| 7 | Mark(s) |
| 8 | Qualifier |

Examples of such terms:

| Script | Latin, Cyrillic, Arabic |
| :--- | :--- |
| Case | capital, small |
| Type | letter, ligature, digit |
| Language | Ukrainian |
| Attribute | final, sharp, subscript, vulgar |
| Designation | customary name, name of letter <br> Mark |
| acute, ogonek, ring above, diaeresis |  |
| Qualifier | sign, symbol |

Examples of names:

## LATIN CAPITAL LETTER A WITH ACUTE

| 1 | 2 | 3 | 6 | 7 |
| :--- | :--- | :--- | :--- | :--- |

## DIGIT FIVE

## 36 <br> LEFT CURLY BRACKET

| 5 | 5 | 6 |
| :--- | ---: | :--- |
| NOTES |  |  |

1 A ligature is a graphic symbol in which two or more other graphic symbols are imaged as single graphic symbol.

2 Where a character comprises a base letter with multiple marks, the sequence of those in the name is the order in which the marks are positioned relative to the base letter, starting with the marks above the letters taken in upwards sequence, and followed by the marks below the letters taken in downwards sequece.

## Rule 7

The letters of the Latin script are represented within their name by their basic graphic symbols (A, B, C, ...). The letters of all other scripts are represented by their transcription in the language of the first published International Standard.

Examples:
K LATIN CAPITAL LETTER K

## ○ CYRILLIC CAPITAL LETTER YU

## Rule 8

In principle when a character of a given script is used in more than one language, no language name is specified. Exceptions are tolerated where an ambiguity would otherwise result.

Examples:

## И

I

## CYRILLIC CAPITAL LETTER BYELORUSSIAN-UKRAINIAN I

## Rule 9

Letters that are elements of more than one script are considered different even if their shape is the same; they have different names.

## Examples:

A LATIN CAPITAL LETTER A
A GREEK CAPITAL LETTER ALPHA
A CYRILLIC CAPITAL LETTER A

## Rule 10

A character of one script used in isolation in another script, for example as a graphic symbol in relation with physical units of dimension, is considered as a character different from the character of its native script.

Example:
$\mu \quad$ MICRO SIGN

## Rule 11

A number of characters have a traditional name consisting of one or two words. It is not intended to change this usage.
Examples:

| $'$ | APOSTROPHE |
| :--- | :--- |
| $:$ | COLON |
| $@$ | COMMERCIAL AT |
| $\bar{\sim}$ | LOW LINE |
|  | TILDE |

## Rule 12

In some cases, characters of a given script, often punctuation marks, are used in another script for a different usage. In these cases the customary name reflecting the most general use is given to the character. The customary name may be followed in the list of characters of a particular standard by the name in parentheses which this character has in the script specified by this particular standard.
Example:
UNDERTIE (Enotikon)

## Rule 13

The above rules do not apply to ideographic characters. These characters are identified by alphanumeric identifiers specified for each ideographic character (see clause 27).

## Annex M <br> (informative)

## Sources of characters

Several sources and contributions were used for constructing this coded character set. In particular, characters of the following national and international standards are included in this part of ISO/IEC 10646.

ISO 233:1984, Documentation - Transliteration of Arabic characters into Latin characters.

ISO/IEC 646:1991, Information technology - ISO 7bit coded character set for information interchange.

ISO 2033:1983, Information processing - Coding of machine-readable characters (MICR and OCR).

ISO 2047:1975, Information processing - Graphical representations for the control characters of the 7-bit coded character set.

ISO 5426:1983, Extension of the Latin alphabet coded character set for bibliographic information interchange.

ISO 5427:1984, Extension of the Cyrillic alphabet coded character set for bibliographic information interchange.

ISO 5428:1984, Greek alphabet coded character set for bibliographic information interchange.

ISO 6438:1983, Documentation - African coded character set for bibliographic information interchange.

ISO 6861, Information and documentation - Cyrillic alphabet coded character set for historic Slavonic languages and European non-Slavonic languages written in a Cyrillic script for bibliographic information interchange.

ISO 6862, Information and documentation Mathematical coded character set for bibliographic information interchange.

ISO 6937:1993, Information technology - Coded graphic character sets for text communication - Latin alphabet.

ISO/IEC 8859, Information technology - 8-bit singlebyte coded graphic character sets
$\begin{array}{ll}\text {-Part } & \text { 1. Latin alphabet No. } 1 \text { (1998). } \\ \text {-Part } & \text { 2. Latin alphabet No. } 2 \text { (1999). }\end{array}$
-Part 3. Latin alphabet No. 3 (1999).
-Part 4. Latin alphabet No. 4 (1998).
-Part 5. Latin/Cyrillic alphabet (1999)
-Part 6. Latin/Arabic alphabet (1999)
-Part 7. Latin/Greek alphabet (1987)
-Part 8. Latin/Hebrew alphabet (1999)
-Part 9. Latin alphabet No. 5 (1999)
-Part 10. Latin alphabet No. 6 (1998).
ISO 8879:1986, Information processing - Text and office systems - Standard Generalized Markup Language (SGML).

ISO 8957:1993, Information and documentation Hebrew alphabet coded character sets for bibliographic information interchange.

ISO 9036:1987, Information processing - Arabic 7-bit coded character set for information interchange.

ISO/IEC 10367:1991, Information technology Standardized coded graphic character sets for use in 8-bit codes.

ISO/IEC TR 15285:1998, Information technology An operational model for characters and glyphs.

ISO international register of character sets to be used with escape sequences. (registration procedure ISO 2375:1985) .

ANSI X3.4-1986 American National Standards Institute. Coded character set - 7-bit American national standard code for information interchange.

ANSI X3.32-1973 American National Standards Institute. American national standard graphic representation of the control characters of American national standard code for information interchange.

ANSI Y10.20-1988 American National Standards Institute. Mathematic signs and symbols for use in physical sciences and technology.

ANSI Y14.5M-1982 American National Standard. Engineering drawings and related document practices, dimensioning and tolerances.

ANSI Z39.47-1985 American National Standards Institute. Extended Latin alphabet coded character set for bibliographic use.

ANSI Z39.64-1989 American National Standards Institute. East Asian character code for bibliographic use.

ASMO 449-1982 Arab Organization for Standardization and Methodology. Data processing -7-bit coded character set for information interchange.

GB2312-80 Code of Chinese Graphic Character Set for Information Interchange: Jishu Biaozhun Chubanshe (Technical Standards Publishing).

NOTE - For additional sources of the CJK unified ideographs in this part of ISO/IEC 10646 refer to clause 27.
GBK (Guo Biao Kuo) Han character internal code extension specification: Jishu Biaozhun Chubanshe (Technical Standards Publishing)

LTD 37(1610)-1988 Indian standard code for information interchange.

JIS X 0201-1976 Japanese Standards Association. Jouhou koukan you fugou (Code for Information Interchange).

JIS X 0208-1990 Japanese Standards Association. Jouhou koukan you kanji fugoukei (Code of the Japanese Graphic Character Set for Information Interchange).
JIS X 0212-1990 Japanese Standards Association. Jouhou koukan you kanji fugou-hojo kanji (Code of the supplementary Japanese graphic character set for information interchange).

KS C 5601-1992 Korean Industrial Standards Association. Jeongbo gyohwanyong buho (Code for Information Interchange).

SI 1311.2-1996 The Standards Institution of Israel Information Technology. ISO 8-bit coded character set for information interchange with Hebrew points and cantillation marks.

TIS 620-2533:1990 Thai Industrial Standard for Thai Character Code for Computer.

Esling, John. Computer coding of the IPA: supplementary report. Journal of the International Phonetic Association, 20:1 (1990), p. 22-26.

International Phonetic Association. The IPA 1989 Kiel Convention Workgroup 9 report: Computer Coding of IPA Symbols and Computer Representation of Individual Languages. Journal of the International Phonetic Association, 19:2 (1989), p. 81-82.

International Phonetic Association. The International Phonetic Alphabet (revised to 1989).
Knuth, Donald E. The TeXbook. - 19th. printing, rev.- Reading, MA : Addison-Wesley, 1990.

Pullum, Geoffrey K. Phonetic symbol guide. Geoffrey K. Pullum and William A. Ladusaw. - Chicago : University of Chicago Press, 1986.

Pullum, Geoffrey K. Remarks on the 1989 revision of the International Phonetic Alphabet. Journal of the International Phonetic Association, 20:1 (1990), p. 33-40.

Selby, Samuel M. Standard mathematical tables. 16th ed. - Cleveland, OH : Chemical Rubber Co., 1968. Shepherd, Walter.

Shepherd, Walter. Shepherd's glossary of graphic signs and symbols. Compiled and classified for ready reference. - New York: Dover Publications, [1971].

Shinmura, Izuru. Kojien - Dai 4-han. - Tokyo : Iwanami Shoten, Heisei 3 [1991].
The Unicode Consortium. The Unicode Standard. Worldwide Character Encoding Version 1.0, Volume One. - Reading, MA : Addison-Wesley, 1991.

## Annex N <br> (informative)

## External references to character repertoires

## N. 1 Methods of reference to character repertoires and their coding

Within programming languages and other methods for defining the syntax of data objects there is commonly a need to declare a specific character repertoire from among those that are specified in ISO/IEC 10646. There may also be a need to declare the corresponding coded representations applicable to that repertoire.

For any character repertoire that is in accordance with ISO/IEC 10646 a precise declaration of that repertoire should include the following parameters: identification of ISO/IEC 10646,

- the adopted subset of the repertoire, identified by one or more collection numbers,
- the adopted implementation level (1, 2 or 3),
- the adopted coded representation form (4-octet or 2-octet).

One of the methods now in common use for defining the syntax of data objects is Abstract Syntax Notation 1 (ASN.1) specified in ISO/IEC 8824. The corresponding coded representations are specified in ISO/IEC 8825. When this method is used the forms of the references to character repertoires and coding are as indicated in the following clauses.

## N. 2 Identification of ASN. 1 character abstract syntaxes

The set of all character strings that can be formed from the characters of an identified repertoire in accordance with ISO/IEC 10646 is defined to be a "character abstract syntax" in the terminology of ISO/IEC 8824. For each such character abstract syntax, a corresponding object identifier value is defined to permit references to be made to that syntax when the ASN. 1 notation is used.
ISO/IEC 8824 annex B specifies the form of object identifier values for objects that are specified in an ISO standard. In such an object identifier the features and options of this part of ISO/IEC 10646 are identified by means of numbers (arcs) which follow the arcs "10646" and "1" which identify the part one of ISO/IEC 10646.

The first such arc identifies the adopted implementation level, and is either:

- level-1 (1), or
- level-2 (2), or
- level-3 (3).

The second such arc identifies the repertoire subset, and is either:

- all (0), or
- collections (1).

Arc (0) identifies the entire collection of characters specified in this part of ISO/IEC 10646. No further arc follow this arc.

NOTE - This collection includes private groups and planes, and is therefore not fully-defined. Its use without additional prior agreement is deprecated.

Arc (1) is followed by one or a sequence of further arcs, each of which is a collection number from annex A, in ascending numerical order. This sequence identifies the subset consisting of the collections whose numbers appear in the sequence.

NOTE - As an example, the object identifier for the subset comprising the collections BASIC LATIN, LATIN-1 SUPPLEMENT, and MATHEMATICAL OPERATORS, at implementation level 1, is:
\{iso standard 106461 level-1 (1) collections (1) 12 39\}
ISO/IEC 8824 also specifies object descriptors corresponding to object identifier values. For each combination of arcs the corresponding object descriptors are as follows:

10 : "ISO 10646 part-1 level-1 unrestricted"
20 : "ISO 10646 part-1 level-2 unrestricted"
30 : "ISO 10646 part-1 level-3 unrestricted"
For a single collection with collection name "xxx".
11 : "ISO 10646 part-1 level-1 xxx"
21 : "ISO 10646 part-1 level-2 xxx"
31 : "ISO 10646 part-1 level-3 xxx"
For a repertoire comprising more than one collection, numbered $\mathrm{m} 1, \mathrm{~m} 2$, etc.
1 1: "ISO 10646 part-1 level-1 collections m1,m2, m3, .... "

2 1: "ISO 10646 part-1 level-2 collections m1,m2, m3, .... "

31 : "ISO 10646 part-1 level-3 collections m1,m2, m3, .... "
NOTE - All spaces are single spaces.

## N. 3 Identification of ASN. 1 character transfer syntaxes

The coding method for character strings that can be formed from the characters in accordance with ISO/IEC 10646 is defined to be a "character transfer syntax" in the terminology of ISO/IEC 8824. For each such character transfer syntax, a corresponding object identifier value is defined to permit references to be made to that syntax when the ASN. 1 notation is used.

In an object identifier in accordance with ISO/IEC 8824 annex B, the coded representation form specified in this part of ISO/IEC 10646 is identified by means of numbers (arcs) which follow the arcs "10646" and "1" which identify this part of ISO/IEC 10646.

The first such arc is:

- transfer-syntaxes (0).

The second such arc identifies the form and is either:

- two-octet-BMP-form (2), or
- four-octet-form (4), or
- UTF16-form (5), or
- UTF8-form (8).

NOTE - As an example, the object identifier for the two-octet coded representation form is:
\{iso standard 106461 transfer-syntaxes (0) two-octet-BMPform (2)\}

The corresponding object descriptors are:

- "ISO 10646 part-1 form 2"
- "ISO 10646 part-1 form 4"
- "ISO 10646 part-1 utt-16"
- "ISO 10646 part-1 utf-8".


## Annex P <br> (Informative)

## Additional information on characters

This Annex contains additional information on some of the characters specified in clause 26 of this International Standard. This information is intended to clarify some feature of a character, such as its naming or usage, or its associated graphic symbol.

Each entry in this Annex consists of the name of a character and its code position in the two-octet form, followed by the related additional information. Entries are arranged in ascending sequence of code position.

When an entry for a character is included in this Annex an * symbol appears immediately following its name in the corresponding table in clause 26 of this International Standard.

## Group 00, Plane 00 (BMP)

00AB LEFT-POINTING DOUBLE ANGLE QUOTATION MARK This character may be used as an Arabic opening quotation mark, if it appears in a bidirectional context as described in clause 19. The graphic symbol associated with it may differ from that in the table for Row 00.

00BB RIGHT-POINTING DOUBLE ANGLE QUOTATION MARK This character may be used as an Arabic closing quotation mark, if it appears in a bidirectional context as described in clause 19. The graphic symbol associated with it may differ from that in the table for Row 00.

## $00 C 6$ LATIN CAPITAL LETTER AE (ash)

In the first edition of this International Standard the name of this character was:

LATIN CAPITAL LIGATURE AE
00E6 LATIN SMALL LETTER AE (ash)
In the first edition of this International Standard the name of this character was:

LATIN SMALL LIGATURE AE
0189 LATIN CAPITAL LETTER AFRICAN D
This character is the capital letter form of:
0256 LATIN SMALL LETTER D WITH TAIL
019F LATIN CAPITAL LETTER O WITH MIDDLE TILDE
This character is the capital letter form of:
0275 LATIN SMALL LETTER BARRED O
01A6 LATIN LETTER YR
This character is the capital letter form of: 0280 LATIN LETTER SMALL CAPITAL R

01E2 LATIN CAPITAL LETTER AE WITH MACRON (ash)
In the first edition of this International Standard the name of this character was:

LATIN CAPITAL LIGATURE AE WITH MACRON
01E3 LATIN SMALL LETTER AE WITH MACRON (ash)
In the first edition of this International Standard the name of this character was:

LATIN SMALL LIGATURE AE WITH MACRON
01FC LATIN CAPITAL LETTER AE WITH ACUTE (ash)
In the first edition of this International Standard the name of this character was:

LATIN CAPITAL LIGATURE AE WITH ACUTE
01FD LATIN SMALL LETTER AE WITH ACUTE (ash)
In the first edition of this International Standard the name of this character was:

LATIN SMALL LIGATURE AE WITH ACUTE
0218 LATIN CAPITAL LETTER S WITH COMMA BELOW
This character is intended for use only in those cases where it is necessary to make a distinction from the letter with cedilla. Both forms of the letter may be found in a single document written in a single language, e.g. Romanian or Turkish.

In ISO/IEC 8859-2 only a single (8-bit) coded character is provided, LATIN CAPITAL LETTER S WITH CEDILLA, which maps to 015E in ISO/IEC 10646 by default, and may map by mutual agreement between sender and receiver to this letter with comma below. See ISO/IEC 8859-2 for further information on the use of that standard.

