IRG N949

ISO/IEC JTC 1/SC 2/WG 2/IRG
Ideographic Rapporteur Group
Secretariat: China

Title: Clarification to ISO/IEC 10646 Annex S on CJK Unification

Doc. Type: National Body Contribution

Source: James Seng - Singapore, Chen Zhuang - China, Selena - TCA,

on the behalf of the Ext C Adhoc Group

Status: work in progress

Date: 19th Nov 2002

Distribution: ISO/IEC JTC 1/SC2/WG2 & IRG

Reference: IRG N941

This is a work-in-process document started during the Ext C Adhoc Group in IRG #20 Hanoi. This document is subjected to further changes and modifications and must not be used as a normative reference in other work.

The document clarifies how Annex S do its CJK Unification. It also documents the current practices of IRG on submission and reviewing methodology of the ideographs.

This document is in addition to the Annex S of ISO/IEC 10646. It is not intended to be a replacement to Annex S. If there is any conflict between this document and Annex S, Annex S takes precedence.

1. Definitions

CJK Ideograph/Ideograph
Component
Nodes
Unification
Node
Element
Cognate character

2. Introduction

Annex S of ISO/IEC 10646 describe the “Procedure for the Unification and arrangement of CJK Ideographs”. The goal for this document is to clarify the procedure and process used for CJK Ideographs Unification.

3. Submission Procedure

3.1. Submission criteria

[TBD] In general, non-ideograph must not be submitted for CJK Ideographs. CJK Radical, Symbols, Strokes and other CJK-related or non-CJK characters should be proposed into their respective blocks in the ISO/IEC 10646 directly to the WG2. However, exception may be made on a case-by-case basis. [/TBD]

Ideographs from handwritten source are not accepted. And any handwritten variants should not be submitted. [Question: How do we define "handwritten variants?"]

The radical of the ideograph should be well-defined. Otherwise, the ideograph will be rejected unless the submitter provides additional information on the new radical.

In mainland China, there is a difference between stroke normalization, and simplification including derived simplification. And for many applications, there is a need to distinguish characters in its simplified or traditional form in display or other processes, for other applications, to treat the simplified and traditional form as the same. Therefore, traditional or simplified ideographs will not be rejected for submission.

However, theoretical ideographs that are created from existing ideographs using mainland China Simplification Rules will not be a consideration for submission. Instead, all ideographs must provide documented proof on its usage.

[Question; Who can submit?]

3.2. Chief Editor for submission

[To be clarified in IRG#20] For any submission of ideographs, there must be a Chief Editor for the submission who is responsible for (a) the quality of the proposal (b) first-point of contact for any questions and issues of the submission.

3.4. Mistakes in submission

To ensure the quality of the submission and to reduce the work load for the reviewers, any submission that has more than 5% mistakes of its ideographs will be rejected.

3.4. Submission information

[TBD] Any submission for an ideograph should include the glyph, kangxi code, radical, first stroke, stroke count and source. [/TBD]

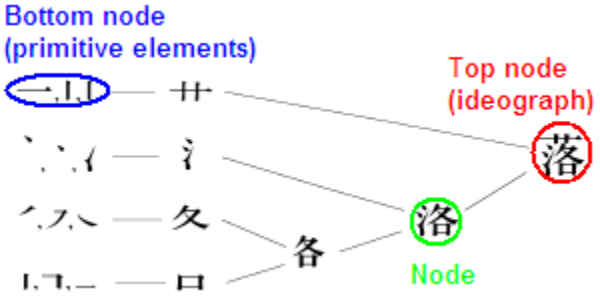
It is possible to create characters using China Simplification rules. With this rule, there are Therefore, there are many theoretical possible Simplified Chinese Ideographs or Traditional Chinese Ideographs

4. Unification Procedure

4.1. Abstract Shape (Annex S1.3)

The basic unification procedure is to check if the two ideographs have the same abstract shape subjected to Unification Rules.

An ideograph can be expressed by its components and the components structure. An ideograph can therefore be expressed by a component tree, where the top node is the ideograph itself, and the bottom nodes are the primitive elements.



