Annex S
(informative)
Procedure for the unification and arrangement of CJK ideographs

The graphic character collections of CJK Unified ideographs in this document are specified in Clause 34. They are derived from many more ideographs which are found in various different national and regional standards for coded character sets (the "sources").

Annex S describes how the ideographs in this standard are derived from the sources by applying a set of unification procedures. It also describes how the ideographs in this standard are arranged in the sequence of consecutive code points to which they are assigned.

The source references for CJK Unified ideographs are specified in Clause 24.

Within the context of this document 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 point 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”.

NOTE – The unification process does not apply to the following collections of ideographic characters:

CJK RADICALS SUPPLEMENT (2E80 - 2EFF)

KANGXI RADICALS (2F00 - 2FDF)

CJK COMPATIBILITY IDEOGRAPHS (F900 - FAFF with the exception of FA0E, FA0F, FA11, FA13, FA14, FA1F, FA21, FA23, FA24, FA27, FA28 and FA29)

CJK COMPATIBILITY IDEOGRAPHS SUPPLEMENT (2F800-2FA1F).

S.1 Unification procedure

S.1.1 Scope of unification

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

EXAMPLE

士,土

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.

S.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.

S.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;

S.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 S.1.

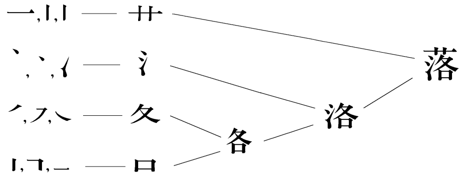


Figure S.1 - Component structure

S.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 top level, as shown in Figure S.2.

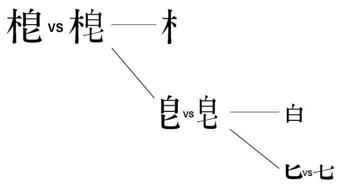


Figure S.2 - The most superior node of a component

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.

S.1.4 Examples of differences of abstract shapes

To illustrate rules derived from a) to c) in S.1.3.2, some typical examples of ideographs that are not unified, owing to differences of abstract shapes, are shown below.

S.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.

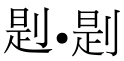
崖•厓, 肱•厷, 降•夅

S.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.

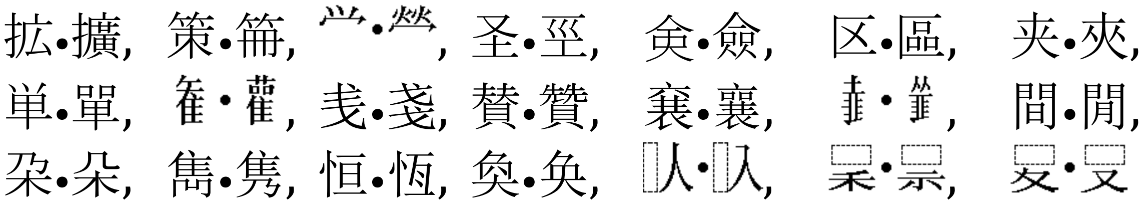
峰•峯, 荊•荆

When the ideographs consists of two horizontally aligned components, a difference of the last stroke of the left-hand component going beneath the right-hand component should not warrant separate encoding, as in the case of the source glyphs for U+34F3:



S.1.4.3 Different structure of a corresponding component

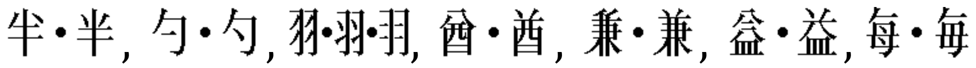
The examples below illustrate rule c). The structure of one (or more) corresponding components within the two ideographs in each pair is different.



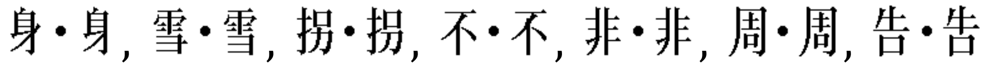
S.1.5 Differences of actual shapes

To illustrate the classification described in S.1.2, some typical examples of ideographs that are unified are shown below. The two or more ideographs in each group below have different actual shapes, but they are considered to have the same abstract shape and are therefore unified on their own or as component in larger ideographs. The differences are classified according to the following examples.

