Annex S
(informative)

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”.

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).

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.

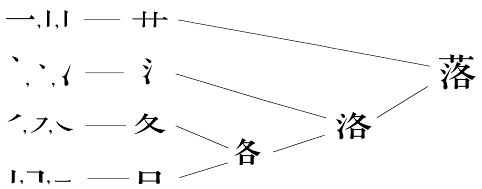


Figure S.1 - Component structure

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 most superior node, as shown in Figure S.2.

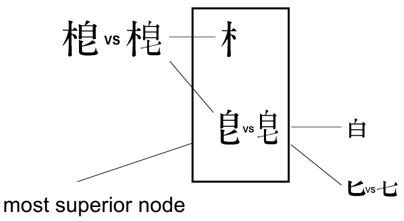


Figure S.2 - The most superior node of a component

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

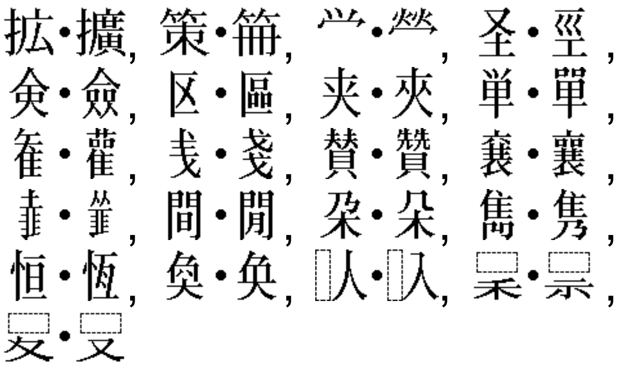
S.1.4.2 Different relative positions of components

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

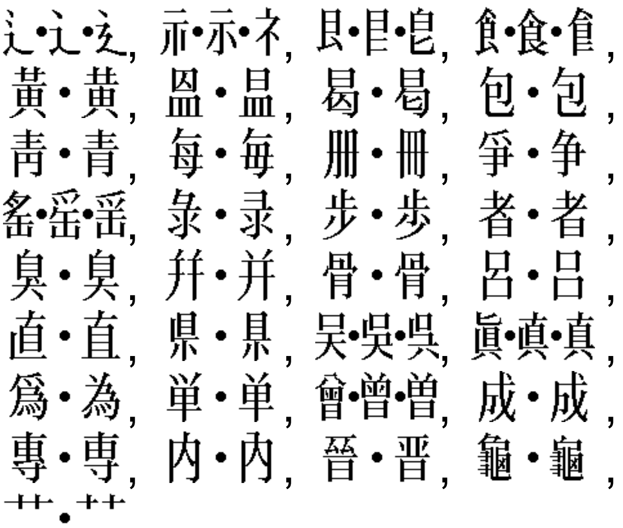
S.1.4.3 Different structure of a corresponding component

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



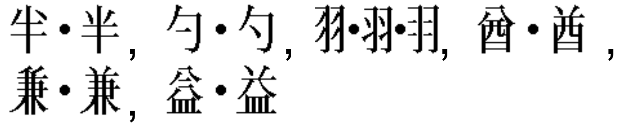
S.1.5 Differences of actual shapes

To illustrate the classification described in S.1.2, some typical examples of ideographs that are unified are shown below. The two or 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



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

奥•奥, 酉•酉, 児•児, 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, 廴•廴

h) Differences in "rooftop" modification

八•八, 穴•穴

j) Combinations of the above differences

刃•刃•刃

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.

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 C 5601-1989, KS C 5657-1991

(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.

S.2. Arrangement procedure

S.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.

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	汉语大字典	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 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.

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 code separation examples

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

NOTE - 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.

丟丢	4E1F 4E22	T	兎兔	514E 5154	TJ	劒劔	5292 5294	J	喩喻	55A9 55BB	T
么幺	4E48 5E7A	GT	兖兗	5156 5157	T	勻匀	52FB 5300	T	嘘噓	5618 5653	T
争爭	4E89 722D	GTJ	冊册	518A 518C	TJ	单単	5355 5358	T	嚏嚔	568F 5694	GTJ
仞仭	4EDE 4EED	J	净淨	51C0 51C8	G	即卽	5373 537D	TK	囯国	56EF 56FD	T
併倂	4F75 5002	T	凢凣	51E2 51E3	T	卷巻	5377 5DFB	TJ	圈圏	5708 570F	TJ
侣侶	4FA3 4FB6	T	刃刄	5203 5204	TJ	叁参	53C1 53C2	GT	圎圓	570E 5713	T
俁俣	4FC1 4FE3	TJK	刊刋	520A 520B	TJ	參叄	53C3 53C4	T	圖圗	5716 5717	T
俞兪	4FDE 516A	T	删刪	5220 522A	T	吕呂	5415 5442	T	坙巠	5759 5DE0	T
俱倶	4FF1 5036	T	别別	5225 522B	T	吞呑	541E 5451	T	埒埓	57D2 57D3	J
値值	5024 503C	T	券劵	5238 52B5	TJ	吳吴呉	5433 5434 5449	TJ	塈墍	5848 588D	T
偷偸	5077 5078	T	刹剎	5239 524E	T	吶呐	5436 5450	T	塡填	5861 586B	TJ
偽僞	507D 50DE	TJ	剏剙	524F 5259	T	吿告	543F 544A	T	增増	5897 589E	T
兌兑	514C 5151	T	剝剥	525D 5265	T	唧喞	5527 559E	T	壮壯	58EE 58EF	GTJ