[Question: There component tree are not unique. There are many ways to break it down]

Examine the every node of an ideograph, starting from the bottom nodes (i.e. primitive elements) and compare the corresponding node of the other ideograph and then move upwards.

The nodes are considered *same* if

a) the abstract shapes and the relative position of the components are exactly the same; or
b) the abstract shapes of the nodes are unified according to Unification Rules in Appendix A; or
c) all the immediate child nodes are *same*

The ideographs are considered *unified* if the top nodes are *same*, unless

a) the two ideographs have different number of components, then they are *dis-unified*. For example,

崖•厓, 肱•厷, 降•夅

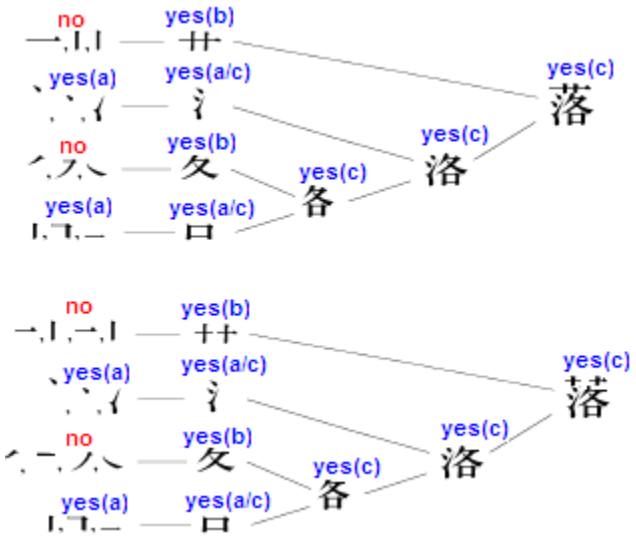
b) the component structures or the position of the components of the two ideographs are different, even if they have the same components, then they are *dis-unified*. For example

峰•峯, 荊•荆

If any of the components *maybe unified* according to the Unification Rules specified in A5, then ideographs must be bought up for further discussion on a case-by-case basis by the reviewers.

For example,

a) 落 and 落,



Therefore, 落 and 落 are *unified*.

b) [Will provide a counter-example of when it will dis-unified]

4.2. Exception to the unification rules

Even if two ideographs are determined to be unified as according to their abstract shape (as described in 4.1), there are exceptions whereby they *maybe dis-unified*

a) if the two ideographs are unrelated in historical derivation (non-cognate characters)

Stroke count difference is not a consideration for dis-unification.

Source separate is *no longer* a consideration for dis-unification.

4.3. Exception rules to the ideographs that are not unified

Even if two ideographs are determined to be not unified as according to their abstract shape (as described in 4.1), there are exceptions whereby they *maybe unified*

a) if the two ideographs are very similar and considered related in historical derivation (cognate characters)

Appendix A – Unification Rules

[TBD] Any modification to the Unification Rules must be a consensus of the IRG members. [/TBD]

[Note: Should put this into A5] We do not attempt unified simplified and traditional ideographs. However, stroke normalization are considered for unification and they should be unified. [/Note]

A1. Basic Unification Rules

Examine the ideographs for font and glyph design differences. All font or glyph design differences are *unified*. For example,
[Need some examples]

Any ideographs or components are *unified* if they are listed in A2.

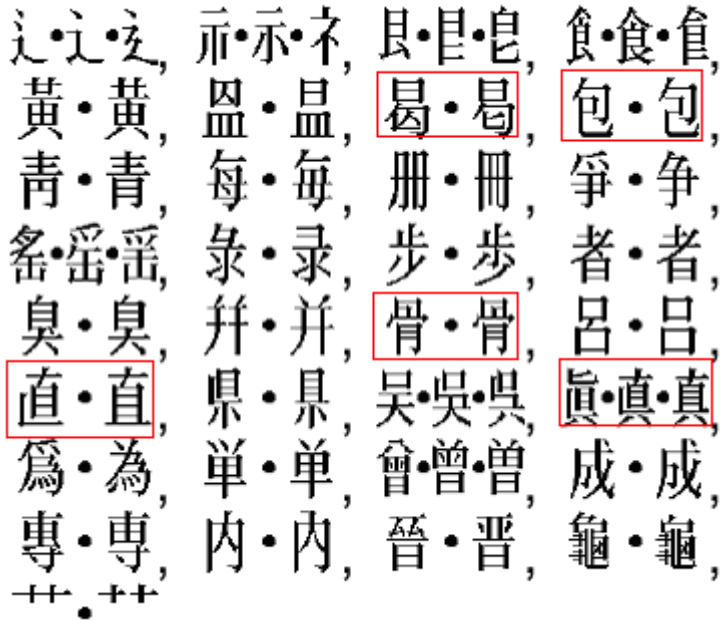
Any components are *unified* if they are listed in A3.

