

JTC1/SC2/WG2 **N1333**

**ISO/IEC JTC1/SC2/WG2 & WG2/IRG
Universal Multi-Octet Coded Character Set
(UCS)**

JTC1/SC2/WG2/IRG **N318**

Date:1996-03-20

TITLE:	Revised Draft Text of Annex on CJK Ideographs
Source:	Tateo Koike (IRG Editor of Annex)
Status:	IRG Contribution
Action:	To accept this draft text by SC2/WG2
Distribution:	ISO/IEC JTC1/SC2/WG2 and WG2/IRG

Attached text is a revised version of Annex on CJK Ideographs that considered the pDAM8 voting comments by IRG members.

Annex (informative)

Procedure^{for the} of unification and arrangement of the CJK Ideographs

The graphic character collection CJK UNIFIED IDEOGRAPHS in this Edition of ISO/IEC 10646-1 contains 20,902 ideographs (see clause 26). They are derived from over 54,000 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 procedure. 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 below in four groups according to their origins. The groups are identified as the G-, T-, J-, and K-sources.

- G-source: GB2312-80, GB12345-90,
GB7589-87*, GB7590-87*,
GB8565-88*,
General Purpose Hanzi List for
Modern Chinese Language*
- T-source: TCA-CNS 11643-1986/1st plane,
TCA-CNS 11643-1986/2nd plane,
TCA-CNS 11643-1986/14th plane*
- J-source: JIS X 0208-1990, JIS X 0212-1990
- K-source: KS C 5601-1989, KS C 5657-1991

(Standards indicated with an "*" do not have all of their ideographic contents included in the source)

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.

1. Source separation rule

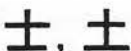
To preserve round-trip integrity, ideographs separately encoded in any of these sources will not be unified.

This rule does not prevent unification of ideographs across sources (using the procedure below). This rule does not prevent unification of the same ideograph within two standards used in a single source (e.g., GB2312-80 and GB 12345-90). Furthermore, this rule does not prevent unification of separately encoded ideographs within other sources.

2. Unification procedure

2.1 Scope of unification

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

Example: 

The shape difference of the sample above ideographs is the stroke length of second horizontal line. This is considered an actual shape difference. But these ideographs have different meaning. The meaning of former is "Soldier" and latter is "Earth".

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

2.2 Two level classification

The two-level classification differentiates between (a) abstract shapes and (b) actual shapes determined by particular typeface. Variant forms are identified based on the difference of abstract shapes.

2.3 Procedure

Apply the unification procedure in this order;

- Analysis of component structures;
- Analysis of component features;

2.3.1 Analysis of component structure

The component structure of each ideograph is examined. The component is a geometrical combination of primitive elements. Alternative ideograph can be configured from components and other components. Some components can be combined to be a component with more complicated structure. As ideograph, therefore, can be defined a component tree, where top node is ideograph itself, and the bottom nodes are the primitive elements. This is shown in Figure 1.

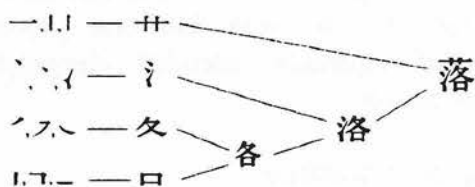


Figure 1 - Component structure

2.3.2 Analysis of component features

In this procedure, such components located at the most superior node are employed as shown in Figure 2.

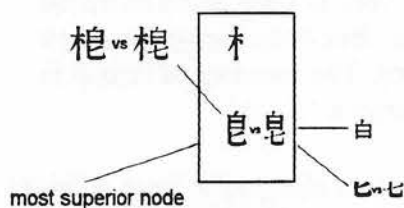


Figure 2 - The most superior node of component

- The following features of each ideograph to be compared are examined:

- number of components
- relative position of components in each complete ideograph
- structure of corresponding component

- If one or more features (a to c above) are different between the ideographs undergoing comparison, the ideographs are considered to have different abstract shapes and are therefore not unified.

- If all the features (a to c above) are not different between the ideographs, the ideographs are considered to have the same abstract shape and are therefore unified.