壽夀	T	58FD 5900
夐敻	T	5910 657B
夲本	GTJ	5932 672C
奥奧	J	5965 5967
奨奬獎	TJ	5968 596C 734E
妆妝	GT	5986 599D
妍姸	T	598D 59F8
姍姗	T	59CD 59D7
姫姬	GT	59EB 59EC
娛娯娱	T	5A1B 5A2F 5A31
婕媫	T	5A55 5AAB
婾媮	T	5A7E 5AAE
媪媼	TK	5AAA 5ABC
媯嬀	T	5AAF 5B00
嬎嬔	T	5B0E 5B14
嬤嬷	GT	5B24 5B37
孳孶	T	5B73 5B76
宫宮	T	5BAB 5BAE
寛寬	T	5BDB 5BEC
寜寧	T	5BDC 5BE7
寝寢	GTJ	5BDD 5BE2
専專	J	5C02 5C08
将將	GTJ	5C06 5C07
尓尔	T	5C13 5C14
尙尚	T	5C19 5C1A
尪尫	T	5C2A 5C2B
尶尷	T	5C36 5C37
屏屛	T	5C4F 5C5B
峥崢	GT	5CE5 5D22
巓巔	T	5DD3 5DD4
帡帲	T	5E21 5E32
带帶	TJ	5E2F 5E36
并幷	T	5E76 5E77
廄廏	T	5EC4 5ECF
弑弒	T	5F11 5F12
強强	T	5F37 5F3A
弹弾	T	5F39 5F3E
彐彑	TJ	5F50 5F51
彔录	T	5F54 5F55
彙彚	T	5F59 5F5A
彛彜	J	5F5B 5F5C
彝彞	T	5F5D 5F5E
彥彦	T	5F65 5F66
徳德	T	5FB3 5FB7
徴徵	T	5FB4 5FB5
恵惠	TJ	6075 60E0
悅悦	T	6085 60A6
悞悮	T	609E 60AE
悳惪	T	60B3 60EA
愠慍	T	6120 614D
愼慎	TJ	613C 614E
戩戬	GT	6229 622C
戯戱	T	622F 6231
戶户戸	T	6236 6237 6238
戻戾	T	623B 623E
抛拋	T	629B 62CB
抜拔	TJ	629C 62D4
挩捝	T	6329 635D
挿插揷	TJ	633F 63D2 63F7
捏揑	TJ	634F 63D1
捜搜	TJ	635C 641C
掲揭	T	63B2 63ED
揺搖摇	TJ	63FA 6416 6447
揾搵	T	63FE 6435

撃擊 6483 64CA	TJ	槇槙 69C7 69D9	J	淥渌 6DE5 6E0C	T	產産 7522 7523	T
敎教 654E 6559	T	様樣 69D8 6A23	TJ	淸清 6DF8 6E05	T	痩瘦 75E9 762	J
敓敚 6553 655A	T	横橫 6A2A 6A6B	T	渇渴 6E07 6E34	T	皡皥 76A1 76A5	T
既旣 65E2 65E3	T	步歩 6B65 6B69	T	温溫 6E29 6EAB	T	眞真 771E 771F	TJ
昂昻 6602 663B	T	歲歳 6B72 6B73	T	溈潙 6E88 6F59	T	眾衆 773E 8846	TJK
晚晩 665A 6669	T	歿殁 6B7F 6B81	T	溉漑 6E89 6F11	T	研硏 7814 784F	T
暨曁 66A8 66C1	T	殻殼 6BBB 6BBC	GTJ	滚滾 6EDA 6EFE	T	祿禄 797F 7984	TJ
曽曾 66FD 66FE	J	毀毁 6BC0 6BC1	T	潛濳 6F5B 6FF3	GTJK	禿秃 79BF 79C3	T
枴柺 67B4 67FA	T	毎每 6BCE 6BCF	T	瀨瀬 7028 702C	T	税稅 7A05 7A0E	T
查査 67E5 67FB	T	氲氳 6C32 6C33	T	為爲 70BA 7232	GTJ	穂穗 7A42 7A57	TJ
柵栅 67F5 6805	T	汚污 6C5A 6C61	T	焭煢 712D 7162	GTJK	筝箏 7B5D 7B8F	GJ
梲棁 68B2 68C1	T	沒没 6C92 6CA1	TJ	煕熙 7155 7199	J	箳簈 7BB3 7C08	T
楡榆 6961 6986	T	浄淨 6D44 6DE8	TJ	煴熅 7174 7185	T	篡簒 7BE1 7C12	T
概槪 6982 69EA	T	涉渉 6D89 6E09	T	状狀 72B6 72C0	GT	粤粵 7CA4 7CB5	T
榅榲 6985 69B2	T	涗涚 6D97 6D9A	T	瑤瑶 7464 7476	TJ	絕絶 7D55 7D76	T
榝樧 699D 6A27	T	涙淚 6D99 6DDA	T	瓶甁 74F6 7501	T	綠緑 7DA0 7DD1	T