Any ideographs are *unified* if they are listed in A4.

If the ideographs or components have slight differences described listed in A5, they *maybe unified* (unless they are listed in the examples in A5, then they are *unified*). These ideographs must be bought up for further discussions on a case-by-case basis by the reviewers.

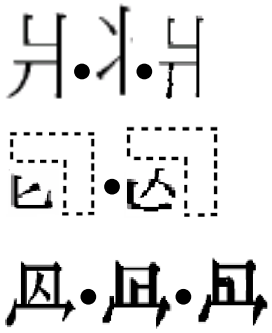
A2. Unified Ideographs

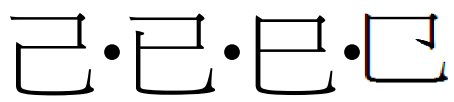
The groups of two or three ideographs below are also considered *unified*, regardless whether they are been used as an ideograph or as a component of an ideograph.

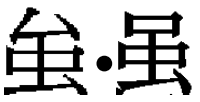


A3. Unified Components only

The groups of two or three ideographs below are considered *unified* only if they are used as a component of an ideograph.



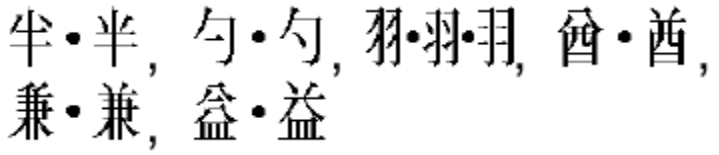


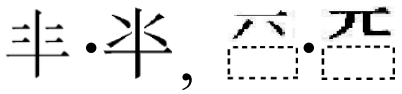


A4. Unified Ideographs only

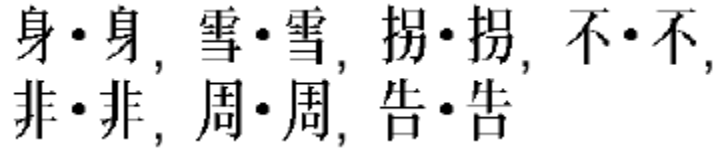
A5. Ideographs or Components that maybe unified

a) Differences between rotated strokes and dots *maybe unified*. For example,

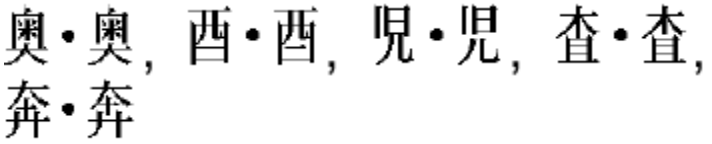




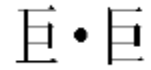
b) Differences between overshoot at the stroke initiation and/or termination *maybe unified*. For example,



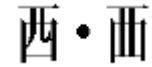
c) Differences in contact of stroke are *unified*. For example,



d) Differences in protrusion at the folded corner of strokes *maybe unified*. For example,

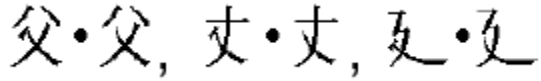


e) Differences in bent strokes *maybe unified*. For example,

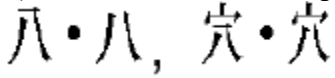


f) Differences in folding back at the stroke termination *maybe unified*. For example,

