L2/24-126

Source:	CheonHyeong Sim (沈天珩)
Title:	Comments on L2/24−125
Date:	2024−04−25
Action:	To be considered by UTC and CJK/Unihan Working Group

First of all, I would like to say that I support to add such a new block and encode some so−called ゲバ文字. The necessity to digitize them does exist. But some of the characters in L2/24−125 may be a little bit questionable. I would like to devide my comments into five sections:

- Single−source characters
- Cursive forms
- Similar shapes with Han Ideographs
- Whether to use combining marks
- About the script properties

Single−source characters

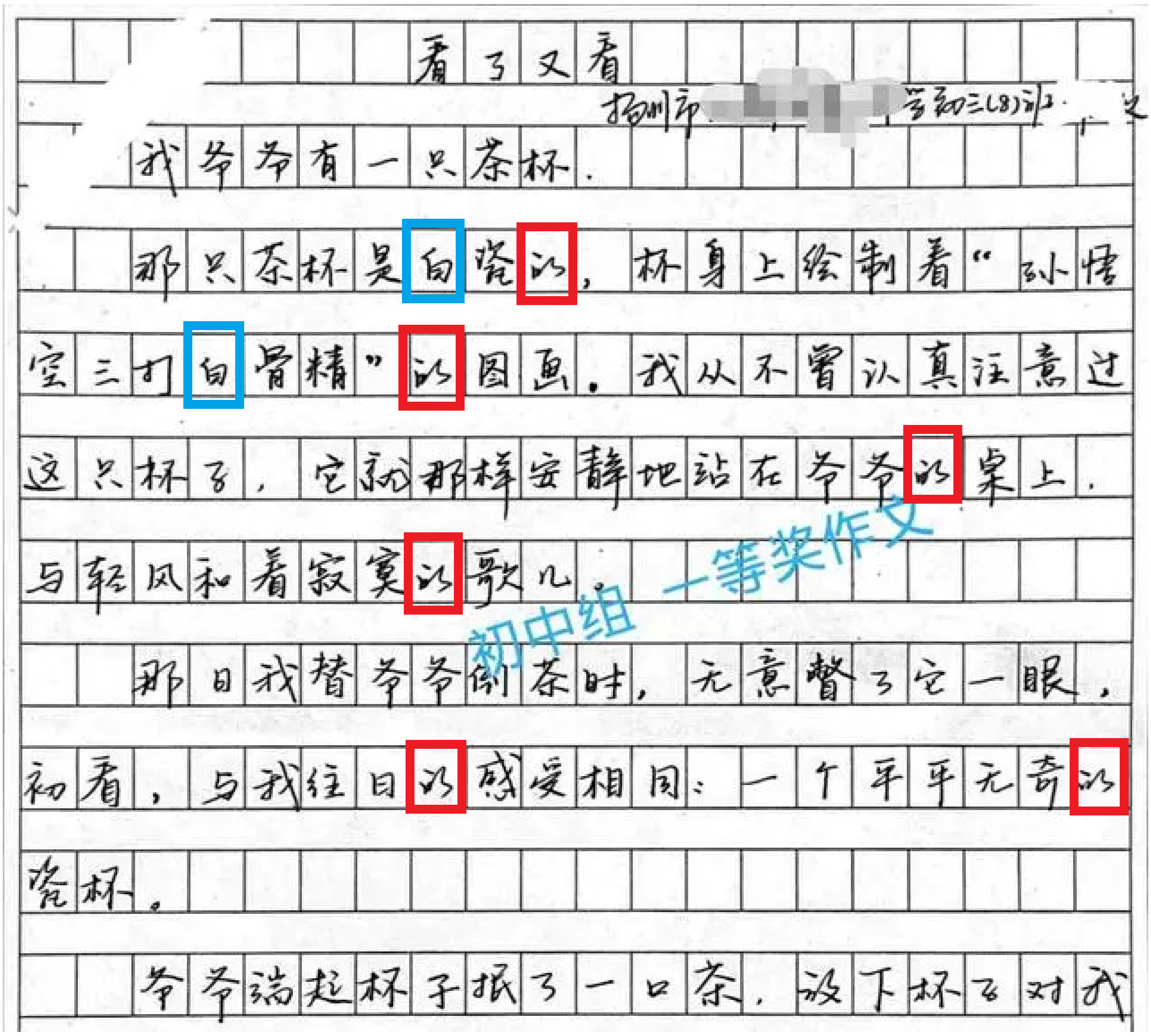
29 out of 51 characters (56.86%) in that proposal have a gray background, which means they could only be finded in a single source. These characters are the most questionable ones. Everyone owns her/his right of publication, and if someone selfly created a ゲバ文字 in her/his published book, will we really accept it to be included in Unicode?