0219 LATIN SMALL LETTER S WITH COMMA BELOW
This character is intended for use only in those cases where it is necessary to make a distinction from the letter with cedilla. Both forms of the letter may be found in a single document written in a single language, e.g. Romanian or Turkish.
In ISO/IEC 8859-2 only a single (8-bit) coded character is provided, LATIN SMALL LETTER S WITH CEDILLA, which maps to 015F in ISO/IEC 10646 by default, and may map by mutual agreement between sender and receiver to this letter with comma below. See ISO/IEC 8859-2 for further information on the use of that standard.

021A LATIN CAPITAL LETTER T WITH COMMA BELOW
This character is intended for use only in those cases where it is necessary to make a distinction from the letter with cedilla. Both forms of the
letter may be found in a single document written in a single language, e.g. Romanian.
In ISO/IEC 8859-2 only a single (8-bit) coded character is provided, LATIN CAPITAL LETTER T WITH CEDILLA, which maps to 0162 in ISO/IEC 10646 by default, and may map by mutual agreement between sender and receiver to this letter with comma below. See ISO/IEC 8859-2 for further information on the use of that standard.

021B LATIN SMALL LETTER T WITH COMMA BELOW
This character is intended for use only in those cases where it is necessary to make a distinction from the letter with cedilla. Both forms of the letter may be found in a single document written in a single language, e.g. Romanian.
In ISO/IEC 8859-2 only a single (8-bit) coded character is provided, LATIN SMALL LETTER T WITH CEDILLA, which maps to 0163 in ISO/IEC 10646 by default, and may map by mutual agreement between sender and receiver to this letter with comma below. See ISO/IEC 8859-2 for further information on the use of that standard.

## 0280 LATIN LETTER SMALL CAPITAL R

This character is the small letter form of:
01A6 LATIN LETTER YR
0596 HEBREW ACCENT TIPEHA
This character may be used as a Hebrew accent tarha.

## 0598 HEBREW ACCENT ZARQA

This character may be used as a Hebrew accent zinorit.

## 05A5 HEBREW ACCENT MERKHA

This character may be used as a Hebrew accent yored.

05A8 HEBREW ACCENT QADMA
This character may be used as a Hebrew accent azla.

05AA HEBREW ACCENT YERAH BEN YOMO
This character may be used as a Hebrew accent galgal.

05BD HEBREW POINT METEG
This character may be used as a Hebrew accent sof pasuq or siluq.

05C0 HEBREW PUNCTUATION PASEQ
This character may be used as a Hebrew accent legarme.
05C3 HEBREW PUNCTUATION SOF PASUQ
This character may be used as a Hebrew punctuation colon.

06AF ARABIC LETTER GAF
The symbol for a Hamza (see position 0633) may appear in the centre of the graphic symbol associated with this character.

06D0 ARABIC LETTER E
This character may be used as an Arabic letter Sindhi bbeh.

OFAD TIBETAN SUBJOINED LETTER WA
The graphic symbol for this character occurs in two alternative forms, a full form and a short form (known as wa.zur (wazur)). The short form of the letter is shown in the table, since it occurs more frequently.

OFB1 TIBETAN SUBJOINED LETTER YA
The graphic symbol for this character occurs in two alternative forms, a full form and a short form (known as ya.btags (ya ta)). The short form of the letter is shown in the table, since it occurs more frequently.

OFB2 TIBETAN SUBJOINED LETTER RA
The graphic symbol for this character occurs in two alternative forms, a full form and a short form (known as ra.btags (ra ta)). The short form of the letter is shown in the table, since it occurs more frequently.

## OF6A TIBETAN LETTER FIXED-FORM RA

This character has the same graphic symbol as that shown in the table for:

OF62 TIBETAN LETTER RA
It may be used when the graphic symbol is required to remain unchanged regardless of context.

1100 HANGUL CHOSEONG KIYEOK .....
1112 HANGUL CHOSEONG HIEUH
The Latin letters shown in parenthesis after the names of the characters in the range hex 1100 to 1112 (except 110B) are transliterations of these Hangul characters. These transliterations are used in the construction of the names of the Hangul syllables that are allocated in code positions hex AC00 to D7A3 in this International Standard.

11A8 HANGUL JONGSEONG KIYEOK .....
11C2 HANGUL JONGSEONG HIEUH
The Latin letters shown in parenthesis after the names of the characters in the range hex 11A8 to 11 C 2 are transliterations of these Hangul characters. These transliterations are used in the construction of the names of the Hangul syllables that are allocated in code positions hex AC00 to D7A3 in this International Standard.

234A APL FUNCTIONAL SYMBOL DOWN TACK UNDERBAR
The relation between the name of this character and the orientation of the "tack" element in its
graphical symbol is inconsistent with that of other characters in this International Standard, such as:

22A4 DOWN TACK and 22A5 UP TACK
234E APL FUNCTIONAL SYMBOL DOWN TACK JOT Information for the character at 234A applies.
2351 APL FUNCTIONAL SYMBOL UP TACK OVERBAR Information for the character at 234A applies.

2355 APL FUNCTIONAL SYMBOL UP TACK JOT Information for the character at 234A applies.
2361 APL FUNCTIONAL SYMBOL UP TACK DIAERESIS Information for the character at 234A applies.

FA1F CJK COMPATIBILITY IDEOGRAPH-FA1F
This character should be considered as an extension to the block of characters CJK UNIFIED IDEOGRAPHS EXTENSION A (see clause 27). It is not a duplicate of a character already allocated in the blocks of CJK Unified Ideographs, unlike many other characters in the block CJK COMPATIBILITY IDEO-GRAPHS. The source of this character, shown as described in clause 27, is:

| $\mathbf{C}$ | $\mathbf{J}$ | $\mathbf{K}$ | $\mathbf{V}$ |
| :---: | :---: | :---: | :---: |
| G- Hanzi- T | Kanji <br> A-264B <br>  <br>  <br>  <br>  <br> A-0643 |  |  |
|  |  |  |  |

FA23 CJK COMPATIBILITY IDEOGRAPH-FA23
This character should be considered as an extension to the block of characters CJK UNIFIED IDEOGRAPHS EXTENSION A (see clause 27). It is not a duplicate of a character already allocated in the blocks of CJK Unified Ideographs, unlike many other characters in the block CJK COMPATIBILITY IDEOGRAPHS. The sources of this character, shown as described in clause 27, are:

| C | $\mathbf{J}$ | $\mathbf{K}$ | $\mathbf{V}$ |
| :---: | :---: | :---: | :---: |
| G- Hanzi- T | Kanji | Hanja | ChuNom |
| F-3862 | A-2728 |  |  |
| F-2466 | A-0708 |  |  |

FFE3 FULLWIDTH MACRON
This character is the full-width form of the character: OOAF MACRON. It may also be used as the full-width form of the character:

203E OVERLINE

## Annex Q <br> (informative)

## Code mapping table for Hangul syllables

This Annex provides a cross-reference between the Hangul syllables (and code positions) that were specified in the First Edition of this International Standard and their amended code positions as now specified in this edition.

In the First Edition of this International Standard 6656 Hangul syllables were allocated to consecutive code positions in the range hexadecimal 3400 to 4DFF. These Hangul syllables are now re-allocated non-consecutively to code positions in the larger range hexadecimal AC00 to D7A3.

For each Hangul syllable in the First Edition its code position provides an index to a cell in Table Q. 1 which appears on the following pages. The first three hexadecimal digits of the code position identify a row in the table, and the final hexadecimal digit
identifies a column in the table. The cell at the identified row and column position contains the code position (in hexadecimal) to which the Hangul syllable is now allocated.

## Example:

In the table for Row 38 (Table 67) of the First Edition of this International Standard

HANGUL SYLLABLE SIOS O RIEUL is found at position 389D. In row 389, column D, of Table Q. 1 the entry C194 is found. This entry indicates that this Hangul syllable is now allocated to code position C194.

NOTE - The name shown for the Hangul syllable at C194 is: HANGUL SYLLABLE SOL.
This is because the names of Hangul syllables are now constructed from the Latin transliterations shown in the tables for Row 11 (see also 26.2 and Annex P).

Table Q. 1 - Code mapping for Hangul syllables

|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 340 | ACOO | AC01 | AC04 | AC07 | AC08 | AC09 | AC0A | AC10 | AC11 | AC12 | AC13 | AC14 | AC15 | AC16 | AC17 | AC19 |
| 341 : | AC1A | AC1B | AC1C | AC1D | AC20 | AC24 | AC2C | AC2D | AC2F | AC30 | AC31 | AC38 | AC39 | AC3C | AC40 | AC4B |
| 342 | AC4D | AC54 | AC58 | AC5C | AC70 | AC71 | AC74 | AC77 | AC78 | AC7A | AC80 | AC81 | AC83 | AC84 | AC85 | AC86 |
| 343 | AC | AC8A | AC8 | AC8C | AC90 | AC94 | AC9C | AC9D | AC9F | ACAO | ACA1 | ACA8 | ACA9 | ACAA | ACAC | ACAF |
| 344 | ACB0 | ACB8 | ACB9 | ACBB | ACBC | ACBD | ACC1 | ACC4 | ACC8 | ACCC | ACD5 | ACD7 | ACEO | ACE1 | ACE4 | ACE7 |
| 345 | ACE8 | ACEA | ACEC | ACEF | ACFO | ACF1 | ACF3 | ACF5 | ACF6 | ACFC | ACFD | AD00 | AD04 | AD06 | ADOC | ADOD |
| 34 | dof | AD11 | AD18 | AD1C | AD20 | AD29 | AD2C | AD2D | AD34 | AD35 | AD38 | AD3C | AD44 | AD45 | AD47 | d49 |
| 347 : | AD50 | AD54 | AD58 | AD61 | AD63 | AD6C | AD6D | AD70 | AD73 | AD74 | AD75 | AD76 | AD7B | AD7C | AD7D | AD7F |
| 348 | AD81 | AD82 | AD88 | AD89 | AD8C | AD90 | AD9C | AD9D | ADA4 | ADB7 | ADCO | ADC1 | ADC4 | ADC8 | ADDO | ADD1 |
| 349 | ADD3 | ADDC | ADEO | ADE4 | ADF8 | ADF9 | ADFC | ADFF | AEOO | AE01 | AE08 | AE09 | AEOB | AEOD | AE14 | AE30 |
| 34 A | AE31 | AE34 | AE37 | AE38 | AЕ3A | AE40 | AE41 | AE43 | AE45 | AE46 | AE4A | AE4C | AE4D | AE4E | AE50 | AE54 |
| 34B | AE56 | AE5C | AE5D | AE5F | AE60 | AE61 | AE65 | AE68 | AE69 | AE6C | AE70 | AE78 | AE79 | AE7B | AE7C | AE7D |
| 34 C | AE84 | AE85 | AE8C | AEBC | AEBD | AEBE | AECO | AEC4 | AECC | AECD | AECF | AEDO | AED | AED8 | AED9 | AEDC |
| 34D | AEE8 | AEEB | AEED | AEF4 | AEF8 | AEFC | AF07 | AF08 | AFOD | AF10 | AF2C | AF2D | AF30 | AF32 | AF34 | AF3C |
| 34E: | AF3D | AF3F | AF41 | AF42 | AF43 | AF48 | AF49 | AF50 | AF5C | AF5D | AF64 | AF65 | AF79 | AF80 | AF84 | AF88 |
| 34F: | AF90 | AF91 | AF95 | AF9C | AFB8 | AFB9 | AFBC | AFCO | AFC7 | AFC8 | AFC9 | AFCB | AFCD | AFCE | AFD4 | AFDC |
| 350 | AFE8 | AFE9 | AFFO | FF1 | AFF4 | AFF8 | B000 | B001 | B004 | B00C | B010 | B014 | B01C | B01D | B028 | B044 |
| 351 : | B045 | B048 | B04A | B04C | B04E | B053 | B054 | B055 | B057 | B059 | B05D | B07C | B07D | B080 | B084 | B08C |
| 352 | B08D | B08F | B091 | B098 | B099 | B09A | B09C | B09F | B0AO | B0A1 | B0A2 | B0A8 | B0A9 | B0AB | B0aC | B0AD |
| 353 : | BOAE | B0AF | B0B1 | вов3 | вов4 | B5 | вов8 | BовС | BoC4 | BOC5 | B0C7 | B0C8 | Boc9 | BODO | B0D1 | B0D4 |
| 354 | B0D8 | BOEO | B0E5 | B108 | B109 | B10B | B10C | B110 | B112 | B113 | B118 | B119 | B11B | B11C | B11D | B123 |
| 355 | B124 | B125 | B128 | B12C | B134 | B135 | B137 | B138 | B139 | B140 | B141 | B144 | B148 | B150 | B151 | B154 |
| 356 : | B155 | B158 | B15C | B160 | B178 | B179 | B17C | B180 | B182 | B188 | B189 | B18B | B18D | B192 | B193 | B194 |
| 357 | B198 | B19C | B1A8 | B1CC | B1D0 | B1D4 | B1DC | B1DD | B1DF | B1E8 | B1E9 | B1EC | B1F0 | B1F9 | B1FB | B1FD |
| 358 | B204 | B205 | B208 | B20B | B20C | B214 | B215 | B217 | B219 | B220 | B234 | B23C | B258 | B25C | B260 | B268 |
| 359 | B269 | B274 | B275 | B27C | B284 | B285 | B289 | 290 | B291 | B294 | B298 | B299 | B29A | B2A0 | B2A1 | в2А 3 |
| 35A | B2A5 | B2A6 | B2AA | B2AC | B2B0 | B2B4 | B2C8 | B2C9 | B2CC | B2D0 | B2D2 | B2D8 | B2D9 | B2D | B2DD | B2E2 |
| 35B | B2E4 | B2E5 | B2E6 | B2E8 | B2EB | B2EC | B2ED | B2EE | B2EF | B2F3 | B2F4 | B2F5 | B2F7 | B2F8 | B2F9 | B2FA |
| 35 C | B2FB | B2FF | B300 | B301 | B304 | B308 | B310 | B311 | B313 | B314 | B315 | B31C | B354 | B355 | B356 | B358 |
| 35D | B35B | B35C | B3 | B35F | B364 | B | B367 | 369 | B36B | B36E | B370 | B37 | B374 | B378 | B38 | B381 |
| 35E: | B383 | B384 | B385 | B38C | В390 | B394 | B3A0 | B3A1 | B3A8 | B3AC | B3C4 | B3C5 | B3C8 | ВзСВ | B3CC | B3CE |
| 35F : | B3D0 | B3D4 | B3D5 | B3D7 | B3D9 | B3DB | B3DD | BЗE0 | B3E4 | BЗЕ8 | B3FC | B410 | B418 | B41C | B420 | B428 |
| 360 | B429 | B42B | B434 | B450 | B451 | B454 | B458 | B460 | B461 | B463 | B465 | B46C | B480 | B488 | B49D | B4A4 |
| 361 : | B4A8 | B4AC | B4B5 | B4B7 | B4B9 | B4C0 | B4C4 | B4C8 | B4DO | B4D5 | B4DC | B4DD | B4E0 | B4 | B4 | B4E6 |
| 362 : | B4EC | B4ED | B4EF | B4F1 | B4F8 | B514 | B515 | B518 | B51B | B51C | B524 | B525 | B527 | B528 | B529 | B52A |
| 363 : | B530 | B531 | B534 | B538 | B540 | B541 | B543 | B544 | B545 | B54B | B54C | B54D | B550 | B554 | B55C | B55D |
| 364 | B55F | B560 | B561 | B5A0 | B5A1 | B5A4 | B5A8 | B5AA | B5AB | B5B0 | B5B1 | в5в3 | B5B4 | B5B5 | B5BB | B5BC |
| 365 : | B5BD | B5C0 | B5C4 | B5CC | B5CD | B5CF | B5D0 | B5D1 | B5D8 | B5EC | B610 | B611 | B614 | B618 | B6 | B62C |
| 366 : | B634 | B648 | B66 | B66 | B69C | B69D | B6A0 | B6A4 | B6AB | B6AC | B6B1 | B6D4 | B6F0 | B6F4 | B6F8 | B700 |
| 367 : | B701 | B705 | B728 | B729 | B72C | B72F | B730 | B738 | B739 | B73B | B744 | B748 | B74C | B754 | B755 | B760 |
| 368 : | B764 | B768 | B770 | B771 | B773 | B775 | B77C | B77D | B780 | B784 | B78C | B78D | B78F | B790 | B791 | B792 |
| 369 : | B796 | B797 | B798 | B799 | B79C | B7a0 | B7A8 | B7A9 | b7ab | B7AC | B7AD | B7B | B7B5 | B7B | B7C7 | B7C9 |
| 36A | B7EC | B7ED | B7F0 | B7F4 | B7FC | B7FD | B7FF | B800 | B801 | B807 | B808 | B809 | B80C | B810 | B818 | B819 |
| 36B | B81B | B81D | B824 | B825 | B828 | B82C | B834 | B835 | B837 | B838 | B839 | B840 | B844 | B851 | B853 | B85C |
| 36 C | B85D | B860 | B864 | B86C | B86D | B86F | B871 | B878 | B87C | B88D | B8A8 | B8B0 | B8B4 | B8B8 | B8C0 | B8C1 |
| 36D | B8C3 | B8C5 | B8CC | B8D0 | B8D4 | B8DD | B8DF | B8E1 | B8E8 | B8E9 | B8EC | B8F0 | B8F8 | B8F9 | B8FB | B8FD |
| 36E: | B904 | B918 | B920 | B93C | B93D | B940 | B944 | B94C | B94F | B951 | B958 | B959 | B95C | B960 | B968 | B969 |
| 36F: | B96B | B96D | B974 | B975 | B978 | B97C | B984 | B985 | B987 | B989 | B98A | B98D | B98E | B9AC | B9AD | B980 |