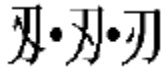
g) Differences in accent at the stroke initiation *maybe unified*. For example,



h) Differences in “rooftop” modification *maybe unified*. For example,



i) Combinations of the above differences *maybe unified*. For example,

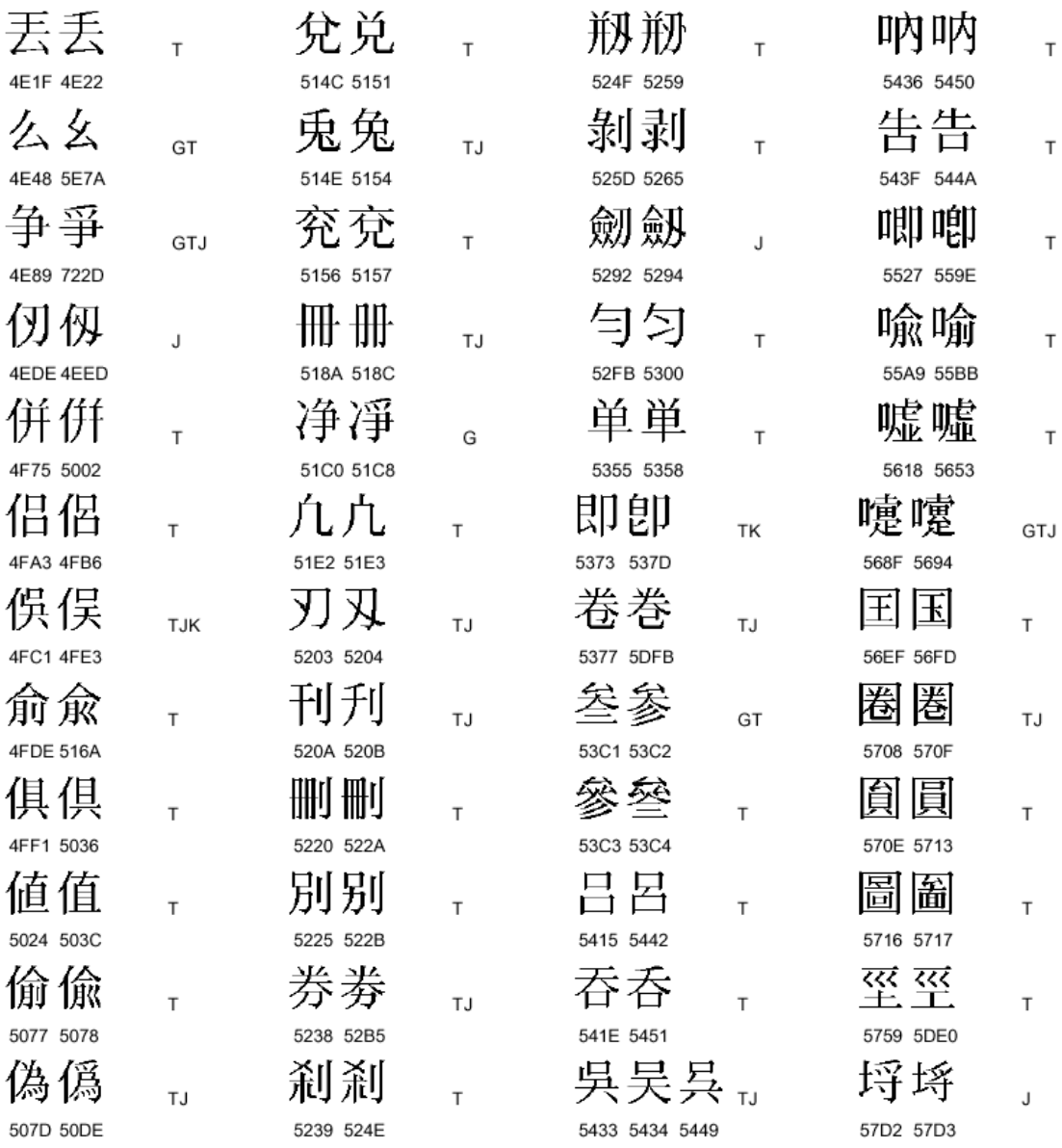


A6. For further discussion

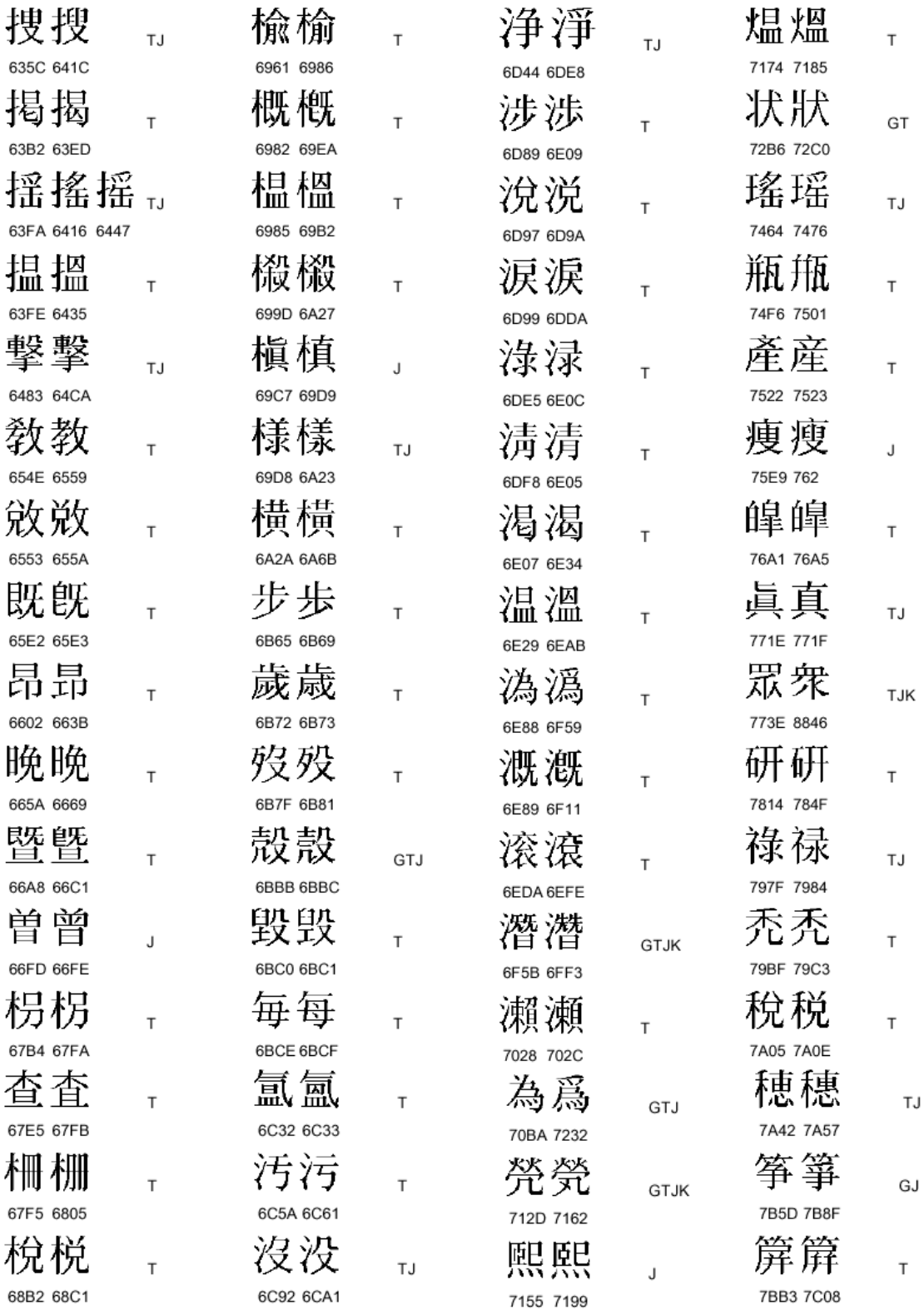
Appendix B – Examples of Source Separation