緒緖 7DD2 7DD6	T	蒍蔿 848D 853F	T	輜輺 8F1C 8F3A	J	陧隉 9667 9689	G
緣縁 7DE3 7E01	T	蕰薀 8570 8580	T	輼轀 8F3C 8F40	T	靑青 9751 9752	T
緼縕 7DFC 7E15	T	薫薰 85AB 85B0	T	达迖 8FBE 8FD6	T	静靜 9759 975C	GTJ
繈繦 7E48 7E66	T	藴蘊 85F4 860A	T	迸逬 8FF8 902C	TJ	靭靱 976D 9771	J
羮羹 7FAE 7FB9	TJ	虚虛 865A 865B	T	遙遥 9059 9065	J	頹頽 9839 983D	T
翶翺 7FF6 7FFA	T	蛻蜕 86FB 8715	T	邢郉 90A2 90C9	T	顏顔 984F 9854	TJ
胼腁 80FC 8141	T	衛衞 885B 885E	TJK	郎郞 90CE 90DE	T	顚顛 985A 985B	J
脫脱 812B 8131	T	衮袞 886E 889E	TK	郷鄉鄕 90F7 9109 9115	T	飮飲 98EE 98F2	J
腽膃 817D 8183	T	装裝 88C5 88DD	GJK	醖醞 9196 919E	T	餅餠 9905 9920	TJ
舃舄 8203 8204	GT	訮詽 8A2E 8A7D	T	醤醬 91A4 91AC	J	馱駄 99B1 99C4	TJK
舍舎 820D 820E	TJ	說説 8AAA 8AAC	T	鈃銒 9203 9292	T	駢騈 99E2 9A08	TK
舖舗 8216 8217	J	諌諫 8ACC 8AEB	TJ	銳鋭 92B3 92ED	T	骩骫 9AA9 9AAB	T
荘莊 8358 838A	TJ	謠謡 8B20 8B21	J	錄録 9304 9332	T	高髙 9AD8 9AD9	T
菑葘 83D1 8458	TJ	豜豣 8C5C 8C63	T	錬鍊 932C 934A	TK	髪髮 9AEA 9AEE	TJ
蒀蒕 8480 8495	T	走赱 8D70 8D71	TJ	鎭鎮 93AD 93AE	TJ	鬬鬭 9B2C 9B2D	T
蒋蔣 848B 8523	GJ	軿輧 8EFF 8F27	T	閱閲 95B1 95B2	T	鰛鰮 9C1B 9C2E	TJ

鳯鳳 9CEF 9CF3	T	鷆鷏 9DC6 9DCF	J	麼麽 9EBC 9EBD	T	黑黒 9ED1 9ED2	T
鶇鶫 9D87 9DAB	J	麪麫 9EAA 9EAB	T	黃黄 9EC3 9EC4	T		

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

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

冑胄 5191 80C4	non cognate	寳寶 5BF3 5BF6	S.1.4.3	朐胊 6710 80CA	non cognate	稲稻 7A32 7A3B	S.1.4.3
冲沖 51B2 6C96	S.1.4.3	廰廳 5EF0 5EF3	S.1.4.1	朓脁 6713 8101	non cognate	翱翶 7FF1 7FF6	S.1.4.3
决決 51B3 6C7A	S.1.4.3	懐懷 61D0 61F7	S.1.4.1	朘脧 6718 8127	non cognate	耇耈耉 8007 8008 8009	S.1.4.3
况況 51B5 6CC1	S.1.4.3	敠敪 6560 656A	S.1.4.3	朣膧 6723 81A7	non cognate	聴聼聽 8074 807C 807D	S.1.4.1
垛垜 579B 579C	S.1.4.3	朌肦 670C 80A6	non cognate	朵朶 6735 6736	S.1.4.3	荆荊 8346 834A	S.1.4.2
孼孽 5B7C 5B7D	S.1.4.2	朏胐 670F 80D0	non cognate	灔灧 7054 7067	S.1.4.3	躱躲 8EB1 8EB2	S.1.4.3