Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по Стандартизации

Doc Type:	Working Group Document
Title:	Feedback on the encoding model of Han ligatures
Source:	Kushim JIANG (姜兆勤)
Status:	Individual Contribution
Action:	For consideration by China, SAT and IRG
Date:	2023–10–04

The issues related to the so–called "Han ligature" contains following materials:
和所谓 Han ligature 相关的议题包含如下材料：

- Online Review Tool, WS2021–01615 <hc.jsecs.org/irg/ws2021/app/?id=01615>.
- Online Review Tool, WS2021–03670 <hc.jsecs.org/irg/ws2021/app/?id=03670>.
- Eiso CHAN (2017). *Do we need a new block for ideographs now?* [Chinese]. Zhihu Zhuanlan.
- Eiso CHAN (2021). *Feedback on IRG N2492 and the preliminary encoding method of early Chinese organic chemical character, Sanskrit transcription, Tibetan transcription, Tangut transcription and Jianzi Musical Notation*. IRG N2492 Eiso Feedback = L2/21–165.
- Tianheng SHEN (2023). *About the Encoding Model on Han Ligatures*. IRG N2581 Tianheng Feedback.

We first identify a number of concepts that are related to the so–called "Han ligature", and then make recommendations for various situations and WS2021–01615 & WS2021–03670 situation.
为此，我们首先辨析和所谓 Han ligature 相关的若干概念，并给出建议，将这些工作运用于 WS2021–01615 和 WS2021–03670 的分析上。

1 Concepts related to "Han ligature"

Han Ligature. Ligature (Unicode glossary) (from Latin *ligātus*) is a writing and typographical concept, where the forms (of letters) are connected to each other by natural cursive penmanship, and the total width changes slightly. The product of the connection sometimes has exactly the same glottographic function as the original letters (stylistic ligatures); or sometimes has a completely different glottographic function (orthographic ligatures), thus becomes part of the orthography.
连字。连字是西文书写领域的概念，也是西文文字设计的概念，其中字形因自然连写结合在一起，且总字宽变化细微。其结合的产物，有时与原有的字母具有完全相同的对语功能（样式性连字）；有时则具有截然不同的对语功能（正字法性连字），从而成为正字法的一部分。本文的"连字"专门用来对译 ligature。

If one were to transpose this situation directly to Han, then the Han ligature would be the product of a fusion of strokes between the two characters (e.g., joining a stop terminal to another begin terminal, sharing stroke, etc., since cursiveness cannot be defined in Kaishu), and the total width (doublewidth) changes slightly. We couldn't find any such examples.
若要将这一情况直接照搬至广义汉字领域，则真正的 Han ligature 即是两广义汉字之间通过笔画融合的产物（如一收笔接入另一起笔、笔画共用等，因楷书中无法定义连写性），且总字宽（二字宽）变化细微。我们找不到这样的例子。

Han Conjunct. Conjunct form (Unicode glossary) (from Latin *conjūnctus*) is widely used in Unicode Standard to describe the form structure of Brahmic scripts, where the consonant forms are combined in various ways, and the total width changes significantly (roughly equivalent to the width of single form). The glottographic function of the products also vary, sometimes inheriting the original glottographic function, others not.
合字。合字是 Unicode 标准中常用于描述婆罗米诸文种的字形结构的概念，其中辅音字形通过多种方式结合在一起形成整体，且总字宽变化较大（仅大致相当于原来的单个字形的字宽）。其结合产物的对语功能也各异，有时继承原有的对语功能，有的则不然。

If one were to transpose this situation directly to Han, then the Han conjunct would be the product of a combination of several characters in various ways, and the total width changes significantly (fullwidth only). An example of this is the scheme to record Sanskrit and Tibetan in *Tongwen Yuntong* (**Example 1**, see **Figure 1**). Eiso *et al.* will give a more detailed description of their study later, so only a brief note will be given here that we see the same text structure in the character structure of this scheme as in the Brahmic scripts. Other examples may include 𢧐 to *geul* 글, 𢧑 to *dol* 돌, 𢧒 to *nol* 놀, 𢧓 to *seol* 설.
若要将这一情况直接照搬至广义汉字领域，则真正的 Han conjunct 即是广义汉字的字形通过多种方式结合在一起形成整体，且总字宽变化明显（仅为原来的单字字宽）。《同文韵统》使用汉字记录梵语和藏语的方案便是一个例子（例 1，参见 **Figure 1**）。Eiso 等人将在之后给出更加详细的研究介绍，因此这里只简要说明，我们在这份方案中看到了与婆罗米诸文种如出一辙的文本结构。“𢧐”“𢧑”“𢧒”“𢧓”可能也可以归入此类。