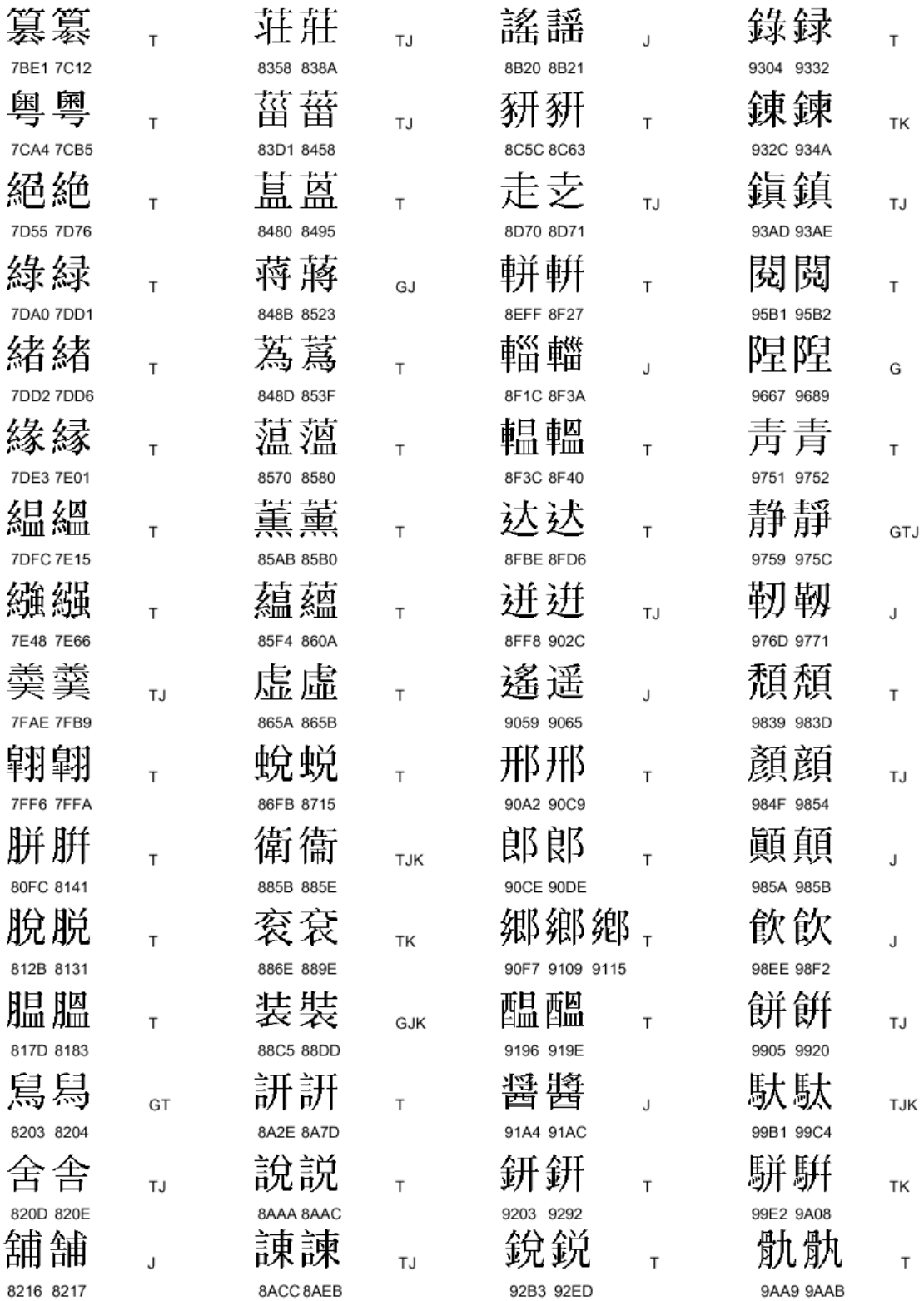
The following are source separated examples extracted from Annex S. Source separation is no longer a consideration for *dis-unification*. The list may not be exhaustive.

However, these are useful examples of what ideographs should be *unified* but was not for historical reasons.