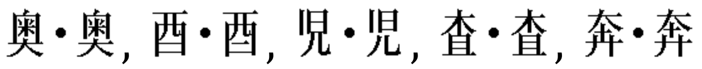
a) Differences in rotated strokes/dots



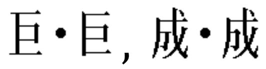
b) Differences in overshoot at the stroke initiation and/or termination



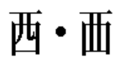
c) Differences in contact of strokes



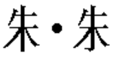
d) Differences in protrusion at the folded corner of strokes



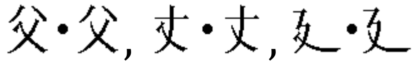
e) Differences in bent strokes



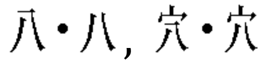
f) Differences in folding back at the stroke termination



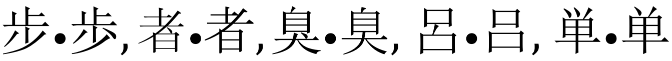
g) Differences in accent at the stroke initiation



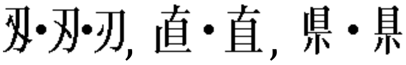
h) Differences in "rooftop" modification



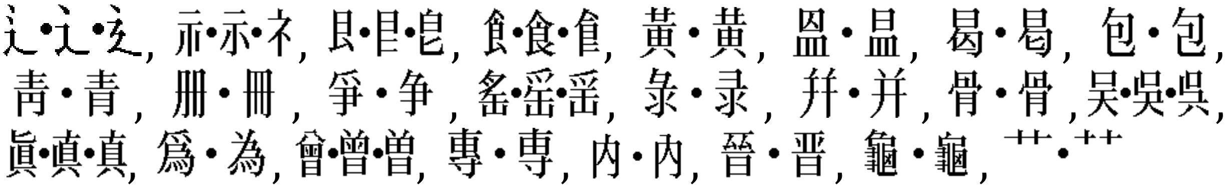
i) Addition or omission of a minor stroke



j) Combinations of the above differences



k) Miscellaneous



NOTE – Some of the group items are unified when used as components in more complex ideographs, but are not unified themselves for other reasons, such as the source separation rule.

These differences in actual shapes of a unified ideograph are presented in the corresponding source columns for each code point entry in the code charts in Clause 34 of this document.

S.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*
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 X 1001:2004 (previously KS C 5601-1989), KS X 1002:2001 (previously KS C 5657-1991)

NOTE 1 – The characters from the J source: JIS X 0212-1990 encoded in this document are listed in the collection 372 JAPANESE IDEOGRAPHICS SUPPLEMENT.

NOTE 2 – 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.

The source separation rule described in S.1.6 only applies to the CJK UNIFIED IDEOGRAPHS block specified in the Basic Multilingual Plane.

NOTE 3 – CJK Compatibility ideographs are created following a rule very similar to the source separation rule. However, the end result is the combination of a single CJK Unified ideograph and one or several CJK Compatibility ideographs. When the source separation rule is applied, all 'similar' source CJK ideographs result in separate CJK Unified ideographs.

S.2 Arrangement procedure

S.2.1 Scope of arrangement

The arrangement of the CJK Unified ideographs in the code charts of Clause 34 of this document is based on the filing order of ideographs in the following dictionaries.

Priority	Dictionary	Edition
1	Kangxi Dictionary 康熙字典	Beijing 7th edition
2	Daikanwa Jiten 大漢和辞典	9th edition
3	Hanyu Dazidian 漢語大字典	1st edition
4	Daejaweon 大字源	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.

S.2.2 Procedure

S.2.2.1 Ideographs found in the dictionaries

a) If an ideograph is found in the Kangxi Dictionary, it is positioned in the code chart 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.

S.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.

S.3 Source separation examples

The pairs (or triplets) of ideographs shown below are exceptions to the unification rules described in S.1 . They are not unified because of the source separation rule described in S.1.6.