Of course, I am not doubting that those characters in the figures were selfly created by the writers, but some extra evidences may be needed to prove that they are indeed being used or have indeed been used in practice. For example, Fig.11 shows lots of characters which are considered to be 一九六五～一九七五年度頃の略字 (abbreviated characters between 1965 and 1975), then we may need some materials during those ten years, either printed or hand−written, as a circumstantial evidence, otherwise we may probably not accept them to be encoded.

Cursive forms

Generally, Unicode encodes **characters** but not **glyphs**. Although Hiragana and Katakana are also derived from Han Ideographs, they are considered as different scripts from Han script as a common understanding. However, U+1AF90..U+1AF92 (perspectively the cursive form of 御, 前 and 揃) in that proposal do not look like a new script but only the cursive form of Han script. From the figures we could know that they are only used in hand−written texts but no printed text uses such forms. Moreover, U+1AF90 is marked as "some variants exist" which shows that even the glyph is not stable. No matter how common the usage is, whether to encode such a hand−written only stylistic variant may need further investigation.

In addition, many people also write the character 的 in a cursive form in Chinese texts, because it has too many strokes as an extremely commonly used character. I have seen many such examples in my daily life, but I do not deliberately take photos, so I found some examples on the Internet instead, as shown below:



看了又看

扬州市[redacted]学初三(8)班 文

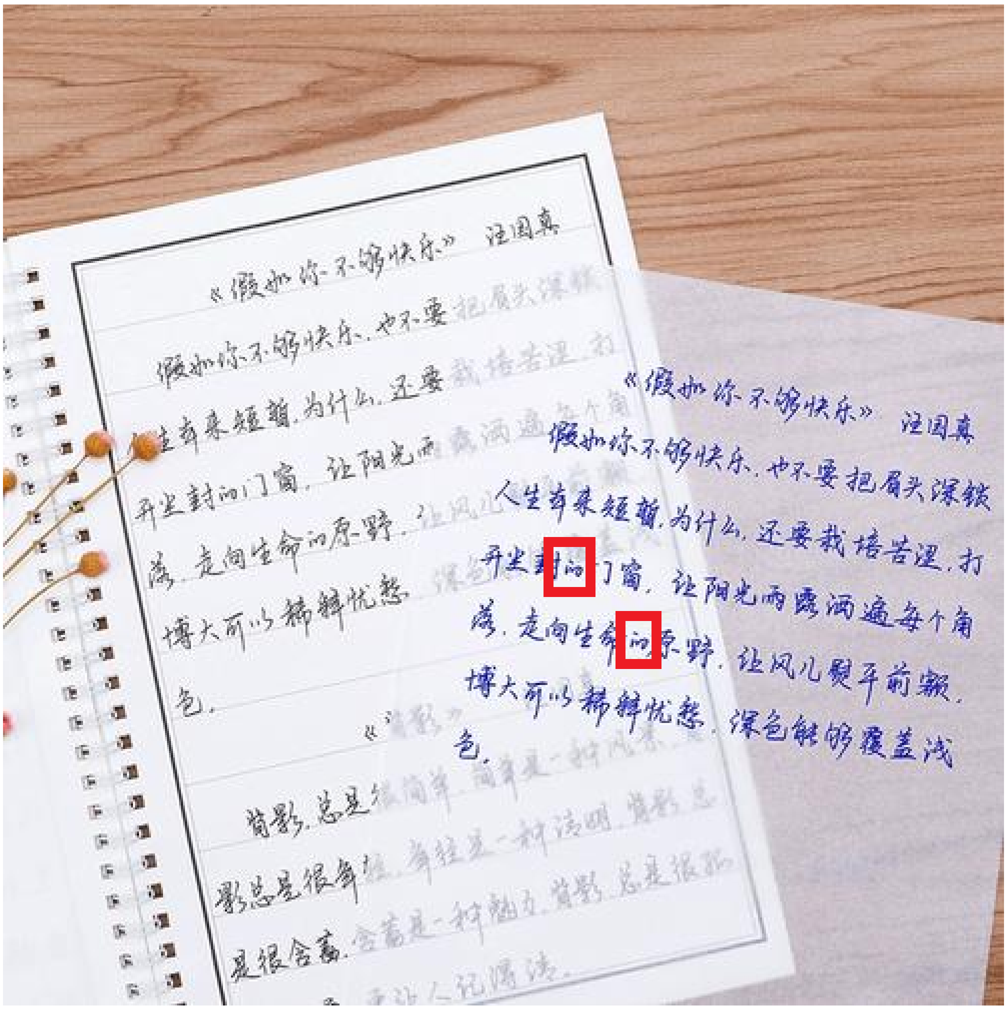
我爷爷有一只茶杯。

那只茶杯是白瓷的，杯身上绘制着“孙悟空三打白骨精”的图画。我从不曾认真注意过这只杯子，它就那样安静地站在爷爷的桌上，与轻风和着寂寞的歌儿。

那日我替爷爷倒茶时，无意瞥了它一眼，初看，与我往日的感受相同：一个平平无奇的瓷杯。

爷爷端起杯子抿了一口茶，放下杯子对我

Fig.1 https://baijiahao.baidu.com/s?id=1671455430525460287



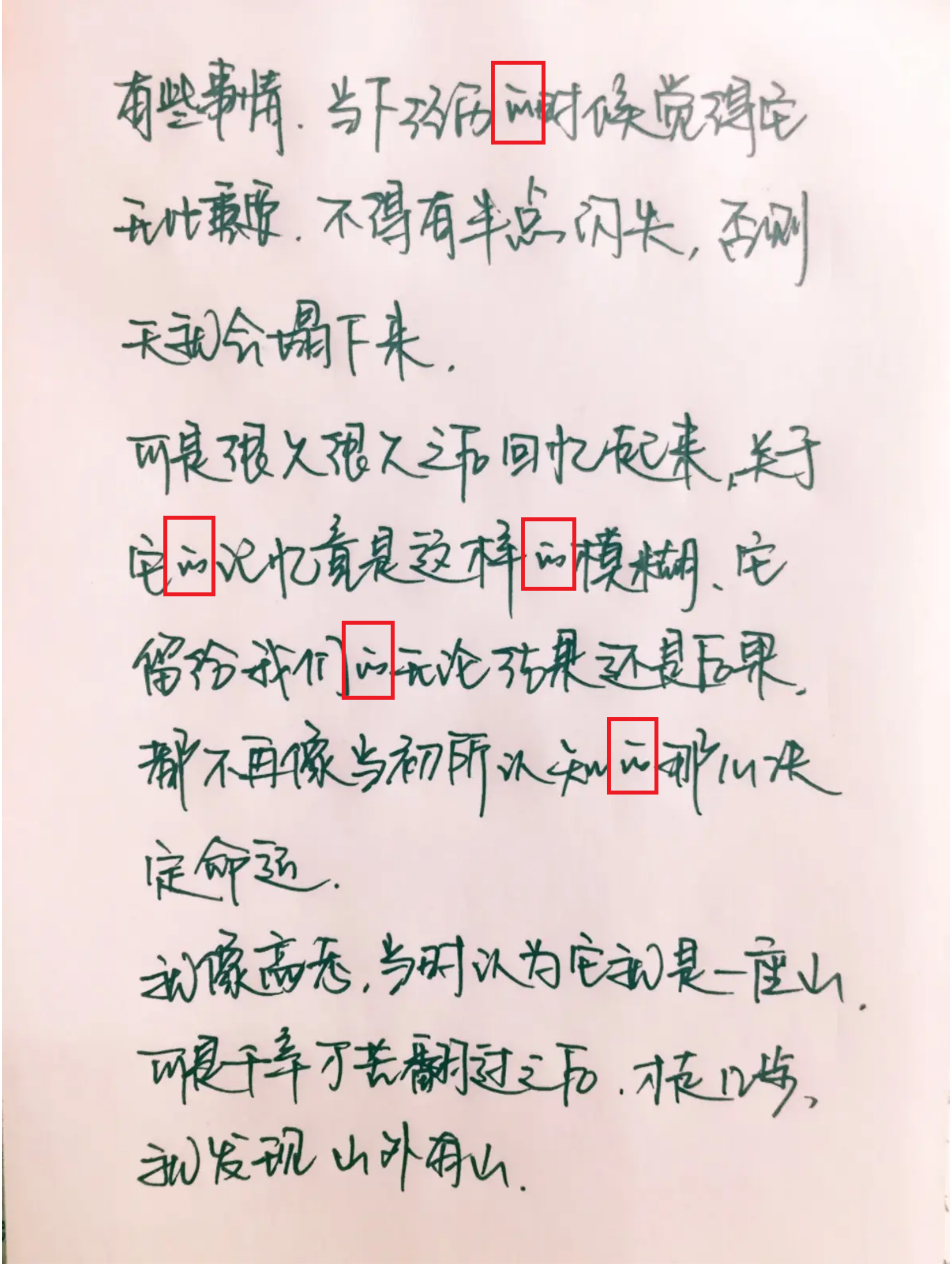
《假如你不够快乐》 汪国真

假如你不够快乐，也不要把眉头深锁
人生本来短暂，为什么，还要栽培苦涩，打
开尘封的门窗，让阳光雨露洒遍每个角
落，走向生命的原野，让风儿熨平前额，
博大可以稀释忧愁，深色能够覆盖浅
色。

《背影》

背影，总是很简单，简单是一种风景，背
影总是很年轻，年轻是一种清明，背影总
是很含蓄，含蓄是一种魅力，背影，总是很孤
更让人记得清

Fig.2 https://image.baidu.com/search/detail?word=...