# Table Q. 1 (continued) 

| 370 | B9B4 | B9BC | B9BD | B9BF | B9C1 | B9C8 | B9C9 | B9CC | B9CE | B9CF | B9D0 | B9D1 | B9D2 | B9D8 | B9D9 | B9DB |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 371 | B9DD | B9DE | B9E1 | B9E3 | B9E4 | B9E5 | B9E8 | B9EC | B9F4 | B9F5 | B9F7 | B9F8 | B9F9 | B9FA | ba00 | BA01 |
| 372 | BA08 | BA15 | BA38 | BA39 | ВАЗС | BA40 | BA42 | BA48 | BA49 | BA4B | BA4D | BA4E | BA53 | BA54 | BA55 | BA58 |
| 373 | BA5C | BA64 | BA65 | BA67 | BA68 | BA69 | Ba70 | BA71 | BA74 | BA78 | BA83 | BA84 | BA85 | BA87 | BABC | AA8 |
| 374 | BAA9 | BAAB | BAAC | BAB0 | BAB2 | BAB8 | BAB9 | BABB | BABD | BAC4 | BAC8 | BAD8 | BAD9 | BAFC | BB00 | BB04 |
| 375 | BBOD | BBOF | BB11 | BB18 | BB1C | BB20 | BB29 | BB2B | BB34 | BB35 | BB36 | BB38 | ввзв | BB3C | BB3D | BB3E |
| 376 | BB44 | BB45 | BB47 | BB49 | 4D | BB4F | 350 | B54 | BB58 | BB6 | BB63 | BB6C | BB | BB8C | 0 | BBA4 |
| 377 | BBA8 | BBAC | BBB4 | BBB7 | BBCO | BBC4 | BBC8 | BBDO | BBD3 | BBF8 | BBF9 | BBFC | BBFF | BCOO | BC02 | BC08 |
| 378 | BC09 | BCOB | BCOC | BCOD | BCOF | BC11 | BC14 | BC15 | BC16 | BC17 | BC18 | BC1B | BC1C | BC1D | BC1E | BC1F |
| 37 | BC24 | BC25 | BC27 | BC29 | BC2D | BC30 | BC31 | BC34 | BC38 | BC40 | BC41 | BC43 | BC44 | BC45 | BC49 | BC4C |
| 37A | BC4D | BC50 | BC5D | BC84 | BC85 | BC88 | BC8B | BC8C | BC8E | BC94 | BC95 | BC97 | BC99 | BC9A | BCAO | BCA1 |
| 37B | BCA4 | BCA7 | BCA8 | BCB0 | BCB1 | BCB3 | BCB4 | BCB5 | BCBC | BCBD | BCCO | BCC4 | BCCD | BCCF | BCDO | BCD1 |
| 37 | BCD5 | BCD8 | BCDC | BCF4 | BCF5 | BCF6 | BCF8 | BCFC | BD04 | BD05 | BD07 | BD09 | BD10 | BD14 | BD24 | BD2C |
| 37D | BD40 | BD48 | BD49 | BD4C | BD50 | BD58 | BD59 | BD64 | BD68 | BD80 | BD81 | BD84 | BD87 | BD88 | BD89 | BD8A |
| 37 E | BD90 | BD91 | BD93 | BD95 | BD99 | BD9A | BD9C | BDA4 | BDBO | BDB8 | BDD4 | BDD5 | BDD8 | BDDC | BDE9 | BDFO |
| 37 F | BDF4 | BDF8 | BE00 | BE03 | BE05 | BEOC | BEOD | BE10 | BE14 | BE1C | BE1D | BE1F | BE44 | BE45 | BE48 | BE4C |
| 380 | BE4E | BE54 | BE55 | BE57 | BE59 | BE5A | BE5B | BE60 | BE61 | BE64 | BE68 | BE6A | BE70 | BE71 | BE73 | BE74 |
| 381 | BE75 | BE7B | BE7C | BE7D | BE80 | BE84 | BE8C | BE8D | BE8F | BE90 | BE91 | BE98 | BE99 | BEA8 | BEDO | BED1 |
| 382 | BED4 | BED7 | BED8 | BEEO | BEE3 | BEE4 | BEE5 | BEEC | BF01 | BF08 | BF09 | BF18 | BF19 | BF1B | BF1C | BF1D |
| 383 | BF40 | BF41 | BF44 | 48 | BF50 | 51 | BF55 | BF94 | BFBO | BFC5 | BFCC | BFC | BFDO | BFD4 | BFDC | BFDF |
| 384 | BFE1 | C03C | C051 | C 058 | C05C | C060 | C068 | C069 | C090 | C091 | C094 | C098 | COAO | COA1 | COA4 | COA5 |
| 385 | COAC | COAD | COAF | COBO | COB3 | COB4 | C0B5 | COB6 | COBC | COBD | COBF | COCO | $\mathrm{COC1}$ | COC5 | COC8 | cocs |
| 386 | COCC | CODO | COD8 | COD9 | CODB | CODC | CODD | COE4 | COE5 | COE8 | COEC | C0F4 | C0F5 | C0F7 | C0F9 | C100 |
| 387 | C104 | C108 | C110 | C 115 | C11C | C11D | C11E | C11F | C120 | C123 | C124 | C126 | C127 | C12C | C12D | C12F |
| 38 | C130 | C131 | C136 | C138 | C139 | C13C | C140 | C148 | C149 | C14B | C14C | C14D | C154 | C155 | C158 | C15C |
| 389 | C1 | C165 | C167 | C168 | C1 | C170 | 74 | C178 | C185 | C18C | C18 | C18 | C19 | C19 | C196 | C19C |
| 38A | C19D | C19F | C1A1 | C1A5 | C1A8 | C1A9 | C1AC | C1B0 | C1BD | C1C4 | C1C8 | C1CC | C1D4 | C1D7 | C1D8 | C1E0 |
| 38B | C1E4 | C1E8 | C1F0 | C1F1 | C1F3 | C1FC | C1FD | C200 | C204 | C20C | C20D | C20F | C211 | C218 | C219 | C21C |
| 38 C | C21F | C220 | C228 | C229 | C | C22D | C22F | C231 | C232 | C234 | C248 | C250 | C25 | C254 | C258 | C260 |
| 38 | C265 | C26C | C26D | C270 | C274 | C27C | C27D | C27F | C281 | C288 | C28 | C290 | C29 | C29 | C29 | C2A4 |
| 38E | C2A5 | C2A8 | C2AC | C2AD | C2B4 | C2B5 | C2B7 | C2B9 | C2DC | C2DD | C2E0 | C2E3 | C2E4 | C2EB | C2EC | C2ED |
| 38 | C2EF | C2F1 | C2F6 | C2F8 | C2F9 | C2FB | C2FC | C300 | C308 | C309 | C30C | C30D | C313 | C314 | C315 | C318 |
| 39 | C31C | C324 | C325 | C328 | C329 | C345 | C368 | C369 | C36C | C370 | C372 | C378 | C379 | C37C | C37D | C384 |
| 391 | C388 | C38C | C3C0 | C3D8 | C3D9 | C3DC | C3DF | C3E0 | C3E2 | CЗE8 | СЗЕ9 | C3ED | C3F4 | C3F5 | C3F8 | C408 |
| 39 | C410 | C424 | C42C | C430 | C434 | C43C | C43D | C448 | C464 | C465 | C468 | C46C | C474 | C475 | C479 | C480 |
| 393 | C494 | C49C | C4B8 | C4BC | C4E9 | C4FO | C4F1 | C4F4 | C4F8 | C4FA | C4FF | C500 | C501 | C50C | C510 | C514 |
| 394 | C51C | C528 | C529 | C52C | C530 | C538 | C539 | C53B | C53D | C544 | C545 | C548 | C549 | C54A | C54C | C54D |
| 395 | C54E | C553 | C554 | C555 | C557 | C558 | C559 | C55D | C55E | C560 | C561 | C564 | C568 | C570 | C571 | C573 |
| 396 | C574 | C575 | C57C | C57D | C580 | C584 | C587 | C58C | C58D | C58F | C591 | C595 | C597 | C598 | C59C | C5A0 |
| 397 | C5A9 | C5B4 | C5B5 | C5B8 | C5B9 | C5BB | C5BC | C5BD | C5BE | C5C4 | C5C5 | C5C6 | C5C7 | C5C8 | C5C9 | C5CA |
| 398 | C5CC | C5CE | C5DO | C5D1 | C5D4 | C5D8 | C5E0 | C5E1 | C5E3 | C5E5 | C5EC | C5ED | C5EE | C5F0 | C5F4 | C5F6 |
| 399 | C5F7 | C5FC | C5FD | C5FE | C5FF | C600 | C601 | C605 | C606 | C607 | C608 | C60C | C610 | C618 | C619 | C61B |
| 39A | C61C | C624 | C625 | C628 | C62C | C62D | C62E | C630 | C633 | C634 | C635 | C637 | C639 | С63B | C640 | C641 |
| 39B | C644 | C648 | C650 | C651 | C653 | C654 | C655 | C65C | C65D | C660 | C66C | C66F | C671 | C678 | C679 | C67C |
| 39 C | C680 | C688 | C689 | C68B | C68D | C694 | C695 | C698 | C69C | C6A4 | C6A5 | C6A7 | C6A9 | C6B0 | C6B1 | C6B4 |
| 39D | C6B8 | C6B9 | C6BA | C6CO | C6C1 | C6C3 | C6C5 | C6CC | C6CD | C6DO | C6D4 | C6DC | C6DD | C6E0 | C6E1 | C6E8 |
| 39E | C6E9 | C6EC | C6F0 | C6F8 | C6F9 | C6FD | C704 | C705 | C708 | C70C | C714 | C715 | C717 | C719 | C720 | C721 |
| 39F | C724 | C728 | C730 | C731 | C733 | C735 | C737 | C73C | C73D | C740 | C744 | C74A | C74C | C74D | C74F | C751 |

Table Q. 1 (continued)

| 3A0 | 52 | C753 | C754 | C755 | C756 | C757 | C758 | C75C | C760 | C768 | C76B | C774 | C775 | C778 | C77C | C77D |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 3 A | C77E | C783 | C784 | C785 | C787 | C788 | C789 | C78A | C78E | C790 | C791 | C794 | C796 | C797 | C798 | C79A |
| 3 A 2 | C7A0 | C7A1 | C7A3 | C7A4 | C7A5 | C7A6 | C7AC | C7AD | С7B0 | C7B4 | C7BC | C7BD | C7BF | C7C0 | C7C1 | C7C8 |
| 3A3 | C7C9 | C7CC | C7CE | C7D0 | C7D8 | C7DD | C7E4 | C7E8 | C7EC | C800 | C801 | C804 | C808 | C80A | C810 | C811 |
| 3A | C813 | C815 | C816 | C81C | C81D | C820 | C824 | C82 | C82 | C82F | C8 | C8 | C83C | C840 | 48 | 49 |
| 3 A 5 | C84C | C84D | C854 | C870 | C871 | C874 | C878 | C87A | C880 | C881 | C883 | C88 | C886 | C887 | C88B | C88C |
| 3A | C88D | C894 | C89D | C89F | C8A1 | C8A8 | C8BC | C8BD | C8C4 | C8C8 | C8CC | C8D4 | C8D5 | C8D7 | C8D9 | C8E0 |
| 3A7 | C8E1 | C8E4 | C8F5 | C8FC | C8FD | 00 | C904 | 905 | C906 | C90C | c90 | C90F | C9 | C918 | 2 C | 3 |
| 3 A 8 | C950 | C951 | C954 | C958 | C960 | C961 | C963 | C96C | C970 | C974 | C97C | C988 | C989 | C98C | C990 | C998 |
| 3A9 | C999 | C99B | C99D | c9C0 | C9C1 | C9C4 | C9C7 | c9C8 | c9CA | c9D0 | C9D1 | C9D3 | C9D5 | C9D6 | c9D9 | c9DA |
| 3AA | C9DC | C9DD | C9E0 | C9E2 | C9E4 | C9E7 | C9EC | C9ED | C9EF | C9F0 | C9F1 | C9F8 | C9F9 | C9FC | ca00 | ca08 |
| 3 AB | ca09 | Саов | caoc | CAOD | CA14 | CA18 | CA29 | CA4C | CA4D | CA50 | CA54 | CA5C | CA5D | CA5F | Ca60 | CA61 |
| 3 AC | CA68 | CA7D | CA84 | CA98 | CABC | CABD | CACO | CAC4 | CACC | CACD | CACF | CAD1 | CAD3 | CAD8 | CAD9 | caeo |
| 3A | CAEC | CAF4 | CB08 | CB10 | CB14 | CB18 | CB20 | CB21 | CB41 | CB48 | CB49 | CB4C | CB5 | CB58 | CB59 | B5D |
| 3AE | CB64 | CB78 | CB79 | CB9C | CBB8 | CBD4 | CBE4 | CBE7 | CBE9 | CCOC | CCOD | CC10 | CC14 | CC1C | CC1D | CC21 |
| 3AF | CC22 | CC27 | CC28 | CC29 | CC2C | CC2E | CC30 | CC38 | CC39 | сСзв | СС3C | CC3D | ССЗЕ | CC44 | CC45 | CC48 |
| зво | CC4C | CC54 | CC55 | CC57 | CC58 | CC59 | CC60 | 64 | CC66 | CC68 | cc70 | CC75 | CC98 | CC99 | cc9C | ccao |
| 3B1 | CCA8 | CCA9 | CCAB | CCAC | CCAD | CCB4 | CCB5 | CCB8 | CCBC | CCC4 | ccc5 | CCC7 | CCC9 | CCDO | CCD4 | CCE4 |
| $3 \mathrm{B2} 2$ | CCEC | CCFO | CD01 | CD08 | CD09 | CDOC | CD10 | CD18 | CD19 | CD1B | CD1D | CD24 | CD28 | CD2C | CD39 | CD5C |
| 3B | 60 | CD64 | CD6C | CD6D | CD6F | CD71 | CD78 | CD88 | CD94 | CD95 | CD98 | CD9C | CDA4 | CDA5 | CDA7 | CDA9 |
| 3B4 | CDB0 | CDC4 | CDCC | CDDO | CDE8 | CDEC | CDFO | CDF8 | CDF9 | CDFB | CDFD | CE04 | CE08 | CEOC | CE14 | CE19 |
| 3B5 | CE20 | CE21 | CE24 | CE28 | СЕ30 | CE31 | СЕЗ3 | CE35 | CE58 | CE59 | CE5C | CE5F | CE60 | CE61 | CE68 | CE69 |
| 3B6 | CE6B | CE6D | CE74 | CE75 | CE78 | CE7C | CE84 | CE85 | CE87 | CE89 | CE90 | CE91 | CE94 | CE98 | CEAO | CEA1 |
| $3 \mathrm{B7}$ | CEA3 | CEA4 | CEA5 | CEA | CEAD | CE | CE | CEE5 | CEE8 | CEEB | Ceec | CEF4 | CEF5 | CEF7 | CEF8 | CEF9 |
| 3B8 | CFOO | CF01 | CF04 | CF08 | CF10 | CF11 | CF13 | CF15 | CF1C | CF20 | CF24 | CF2C | CF2D | CF2F | CF30 | CF31 |
| 3B9 | CF38 | CF54 | CF55 | CF58 | CF5C | CF64 | CF65 | CF67 | CF69 | CF70 | CF71 | CF74 | CF78 | CF80 | CF85 | CF8C |
| 3B | A1 | CFA8 | CFBO | CFC4 | CFEO | CFE1 | CFE4 | CFE8 | CFFO | CFF1 | CFF | CFF | CFFC | D00 | D00 | D011 |
| 3BB | D018 | D02D | D034 | D035 | D038 | D03C | D044 | D045 | D047 | D049 | D050 | D054 | D058 | D060 | D06C | D06D |
| 3BC | D070 | D074 | D07C | D07D | D081 | DOA4 | D0A5 | D0A8 | Doac | D0B4 | D0B5 | DOB7 | D0B9 | DOCO | DOC1 | DOC4 |
| 3B | D0C8 | DOC9 | DODO | DoD1 | DOD3 | DOD4 | D0D5 | DODC | DODD | DOEO | DOE4 | DOEC | DOED | DOEF | DOF | D0F1 |
| 3B | D0F8 | D10D | D130 | D131 | D134 | D138 | D13A | D140 | D141 | D143 | D144 | D145 | D14C | D14D | D150 | D154 |
| 3BF | D15C | D15D | D15F | D161 | D168 | D16C | D17C | D184 | D188 | D1A0 | D1A1 | D1A4 | D1A8 | D1B0 | D1B1 | D1B3 |
| 3 CO | D1B5 | D1BA | D1BC | D1C0 | $1{ }^{10}$ | D1F4 | D1F8 | D207 | D209 | D210 | D22C | D22D | D230 | D234 | D23C | D23D |
| $3 C$ | D23F | D241 | D248 | D25C | 264 | D280 | D281 | D284 | D288 | D290 | D291 | D295 | D29C | D2A0 | D2A4 | D2AC |
| 3 C 2 | D2B1 | D2B8 | D2B9 | D2BC | D2BF | D2C0 | D2C2 | D2C8 | D2C9 | D2CB | D2D4 | D2D8 | D2DC | D2E4 | D2E5 | D2F0 |
| $3 C 3$ | D2F1 | D2F4 | D2F8 | D300 | D301 | D303 | D305 | D30C | D30D | D30E | D310 | D314 | D316 | D31C | D31D | D31F |
| 364 | D320 | D321 | D325 | D328 | D329 | D32C | D330 | D338 | D339 | Dз3в | D33C | D33D | D344 | D345 | D37C | D37D |
| 3 C 5 | D380 | D384 | D38C | D38D | D38F | D390 | D391 | D398 | D399 | D39C | D3A0 | D3A8 | D3A9 | DзAB | D3AD | D3B4 |
| 3C6 | D3B8 | D3BC | D3C4 | D3C5 | D3C8 | D3C | D3D0 | D3D | D3E | D3 | D3EC | D3ED | D3F | D3 | D3FC | D3FD |
| $3 C 7$ | D3FF | D401 | D408 | D41D | D440 | D444 | D45C | D460 | D464 | D46D | D46F | D478 | D479 | D47C | D47F | D480 |
| 3 C 8 | D482 | D488 | D489 | D48B | D48D | D494 | D4A9 | D4CC | D4D0 | D4D4 | D4DC | D4DF | D4E8 | D4EC | D4F0 | D4F8 |
| $3 \mathrm{C9}$ | D4FB | D4FD | D504 | D508 | D50C | D514 | D515 | D517 | D53C | D53D | D540 | D544 | D54C | D54D | D54F | D551 |
| 3CA | D558 | D559 | D55C | D560 | D565 | D568 | D569 | D56B | D56D | D574 | D575 | D578 | D57C | D584 | D585 | D587 |
| 3CB | D588 | D589 | D590 | D5A5 | D5C8 | D5C9 | D5CC | D5D0 | D5D2 | D5D8 | D5D9 | D5DB | D5DD | D5E4 | D5E5 | D5E8 |
| 3 CC | D5EC | D5F4 | D5F5 | D5F7 | D5F9 | D600 | D601 | D604 | D608 | D610 | D611 | D613 | D614 | D615 | D61C | D620 |
| 3 CD | D624 | D62D | D638 | D639 | D63C | D640 | D645 | D648 | D649 | D64B | D64D | D651 | D654 | D655 | D658 | D65C |
| 3CE | D667 | D669 | D670 | D671 | D674 | D683 | D685 | D68C | D68D | D690 | D694 | D69D | D69F | D6A1 | D6A8 | d6ac |
| 3CF | D6B0 | D6B9 | D6BB | D6C4 | D6C5 | D6C8 | D6CC | D6D1 | D6D4 | D6D7 | D6D9 | D6E0 | D6E4 | D6E8 | D6F0 | D6F5 |