2.4 Examples of differences of abstract shapes (not unified)

To illustrate the above rules (a to c in 2.3.2), some typical examples of differences of abstract shapes are shown:

2.4.1 Examples of different number of components (a)

崖·厓, 肱·肱, 降·牟

2.4.2 Examples with the same number of components, but with different relative positions of components (b)

峰·峯, 荊·荊

2.4.3 Examples with different structures of corresponding components (c)

扌·擴, 策·箭, 艹·然, 圣·亓,
 𠂇·僉, 区·區, 夾·夾, 單·單,
 隹·隹, 𠂇·𠂇, 贊·贊, 襄·襄,
 隹·隹, 間·間, 朶·朶, 雋·雋,
 恒·恆, 臭·臭, 人·人, 𠂇·𠂇,
 𠂇·𠂇

2.5 Examples of differences of actual shapes (unified)

The examples shown here are considered to have the same abstract shape, but to have different actual shapes.

讠·讠·讠, 示·示·示, 艮·艮·艮, 食·食·食,
 黄·黄, 盥·盥, 曷·曷, 包·包,
 青·青, 每·每, 册·册, 爭·爭,
 畺·畺, 录·录, 步·步, 者·者,
 臭·臭, 并·并, 骨·骨, 呂·呂,
 直·直, 鼎·鼎, 吳·吳, 眞·眞·眞,
 爲·爲, 單·單, 曾·曾, 成·成,
 專·專, 內·內, 晉·晉, 龜·龜,
 ++, ++

The differences are further classified according to the following examples.

a) Differences in rotated strokes/dots

半·半, 勺·勺, 羽·羽·羽, 酋·酋,
 兼·兼, 益·益

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

身·身, 雪·雪, 拐·拐, 不·不,
 非·非, 周·周, 告·告

c) Differences in contact of strokes

奧·奧, 酉·酉, 兕·兕, 查·查,
 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

父·父, 丈·丈, 乚·乚

h) Differences in "rooftop" modification

八·八, 宀·宀

j) Combinations of the above differences

羽·羽·羽

The differences in actual shapes are illustrated in clause 26 in the four columns representation.

3. Arrangement procedure

3.1 Scope of arrangement

The arrangement of the CJK UNIFIED IDEOGRAPHS is based on the filing order in these dictionaries.

Priority	Dictionary	Edition
1	Kangxi Dictionary 康熙字典	Beijing, 7th edition
2	Daikanwa Jiten 大漢和辭典	9th edition
3	Hanyu Dazidian 汉语大字典	1st edition
4	Daejaweon 大字源	1st edition

Use the dictionaries according to the priority order given in the table above. Priority 1 is highest. If an ideograph is found in one dictionary, do not examine dictionaries of lower priority.

3.2 Procedure

3.2.1 Ideographs found in the dictionaries

a) If an ideograph is found in the Kangxi Dictionary, it follows the Kangxi Dictionary order.

b) If an ideograph is not found in the Kangxi Dictionary but 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 Daejwaeon dictionaries are referred to with a similar procedure.

3.2.2 Ideographs not found in any of the four 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.

4. Source code separation examples

The following list shows exceptions to the unification rules caused by the source code separation.

丟丟 4E1F 4E22	T	兗兗 5156 5157	T
么么 4E48 5E7A	GT	冊冊 518A 518C	TJ
爭爭 4E89 722D	GTJ	淨淨 51C0 51C8	G
仞仞 4EDE 4EED	J	凡凡 51E2 51E3	T
併併 4F75 5002	T	刃刃 5203 5204	TJ
侶侶 4FA3 4FB6	T	刊刊 520A 520B	TJ

俣俣 4FC1 4FE3	TJK	刪刪 5220 522A	T
俞俞 4FDE 516A	T	別別 5225 522B	T
俱俱 4FF1 5036	T	券券 5238 52B5	TJ
值值 5024 503C	T	剎剎 5239 524E	T
偷偷 5077 5078	T	勗勗 524F 5259	T
偽偽 507D 50DE	TJ	剝剝 525D 5265	T
兌兌 514C 5151	T	劒劒 5292 5294	J
兔兔 514E 5154	TJ	勻勻 52FB 5300	T
單單 5355 5358	T	暨暨 5848 588D	T
即即 5373 537D	TK	填填 5861 586B	TJ
卷卷 5377 5DFB	TJ	增增 5897 589E	T
叁叁 53C1 53C2	GT	壯壯 58EE 58EF	GTJ
參參 53C3 53C4	T	壽壽 58FD 5900	T
呂呂 5415 5442	T	簀簀 5910 657B	T
吞吞 541E 5451	T	本本 5932 672C	GTJ
吳吳吳 5433 5434 5449	TJ	奧奧 5965 5967	J
訥訥 5436 5450	T	獎獎獎 5968 596C 734E	TJ
告告 543F 544A	T	妝妝 5986 599D	GT
唧唧 5527 559E	T	妍妍 598D 59F8	T