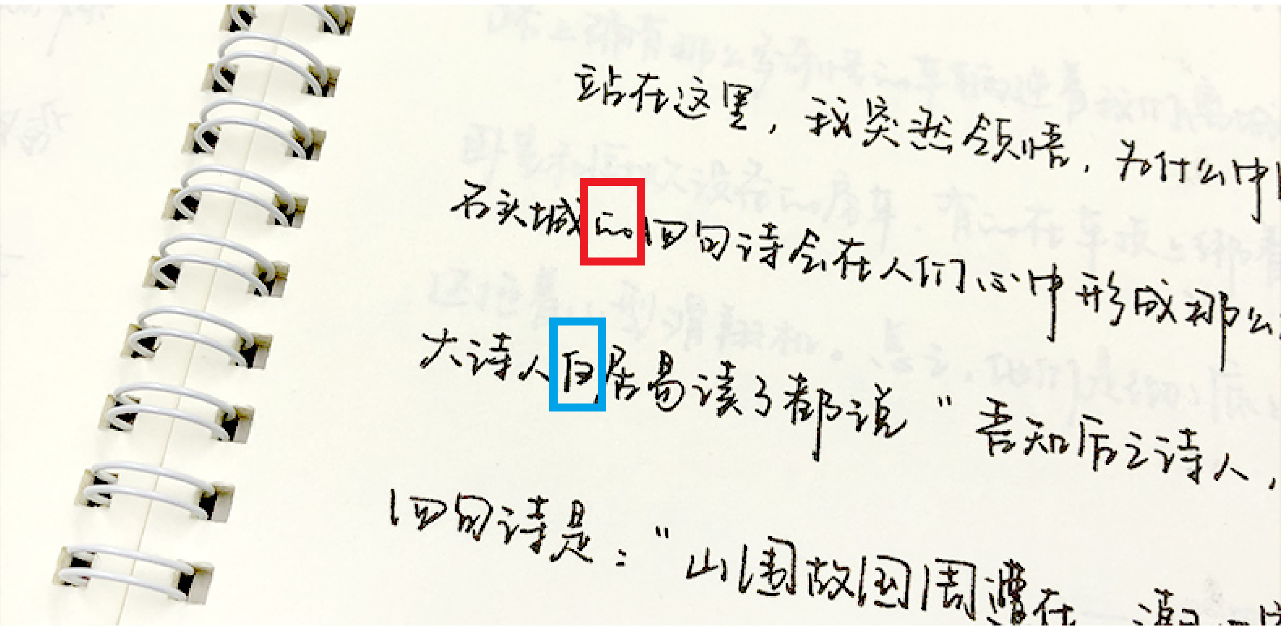


有些事情。当下经历的时候觉得它
无比重要。不得有半点闪失，否则
天就会塌下来。

可是很久很久之后回忆起来，关于
它的记忆竟是这样的模糊、它
留给我们的无论结果还是后果，
都不再像当初所认知的那么决
定命运。

就像高考，当时认为它就是一座山。
可是千辛万苦翻过之后。才走几步，
我发现山外有山。

Fig.3 https://image.baidu.com/search/detail?word=...



站在这里，我突然领悟，为什么中

石头城的四句诗会在人们心中形成那么

大诗人白居易读了都说"吾知后之诗人，

四句诗是："山围故国周遭在，潮

Fig.4 https://image.baidu.com/search/detail?word=...