Table Q. 1 (continued)

|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 3D0 | D6FC | D6FD | D700 | D704 | D711 | D718 | D719 | D71C | D720 | D728 | D729 | D72B | D72D | D734 | D735 | D738 |
| 3D | D73C | D7 | D7 | D7 | 50 | D751 | 54 | 756 | D757 | 758 | D759 | 760 | D761 | D763 | D765 | D7 |
| 3 D 2 | D76C | D770 | D774 | D77C | D77 | D781 | D788 | D78 | D78C | D790 | D798 | D79 | D79B | D79D | AC02 | ACOB |
| 3 D 3 | ACOC | AC22 | AC23 | AC32 | AC35 | AC36 | AC3F | AC41 | AC47 | AC48 | AC49 | AC4C | AC64 | AC65 | AC73 | AC75 |
| 3 D 4 | AC79 | AC87 | AC8D | AC93 | ACA5 | ACA7 | ACB1 | ACB4 | ACB7 | ACBE | ACBF | ACC2 | ACC5 | ACCB | ACD4 | ACD8 |
| 3D | ACD | ACE9 | ACEB | ACEE | ACF7 | ACF9 | ACFA | ACFB | AD03 | AD10 | AD | AD1F | AD | AD2 | AD2 | AD3B |
| 3D6 | AD3E | AD48 | AD51 | AD57 | AD60 | AD65 | AD78 | AD83 | AD86 | AD8F | AD99 | AD9B | ADA5 | ADA8 | ADA | da |
| 3 D 7 | ADB4 | ADB5 | ADB8 | ADB9 | ADC7 | ADCA | ADD4 | ADD5 | ADDD | ADE8 | ADEC | ADED | ADEF | ADF1 | ADF | AE02 |
| 3 D 8 | AE03 | AE04 | AE07 | AEOE | AEOF | AE11 | AE12 | AE13 | AE15 | AE18 | AE1C | AE2 | AE24 | AE25 | AE27 | 29 |
| $3 \mathrm{D9}$ | AE39 | AE3C | AE44 | AE47 | AE49 | AE4B | AE53 | AE | AE63 | AE6A | AE6F | AE73 | AE76 | AE81 | AE88 | AE |
| 3DA | AE8D | AE97 | AE99 | AEAO | AEB5 | AEC2 | AEC3 | AED2 | AED3 | AED5 | AEDA | AEDF | AEEO | AEE9 | AEEC | AEF1 |
| 3DB | AEF5 | AEFB | AF04 | AF05 | AF09 | AF17 | AF25 | AF33 | AF36 | AF38 | AF3B | AF45 | AF47 | AF4C | AF4 | AF58 |
| 3 CC | AF59 | AF5B | AF68 | AF6B | AF6C | AF74 | AF75 | AF78 | AF81 | AF87 | AF93 | AF94 | AFAO | AFA3 | AFA4 | AFAC |
| 3DD | AFAD | AFB2 | AFBF | AFC1 | AFCF | AFD5 | AFD8 | AFDB | AFE4 | AFE7 | AFF7 | B003 | B005 | B00D | B013 | B01F |
| 3D | B021 | B02C | B030 | B038 | B039 | B04B | B04D | B05B | B05F | B060 | B061 | B067 | B068 | B06B | B073 | B075 |
| 3DF | B083 | B08B | B090 | B095 | B09B | B0A4 | BOAA | B0B0 | вов2 | BobB | B0C | B0D7 | B0E1 | B0E3 | B0E6 | BoE9 |
| 3E0 | B0EC | B101 | B10A | B10F | B117 | B11E | B120 | B121 | B122 | B12B | B13C | B13D | B13 | B13F | B143 | B147 |
| 3 E 1 | B14B | B153 | B159 | B15A | B15B | B15D | B163 | B164 | B16C | B16D | B16F | B171 | B17A | B17B | B17E | B17F |
| 3E | B18 | B190 | B1 | B195 | B1 | B1A4 | B1A5 | B1A7 | B1A9 | B180 | B1 | B188 | C4 | B1CD | B1D | B1E0 |
| 3E3 | B1E1 | B1E6 | B1EF | B1F8 | B20D | B210 | B213 | B21B | B21E | B221 | B224 | B227 | B228 | B230 | B231 | B233 |
| 3E4 | B235 | B23D | B240 | B243 | B24 | B24C | B24D | B24F | B250 | B251 | B259 | B25 | B26B | B26D | B26F | B278 |
| 3 E5 | B27B | B28 | B28 | B28B | B297 | B29C | B2A7 | B2A | B2A | B2B | B2BC | B2B | B2B | B2C | B2C | B2CF |
| 3 E6 | B2D1 | B2D3 | B2 | B2 | B2 | B2E3 | B2F0 | B2F2 | B2F6 | B2FC | B2FD | B307 | B319 | B31D | B320 | B324 |
| 3E7 | B327 | B32C | B32D | B32F | B331 | B338 | Вз3С | B34D | B359 | B366 | B368 | B36A | B36D | B36F | B377 | B386 |
| 3E8 | B38A | 38 | B38F | B39 | 98 | B39C | B39D | B39F | ЗЗА | взво | B3B8 | В3в9 | В3 | B3B | B3C | B3C7 |
| 3Е9 | B3CF | B3D3 | B3DA | B3DC | B3DE | B3DF | B3E1 | B3F0 | B3F1 | B3F3 | B3F4 | B3F5 | B400 | B403 | B404 | B40C |
| 3EA | B40D | B40F | B419 | B41F | B424 | B42C | B42D | B435 | B438 | B43B | B43C | B444 | B445 | B447 | B449 | B44F |
| 3 B | B457 | B459 | B45A | B45B | B46A | B4 | B470 | B4 | B474 | B47 | B47 | B47 | B481 | B48 | B48 | B48F |
| 3EC | B490 | B498 | B499 | B49B | B49C | B4A5 | B4AB | B4B4 | B4B8 | B4C1 | B4 | B4D | B4E | B4E7 | B4E | B4F9 |
| 3ED | B4FC | B4FF | B500 | B508 | B509 | B50B | B50D | B52B | B52D | B52E | B52F | B532 | B537 | B539 | B53A | B53B |
| 3EE | B53F | B54E | B553 | B567 | B568 | B569 | B56C | B57D | B584 | B588 | B5A7 | B5AF | B5C3 | B5D9 | B5DC | B5DF |
| 3 E | B5E8 | B5 | B5 | B5ED | B5F4 | B5F8 | B605 | B612 | B617 | B619 | B61A | B6 | B62 | B62 | B6 | B62D |
| 3 FO | B630 | B641 | B649 | B64C | B64F | B650 | B658 | B659 | B65B | B65C | B665 | B66B | B66C | B674 | B675 | B677 |
| 3 F 1 | B678 | B679 | B680 | B681 | B6A3 | B6A6 | B6A | B6AD | B6A | B6B | B6B8 | B6BC | B6B | B6C | B6D8 | B6DB |
| 3F2 | B6DC | B6E4 | B6E5 | B6E8 | B6E9 | B6F7 | B703 | B70C | B70D | B714 | B71C | B721 | B732 | B733 | B737 | B73D |
| 3F3 | B759 | B761 | B767 | B77B | B783 | B788 | B793 | B794 | B795 | B79F | B7B0 | B7B1 | B7B2 | B7BB | B7BC | B7C4 |
| $3 F 4$ | B7C5 | B7D0 | B7E3 | B7F2 | B7F3 | B7FE | B802 | B804 | B806 | B80F | B81C | B821 | B82 | B823 | B82B | B830 |
| 3 F 5 | B83C | B83E | B841 | B847 | B848 | B850 | B855 | B863 | B868 | B86B | B874 | B876 | B877 | B879 | B880 | B888 |
| 356 | B889 | B88B | B88C | B894 | B898 | B89B | B89C | B8A7 | B8B1 | B8B5 | B8B7 | B8C4 | B8CD | B8D3 | B8DC | B8EF |
| $3 F 7$ | B8F3 | B900 | B902 | B905 | B908 | B90B | B90C | B914 | B915 | B917 | B919 | B921 | B924 | B928 | B930 | B931 |
| 358 | B933 | B934 | B935 | B943 | B94D | B950 | B95F | B97B | B97D | B980 | B98B | B98 | B990 | B991 | B994 | B998 |
| 3F9 | B99E | B9A0 | B9A1 | B9A3 | B9A5 | B9B3 | B9BE | B9C0 | B9C4 | B9C6 | B9CA | B9D4 | B9DC | B9DF | B9E0 | B9E2 |
| 3 FA | B9EB | B9ED | B9FB | B9FD | B9FE | BA04 | BA10 | BA11 | BA13 | BA18 | BA1C | BA3B | BA3F | BA41 | BA4C | BA4F |
| 3FB | BA5B | BA60 | BA6A | BA6B | BA6D | BA77 | BA7A | BA80 | BA81 | BA89 | BA8D | BA90 | BA93 | BA94 | BA9C | BA9D |
| 3FC | BA9F | BAA1 | BAA3 | baA5 | baA6 | BAAF | BAB1 | BAB4 | BABF | BAC3 | BAC5 | BACB | BACC | BAD4 | BAD5 | BAD7 |
| 3FD | BAEO | BAE4 | BAE8 | BAF1 | BAF4 | BAFD | BB03 | BBOC | BB19 | BB1F | BB28 | BB2D | BB3A | BB40 | BB4B | BB51 |
| 3 FE | BB57 | BB60 | BB64 | BB65 | BB6D | BB70 | BB74 | BB7C | BB7D | BB7F | BB80 | BB81 | BB89 | BB8A | BB98 | BB99 |
| 3FF | BB9B | BB9C | BB9D | BBA5 | BB | BBB | B9 | BBC | BBC | BBC9 | BBCC | BBCF | BBD1 | BBD5 | BBD9 | C |

Table Q. 1 (continued)