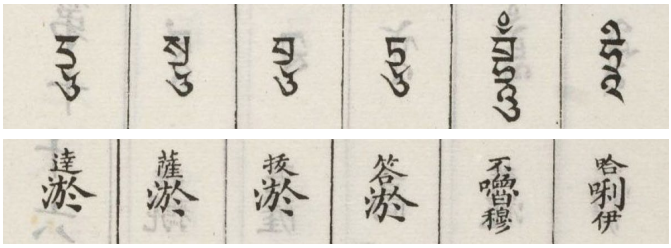


Figure 1 *Tongwen Yuntong*, volume 2, page 32
(the first row shows Sanskrit in Tibetan, the second row shows Han characters)

Han Unity (*Tuánjiézì*, see Figure 2). In folk life, multiple characters that refer to an auspicious word are combined into a single unity form. The characters that make up this form make a complex arrangement, and there may also be shared strokes. Functionally, Han unity completely inherits the function of the original characters, and this function is realized by interpreting the original characters of Han unity.
团结字（参见 **Figure 2**）。民俗生活中，表示吉祥含义的多字词融合成一个形，称为团结字。组成这个形的字之间形成复杂的排列结构，还可能存在共用笔画的情况。从功能上看，团结字完全继承原有汉字的功能，且这种功能是通过解读团结字的原有成分来实现的。

Figure 2 *Tuánjiézì*, from *Beijing Encyclopedia*, page 212
(黄金萬両 *huángjīn wànliǎng*, "ten thousand taels of gold"; 招財進寶 *zhāocái jìnbǎo*, "inviting wealth", 日進斗金 *rì jìn dǒujīn*, "make a lot of money in a day"; 大元寶 *dàyuánbǎo*, "big sycee")

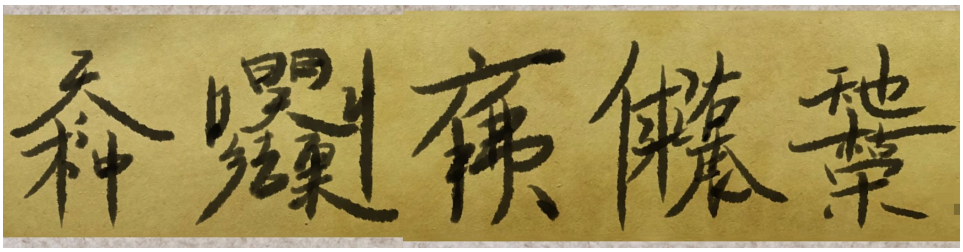


Figure 3 *Tuánjiézì*, from *Sìfēng Pàncí*
今天种 *jīntiān zhòng*, "sow today"; 叫明天结果 *jiào míngtiān jiéguǒ*, "let (it) bear fruit tomorrow", 夜拜佛 *yè bàifó*, "worship the Buddha in the night"; 求晨来保佑 *qiú chén lái bǎoyòu*, "let (it) bless (self) in the morning", 天地枯荣 *tiāndì kūróng*, "heaven and earth wither and flourish (again and again)"

Han Combination (*Héwén*). Han combination is a common phenomenon in old Hanzi and modern Han, in which multiple characters are combined into a single form. The characters that make up this form make a complex arrangement, and there may also be shared strokes. Functionally, Han combination completely inherits the function of the original characters.
合文。合文是古代汉字和现代汉字中的常见现象，其中多字词融合成一个形。组成这个形的字之间形成复杂的排列结构，还可能存在共用笔画的情况。从功能上看，合文完全继承原有汉字的功能。