NOTE 1 – The particular source group (or groups) that causes the source 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 groups that correspond to these letters are identified in S.1.6.

NOTE 2 – Seven pairs described below have J sources which were originally part of the JIS X 212-1990 level-3, therefore covered by the Source separation rule, which are also part of the JIS X 213:2004 with source identified as J13. The pairs (J13 character code points emphasized) are: **U+5861**-U+586B, U+6483-**U+64CA**, U+75E9-**U+7626**, **U+83D1**-U+8458, U+848B-**U+8523**, U+91A4-**U+91AC**, and **U+985A**-U+985B.

丟 丢 4E1F 4E22	T	仞 仭 4EDE 4EED	J	俁 俣 4FC1 4FE3	TJK	値 值 5024 503C	T
么 幺 4E48 5E7A	GT	併 倂 4F75 5002	T	俞 兪 4FDE 516A	T	偷 偸 5077 5078	T
争 爭 4E89 722D	GTJ	侣 侶 4FA3 4FB6	T	俱 倶 4FF1 5036	T	偽 僞 507D 50DE	TJ

兌 兑	514C 5151	T	剝 剥	525D 5265	T	唧 喞	5527 559E	T	壮 壯	58EE 58EF	GTJ
兎 免	514E 5154	TJ	劒 劔	5292 5294	J	喩 喻	55A9 55BB	T	壽 夀	58FD 5900	T
兖 兗	5156 5157	T	勻 匀	52FB 5300	T	嘘 噓	5618 5653	T	夐 敻	5910 657B	T
冊 册	518A 518C	TJ	单 単	5355 5358	T	嚏 嚔	568F 5694	GTJ	夲 本	5932 672C	GTJ
净 淨	51C0 51C8	G	即 卽	5373 537D	TK	囯 国	56EF 56FD	T	奥 奧	5965 5967	J
凢 凣	51E2 51E3	T	卷 巻	5377 5DFB	TJ	圈 圏	5708 570F	TJ	奨 奬 獎	5968 596C 734E	TJ
刃 刄	5203 5204	TJ	叁 参	53C1 53C2	GT	圎 圓	570E 5713	T	妆 妝	5986 599D	GT
刊 刋	520A 520B	TJ	參 叄	53C3 53C4	T	圖 圗	5716 5717	T	妍 姸	598D 59F8	T
删 刪	5220 522A	T	吕 呂	5415 5442	T	坙 巠	5759 5DE0	T	姍 姗	59CD 59D7	T
别 別	5225 522B	T	吞 呑	541E 5451	T	埒 埓	57D2 57D3	J	姫 姬	59EB 59EC	GT
券 劵	5238 52B5	TJ	吳 吴 呉	5433 5434 5449	TJ	塈 墍	5848 588D	T	娛 娯 娱	5A1B 5A2F 5A31	T
刹 剎	5239 524E	T	吶 呐	5436 5450	T	塡 填	5861 586B	TJ	婕 媫	5A55 5AAB	T
剏 剙	524F 5259	T	吿 告	543F 544A	T	增 増	5897 589E	T	婾 媮	5A7E 5AAE	T