|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 400 | BD | BBEO | BBE4 | BBEC | BBED | BBEF | BBF1 | BBF2 | BC01 | BC04 | BCOE | BC10 | BC20 | BC23 | BC28 | BC2B |
| 401 | BC2C | BC2F | BC37 | BC46 | BC54 | BC5C | BC5F | BC61 | BC67 | BC68 | BC70 | BC77 | BC7D | BC86 | BC8D | BC90 |
| 40 | BC98 | BC9C | BC9D | BCA2 | BCB8 | BCB9 | BCC3 | BCC8 | BCCC | BC | BCD3 | BCD4 | 9 | BCEO | E8 | E9 |
| 403 | BCEB | BCED | BCF | BCFB | BCFD | BCFF | BDOA | BDOB | BDOD | BDOF | BD11 | BD17 | BD18 | BD20 | BD21 | BD23 |
| 404 | BD25 | BD30 | BD33 | BD34 | BD4A | BD4F | BD5B | BD5C | BD5D | BD65 | BD6C | BD74 | BD75 | BD77 | BD79 | BD82 |
| 405 | BD8B | BD8E | BD96 | BD97 | BD98 | BD9B | BD9D | bDaO | BDA3 | BDAC | BDAD | BDAF | BDB1 | BDB4 | BDB9 | BDBC |
| 406 | BDBF | BDCO | BDC8 | BDC9 | BDCB | BDCC | BDCD | BDDB | BDE4 | BDE5 | BDE7 | BD | BDF7 | BE01 | BE13 | BE15 |
| 407 | BE17 | BE18 | BE1B | BE21 | BE23 | BE25 | BE27 | BE28 | BE29 | BE2C | BE30 | BE38 | ВЕ39 | ВЕЗВ | BE3D | BE4B |
| 408 | BE58 | BE5C | BE5D | BE5F | BE62 | BE67 | BE69 | BE76 | BE79 | BE7E | BE83 | BE9C | BEB4 | BEB8 | BEE1 | BEE6 |
| 409 | beed | BEFO | BEF3 | BEF4 | BEFC | BEFD | BE | BFOC | BFOF | BF10 | BF1F | BF21 | BF24 | 37 | F38 | 39 |
| 40A | BF47 | BF53 | BF5B | BF5C | BF60 | BF63 | BF | BF7F | BFA4 | BFA5 | BFAC | BFCO | BFC1 | BFD3 | BFD5 | BFDD |
| 40B | BFE8 | BFEF | C004 | C020 | C021 | C 043 | C044 | C 059 | C05F | C06B | C074 | C 097 | COA3 | COA6 | COA 7 | COAB |
| 40 C | COAE | C0B7 | COB8 | COBA | COBB | COC2 | COC3 | coc4 | COC6 | C0C7 | COCB | COCF | COE3 | COEB | C0F8 | COFB |
| 40D | COFE | COFF | C125 | C128 | C12A | C134 | C13F | C14E | C151 | C152 | C157 | C15B | C15F | C160 | C171 | C173 |
| 40E : | C180 | C181 | C183 | C184 | C193 | C195 | 197 | C198 | C1A3 | C1A6 | C1B8 | C1B9 | C1BB | C1BC | C1C5 | C1CB |
| 40F: | D5 | C1D9 | C1E1 | C1E7 | $1 F 4$ | C1F5 | 203 | C216 | C221 | C224 | C227 | C22E | C233 | C235 | C238 | С23B |
| 410 | C23C | C244 | C245 | C247 | C249 | C257 | C261 | C263 | C264 | C273 | C286 | C28C | C28F | C299 | C2AB | C2AE |
| 411 | C2AF | С2во | С2в2 | С2B3 | C2BA | C2BB | C2BE | C2CO | C2C1 | C2C4 | C2C8 | C2DO | C2D1 | C2D3 | C2D5 | C2de |
| 412 : | C2E2 | C2E5 | C2E | C2E8 | C2FO | C2F | C2F4 | 2 FF | 301 | C302 | сзов | C30E | 311 | C31B | C327 | C32F |
| 413 | С330 | C331 | C334 | C337 | C338 | C340 | C341 | C343 | C34C | C350 | C354 | C35C | C361 | С36A | C36F | C37B |
| 414 | C382 | С385 | С38B | С394 | C395 | C397 | C398 | C399 | C39D | СЗa0 | C3A1 | C3A4 | С3A7 | СЗА8 | сзво | С3В1 |
| 415 | B3 | С3B | С3B5 | 3BC | C3BD | C3CC | C3CD | C3CF | 3D0 | C3D1 | C3E | C3F1 | FB | C3F | C404 | C405 |
| 416 | C407 | C40 | C411 | C4 | C417 | C418 | C423 | C42D | C433 | C43F | C440 | C441 | C4 | C44C | C44F | C450 |
| 417 | C458 | C459 | C45B | C45D | C46B | C477 | C47E | C481 | C484 | C487 | C488 | C490 | C491 | C493 | C495 | C49D |
| 418 | C4AO | C4A3 | C4A4 | C4A | C4AD | C4A | C4B0 | C4B1 | C4B9 | CABF | C4CO | C4C8 | C4C9 | C4CB | C4CD | C4D3 |
| 419 | C4D4 | C4D5 | C | C4DB | C4DC | C4E4 | C4E5 | 4E7 | C4F7 | C503 | C50 | C50D | C5 | C5 | C521 | C52F |
| 41A | C531 | C53C | C53F | C540 | C543 | C54B | C54F | C552 | C556 | C55A | C55B | C55F | C567 | C579 | C57A | C57E |
| 41 B | C583 | C590 | C592 | C594 | C599 | C59F | C5A8 | C5AB | C5AC | C5AD | C5B6 | C5BA | C5BF | C5CF | C5D7 | C5E4 |
| 41 C | C5E | C5E | C5F1 | C5F3 | C5F8 | C604 | C609 | 607 | 61 D | C620 | C626 | C62 | C62B | C62F | C63 | C63A |
| 41D | C63D | C63E | C647 | C658 | C659 | C663 | C664 | C66D | C670 | C67F | C682 | C68C | C692 | C69B | C69D | C6AC |
| 41E: | C6B7 | C6BC | C6C2 | C6C6 | C6C7 | C6C9 | C6D2 | C6D3 | C6DF | C6E4 | C6E5 | C6EF | C6FB | C6FC | C701 | C70B |
| 41 F | C70E | C71 | C718 | C71C | C71D | C727 | C736 | C738 | C739 | C743 | C745 | C746 | C747 | C74E | C759 | C75F |
| 420 | C766 | C769 | C76D | C77B | C780 | C782 | C786 | C78B | C78C | C78D | C78F | C793 | C799 | C7A7 | C7A9 | C7AA |
| 421 : | C7AB | C7B2 | C7B3 | C7C2 | C7CF | C7D9 | C7DB | C7EB | C7F4 | C7F5 | C7F7 | C7F9 | C802 | C806 | C807 | C809 |
| 422 : | C814 | C819 | C81B | C823 | C830 | C832 | C839 | C83F | C841 | C842 | C843 | C847 | C84B | C84E | C851 | C853 |
| 423 | C855 | C858 | C85C | C864 | C865 | C867 | C869 | C877 | C890 | C892 | C893 | C895 | C89C | C8AO | C8A9 | C8AC |
| 424 | C8AF | C8B0 | C8B8 | C8BB | C8C5 | C8CB | C8D8 | C8E7 | C8E8 | C8FO | C8F | C8F3 | C8FB | C903 | C908 | C917 |
| 425 : | C919 | C91C | C91F | C920 | C 928 | C 929 | C92B | C92D | C935 | C938 | С93в | C93C | C944 | C 945 | C947 | C948 |
| 426 : | C949 | C957 | C965 | C96D | C97D | C97F | C981 | C98F | C991 | C992 | C994 | C99E | C9A4 | C9A5 | C9A8 | c9ac |
| 427 : | c9B4 | С9B5 | C9B7 | с9в9 | C9CF | C9D2 | C9D4 | C9D7 | C9DB | cade | C9E3 | C9E8 | C9F7 | C9FF | CA15 | CA1A |
| 428 : | CA1C | CA24 | CA25 | CA27 | CA2D | CA30 | CA53 | CA57 | CA58 | CA5B | CA67 | CA69 | CA6C | CA6F | CA70 | CA78 |
| 429 | CA79 | CA7B | CA7C | CA81 | CA85 | CA88 | CA8B | CA8C | CA94 | CA95 | CA97 | CA99 | CAAO | CAA | CAA4 | CAA8 |
| 42A | CABO | CAB1 | CAB3 | CAB5 | CABE | CAC3 | CAC6 | CAD2 | CAD7 | CADC | CADF | CAE8 | CAE9 | CaEb | CAED | CAF5 |
| 42B | CAF8 | CAFB | CAFC | CB11 | CB17 | CB23 | CB24 | CB25 | CB27 | CB2C | CB2D | CB30 | CB34 | CB3C | CB3D | CB3 |
| 42 C | CB4A | CB4F | CB52 | CB5B | CB65 | CB68 | CB6B | CB6C | CB74 | CB75 | CB77 | CB80 | CB81 | CB84 | CB87 | CB88 |
| 42D | CB90 | CB91 | CB93 | CB95 | CB9D | CBAO | CBA3 | CBA4 | CBAC | CBAD | CBAF | CBB1 | CBB9 | CBBC | CBCO | CBC8 |
| 42E: | CBC9 | СвСв | CBCD | CBD5 | CBD8 | CBDB | CBDC | CBE5 | CBEA | CBFO | CBF1 | CBF4 | CBF8 | CCOO | CC01 | CCO3 |
| 42F : | CC05 | CC06 | CC13 | CC1F | CC26 | CC2F | CC31 | CC3F | CC42 | CC4B | CC5B | CC5E | CC61 | CC71 | CC73 | CC7A |

Table Q. 1 (continued)

| 430 | CC7C | CC91 | CC9F | CCA7 | CCAE | CCB2 | CCBB | CCC3 | CCC8 | CCD1 | CCD7 | CCD8 | CCEO | CCE1 | CCE3 | CCE5 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 431 | CCED | CCF4 | CCFC | CCFD | CCFF | CDOF | CD25 | CD2B | CD34 | CD35 | CD37 | CD40 | CD53 | CD54 | CD5D | CD63 |
| 432 | CD70 | CD79 | CD7C | CD80 | CD89 | CD8B | CD8D | CD9B | CDB1 | CDB4 | CDB7 | CDB8 | CDCO | CDC1 | CDC3 | CDC5 |
| 433 | CDCD | CDD4 | CDDC | CDDD | CDDF | CDEO | CDE1 | CDE9 | CDEF | CE05 | CE15 | CE17 | CE27 | CE29 | CE2C | СЕЗС |
| 434 | CE3D | CE40 | CE44 | CE4C | CE4D | CE4F | CE51 | CE62 | CE6C | CE6E | CE70 | CE72 | CE7B | CE88 | CE8D | CE8E |
| 435 | CE97 | CEA9 | CEAA | CEBO | CEB4 | cebc | cebd | CEBF | CEC8 | CECC | CEFB | CEFD | CEFE | CF07 | CF14 | CF19 |
| 436 | CF1A | CF1D | CF23 | CF39 | CF3C | CF40 | CF48 | CF49 | CF4B | CF4C | CF4D | CF5B | CF6B | CF6E | CF81 | CF83 |
| 437 | CF8D | CF90 | CF93 | CF94 | CF9C | CF9D | CF9F | CFA9 | CFAC | CFB8 | CFB9 | CFBB | CFBD | CFC5 | CFC8 | CFCC |
| 438 | CFD4 | CFD5 | CFD7 | CFD9 | CFE7 | CFFA | CFFD | D003 | Dooc | D00D | D010 | D019 | D01C | D020 | D028 | D029 |
| 439 | D02B | D02C | D03B | D051 | D061 | D063 | D065 | D068 | D072 | D076 | D07F | D088 | D089 | D08C | D090 | D098 |
| 43A | D099 | D09B | D09D | DoAB | D0B8 | DOBE | D0C7 | DOCA | D0CF | DODA | DOE2 | DOE3 | DOE5 | D0F6 | D0F9 | DOFC |
| 43B | D100 | D108 | D109 | D10B | D114 | D118 | D11A | D137 | D139 | D13B | D153 | D160 | D166 | D169 | D16F | D170 |
| 43 C | D178 | D179 | D17B | D17D | D185 | D18C | D194 | D197 | D199 | D1A7 | D187 | D1B8 | D189 | D1BD | D1C4 | D1CC |
| 43D : | D1CD | D1CF | D1D1 | D1EC | D1ED | D1F5 | D1FB | D1FC | D204 | D205 | D208 | D211 | D214 | D218 | D220 | D221 |
| 43 E | D223 | D225 | D233 | D236 | D249 | D24C | D24F | D250 | D254 | D258 | D259 | D25B | D25D | D265 | D268 | D26B |
| 43F | D26C | D274 | D275 | D277 | D278 | D279 | D287 | D293 | D29D | D2AD | D2AF | D2BB | D2C1 | D2C4 | D2CD | D2D5 |
| 440 | D2E0 | D2E7 | D2E9 | D2F7 | D308 | D30A | D313 | D323 | D326 | D32F | D341 | D348 | D34C | D354 | D355 | D357 |
| 441 | D359 | D360 | D383 | D38E | D395 | D39F | D3AC | D3B2 | D3B5 | D389 | D3BB | D3C7 | D3D1 | D3D4 | D3D7 | D3E0 |
| 442 | D3E5 | D3EE | D3F3 | D404 | D405 | D406 | D409 | D40C | D410 | D418 | D419 | D41B | D424 | D441 | D448 | D450 |
| 443 | D451 | D453 | D455 | D45D | D463 | D46C | D471 | D495 | D498 | D49B | D49C | D4A4 | D4A5 | D4A7 | D4A8 | D4B0 |
| 444 | D4B1 | D4B4 | D4B8 | D4C0 | D4C1 | D4C3 | D4C5 | D4CD | D4DD | D4E1 | D4E9 | D4EF | D4F9 | D505 | D50B | D510 |
| 445 | D519 | D520 | D521 | D524 | D528 | D530 | D531 | D533 | D535 | D543 | D555 | D556 | D55F | D561 | D563 | D564 |
| 446 | D567 | D56C | D56E | D571 | D57B | D58D | D591 | D594 | D598 | D5A0 | D5A1 | D5A3 | D5AB | D5AC | D5C0 | D5CF |
| 447 | D5D1 | D5D7 | D5DC | D5E1 | D5E2 | D5EB | D5F8 | D5FE | D607 | D61D | D62C | D62F | D630 | D631 | D63D | D63F |
| 448 | D641 | D644 | D647 | D65B | D660 | D664 | D665 | D666 | D668 | D677 | D678 | D680 | D681 | D684 | D693 | D69C |
| 449 | D6A9 | D6AF | D6B8 | D6BD | D6CB | D6CD | D6CE | D6D2 | D6D3 | D6D5 | D6DC | D6DD | D6E1 | D6F1 | D6F3 | D6F4 |
| 44A | D703 | D70C | D70D | D70F | D710 | D71F | D73A | D73B | D743 | D745 | D74D | D755 | D75D | D75 | D768 | D76A |
| 44 B | D76B | D76D | D773 | D77F | D78F | D797 | D79C | D7A0 | BBC3 | D63B | BF59 | BFE5 | CB94 | C6D8 | AC03 | AC05 |
| 44C : | AC06 | ACOD | ACOE | ACOF | AC18 | AC1E | AC1F | AC21 | AC25 | AC26 | AC27 | AC28 | AC29 | AC2A | AC2B | AC2E |
| 44 | AC33 | AC34 | AC37 | AС3A | АСЗВ | AC3D | AC3E | AC42 | AC43 | AC44 | AC45 | AC46 | AC4A | AC4E | AC4 | AC50 |
| 44E | AC51 | AC52 | AC53 | AC55 | AC56 | AC57 | AC59 | AC5A | AC5B | AC5D | AC5E | AC5F | AC60 | AC61 | AC62 | AC63 |
| 44F | AC66 | AC67 | AC68 | AC69 | AC6A | AC6B | AC6C | AC6D | AC6E | AC6F | AC72 | AC76 | AC7B | AC7C | AC7D | AC7E |
| 450 | AC7F | AC82 | AC88 | AC8E | AC8F | AC91 | AC92 | AC95 | AC96 | AC97 | AC98 | AC99 | AC9A | AC9B | AC9E | ACA2 |
| 451 | ACA3 | ACA4 | ACA6 | ACAB | ACAD | ACAE | ACB2 | ACB3 | ACB5 | ACB6 | ACBA | ACCO | ACC3 | ACC6 | ACC7 | ACC9 |
| 452 | ACCA | ACCD | ACCE | ACCF | ACDO | ACD1 | ACD2 | ACD3 | ACD6 | ACDA | ACDB | ACDC | ACDD | ACDE | ACDF | ACE2 |
| 453 | ACE3 | ACE5 | ACE6 | ACED | ACF2 | ACF4 | ACF8 | ACFE | ACFF | AD01 | AD02 | AD05 | AD07 | AD08 | AD09 | ADOA |
| 454 | ADOB | ADOE | AD12 | AD13 | AD14 | AD15 | AD16 | AD17 | AD1A | AD1B | AD1D | AD1E | AD21 | AD23 | AD24 | AD25 |
| 455 | AD26 | AD27 | AD2A | AD2E | AD2F | AD30 | AD31 | AD32 | AD33 | AD36 | AD37 | AD39 | AD3A | AD3D | AD3F | AD40 |
| 456 | AD41 | AD42 | AD43 | AD46 | AD4A | AD4B | AD4C | AD4D | AD4E | AD4F | AD52 | AD53 | AD55 | AD56 | AD59 | AD5A |
| 457 | AD5B | AD5C | AD5D | AD5E | AD5F | AD62 | AD64 | AD66 | AD67 | AD68 | AD69 | AD6A | AD6B | AD6E | AD6F | AD71 |
| 458 | AD72 | AD77 | AD79 | AD7A | AD7E | AD80 | AD84 | AD85 | AD87 | AD8A | AD8B | AD8D | AD8E | AD91 | AD92 | AD93 |
| 459 | AD94 | AD95 | AD96 | AD97 | AD98 | AD9A | AD9E | AD9F | ADA0 | ADA1 | ADA2 | ADA3 | ADA6 | ADA7 | ADA9 | ADAA |
| 45A | ADAD | ADAE | ADAF | ADBO | ADB1 | ADB2 | ADB3 | ADB6 | ADBA | ADBB | ADBC | ADBD | ADBE | ADBF | ADC2 | ADC3 |
| 45B | ADC5 | ADC6 | ADC9 | ADCB | ADCC | ADCD | ADCE | ADCF | ADD2 | ADD6 | ADD7 | ADD8 | ADD9 | ADDA | ADDB | ADDE |
| 45C : | ADDF | ADE1 | ADE2 | ADE3 | ADE5 | ADE6 | ADE7 | ADE9 | ADEA | ADEB | adee | ADFO | ADF2 | ADF3 | ADF4 | ADF5 |
| 45D : | ADF6 | ADF7 | ADFA | ADFB | ADFD | AE05 | AE06 | AEOA | AEOC | AE10 | AE16 | AE17 | AE19 | AE1A | AE1B | AE1D |
| 45E | AE1E | AE1F | AE21 | AE22 | AE23 | AE26 | AE28 | AE2A | AE2B | AE2C | AE2D | AE2E | AE2F | AE32 | AE33 | AE35 |
| 45F | AE36 | АЕЗВ | AE3D | АЕЗЕ | AE3F | AE42 | AE48 | AE4F | AE51 | AE52 | AE55 | AE57 | AE58 | AE59 | AE5A | AE5B |