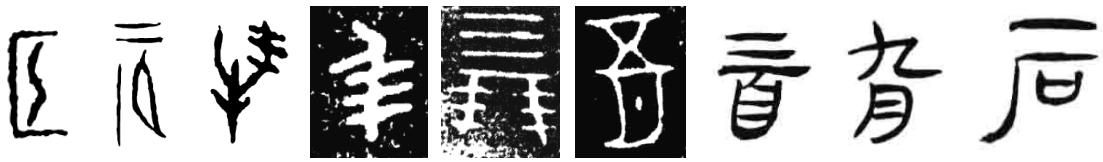


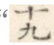
Figure 3 *Héwén* in Oracle–Bone script (three examples), Western Zhou Bronze script (three examples) and Zhangjiashan Han Bamboo slip script (three examples)
匚乙 *bàoyǐ* (person name), 十二月 *shíèryuè* "the twelfth month", 三牛 *sānniú* "three cattle"
三千 *sānqiān* "three thousand", 四朋 *sìpéng* "four péng (of bèi)", 五月 *wǔyuè* "the fifth month"
二百 èrbǎi "two hundred", 九月 jiǔyuè "the ninth month", 一石 yīshí (now yīdàn) "one picul"

Figure 3 shows ancient examples, and modern examples include *qiānwǎ* "kilowatt", *límǐ* "centimeter", *hǎilǐ* "mile", *jiālún* "gallon", *hétóng* "contract", *túshū* "book", *túshūguǎn* "library", *bówùguǎn* "museum". Note that in some examples, each character does not appear as a whole, but as a representative component of it. The examples in Japanese include *kabu* "singing and dancing", *miyage* "local specialty", *mane* "imitation". Other

examples may include 麿 to *maro* (麻 *ma* and 呂 *ro*) (used in person name), 𠮷 to *shaka* “Śākya, शाक्य” (尺 *sha* and 加 *ka*), 嚳 to *bitō* (毗 *bi* and 登 *tō*) (used in person name), because these components belong to the *Man’yōgana* system, these characters are also characterized by *Qièshēnzì*.
Figure 3 给出前现代汉字中的例子，现代的例子还包括“瓩”“糎”“浬”“嚣”，以及“啇”“圕”“圕”“𡈽”等。注意到，一些例子中每个字不以整体而出现，而是以其中的代表性构件而出现。日文中的例子包括“䯁”“壥”“𠍱”等。“麿”“𠮷”“嚳”也可能纳入这类字中，由于这些构件属于万叶假名系统，这些字也带有切身字的特征。

Syssemantograph. The form that uses syssemantic component(s) for its construction (cognitively) is called a syssemantograph, in which multiple characters are combined into a single form. For example, in *Shuowen*, 信 *xìn*, 圂 *hùn*, 昶 *chǎng* and 曇 *tán* are regarded as syssemantographs. Other examples include *wāi* “crooked”, *jiān* “pointed”, *lū* “coarse”, *zǎi* “child”. The pronunciation of the syssemantograph is usually completely unrelated to the pronunciation of its components. The examples in Japanese include *segare* “son”, *kogarashi* “autumn wind”, *nagi* “wind stops”, *tōge* “the highest point of the trail”, *shizuku* “drip”, *oroshi* “wind blowing from above”.
会意字。使用会意作为构字理据的字形称为会意字。如《说文解字》认为“信”“从人从言，会意”，“圂”“象豕在囗中，会意”，“昶”“从日、永，会意”，“曇”“从日、雲，会意”等等。其他例子还有“歪”“尖”“𠅤”“𡥧”等。会意字的读音通常与构件完全无关。日文中的例子包括“躮”“凩”“凪”“峠”“雫”“颪”等。

***Qièshēnzì*.** The form that records syllables in other languages by indicating the composition of the parts of the syllable are called *Qièshēnzì*. Eiso (2022, IRG N2578) has introduced this kind of character before. For example, 𨘋 is used to record *dhya* ध्य, 𩋎 to *maṃ* मं, 𡲼 to *viṃ* विं, 𦚰 to *vyai* व्यै, 𠐬 to *gi* ги, 𤁕 to *rin* рин, 聕 to *he* хе, 𪍔 to *vin* вин.
切身字。通过指示音节各部分的构成，来记录其他的语言中的音节的字称为切身字。Eiso 曾在 IRG N2578 第 3 页起介绍过这种字。如“𨘋”“𩋎”“𡲼”“𦚰”以及“𠐬”“𤁕”“聕”“𪍔”。