媪 媼	5AAA 5ABC	TK
媯 嬀	5AAF 5B00	T
嬎 嬔	5B0E 5B14	T
嬤 嬷	5B24 5B37	GT
孳 孶	5B73 5B76	T
宫 宮	5BAB 5BAE	T
寛 寬	5BDB 5BEC	T
寜 寧	5BDC 5BE7	T
寝 寢	5BDD 5BE2	GTJ
専 專	5C02 5C08	J
将 將	5C06 5C07	GTJ
尓 尔	5C13 5C14	T
尙 尚	5C19 5C1A	T
尪 尫	5C2A 5C2B	T
尶 尷	5C36 5C37	T
屏 屛	5C4F 5C5B	T
峥 崢	5CE5 5D22	GT
巓 巔	5DD3 5DD4	T
帡 帲	5E21 5E32	T
帯 帶	5E2F 5E36	TJ
并 幷	5E76 5E77	T
廄 廏	5EC4 5ECF	T
弑 弒	5F11 5F12	T
強 强	5F37 5F3A	T
弹 弾	5F39 5F3E	T
彐 彑	5F50 5F51	TJ
彔 录	5F54 5F55	T
彙 彚	5F59 5F5A	T
彛 彜	5F5B 5F5C	J
彝 彞	5F5D 5F5E	T
彥 彦	5F65 5F66	T
徳 德	5FB3 5FB7	T
徴 徵	5FB4 5FB5	T
恵 惠	6075 60E0	TJ
悅 悦	6085 60A6	T
悞 悮	609E 60AE	T
悳 惪	60B3 60EA	T
愠 慍	6120 614D	T
愼 慎	613C 614E	TJ
戩 戬	6229 622C	GT
戯 戱	622F 6231	T
戶 户 戸	6236 6237 6238	T
戻 戾	623B 623E	T
抛 拋	629B 62CB	T
抜 拔	629C 62D4	TJ
挩 捝	6329 635D	T
挿 插 揷	633F 63D2 63F7	TJ
捏 揑	634F 63D1	TJ
捜 搜	635C 641C	TJ
掲 揭	63B2 63ED	T
揺 搖 摇	63FA 6416 6447	TJ
揾 搵	63FE 6435	T