喻喻
55A9 55BB
嘘嘘
5618 5653
噓噓
568F 5694
囯囯
56EF 56FD
圈圈
5708 570F
圓圓
570E 5713
圖圖
5716 5717
垚垚
57D2 57D3
孳孳
5B73 5B76
宮宮
5BAB 5BAE
寬寬
5BDB 5BEC
寧寧
5BDC 5BE7
寢寢
5BDD 5BE2
專專
5C02 5C08
將將
5C06 5C07
尔尔
5C13 5C14
尙尙
5C19 5C1A
尙尙
5C19 5C1A

姍姍
59CD 59D7
姪姪
59EB 59EC
娛娛娛
5A1B 5A2F 5A31
婕婕
5A55 5AAB
媮媮
5A7E 5AAE
媼媼
5AAA 5ABC
嫫嫫
5AAF 5B00
嫫嫫
5B0E 5B14
嫫嫫
5B24 5B37
彈彈
5F39 5F3E
彡彡
5F50 5F51
彡彡
5F54 5F55
彡彡
5F59 5F5A
彡彡
5F5B 5F5C
彡彡
5F5D 5F5E
彡彡
5F65 5F66
德德
5FB3 5FB7
徵徵
5FB4 5FB5
惠惠
5FB4 5FB5

5C2A 5C2B
檻檻
5C36 5C37
屏屏
5C4F 5C5B
崢崢
5CE5 5D22
巔巔
5DD3 5DD4
幘幘
5E21 5E32
帶帶
5E2F 5E36
并并
5E76 5E77
廐廐
5EC4 5ECF
絨絨
5F11 5F12
強強
5F37 5F3A
拔拔
629C 62D4
撚撚
6329 635D
插插插
633F 63D2 63F7
捏捏
634F 63D1
搜搜
635C 641C
揭揭
63B2 63ED
搖搖搖
63FA 6416 6447
搵搵
63FE 6435

6075 60E0
悅悅
6085 60A6
悞悞
609E 60AE
惠惠
60B3 60EA
愠愠
6120 614D
愼愼
613C 614E
戩戩
6229 622C
戲戲
622F 6231
戶戶戶
6236 6237 6238
戾戾
623B 623E
拋拋
629B 62CB
榆榆
6961 6986
概概
6982 69EA
榼榼
6985 69B2
檣檣
699D 6A27
檣檣
69C7 69D9
樣樣
69D8 6A23
橫橫
6A2A 6A6B
步步
6B65 6B69

擊擊 6483 64CA
教教 654E 6559
敫敫 6553 655A
既既 65E2 65E3
昂昂 6602 663B
晚晚 665A 6669
暨暨 66A8 66C1
曾曾 66FD 66FE
柺柺 67B4 67FA
查查 67E5 67FB
柵柵 67F5 6805
稅稅 68B2 68C1
淥淥 6DE5 6E0C
清清 6DF8 6E05
渴渴 6E07 6E34
溫溫 6E29 6EAB
漚漚 6E88 6F59
漑漑 6E89 6F11
滾滾 6EDA 6EFE

- TJ
- T
- T
- T
- T
- T
- T
- J
- T
- T
- T
- T
- T
- T
- T
- T
- T
- T
- T

歲歲 6B72 6B73
歿歿 6B7F 6B81
殼殼 6BBB 6BBC
毀毀 6BC0 6BC1
每每 6BCE 6BCF
氫氫 6C32 6C33
汚汚 6C5A 6C61
沒沒 6C92 6CA1
淨淨 6D44 6DE8
涉涉 6D89 6E09
浼浼 6D97 6D9A
淚淚 6D99 6DDA
眾眾 773E 8846
研研 7814 784F
祿祿 797F 7984
禿禿 79BF 79C3
稅稅 7A05 7A0E
穗穗 7A42 7A57
箏箏 7B5D 7B8F