Kerned Han. The form by forcing multiple characters into a single body is called a kerned form. This operation is not set at the level of characters and forms, but is created at the level of writing or typesetting. For example, WS2021–03670  ,  (in evidence of WS2021–00593),  (in evidence of IRG N2645, page 141) is created in writing or woodcutting process.
紧排字。将多个汉字置入一个方形字身框中形成的字形称为紧排字。这个操作不是在字符和字形的层面上被设定的，而是在书写或排版的层面上被创造出来的。如“  ”“  ”“  ”都是在书写或雕刻的过程中形成的。

2 Suggestion

In determining encoding decisions for these glyphs, two key points should be captured.
在为这些字形确定编码决策时，应当把握两个要点。

One is the totality or internal separateness of the form. The totality indicates that the characters are fused together and the boundaries are so blurred that it is impossible to say which part of the form belongs to which character. Thus, in grasping the variation of the form, it is only possible to establish variation in terms of the totality or possible components, and it is not possible to say that the part of the form belonging to a particular character undergoes variation. Separateness, on the other hand, suggests that the individual characters are clearly defined, with a clear boundary of labor, and that each influence only a limited part of the overall form without interfering with the functioning of the other parts.

其一是字形的总体性与内部分立性。总体性说明字形融合在一起，边界模糊不清，无法言明字形的哪一部分属于哪个字。因此在把握这些字形的异写性时，只能以总体或可能的构件为对象建立异写性，而无法说其中属于某个字的部分发生变异。分立性则说明各个字界限清晰，分工明确，各自仅影响总体的有限部分，而不介入其他部分的功能。

The other is the related attribute. When a form is represented as a sequence of multiple characters, the existing Unihan Database is no longer applicable for it. Additional database is required to record information on pronunciation, meaning, usage, source, radical, stroke, etc.
其二是与字形相关的属性。当字形被表记为字符序列时，不再适用已有的 Unihan 数据库对其的属性刻画。须要建立额外的数据库记录可能的音义功能、来源、部首笔画等信息。

Therefore, for the forms in the *Tongwen Yuntong*, we tend to build additional encoding, because the characters are clearly discrete from each other and their structures correspond strictly to their glottographic function.
因此，对于《同文韵统》中的字形，我们倾向于建立额外的字符集模型进行刻画。因为字形之间具有明显的分立性，且其结构与对语功能严格对应。

We tend to encode Han unity as a whole, because they obviously have stylistic totality, that is, variation cannot be built into the original characters that compose them.
我们倾向于单独编码团结字，因为它们在造型上具有明显的总体性，即异写性无法建立在组成它们的原有汉字中。

We tend to encode syssemantograph as a whole, because they obviously have functional totality, that is, their pronunciation (and sometimes meaning) varies considerably from their origin.
我们倾向于单独编码会意字，因为它们在功能上具有明显的总体性，即其读音（有时含义）与原本的组成部分差异较大。

We tend to strictly reject encoding kerned Han (related to the two characters), because they do not operate or build at the level of the character set. They should be implemented in typefaces and typographic applications. If these effects are to be stored, additional structure above the character level should be built.
我们倾向于严格拒绝编码紧排字（涉及对 WS2021–01615 和 WS2021–03670 的决策），因为它们不在字符集的层面上运作和建立。应当通过字体和排版应用的实现呈现这些效果。若要存储这些效果，则应在字符层以上建立额外结构。

There is some controversy over the treatment of Han Combination and *Qièshēnzì*, both of which are both total and separate. The former is more total than discrete, so we tend to encode them as a whole; the latter is more discrete than total, so we tend to build encoding model on them. However, since different notation systems will design different structures, there will be difficulties in establishing a universal encoding model on *Qièshēnzì*. Shen's model has this universality, but this leads to a conflict in explaining the principle that IDS may not be used for mandatory shaping.
对于合文与切身字的处理有一些争议，二者兼具总体性与分立性。前者总体性强于分立性，所以我们倾向于整体编码；后者分立性强于总体性，所以我们倾向于建立字符层结构。但由于不同的记音系统会设计不同的结构，所以在建立统一的编码模型上会存在困难。沈的模型具有这种统一性，但这会导致在解释 IDS 不得用于强制性文本成形的原则上存在冲突。

(End of Document)