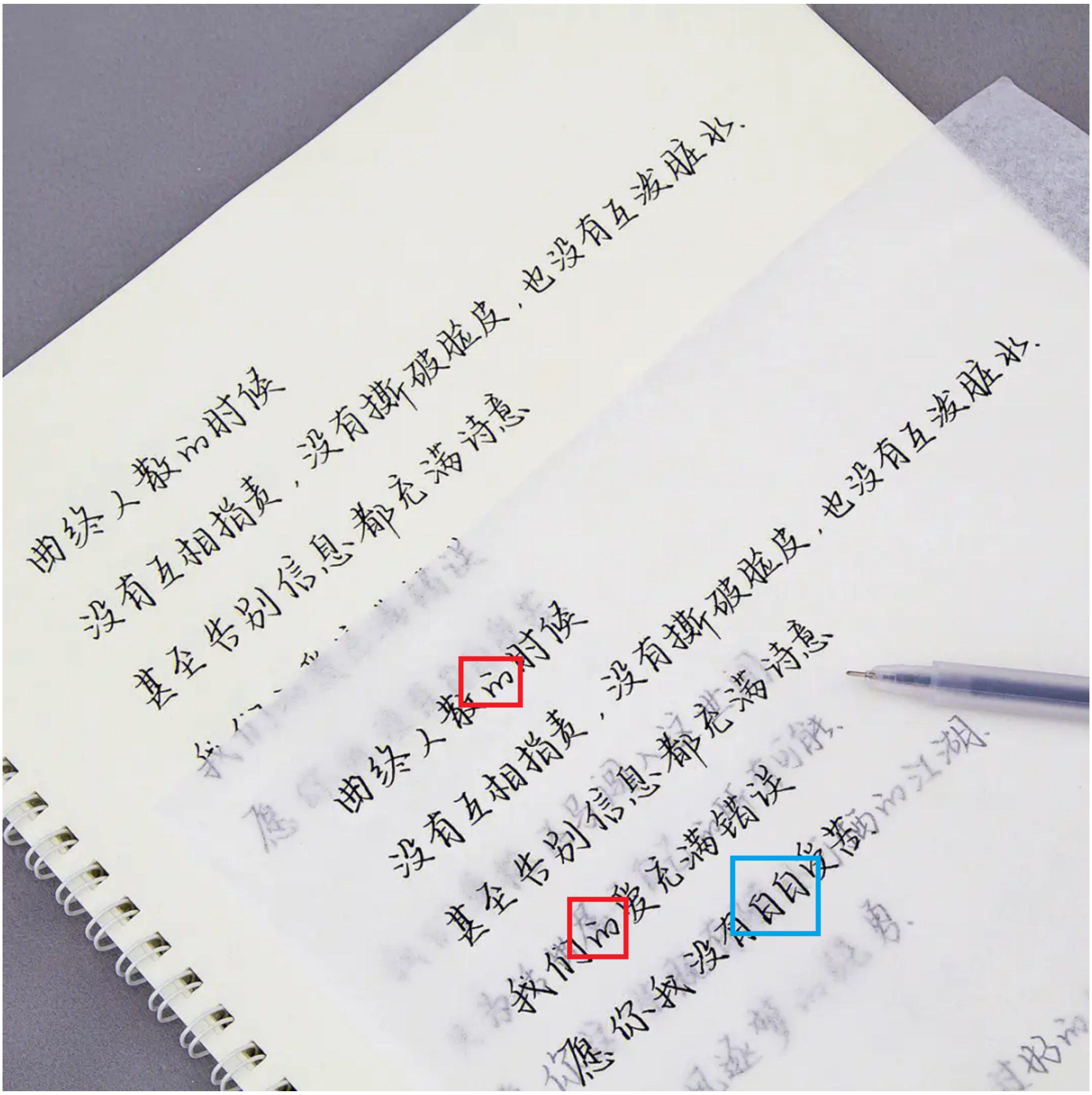


Fig.5 https://image.baidu.com/search/detail?word=...

Even if some of the strokes are joined–up, except for 的, almost all the characters have a relatively clear structure, especially in Fig.5; 的 is so cursive that it becomes only two strokes. In comparison, the left part of 的 (i.e. 白) also appears in Fig.1, Fig.4 and Fig.5, you can see the obvious difference. However, nobody thinks that such a cursive structure is “another character” and should be encoded separately.

Similar shapes with Han Ideographs

U+1AFB5（⿺辶チ）and U+1AFB6（⿺辶卜）in that proposal have the similar shape with U+8FC1（迁）and ⿺辶卜（Currently unencoded, both used in ancient China as a

Liding character and in Japan as a Kokuji). This may confuse the users. Currently, U+8FC1 is supported on all the devices with the default fonts, in comparison, ⿺辶チ may be supported several years after being included in Unicode, if the users do exist, they may probably prefer using U+8FC1 than the newly encoded non-BMP character for years. Besides, people may use the similar shape to deceive people, which is similar to something like “unicode.org” (two cyrillic letter o are used here). This was a true occurrence that a trending topic on Weibo used U+2F0A (⼊) instead of U+5165 (入) to make the search engines unable to match the normally input sequences. Maybe we need a further discussion on whether and how to unify these kind of characters to Han Ideographs.