- T
- T
- GTJ
- T
- T
- T
- T
- TJ
- TJ
- T
- T
- TJ
- TJ
- T
- TJ
- T
- TJ
- TJ
- GJ

潛潛 6F5B 6FF3
瀨瀨 7028 702C
為為 70BA 7232
莞莞 712D 7162
熙熙 7155 7199
煨煨 7174 7185
狀狀 72B6 72C0
瑤瑤 7464 7476
瓶瓶 74F6 7501
產產 7522 7523
瘦瘦 75E9 7626
皞皞 76A1 76A5
眞眞 771E 771F
膾膾 817D 8183
烏烏 8203 8204
舍舍 820D 820E
鋪鋪 8216 8217
莊莊 8358 838A
菑菑

- GTJK
- T
- GTJ
- GTJK
- J
- T
- GT
- TJ
- T
- T
- J
- T
- T
- GT
- TJ
- J
- TJ
- TJ
- TJ

箴箴 7BB3 7C08
篡篡 7BE1 7C12
粵粵 7CA4 7CB5
絕絕 7D55 7D76
綠綠 7DA0 7DD1
緒緒 7DD2 7DD6
緣緣 7DE3 7E01
緼緼 7DFC 7E15
緼緼 7E48 7E66
羹羹 7FAE 7FB9
翺翺 7FF6 7FFA
胼胼 80FC 8141
脫脫 812B 8131
謠謠 8B20 8B21
豸豸 8C5C 8C63
走走 8D70 8D71
駟駟 8EFF 8F27
輜輜 8F1C 8F3A
輜輜

- T
- T
- T
- T
- T
- T
- T
- T
- T
- TJ
- T
- J
- T
- TJ
- T
- J
- T
- T

83D1 8458

藍藍
8480 8495

蔣蔣
848B 8523

薦薦
848D 853F

蒞蒞
8570 8580

薰薰
85AB 85B0

蘊蘊
85F4 860A

虛虛
865A 865B

蛻蛻
86FB 8715

衛衛
885B 885E

袞袞
886E 889E

裝裝
88C5 88DD

訢訢
8A2E 8A7D

說說
8AAA 8AAC

諫諫
8ACC 8AEB

隍隍
9667 9689

青青
9751 9752

靜靜
9759 975C

靱靱
976D 9771

8F3C 8F40

達達
8FBE 8FD6

迸迸
8FF8 902C

遙遙
9059 9065

邢邢
90A2 90C9

郎郎
90CE 90DE

鄉鄉鄉
90F7 9109 9115

醞醞
9196 919E

醬醬
91A4 91AC

鉞鉞
9203 9292

銳銳
92B3 92ED

錄錄
9304 9332

鍊鍊
932C 934A

鎮鎮
93AD 93AE

閱閱
95B1 95B2

高高
9AD8 9AD9

髮髮
9AEA 9AEE

鬪鬪
9B2C 9B2D

鰓鰓
9C1B 9C2E

頽頽
9839 983D

顏顏
984F 9854

顛顛
985A 985B

飲飲
98EE 98F2

餅餅
9905 9920

馱馱
99B1 99C4

駢駢
99E2 9A08

飮飮
9AA9 9AAB

T

TJ

J

J

TJ

TJK

TK

T

鳳鳳
9CEF 9CF3

鸛鸛
9D87 9DAB

鷓鷓
9DC6 9DCF

麪麪
9EAA 9EAB

麼麼
9EBC 9EBD

黃黃
9EC3 9EC4

黑黑
9ED1 9ED2

T

J

J

T

T

T

T

Note:

Following samples are not source code separation pair.

胃胃 5191 80C4	non cognate	胸胸 6710 80CA	non cognate
冲冲 51B2 6C96	2.4.3	眺眺 6713 8101	non cognate
決決 51B3 6C7A	2.4.3	腩腩 6718 8127	non cognate
況況 51B5 6CC1	2.4.3	瞳瞳 6723 81A7	non cognate
垛垛 579B 579C	2.4.3	朵朵 6735 6736	2.4.3
孽孽 5B7C 5B7D	2.4.2	灩灩 7054 7067	2.4.3
寶寶 5BF3 5BF6	2.4.3	稻稻 7A32 7A3B	2.4.3
廳廳 5EF0 5EF3	2.4.1	翹翹 7FF1 7FF6	2.4.3
懷懷 61D0 61F7	2.4.1	考耄考 8007 8008 8009	2.4.3
斂斂 6560 656A	2.4.3	聽聽聽 8074 807C 807D	2.4.1
盼盼 670C 80A6	non cognate	荊荊 8346 834A	2.4.2
跽跽 670F 80D0	non cognate	躲躲 8EB1 8EB2	2.4.3