Table Q. 1 (continued)

| 460 | AE5E | AE64 | AE66 | AE67 | AE6B | AE6D | AE6E | AE71 | AE72 | AE74 | AE75 | AE77 | AE7A | AE7E | AE7F | AE80 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 461 | AE82 | AE83 | AE86 | AE87 | AE89 | AE8A | AE8E | AE8F | AE90 | AE91 | AE92 | AE93 | AE94 | AE95 | AE96 | AE98 |
| 462 | AE9A | AE9B | AE9C | AE9D | AE9E | AE9F | AEA1 | AEA2 | АЕА3 | AEA4 | AEA5 | AEA6 | AEA7 | AEA8 | AEA9 | AEAA |
| 463 | AEAB | AEAC | AEAD | AEAE | AEAF | AEBO | AEB1 | AEB2 | AEB3 | AEB4 | AEB6 | AEB7 | AEB8 | AEB9 | AEBA | AEbB |
| 464 | AEBF | AEC1 | AEC5 | AEC6 | AEC7 | AEC8 | AEC9 | AECA | AECB | AECE | AED4 | AED6 | AED7 | AEDB | AEDD | AEDE |
| 465 | AEE1 | AEE2 | AEE3 | AEE4 | AEE5 | AEE6 | AEE7 | AEEA | AEEE | AEEF | AEFO | AEF2 | AEF3 | AEF6 | AEF7 | AEF9 |
| 466 | AEFA | AEFD | AEFE | AEFF | AFOO | AF01 | AF02 | AF03 | AF06 | AFOA | AFOB | AFOC | AFOE | AFOF | AF11 | AF12 |
| 46 | AF13 | AF14 | AF15 | AF16 | AF18 | AF19 | AF1A | AF1B | AF1C | AF1D | AF1E | AF1F | AF20 | AF21 | 22 | AF23 |
| 468 | AF24 | AF26 | AF27 | AF28 | AF29 | AF2A | AF2B | AF2E | AF2F | AF31 | AF35 | AF37 | AF39 | AF3A | AF3E | AF40 |
| 469 | AF44 | AF46 | AF4A | AF4B | AF4D | AF4E | AF51 | AF52 | AF53 | AF54 | AF55 | AF56 | AF57 | AF5A | AF5E | AF5F |
| 46 | AF60 | AF61 | AF62 | AF63 | AF66 | AF67 | AF69 | AF6A | AF6D | AF6E | AF6F | AF70 | AF71 | AF72 | AF73 | AF76 |
| 46B | AF77 | AF7A | AF7B | AF7C | AF7D | AF7E | AF7F | AF82 | AF83 | AF85 | AF86 | AF89 | AF8A | AF8B | AF8C | AF8D |
| 46 C | AF8E | AF8F | AF92 | AF96 | AF97 | AF98 | AF99 | AF9A | AF9B | AF9D | AF9E | AF9F | AFA1 | AFA2 | AFA5 | AFA6 |
| 46D | AFA7 | AFA8 | AFA9 | AFAA | AFAB | AFAE | AFAF | AFB0 | AFB1 | AFB3 | AFB4 | AFB5 | AFB6 | AFB7 | AFBA | AFBB |
| 46E | AFBD | AFBE | AFC2 | AFC3 | AFC4 | AFC5 | AFC6 | AFCA | AFCC | AFDO | AFD1 | AFD2 | AFD3 | AFD6 | AFD7 | AFD9 |
| 46F | AFDA | AFDD | AFDE | AFDF | AFEO | AFE1 | AFE2 | AFE3 | AFE5 | AFE6 | AFEA | AFEB | AFEC | AFED | AFEE | AFEF |
| 470 | AFF2 | AFF3 | AFF5 | AFF6 | AFF9 | AFFA | AFFB | AFFC | AFFD | AFFE | AFFF | B002 | B006 | B007 | B008 | B009 |
| 471 | B00A | B00B | B00E | B00F | B011 | B012 | B015 | B016 | B017 | B018 | B019 | B01A | B01B | B01E | B020 | B022 |
| 472 | B023 | B024 | B025 | B026 | B027 | B029 | B02A | B02B | B02D | B02E | B02F | B031 | B032 | B033 | B034 | B035 |
| 473 | B036 | B037 | B03A | возв | B03C | B03D | B03E | B03F | B040 | B041 | B042 | B043 | B046 | B047 | B049 | B04F |
| 474 | B050 | B051 | B052 | B056 | B058 | B05A | B05C | B05E | B062 | B063 | B064 | B065 | B066 | B069 | B06A | B06C |
| 475 | B06D | B06E | B06F | B070 | B071 | B072 | B074 | B076 | B077 | B078 | B079 | B07A | B07B | B07E | B07F | B081 |
| 476 | B082 | B085 | B086 | B087 | B088 | B089 | B08A | B08E | B092 | B093 | B094 | B096 | B097 | B09D | B09E | B0A3 |
| 477 | B0A5 | B0a6 | B0A7 | Bob6 | B0B7 | вов9 | Boba | Bobd | вов | B0BF | Boco | B0C1 | Bос | восз | B0C6 | BOCA |
| 478 | BOCB | BOCC | BOCD | BOCF | BOD2 | B0D3 | B0D5 | BOD6 | BOD9 | BODA | BODB | BODC | BODD | BODE | BODF | B0E2 |
| 479 | B0E4 | B0E7 | B0E8 | B0EA | BOEB | BoEd | boee | BoEF | B0FO | B0F1 | B0F2 | B0F3 | B0F4 | B0F5 | B0F6 | B0F7 |
| 47A | B0F8 | B0F9 | BOFA | B0FB | BJC | BOFD | BOFE | B0FF | B100 | B102 | B103 | B104 | B105 | B106 | B107 | B10D |
| 47B | B10E | B111 | B114 | B115 | B116 | B11A | B11F | B126 | B127 | B129 | B12A | B12D | B12E | B12F | B130 | B131 |
| 47 C | B132 | B133 | B136 | B13A | B13B | B142 | B145 | B146 | B149 | B14A | B14C | B14D | B14E | B14F | B152 | B156 |
| 47 | B157 | B15E | B15F | B161 | B162 | B165 | B166 | B167 | B168 | B169 | B16A | B16B | B16 | B170 | B172 | B173 |
| 47 | B174 | B175 | B176 | B177 | B17D | B181 | B183 | B184 | B185 | B186 | B187 | B18A | B18C | B18F | B196 | B197 |
| 47F | B199 | B19A | B19D | B19E | B19F | B1A0 | B1A1 | B1A2 | B1A3 | B1A6 | B1AA | B1AB | B1AC | B1AD | B1AE | B1AF |
| 480 | B1B1 | B1B2 | B1B3 | B1B5 | B1B6 | B1B7 | B1B9 | B1BA | B1BB | B1BC | B1BD | B1BE | B1BF | B1C0 | B1C1 | B1C2 |
| 481 | B1C3 | B1C5 | B1C6 | B1C7 | B1C8 | B1C9 | B1CA | B1CB | B1CE | B1CF | B1D1 | B1D2 | B1D5 | B1D6 | B1D7 | B1D8 |
| 482 | B1D9 | B1DA | B1DB | B1DE | B1E2 | B1E3 | B1E4 | B1E5 | B1E7 | B1EA | B1EB | B1ED | B1EE | B1F1 | B1F2 | B1F3 |
| 483 | B1F4 | B1F5 | B1F6 | B1F7 | B1FA | B1FC | B1FE | B1FF | B200 | B201 | B202 | B203 | B206 | B207 | B209 | B20A |
| 484 | B20E | B20F | B211 | B212 | B216 | B218 | B21A | B21C | B21D | B21F | B222 | в223 | B225 | B226 | B229 | B22A |
| 485 | B22B | B22C | B22D | B22E | B22F | B232 | B236 | B237 | B238 | B239 | B23A | В23в | В23E | B23F | B241 | B242 |
| 486 | B245 | B246 | B247 | B248 | B249 | B2 | B24B | B24 | B25 | B253 | B254 | B255 | B25 | B25 | B25A | B25B |
| 487 | B25D | B25E | B261 | B262 | B263 | B264 | B265 | B266 | B267 | B26A | B26C | B26E | B270 | B271 | B272 | B273 |
| 488 | B276 | B277 | B279 | B27A | B27D | B27E | B27F | B280 | B281 | B282 | B283 | B286 | B288 | B28C | B28D | B28E |
| 489 | B28F | B292 | B293 | B295 | B296 | B29B | B29D | B29E | B29F | B2A2 | B2A4 | B2A8 | B2A9 | B2AE | B2AF | B2B1 |
| 48A | B2B2 | B2B5 | B2B6 | B2B7 | B2B8 | B2B9 | B2BA | B2BB | B2BE | B2C2 | B2C3 | B2C4 | B2C5 | B2C6 | B2C7 | B2CA |
| 48 B | B2CB | B2CD | B2CE | B2D5 | B2D6 | B2D7 | B2DA | B2DC | B2DF | B2E1 | B2E7 | B2E9 | B2EA | B2F1 | B2FE | B302 |
| 48 C | В303 | B305 | B306 | B309 | B30A | Взов | B30С | B30D | ВЗоЕ | B30F | B312 | B316 | B317 | B318 | B31A | B31B |
| 48 D | B31E | B31F | B321 | B322 | B323 | B325 | B326 | B328 | B329 | B32A | В32B | B32E | B330 | B332 | В333 | B334 |
| 48 E | B335 | B336 | B337 | B339 | B33A | B33B | B33D | B33E | B33F | B340 | B341 | B342 | B343 | B344 | B345 | B346 |
| 48 F | B347 | B348 | B349 | B34A | B34B | B34C | B34E | B34F | B350 | B351 | B352 | B353 | B357 | B35A | B35D | B360 |

Table Q. 1 (continued)

| 490 | B361 | B362 | B363 | B36C | B372 | B373 | B375 | B376 | B379 | B37A | B37B | B37C | B37D | B37E | B3 | B382 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 491 | B387 | B388 | B389 | B38B | B38E | B391 | B392 | B395 | B396 | B397 | B399 | B39A | вз9в | B39E | B3A2 | ВЗАЗ |
| 492 | B3A4 | B3A5 | B3A6 | B3A7 | B3AA | ВЗАВ | B3AD | B3AE | B3AF | B3B1 | В3в2 | взвЗ | В3B4 | B3B5 | В3в6 | Взв7 |
| 493 | B3BA | ВзвС | B3BE | B3BF | B3C0 | B3C1 | B3C2 | ВЗСЗ | ВЗС9 | B3CA | B3CD | B3D1 | B3D2 | B3D6 | 8 | ВЗе2 |
| 494 | B3E3 | B3E5 | B3 | B3E7 | B3E | B3EA | взев | B3EC | B3ED | B3EE | B3EF | B3F2 | B3F6 | B3F7 | B3F8 | F9 |
| 495 | B3FA | B3FB | B3FD | B3FE | B3FF | B401 | B402 | B405 | B406 | B407 | B408 | B409 | B40A | B40B | B40 | B411 |
| 496 | B412 | B413 | B414 | B415 | B416 | B417 | B41A | B41B | B41D | B41E | B421 | B422 | B423 | B425 | B426 | B427 |
| 497 | B42A | B42E | B42 | B430 | B431 | B432 | B433 | B436 | B437 | B43 | B43A | B43D | B43E | B4 | B440 | B441 |
| 498 | B442 | B443 | B446 | B448 | B44A | B44B | B44C | B44D | B44 | B452 | B453 | B455 | B456 | B45C | B45D | B45E |
| 499 | B45F | B462 | B464 | B466 | B467 | B468 | B469 | B46B | B46E | B46F | B471 | B472 | B475 | B476 | B477 | B478 |
| 49A | B479 | B47A | B47B | 47E | B482 | B483 | 484 | B485 | B486 | B487 | B48A | B48B | B48D | B48E | B491 | B492 |
| 49B | B493 | B494 | B495 | B496 | B497 | B49A | B49E | B49F | B4A0 | B4A1 | B4A2 | B4A3 | B4A6 | B4A7 | B4A9 | B4AA |
| 49 C | B4AD | B4AE | B4AF | B4B0 | B4B1 | B4B2 | B4B3 | B4B6 | B4BA | B4BB | B4BC | B4BD | B4BE | B4BF | B4C2 | B4C3 |
| 49D | B4C5 | B4C6 | B4C7 | B4C9 | B4CA | B4CB | B4CC | B4C | B4C | B4CF | B4D | B4D4 | B4D6 | B4D7 | B4D8 | B4D9 |
| 49 E | B4DA | B4DB | B4DE | B4DF | B4E1 | B4E2 | B4E9 | B4EA | B4EB | B4EE | B4F0 | B4F2 | B4F3 | B4F4 | B4F5 | B4F6 |
| 49F | B4F7 | B4FA | B4FB | B4FD | B4FE | B501 | B502 | B503 | B504 | B505 | B506 | B507 | B50A | B50C | B50E | B50F |
| 4AD | B510 | B511 | B512 | 513 | B516 | B517 | 519 | B51A | B51 | B51E | B51F | B520 | B52 | B52 | 23 | B526 |
| 4A1 | B52C | B533 | B535 | B536 | B53C | B53D | B53E | B542 | B546 | B547 | B548 | B549 | B54A | B54F | B551 | B552 |
| 4 A 2 | B555 | B556 | B557 | B558 | B559 | B55A | B55B | B55E | B562 | B563 | B564 | B565 | B566 | B56A | B56B | B56D |
| 4A3 | B56E | B56F | B570 | B5 | B572 | B573 | B574 | B575 | B576 | B577 | B578 | B579 | B57A | B57 | B57C | B57E |
| 4A4 | B57F | B580 | B581 | B582 | B583 | B585 | B586 | B587 | B589 | B58A | B58B | B58C | B58D | B58E | B58F | B590 |
| 4A5 | B591 | B592 | B593 | B594 | B595 | B596 | B597 | B598 | B599 | B59A | B59B | B59C | B59D | B59E | B59F | B5A2 |
| 4A6 | B5A3 | B5A5 | B5A6 | B5A9 | B5AC | B5AD | B5AE | B5B2 | B5B6 | B5B7 | B5B8 | B5B9 | B5BA | B5BE | B5BF | B5C1 |
| 4A | - BC 2 | B5C5 | B5C6 | B5C | B5C8 | B5C9 | B5CA | B5CB | B5CE | B5D2 | B5D3 | B5D4 | B5D | B5D6 | B5D7 | B5DA |
| 4A8 | B5DB | B5DD | B5DE | B5E0 | B5E1 | B5E2 | B5E3 | B5E4 | B5E5 | B5E6 | B5E7 | B5EA | B5EE | B5EF | B5F0 | B5F1 |
| 4A9 | B5F2 | B5F3 | B5F5 | B5F6 | B5F7 | B5F9 | B5FA | B5FB | B5FC | B5FD | B5FE | B5FF | B600 | B601 | B602 | B603 |
| 4AA | B604 | B606 | B607 | 608 | B609 | B60A | B60 | B60C | B60D | B6 | B60 | B613 | B615 | B6 | B61 | B61C |
| 4 AB | B61D | B61E | B622 | B624 | B626 | B627 | B628 | B629 | B62A | B62B | B62E | B62F | B631 | B632 | B633 | B635 |
| 4A | B636 | B637 | B638 | B639 | B63A | B63B | B63C | B63D | B63E | B63F | B640 | B642 | B643 | B644 | B645 | B646 |
| 4AD | B64 | B64 | B64B | B64D | B64E | B651 | B652 | B653 | B654 | B655 | B656 | B657 | B65A | B65D | B65E | B65F |
| 4AE | B660 | B661 | B662 | B663 | B666 | B667 | B669 | B66A | B66D | B66E | B66 | B670 | B671 | B672 | B673 | B676 |
| 4AF | B67A | B67B | B67C | B67D | B67E | B67F | B682 | B683 | B684 | B685 | B686 | B687 | B688 | B689 | B68A | B68B |
| 4 BO | B68C | B68D | B68E | B68F | B690 | B691 | B692 | B693 | B694 | B695 | B696 | B697 | B698 | B699 | B69A | B69B |
| $4 \mathrm{B1}$ | B69E | B69F | B6A1 | B6A2 | B6A5 | B6A8 | B6A9 | B6AA | B6AE | B6B0 | B6B2 | B6B3 | B6B4 | B6B6 | B6B7 | B6B9 |
| $4 \mathrm{B2}$ | B6BA | B6BB | B6BD | B6BE | B6C0 | B6C1 | B6C2 | B6C3 | B6C4 | B6C5 | B6C6 | B6C7 | B6C8 | B6C9 | B6CA | B6CC |
| 483 | B6CD | B6CE | B6CF | B6D0 | B6D1 | B6D2 | B6D3 | B6D5 | B6D6 | B6D7 | B6D9 | B6DA | B6DD | B6DE | B6D | B6E0 |
| 4B4 | B6E1 | B6E2 | B6E3 | B6E6 | B6E7 | B6EA | B6EB | B6EC | B6ED | B6EE | B6EF | B6F1 | B6F2 | B6F3 | B6F5 | B6F6 |
| $4 \mathrm{B5}$ | B6F9 | B6FA | B6FB | B6FC | B6FD | B6FE | B6FF | B702 | B704 | B706 | B707 | B708 | B709 | B70A | B70B | B70E |
| 4B6 | B70F | B710 | B711 | B712 | B713 | B715 | B716 | B71 | B718 | B719 | B71 | B71 | B71 | B71 | B71 | B720 |
| $4 \mathrm{B7}$ | B722 | B723 | B724 | B725 | B726 | B727 | B72A | B72B | B72D | B72E | B731 | B734 | B735 | B736 | B73A | B73C |
| 4B8 : | B73E | B73F | B740 | B741 | B742 | B743 | B745 | B746 | B747 | B749 | B74A | B74B | B74D | B74E | B74F | B750 |
| 4B9 : | B751 | B752 | B753 | B756 | B757 | B758 | B75A | B75B | B75C | B75D | B75E | B75F | B762 | B763 | B765 | B766 |
| 4BA | B769 | B76A | B76B | B76C | B76D | B76E | B76F | B772 | B774 | B776 | B777 | B778 | B779 | B77A | B77E | B77F |
| 4BB | B781 | B782 | B785 | B786 | B787 | B789 | B78A | B78B | B78E | B79A | B79B | B79D | B79E | B7A1 | B7A2 | B7A3 |
| 4 BC | B7A4 | B7A5 | B7A6 | B7A7 | B7AA | B7AE | B7AF | B783 | B7B6 | B7B7 | B789 | B7BA | B7BD | B7BE | B7BF | B7C0 |
| 4BD | B7C1 | B7C2 | B7C3 | B7C6 | B7C8 | B7CA | B7CB | B7CC | B7CD | B7C | B7CF | B7D1 | B7D | B7D | B7D4 | B7D5 |
| 4BE : | B7D6 | B7D7 | B7D8 | B7D9 | B7DA | B7DB | B7DC | B7DD | B7DE | B7DF | B7E0 | B7E1 | B7E2 | B7E4 | B7E5 | B7E6 |
| 4BF : | B7E7 | B7E8 | B7E9 | B7EA | B7EB | B7EE | B7EF | B7F1 | B7F5 | B7F6 | B7F7 | B7F8 | B7F9 | B7FA | B7FB | B803 |