堅 堅 T
5848 588D
媼 媼 TK
5AAA 5ABC
崢 崢 GT
5CE5 5D22
徵 徵 T
5FB4 5FB5
塡 填 TJ
5861 586B
嫣 嫣 T
5AAF 5B00
巔 巔 T
5DD3 5DD4
惠 惠 TJ
6075 60E0
增 增 T
5897 589E
嫦 嫦 T
5B0E 5B14
帡 帡 T
5E21 5E32
悅 悅 T
6085 60A6
壯 壯 GTJ
58EE 58EF
嬤 嬤 GT
5B24 5B37
帶 帶 TJ
5E2F 5E36
悞 悞 T
609E 60AE
壽 壽 T
58FD 5900
孳 孳 T
5B73 5B76
并 幷 T
5E76 5E77
悳 悳 T
60B3 60EA
奠 奠 T
5910 657B
宮 宮 T
5BAB 5BAE
廄 廄 T
5EC4 5ECF
愠 慍 T
6120 614D
本 本 GTJ
5932 672C
寬 寬 T
5BDB 5BEC
弒 弒 T
5F11 5F12
愼 慎 TJ
613C 614E
奧 奧 J
5965 5967
寧 寧 T
5BDC 5BE7
強 强 T
5F37 5F3A
戩 戩 GT
6229 622C
獎 奬 獎 TJ
5968 596C 734E
寢 寢 GTJ
5BDD 5BE2
彈 弹 T
5F39 5F3E
戲 戲 T
622F 6231
妝 妝 GT
5986 599D
專 專 J
5C02 5C08
彐 彑 TJ
5F50 5F51
戶 户 戶 T
6236 6237 6238
妍 妍 T
598D 59F8
將 將 GTJ
5C06 5C07
彔 录 T
5F54 5F55
戾 戾 T
623B 623E
姍 姗 T
59CD 59D7
尒 尔 T
5C13 5C14
彙 彙 T
5F59 5F5A
拋 抛 T
629B 62CB
姬 姬 GT
59EB 59EC
尙 尚 T
5C19 5C1A
彝 彝 J
5F5B 5F5C
拔 拔 TJ
629C 62D4
娛 娛 娛 T
5A1B 5A2F 5A31
尷 尷 T
5C2A 5C2B
彝 彝 T
5F5D 5F5E
挩 挩 T
6329 635D
婕 婕 T
5A55 5AAB
尷 尷 T
5C36 5C37
彥 彦 T
5F65 5F66
插 插 插 TJ
633F 63D2 63F7
媮 媮 T
5A7E 5AAE
屏 屏 T
5C4F 5C5B
德 德 T
5FB3 5FB7
揑 揑 TJ
634F 63D1


搜 搜 TJ
635C 641C
揭 揭 T
63B2 63ED
搖 搖 搖 TJ
63FA 6416 6447
揾 揾 T
63FE 6435
擊 擊 TJ
6483 64CA
敎 教 T
654E 6559
敓 敓 T
6553 655A
既 既 T
65E2 65E3
昂 昂 T
6602 663B
晚 晚 T
665A 6669
暨 暨 T
66A8 66C1
曾 曾 J
66FD 66FE
枴 枴 T
67B4 67FA
查 查 T
67E5 67FB
柵 柵 T
67F5 6805
梲 梲 T
68B2 68C1
榆 榆 T
6961 6986
概 概 T
6982 69EA
榲 榲 T
6985 69B2
榖 榖 T
699D 6A27
槇 槙 J
69C7 69D9
樣 樣 TJ
69D8 6A23
橫 橫 T
6A2A 6A6B
步 步 T
6B65 6B69
歲 歲 T
6B72 6B73
歿 歿 T
6B7F 6B81
殼 殼 GTJ
6BBB 6BBC
毀 毀 T
6BC0 6BC1
每 每 T
6BCE 6BCF
氳 氳 T
6C32 6C33
汚 污 T
6C5A 6C61
沒 沒 TJ
6C92 6CA1
淨 淨 TJ
6D44 6DE8
涉 涉 T
6D89 6E09
涗 涗 T
6D97 6D9A
淚 淚 T
6D99 6DDA
淥 淥 T
6DE5 6E0C
淸 清 T
6DF8 6E05
渴 渴 T
6E07 6E34
溫 溫 T
6E29 6EAB
潙 潙 T
6E88 6F59
溉 溉 T
6E89 6F11
滾 滾 T
6EDA 6EFE
潛 潛 GTJK
6F5B 6FF3
瀨 瀨 T
7028 702C
為 爲 GTJ
70BA 7232
煢 煢 GTJK
712D 7162
熙 熙 J
7155 7199
煴 煴 T
7174 7185
狀 狀 GT
72B6 72C0
瑤 瑤 TJ
7464 7476
瓶 瓶 T
74F6 7501
產 産 T
7522 7523
瘦 瘦 J
75E9 762
皡 皡 T
76A1 76A5
眞 真 TJ
771E 771F
眾 衆 TJK
773E 8846
研 研 T
7814 784F
祿 禄 TJ
797F 7984
禿 禿 T
79BF 79C3
稅 稅 T
7A05 7A0E
穗 穗 TJ
7A42 7A57
箏 箏 GJ
7B5D 7B8F
箲 箲 T
7BB3 7C08