撃 擊 TJ
6483 64CA

敎 教 T
654E 6559

敓 敚 T
6553 655A

既 旣 T
65E2 65E3

昂 昻 T
6602 663B

晚 晩 T
665A 6669

暨 曁 T
66A8 66C1

曽 曾 J
66FD 66FE

枴 柺 T
67B4 67FA

查 査 T
67E5 67FB

柵 栅 T
67F5 6805

梲 棁 T
68B2 68C1

楡 榆 T
6961 6986

概 槪 T
6982 69EA

榅 榲 T
6985 69B2

榝 樧 T
699D 6A27

槇 槙 J
69C7 69D9

様 樣 TJ
69D8 6A23

横 橫 T
6A2A 6A6B

步 歩 T
6B65 6B69

歲 歳 T
6B72 6B73

歿 殁 T
6B7F 6B81

殻 殼 GTJ
6BBB 6BBC

毀 毁 T
6BC0 6BC1

每 毎 T
6BCE 6BCF

氲 氳 T
6C32 6C33

汚 污 T
6C5A 6C61

沒 没 TJ
6C92 6CA1

浄 淨 TJ
6D44 6DE8

涉 渉 T
6D89 6E09

涗 涚 T
6D97 6D9A

涙 淚 T
6D99 6DDA

淥 渌 T
6DE5 6E0C

淸 清 T
6DF8 6E05

渇 渴 T
6E07 6E34

温 溫 T
6E29 6EAB

溈 潙 T
6E88 6F59

溉 漑 T
6E89 6F11

滚 滾 T
6EDA 6EFE

潛 濳 GTJK
6F5B 6FF3

瀨 瀬 T
7028 702C

為 爲 GTJ
70BA 7232

焭 煢 GTJK
712D 7162

煕 熙 J
7155 7199

煴 熅 T
7174 7185

状 狀 GT
72B6 72C0

瑤 瑶 TJ
7464 7476

瓶 甁 T
74F6 7501

産 產 T
7522 7523

痩 瘦 J
75E9 7626

皡 皥 T
76A1 76A5

眞 真 TJ
771E 771F

眾 衆	773E 8846	TJK
研 硏	7814 784F	T
祿 禄	797F 7984	TJ
禿 秃	79BF 79C3	T
稅 税	7A05 7A0E	T
穂 穗	7A42 7A57	TJ
筝 箏	7B5D 7B8F	GJ
箳 簈	7BB3 7C08	T
篡 簒	7BE1 7C12	T
粤 粵	7CA4 7CB5	T
絕 絶	7D55 7D76	T
綠 緑	7DA0 7DD1	T
緒 緖	7DD2 7DD6	T
緣 縁	7DE3 7E01	T
緼 縕	7DFC 7E15	T
繈 繦	7E48 7E66	T
羮 羹	7FAE 7FB9	TJ
翶 翺	7FF6 7FFA	T
胼 腁	80FC 8141	T
脫 脱	812B 8131	T
腽 膃	817D 8183	T
舃 舄	8203 8204	GT
舍 舎	820D 820E	TJ
舖 舗	8216 8217	J
荘 莊	8358 838A	TJ
菑 葘	83D1 8458	TJ
蒀 蒕	8480 8495	T
蒋 蔣	848B 8523	GJ
蒍 蔿	848D 853F	T
蕰 薀	8570 8580	T
薫 薰	85AB 85B0	T
藴 蘊	85F4 860A	T
虚 虛	865A 865B	T
蛻 蜕	86FB 8715	T
衛 衞	885B 885E	TJK
衮 袞	886E 889E	TK
装 裝	88C5 88DD	GJK
訮 詽	8A2E 8A7D	T
說 説	8AAA 8AAC	T
諌 諫	8ACC 8AEB	TJ
謠 謡	8B20 8B21	J
豜 豣	8C5C 8C63	T
走 赱	8D70 8D71	TJ
軿 輧	8EFF 8F27	T
輜 輺	8F1C 8F3A	J
輼 轀	8F3C 8F40	T
达 迖	8FBE 8FD6	T
迸 逬	8FF8 902C	TJ
遙 遥	9059 9065	J
邢 郉	90A2 90C9	T
郎 郞	90CE 90DE	T
郷 鄉 鄕	90F7 9109 9115	T

醖 醞 9196 919E	T	陧 隉 9667 9689	G	餅 餠 9905 9920	TJ	鳯 鳳 9CEF 9CF3	T
醤 醬 91A4 91AC	J	靑 青 9751 9752	T	馱 駄 99B1 99C4	TJK	鶇 鶫 9D87 9DAB	J
鈃 銒 9203 9292	T	静 靜 9759 975C	GTJ	駢 騈 99E2 9A08	TK	鷆 鷏 9DC6 9DCF	J
銳 鋭 92B3 92ED	T	靭 靱 976D 9771	J	骩 骫 9AA9 9AAB	T	麪 麫 9EAA 9EAB	T
錄 録 9304 9332	T	頹 頽 9839 983D	T	高 髙 9AD8 9AD9	T	麼 麽 9EBC 9EBD	T
錬 鍊 932C 934A	TK	顏 顔 984F 9854	TJ	髪 髮 9AEA 9AEE	TJ	黃 黄 9EC3 9EC4	T
鎭 鎮 93AD 93AE	TJ	顚 顛 985A 985B	J	鬬 鬭 9B2C 9B2D	T	黑 黒 9ED1 9ED2	T
閱 閲 95B1 95B2	T	飮 飲 98EE 98F2	J	鰛 鰮 9C1B 9C2E	TJ		

S.4 Non-unification examples

In accordance with the unification procedures described in S.1, 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 S.1.1.

NOTE – The reason for non-unification in these examples is different from the source separation rule described in S.1.6.

冑 胄 5191 80C4	non cognate	况 況 51B5 6CC1	S.1.4.3	寳 寶 5BF3 5BF6	S.1.4.3	敠 敪 6560 656A	S.1.4.3
冲 沖 51B2 6C96	S.1.4.3	垛 垜 579B 579C	S.1.4.3	廰 廳 5EF0 5EF3	S.1.4.1	朌 肦 670C 80A6	non cognate
决 決 51B3 6C7A	S.1.4.3	孼 孽 5B7C 5B7D	S.1.4.2	懐 懷 61D0 61F7	S.1.4.1	朏 胐 670F 80D0	non cognate

朐胊	6710 80CA	non cognate
朓脁	6713 8101	non cognate
朘脧	6718 8127	non cognate
朣膧	6723 81A7	non cognate
朵朶	6735 6736	S.1.4.3
灔灧	7054 7067	S.1.4.3
稲稻	7A32 7A3B	S.1.4.3
翱翶	7FF1 7FF6	S.1.4.3
耇耈耉	8007 8008 8009	S.1.4.3
聴聼聽	8074 807C 807D	S.1.4.1
荆荊	8346 834A	S.1.4.2
躱躲	8EB1 8EB2	S.1.4.3