Table Q. 1 (concluded)

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 4C0 : B805 | B80A | B80B | B80D | B80E | B811 | B812 | B813 | B814 | B815 | B816 | B817 | B81A | B81E | B81F | B820 |
| 4C1 : B826 | B827 | B829 | B82A | B82D | B82E | B82F | B831 | B832 | B833 | B836 | B83A | B83B | B83D | B83F | B842 |
| 4C2 : B843 | B845 | B846 | B849 | B84A | B84B | B84C | B84D | B84E | B84F | B852 | B854 | B856 | B857 | B858 | B859 |
| 4C3 : B85A | B85B | B85E | B85F | B861 | B862 | B865 | B866 | B867 | B869 | B86A | B86E | B870 | B872 | B873 | B875 |
| 4C4 : B87A | B87B | B87D | B87E | B87F | B881 | B882 | B883 | B884 | B885 | B886 | B887 | B88A | B88E | B88F | B890 |
| 4C5 : B891 | B892 | B893 | B895 | B896 | B897 | B899 | B89A | B89D | B89E | B89F | B8A0 | B8A1 | B8A2 | B8A3 | B8A4 |
| 4C6 : B8A5 | B8A6 | B8A9 | B8AA | B8AB | B8AC | B8AD | B8AE | B8AF | B8B2 | B8B3 | B8B6 | B8B9 | B8BA | B8BB | B8BC |
| 4C7 : B8BD | B8BE | B8BF | B8C2 | B8C6 | B8C7 | B8C8 | B8C9 | B8CA | B8CB | B8CE | B8CF | B8D1 | B8D2 | B8D5 | B8D6 |
| 4C8 : B8D7 | B8D8 | B8D9 | B8DA | B8DB | B8DE | B8E0 | B8E2 | B8E3 | B8E4 | B8E5 | B8E6 | B8E7 | B8EA | B8EB | B8ED |
| 4C9 : B8EE | B8F1 | B8F2 | B8F4 | B8F5 | B8F6 | B8F7 | B8FA | B8FC | B8FE | B8FF | B901 | B903 | B906 | B907 | B909 |
| 4CA : B90A | B90D | B90E | B90F | B910 | B911 | B912 | B913 | B916 | B91A | B91B | B91C | B91D | B91E | B91F | B922 |
| 4CB : B923 | B925 | B926 | B927 | B929 | B92A | B92B | B92C | B92D | B92E | B92F | B932 | B936 | B937 | B938 | B939 |
| 4CC : B93A | B93B | B93E | B93F | B941 | B942 | B945 | B946 | B947 | B948 | B949 | B94A | B94B | B94E | B952 | B953 |
| 4CD : B954 | B955 | B956 | B957 | B95A | B95B | B95D | B95E | B961 | B962 | B963 | B964 | B965 | B966 | B967 | B96A |
| 4CE : B96C | B96E | B96F | B970 | B971 | B972 | B973 | B976 | B977 | B979 | B97A | B97E | B97F | B981 | B982 | B983 |
| 4CF : B986 | B988 | B98C | B992 | B993 | B995 | B996 | B997 | B999 | B99A | B99B | B99C | B99D | B99F | B9A2 | B9A4 |
| 4D0 : B9A6 | B9A7 | B9A8 | B9A9 | B9AA | B9AB | B9AE | B9AF | B9B1 | B9B2 | B9B5 | B9B6 | B9B7 | B9B8 | B9B9 | B9BA |
| 4D1 : B9BB | B9C2 | B9C3 | B9C5 | B9C7 | В9СВ | B9CD | B9D3 | B9D5 | B9D6 | B9D7 | B9DA | B9E6 | B9E7 | B9E9 | B9EA |
| 4D2 : B9EE | B9EF | B9F0 | B9F1 | B9F2 | B9F3 | B9F6 | B9FC | B9FF | BA02 | BA03 | BA05 | BA06 | BA07 | BA09 | BAOA |
| 4D3 : BAOB | BAOC | BAOD | BAOE | BAOF | BA12 | BA14 | BA16 | BA17 | BA19 | BA1A | BA1B | BA1D | BA1E | BA1F | BA20 |
| 4D4 : BA21 | BA22 | BA23 | BA24 | BA25 | BA26 | BA27 | BA28 | BA29 | BA2A | BA2B | BA2C | BA2D | BA2E | BA2F | ВАЗ0 |
| 4D5 : BA31 | BA32 | BA33 | BA34 | BA35 | BA36 | BA37 | BA3A | BA3D | BA3E | BA43 | BA44 | BA45 | BA46 | BA47 | BA4A |
| 4D6 : BA50 | BA51 | BA52 | BA56 | BA57 | BA59 | BA5A | BA5D | BA5E | BA5F | BA61 | BA62 | BA63 | BA66 | BA6C | BA6E |
| 4D7 : BA6F | BA72 | BA73 | BA75 | BA76 | BA79 | BA7B | BA7C | BA7D | BA7E | BA7F | BA82 | BA86 | BA88 | BA8A | BA8B |
| 4D8 : BA8E | BA8F | BA91 | BA92 | BA95 | BA96 | BA97 | BA98 | BA99 | BA9A | BA9B | BA9E | BAA0 | BAA2 | BAA4 | BAA7 |
| 4D9 : BAAA | BAAD | BAAE | BAB3 | BAB5 | BAB6 | BAB7 | BABA | BABC | BABE | BACO | BAC1 | BAC2 | BAC6 | BAC7 | BAC9 |
| 4DA : BACA | BACD | BACE | BACF | BADO | BAD1 | BAD2 | BAD3 | BAD6 | BADA | BADB | BADC | BADD | BADE | BADF | BAE1 |
| 4DB : BAE2 | BAE3 | BAE5 | BAE6 | BAE7 | BAE9 | BAEA | BAEB | BAEC | BAED | BAEE | BAEF | BAFO | BAF2 | BAF3 | BAF5 |
| 4DC : BAF6 | BAF7 | BAF8 | BAF9 | BAFA | BAFB | BAFE | BAFF | BB01 | BB02 | BB05 | BB06 | BB07 | BB08 | BB09 | BB0A |
| 4DD : BBOB | BB0E | BB10 | BB12 | BB13 | BB14 | BB15 | BB16 | BB17 | BB1A | BB1B | BB1D | BB1E | BB21 | BB22 | BB23 |
| 4DE : BB24 | BB25 | BB26 | BB27 | BB2A | BB2C | BB2E | BB2F | BB30 | BB31 | BB32 | BB33 | BB37 | BB39 | BB3F | BB41 |
| 4DF : BB42 | BB43 | BB46 | BB48 | BB4A | BB4C | BB4E | BB52 | BB53 | BB55 | BB56 | BB59 | BB5A | BB5B | BB5C | BB5D |

## Procedure for the unification and arrangement of CJK Ideographs

The graphic character collections of CJK unified ideographs in ISO/IEC 10646-1 are specified in clause 27. They contain almost 27,500 ideographs, and are derived from over 66,000 ideographs which are found in various different national and regional standards for coded character sets (the "source codes").

This Annex describes how the ideographs in this standard are derived from the source codes by applying a set of unification procedures. It also describes how the ideographs in this standard are arranged in the sequence of consecutive code positions to which they are assigned.
The source code standards are shown in clause 27 in five groups according to their origins. The groups are identified as the G-, T-, J-, K- and V-sources.
For the purposes of ISO/IEC 10646-1 a unification process is applied to the ideographic characters taken from the codes in the source groups. In this process single ideographs from two or more of the source groups are associated together, and a single code position is assigned to them in this standard. The associations are made according to a set of procedures that are described below. Ideographs that are thus associated are described here as "unified".

[^0]
## R.1. Unification procedure

## R.1.1 Scope of unification

Ideographs that are unrelated in historical derivation (non-cognate characters) have not been unified.


Example:
NOTE - The difference of shape between the two ideographs in the above example is in the length of the lower horizontal line. This is considered an actual difference of shape. Furthermore these ideographs have different meanings. The meaning of the first is "Soldier" and of the second is "Soil or Earth".

An association between ideographs from different sources is made here if their shapes are sufficiently similar, according to the following system of classification.

## R.1.2 Two level classification

A two-level system of classification is used to differentiate (a) between abstract shapes and (b) between actual shapes determined by particular typefaces. Variant forms of an ideograph, which can not be unified, are identified based on the difference between their abstract shapes.

## R.1.3 Procedure

A unification procedure is used to determine whether two ideographs have the same abstract shape or different ones. The unification procedure has two stages, applied in the following order:
a) Analysis of component structure;
b) Analysis of component features;

## R.1.3.1 Analysis of component structure

In the first stage of the procedure the component structure of each ideograph is examined. A component of an ideograph is a geometrical combination of primitive elements. Alternative ideographs can be configured from the same set of components. Components can be combined to create a new component with a more complicated structure. An ideograph, therefore, can be defined as a component tree, where the top node is the ideograph itself, and the bottom nodes are the primitive elements. This is shown in Figure R.1.


Figure R. 1 - Component structure

## R.1.3.2 Analysis of component features

In the second stage of the procedure, the components located at corresponding nodes of two
ideographs are compared，starting from the most superior node，as shown in Figure R．2．


Figure R． 2 －The most superior node of a component
The following features of each ideograph to be compared are examined：
a ：the number of components，
b ：the relative position of the components in each complete ideograph，
c ：the structure of corresponding components．
If one or more of the features（a to c above）are different between the ideographs in the comparison， the ideographs are considered to have different abstract shapes and are therefore not unified．
If all of the features（a to c above）are the same between the ideographs，the ideographs are considered to have the same abstract shape and are therefore unified．

## R．1．4 Examples of differences of abstract shapes

To illustrate rules derived from a：to c ：in R．1．3．2， some typical examples of ideographs that are not unified，owing to differences of abstract shapes，are shown below．

## R．1．4．1 Different number of components

The examples below illustrate rule a：since the two ideographs in each pair have different numbers of components．
崖•厓，肱•厷，降•条

## R．1．4．2 Different relative positions of components

The examples below illustrate rule b：．Although the two ideographs in each pair have the same number of components，the relative positions of the components are different．
峰•峯，荊•荆

## R．1．4．3 Different structure of a corresponding component

The examples below illustrate rule c：．The structure of one（or more）corresponding components within the two ideographs in each pair is different．


## R．1．5 Differences of actual shapes

To illustrate the classification described in R．1．2， some typical examples of ideographs that are unified are shown below．The two or three ideographs in each group below have different actual shapes，but they are considered to have the same abstract shape，and are therefore unified．


The differences are further classified according to the following examples．
a）Differences in rotated strokes／dots

$$
\begin{aligned}
& \text { 少•半, 勺•勺, 犲•村•羽, 莤•酋, } \\
& \text { 兼•兼, 盆•益 }
\end{aligned}
$$

b）Differences in overshoot at the stroke initiation and／or termination

c）Differences in contact of strokes


奔•奔
d）Differences in protrusion at the folded corner of strokes
巨•巨
e）Differences in bent strokes

## 西•西

f）Differences in folding back at the stroke termination

## 朱•朱

g）Differences in accent at the stroke initiation

h）Differences in＂rooftop＂modification
式•儿，穴•穴
j）Combinations of the above differences
外加加
These differences in actual shapes of a unified ideograph are presented in the corresponding source columns for each code position entry in the code table in clause 27 of this International Standard．

## R．1．6 Source separation rule

To preserve data integrity through multiple stages of code conversion（commonly known as＂round－trip integrity＂），any ideographs that are separately encoded in any one of the source standards listed below have not been unified．

```
G-source: GB2312-80, GB12345-90,
        GB7589-87*, GB7590-87*,
        GB8565-88*,
            General Purpose Hanzi List for
                Modern Chinese Language*
```

$\begin{array}{ll}\text { T－source：} & \text { TCA－CNS 11643－1986／1st plane，} \\ & \text { TCA－CNS 11643－1986／2nd plane，} \\ & \text { TCA－CNS 11643－1986／14th plane } \\ \text { J－source：} & \text { JIS X 0208－1990，JIS X 0212－1990 } \\ \text { K－source：} & \text { KS C 5601－1989，KS C 5657－1991 }\end{array}$
（A＂＊＂after the reference number of a standard indicates that some of the ideographs included in that standard are not introduced into the unified collection．）

However，some ideographs encoded in two standards belonging to the same source group（e．g． GB2312－80 and GB12345－90 ）have been unified during the process of collecting ideographs from the source group．

## R．2．Arrangement procedure

## R．2．1 Scope of arrangement

The arrangement of the CJK UNIFIED IDEOGRAPHS in the code table of clause 27 of this International Standard is based on the filing order of ideographs in the following dictionaries．

$$
\begin{array}{lll}
\text { Priority } & \text { Dictionary } & \text { Edition } \\
\hline
\end{array}
$$

1 Kangxi Dictionary 康熙字典
2 Daikanwa Jiten 大漢和辞典
3 Hanyu Dazidian 汉语大字典
4 Daejaweon 大字源
Beijing 7th edition 9th edition
1st edition 1st edition

The dictionaries are used according to the priority order given in the table above．Priority 1 is highest．If an ideograph is found in one dictionary，the dictionaries of lower priority are not examined．

## R．2．2 Procedure

## R．2．2．1 Ideographs found in the dictionaries

a）If an ideograph is found in the Kangxi Dictionary， it is positioned in the code table in accordance with the Kangxi Dictionary order．
b）If an ideograph is not found in the Kangxi Dictionary but is found in the Daikanwa Jiten，it is given a position at the end of the radical－stroke group under which is indexed the nearest preceding Daikanwa Jiten character that also appears in the Kangxi dictionary．
c）If an ideograph is found in neither the Kangxi nor the Daikanwa，the Hanyu Dazidian and the Daejaweon dictionaries are referred to with a similar procedure．