篡篡 T
7BE1 7C12
粵粵 T
7CA4 7CB5
絕絶 T
7D55 7D76
綠緑 T
7DA0 7DD1
緒緒 T
7DD2 7DD6
緣縁 T
7DE3 7E01
緼緼 T
7DFC 7E15
繈繦 T
7E48 7E66
羹羹 TJ
7FAE 7FB9
翱翱 T
7FF6 7FFA
胼胼 T
80FC 8141
脫脱 T
812B 8131
膃膃 T
817D 8183
舄舄 GT
8203 8204
舍舎 TJ
820D 820E
舖舗 J
8216 8217
莊荘 TJ
8358 838A
蓾蕾 TJ
83D1 8458
葢蓋 T
8480 8495
蔣蔣 GJ
848B 8523
蕿蔦 T
848D 853F
薀薀 T
8570 8580
薰薫 T
85AB 85B0
蘊蘊 T
85F4 860A
虛虚 T
865A 865B
蛻蛻 T
86FB 8715
衛衛 TJK
885B 885E
衰衰 TK
886E 889E
裝裝 GJK
88C5 88DD
訐訐 T
8A2E 8A7D
說説 T
8AAA 8AAC
諫諫 TJ
8ACC 8AEB
謠謡 J
8B20 8B21
豺豺 T
8C5C 8C63
走走 TJ
8D70 8D71
軿軿 T
8EFF 8F27
輜輜 J
8F1C 8F3A
輼輼 T
8F3C 8F40
达达 T
8FBE 8FD6
迸迸 TJ
8FF8 902C
遙遥 J
9059 9065
郎郎 T
90A2 90C9
鄉鄉鄉 T
90F7 9109 9115
醞醞 T
9196 919E
醬醤 J
91A4 91AC
鈃鈃 T
9203 9292
鋭鋭 T
92B3 92ED
錄録 T
9304 9332
鍊錬 TK
932C 934A
鎭鎮 TJ
93AD 93AE
閱閲 T
95B1 95B2
隉隉 G
9667 9689
靑青 T
9751 9752
靜靜 GTJ
9759 975C
靭靱 J
976D 9771
頹頽 T
9839 983D
顏顔 TJ
984F 9854
顯顕 J
985A 985B
飮飲 J
98EE 98F2
餅餠 TJ
9905 9920
駄駄 TJK
99B1 99C4
駢駢 TK
99E2 9A08
骯骯 T
9AA9 9AAB

Note:

a) Did not explain S1.4.3 on "Different structure of a corresponding component" because it is redundant. Any abstract shape which is different are considered not the *same* unless it is stated otherwise according to Unification Rule Appendix A

b) There should be more much more ambiguous cases of what is unified and what is not. We should go back Ext C to look for more of these examples. Source Unification (Appendix B) also have a lot of cases which we have not managed to capture yet.