## R．2．2．2 Ideographs not found in the dictionaries

If an ideograph is not found in any of the four dictionaries，it is given a position at the end of the
radical－stroke group（after the characters that are present in the dictionaries）and it is indexed under the same radical－stroke count．

## R．3．Source code separation examples

The pairs（or triplets）of ideographs shown below are exceptions to the unification rules described in clause R． 1 of this Annex．They are not unified because of the source code separation rule described in clause R．1．6．

## NOTES

1．The particular source code group（or groups）that causes the source code separation rule to apply is indicated by the letter（G，J，K，or T）that appears to the right of each pair（or triplet）of ideographs．The source code groups that correspond to these letters are identified at the beginning of this Annex．

2．The ideograph pairs are listed below in ascending order by the code position of the first ideograph of each pair．The sequence progresses downwards in the left column as far as each marker（ $v^{------}{ }^{v}$ ），and then continues downwards in the adjoining right column，starting at the previous marker．

|  | T | 䛧 | T |
| :---: | :---: | :---: | :---: |
| 5024 503C |  | 5239 524E |  |
|  | T | 并》并頁 | T |
| 50775078 |  | 524F 5259 |  |
|  | TJ |  | T |
| 507D 50DE |  | 525D 5265 |  |
|  | T | 合别 珨》 | J |
| 514C 5151 |  | 52925294 |  |
|  | TJ | $\Rightarrow N$ | T |
| 514E 5154 |  | 52FB 5300 |  |
| V |  | －．－V |  |
|  | T | 艮龙 追先 | T |
| 53555358 |  | 5848 588D |  |
| 明白 | TK | 渞 首 | TJ |
| 5373 537D |  | 5861 586B |  |
|  | TJ | 堛皕 | T |
| 5377 5DFB |  | 5897 589E |  |
|  | GT | $\text { ソ上 } \ddagger$ | GTJ |
| 53C1 53C2 |  | 58EE 58EF |  |
|  | T | $\stackrel{\text { 学豆 }}{\text { 畐 }}$ | T |
| 53C3 53C4 |  | 58FD 5900 |  |
|  | T |  | T |
| 54155442 |  | 5910 657B |  |
|  | T |  | GTJ |
| 541E 5451 |  | 5932 672C |  |
|  | TJ |  | J |
| 543354345449 |  | 59655967 |  |
|  | T |  | TJ |
| 54365450 |  | 5968 596C 734E |  |
| $\frac{H}{\square} \frac{\mu}{\square}$ | T | 必 出 | GT |
| $543 F 544 \mathrm{~A}$ |  | 5986 599D |  |


| 艮啊 | T | 如奸 | T | 由才直等 | J | 奥率类委 | T |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5527 559E |  | 598D 59F8 |  | 5C02 5C08 |  | 5F5D 5F5E |  |
| 喻 喻 | T | 女陑 彻折 | T | 垱师夺 | GTJ | 齐 妾 | T |
| 55A9 55BB |  | 59CD 59D7 |  | 5C06 5C07 |  | 5F65 5F66 |  |
| 嘘嘘 | T | 姖 如宫 | GT | 多负 | T | 德德 | T |
| 56185653 |  | 59EB 59EC |  | 5C13 5C14 |  | 5FB3 5FB7 |  |
| 噎哄 | GTJ | 如飔 始写 吴 | T | 回 句 | T | 徵徵 | T |
| 568F 5694 |  | 5A1B 5A2F 5A31 |  | 5 C 19 5C1A |  | 5FB4 5FB5 |  |
| 玉 玉 | T | 婕姃 | T | 逃 | T |  | TJ |
| 56EF 56FD |  | 5A55 5AAB |  | 5C2A 5C2B |  | 6075 60EO |  |
| 卷 巻 | TJ | 媮 婨 | T | 監 監 | T | 㤋悦 | T |
| 5708 570F |  | 5A7E 5AAE |  | 5C36 5C37 |  | 6085 60A6 |  |
|  | T | 媪 媼 | TK | 屏 尾 | T | 怾悮 | T |
| 570E 5713 |  | 5AAA 5ABC |  | 5C4F 5C5B |  | 609E 60AE |  |
| 呂 畚 畐 | T | 女為 滑 | T | 山争山学 | GT | 首 真 | T |
| 57165717 |  | 5AAF 5B00 |  | 5CE5 5D22 |  | 60B3 60EA |  |
|  | T | 瓝生 㛯 | T | 㯋真頁 | T | 愠慍 | T |
| 5759 5DE0 |  | 5B0E 5B14 |  | 5DD3 5DD4 |  | 6120 614D |  |
| 哣埒 | J | 脄麼 嫲 | GT | 帡低 | T | 恰 慎 | TJ |
| 57D2 57D3 |  | 5B24 5B37 |  | 5E21 5E32 |  | 613C 614E |  |
|  |  | －－－v |  | 茪 费 | TJ | 厽或亚甙 | GT |
| 变玄 䒺苔 | T | 弹弾 | T | 5E2F 5E36 | TJ | $6229 \text { 622C }$ | GT |
| 5B73 5B76 |  | 5F39 5F3E |  | 并手午 | T | 虚或虚 | T |
| $\begin{aligned} & \text { 官官 } \\ & \text { 吕 } \end{aligned}$ | T | $\square 15$ | TJ | 5E76 5E77 |  | 622F 6231 |  |
| 5BAB5BAE |  | 5F50 5F51 |  | 廊塬 | T | $\text { 曰 } 曰 戸$ | T |
|  | T | 旲灵 | T | 5EC4 5ECF |  | 623662376238 |  |
| 5BDB 5BEC |  | 5F54 5F55 |  | $\begin{gathered} \text { 椟 找式 } \end{gathered}$ | T | $\vec{A} \vec{A}$ | T |
|  | T | 彙彙 | T | 5F11 5F12 |  | 623B 623E |  |
| 5BDC 5BE7 |  | 5F59 5F5A |  | 強 強 | T | 扐打抂 | T |
| 寝箒 | GTJ | 类盆㥹 | J | 5F37 5F3A |  | 629B 62CB |  |
| 5BDD 5BE2 |  | 5F5B 5F5C |  |  |  | －－－－ |  |


| 抜拢 | TJ | 㭘㭘 | T | 百 向 | J | 沼没 | TJ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 629C 62D4 |  | 69616986 |  | 66FD 66FE |  | 6C92 6CA1 |  |
| 挩拀 | T | 概 柏的 | T | 枌枵 | T | 争 浸 | TJ |
| 6329 635D |  | 6982 69EA |  | 67B4 67FA |  | 6D44 6DE8 |  |
| 挿插插 | TJ | 榅 榲 | T | 自 自 | T | 沾少 沾少 | T |
| 633F 63D2 63F7 |  | 6985 69B2 |  | 67E5 67FB |  | 6D89 6E09 |  |
| 捏拍 | TJ | 敉榝 | T | 相 肼 | T | 涗労 | T |
| 634F 63D1 |  | 699D 6A27 |  | 67F5 6805 |  | 6D97 6D9A |  |
| 捜掜 | TJ | 槇槙 | J | 柗棁 | T | 戻 戾 | T |
| 635C 641C |  | $69 \mathrm{C7}$ 69D9 |  | 68B2 68C1 |  | 6D99 6DDA |  |
| 掲揭 | T | 嵄樣 | TJ |  |  | －－－－－V |  |
| 63B2 63ED |  | 69D8 6A23 |  | 堟绿 | T | 品䘏 | TJK |
| 授缷缷 | TJ | 横 橫 | T | 6DE5 6E0C |  | 773E 8846 |  |
| 63FA 64166447 |  | 6A2A 6A6B |  | 凊凊 | T | 布开矿 | T |
| 担搵 | T | 少年 | T | 6DF8 6E05 |  | $7814784 F$ |  |
| 63FE 6435 |  | $6 \mathrm{B65}$ 6B69 |  | 渴潟 | T | 祿禄 | TJ |
| 撃撃 | TJ | 䇝 域 | T | 6E07 6E34 |  | $\begin{aligned} & \text { 797F } 7984 \\ & \mp=\mp \end{aligned}$ |  |
| 6483 64CA |  | 6B72 6B73 |  |  | T | 㒫 杏 | T |
| 雄教 | T | 殒殁 | T | 6E29 6EAB |  | 79BF 79C3 |  |
| 654E 6559 |  | 6B7F 6B81 |  | 妫 䟦 | T | 稅税 | T |
| 分人 兄交 | T | 壳几壳反 | GTJ | 6E88 6F59 |  | 7A05 7A0E |  |
| 6553 655A |  | 6BBB 6BBC |  | 飡泊㑆 | T | 䅦穂 | TJ |
| 睌㿟 | T | 回䛼的 | T | 6E89 6F11 |  | 7A42 7A57 |  |
| 65E2 65E3 |  | 6BC0 6BC1 |  | 淯浚 | T | 唃 踇 | GJ |
| 名召 | T | 䇛 | T | 6EDA 6EFE |  | 7B5D 7B8F |  |
| 6602 663B |  | 6BCE 6BCF |  | 潛洗先 | GTJK | 筸庠 | T |
| 日免日免 | T | $\begin{gathered} 气 气 气 ⿱ ⿴ 囗 十 心 刂 \end{gathered}$ | T | 6F5B 6FF3 |  | 7BB3 7C08 |  |
| 665A 6669 |  | 6C32 6C33 |  | 瀨瀬 | T |  | T |
| $\begin{aligned} & \text { 暨追先 } \\ & \text { E } \end{aligned}$ | T | 汚 污 | T | $7028 \text { 702C }$ |  | 7BE1 7C12 |  |
| 66A8 66C1 |  | 6C5A 6C61 |  |  | GTJ |  | T |
|  |  |  |  | 70BA 7232 |  | 7CA4 7CB5 |  |


| 炎塋 | GTJK | 絕絶 | T |  | TJ | 輼轀 | T |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 712 D 7162 |  | $7 \mathrm{~F} 557 \mathrm{D76}$ |  | 83 D 18458 |  | 8F3C 8F40 |  |
| 医点巸 | J | 綠緑 | T | $\begin{gathered} \text { 点苗盢 } \end{gathered}$ | T | 达达 | T |
| 71557199 |  | 7DA0 7DD1 |  | 84808495 |  | 8FBE 8FD6 |  |
| 煴熅 | T | 緒緒 | T | 将將 | GJ | 迸逆 | TJ |
| 71747185 |  | 7DD2 7DD6 |  | 848B 8523 |  | 8FF8 902C |  |
| 状 呪 | GT | 緣縁 | T | 為蕿 | T | 遙遥 | J |
| $72 \mathrm{B6} 72 \mathrm{CO}$ |  | 7DE3 7E01 |  | 848D 853F |  | 90599065 |  |
| 瑤瑶 | TJ | 緼縕 | T | 蕰薀 | T | 形形 | T |
| 74647476 |  | 7DFC 7E15 |  | 85708580 |  | 90A2 90C9 |  |
| 瓶形瓦 | T | 繈繦 | T | 董薫 | T | 郎良家 | T |
| 74F6 7501 |  | $7 E 48$ 7E66 |  | 85AB 85B0 |  | 90CE 90DE |  |
| 汻産 | T | 美美美 | TJ | 緼縕 | T | 榔鄉貌 | T |
| 75227523 |  | 7FAE 7FB9 |  | 85F4 860A |  | 90F7 91099115 |  |
| 痩瘦 | J | 科的 | T | 虚虛 | T | 酯西䁅 | T |
| 75E9 7626 |  | 7FF6 7FFA |  | 865A 865B |  | 9196 919E |  |
| 腺㿞 | T | 胼䏕 | T | 蚡蜕 | T | 酱醬 | J |
| 76A1 76A5 |  | 80FC 8141 |  | 86FB 8715 |  | 91 A 91 AC |  |
| 旨真 | TJ | 脫脱 | T | 徫偉 | TJK | 鈃鈃† | T |
| 771E 771F |  | 812B 8131 |  | 885B 885E |  | 92039292 |  |
|  |  | －－－－v |  | 岩哀 | TK | 銳銧 | T |
| 腽膃 | T | 謠謡 | J | 886E 889E |  |  |  |
| 817D 8183 |  | $8 \mathrm{B20} 8 \mathrm{~B} 21$ |  | 装裝 | GJK | 錄録 | T |
| 刍闖 | GT | 豜刑开 | T | $88 \mathrm{C5}$ 88DD |  | 93049332 |  |
| 82038204 |  | 8C5C 8C63 |  | 訮訮 | T | 錬鍊 | TK |
| 声 㑒 | TJ | 走圭 | TJ | 8A2E 8A7D |  | 932 C 934 A |  |
| 820D 820E |  | 8D70 8D71 |  | 說説可元吹 | T | 鎮鎮 | TJ |
| 舖舗 | J | 軿䡉 | T | 8AAA 8AAC |  | 93AD93AE |  |
| 82168217 |  | 8EFF 8F27 |  | 諌諫 | TJ | 閱閲 | T |
| 荘莊 | TJ | 輜輜 | J | 8ACC 8AEB |  | 95 B 195 B 2 |  |
| 8358 838A |  | 8F1C 8F3A |  |  |  |  |  |


| 䧑澛 | G | 吕㐭 | T |
| :---: | :---: | :---: | :---: |
| 96679689 |  | 9AD8 9AD9 |  |
| 声 夆 | T | 镸镸 | TJ |
| 97519752 |  | 9AEA 9AEE |  |
| 青争 青単 | GTJ |  | T |
| 9759 975C |  | 9B2C 9B2D |  |
| 虭 勒 | J | 鰛 鰮 | TJ |
| 976D 9771 |  | 9C1B 9C2E |  |
| 禾百秃百 | T | 鳥鳥鳥 | T |
| 9839 983D |  | 9CEF 9CF3 |  |
| 彣貣顔 | TJ | 東鷍 東鷍 | J |
| 984F 9854 |  | 9D87 9DAB |  |
| 顚真百 | J | 鷆 真鳥 | J |
| 985A 985B |  | 9DC6 9DCF |  |
| 会人 創 | J | 䴲正 䴮正 | T |
| 98EE 98F2 |  | 9EAA 9EAB |  |
| 命前 創 | TJ | 麻麻 | T |
| 99059920 |  | 9EBC9EBD |  |
| 馬大 馬太 | TJK | 某 異 | T |
| 99B1 99C4 |  | 9EC3 9EC4 |  |
| 馬并 䮁 | TK | 里 呆 | T |
| 99E2 9A08 |  | 9ED1 9ED2 |  |
| 骨九骨丸 | T |  |  |

## 9AA9 9AAB

In accordance with the unification procedures described in R． 1 of this Annex the pairs（or triplets） of ideographs shown below are not unified．The reason for non－unification is indicated by the reference which appears to the right of each pair（or triplet）．For＂non－cognate＂see R．1．1

NOTE－The reason for non－unification in these examples is different from the source code separation rule described in clause R．1．6
昜 䏛
5191 80C4
non ogonae 朐朐
non cognate
6710 80CA

冲沖 51B2 6C96旭 分 51B3 6C7A况況
sisg 6 信
垛蒅 579B 579C菅辛 豆辛


双双双 6560 656A


670C 80A6
朏朏
670F 80D0

| R．1．4．3 | 此此兆 | non cognate |
| :---: | :---: | :---: |
|  | 67138101 |  |
| R．1．4．3 | 日攽胶 | non cognate |
|  | 67188127 |  |
| R．1．4．3 | 朣时童 | non cognate |
|  | 6723 81A7 |  |
| R．1．4．3 | 农尕 | R．1．4．3 |
|  | 67356736 |  |
| R．1．4．2 | 湱名 灃庎 | R．1．4．3 |
|  | 70547067 |  |
| R．1．4．3 | 和活秘 | R．1．4．3 |
|  | 7 A 32 7A3B |  |
| R．1．4．1 | 科籿的 | R．1．4．3 |
|  | 7FF1 7FF6 |  |
| R．1．4．1 | 者者者者 | R．1．4．3 |
|  | 800780088009 |  |
| R．1．4．3 | 耳皿 盂益 㩊 | R．1．4．1 |
|  | 8074 807C 807D |  |
| non cognate |  | R．1．4．2 |
| non cognate 躳禾射 |  |  |
|  |  | R．1．4．3 |
|  | 8EB1 8EB2 |  |


[^0]:    NOTE - The unification process does not apply to the following collections of ideographic characters in the Basic multilingual Plane:

    - CJK RADICALS SUPPLEMENT (2E80-2EFF)
    - KANGXI RADICALS (2F00-2FDF)
    - CJK COMPATIBILITY IDEOGRAPHS (F900 - FAFF with the exception of FA1F and FA23).