My personal opinion on this question is that, to encode Han-Latin Ideographs and Han-Hiragana Ideographs to the new block, but to encode Han-Katakana Ideographs and Han-Hangul Ideographs to CJKUI. Different from Hiragana which is derived from the cursive style (草書) of Han Ideographs, Katakana is derived from the regular style (楷書) of Han Ideographs. The strokes of the Katakana characters are very close to Han Ideographs; and we have already encoded so many Han-Katakana Ideographs in CJKUI blocks. Same for the Han-Hangul Ideographs that we have already encoded so many Han-Hangul Ideographs in CJKUI blocks, and note that, a T-source ideograph U+20B9D (⿰口丨) which is currently not listed in UTN#43 also comes from Hangul (the whole syllable 미) according to its pronunciation and the usage as a person’s name.

Whether to use combining marks

The only difference between U+1AFAC (⿰言コ) and U+1AFAE (⿰言ゴ) in that proposal is the kana voicing mark (◌゙, U+3099). Do we have to consider to encode ⿰言ゴ as a sequence (i.e. ⿰言コ with the kana voicing mark)?

Firstly, based on the discussion history of U+318D7 (𱣗), we could see that the original evidence contains a combining mark on the top-right corner of the ideograph. That combining mark is a stylistic variant of U+16FF1 (◌𖿱). At the beginning, the glyph in the draft codecharts contained that mark, but it was removed after the discussion – in conclusion, it was encoded as a sequence U+318D7 U+16FF1.

Secondly, the kana voicing mark is not solely used with Hiragana and Katakana. We can see the usage with Bopomofo (to indicate some voiced initials in some dialects), as well as the usage with Han Ideographs:

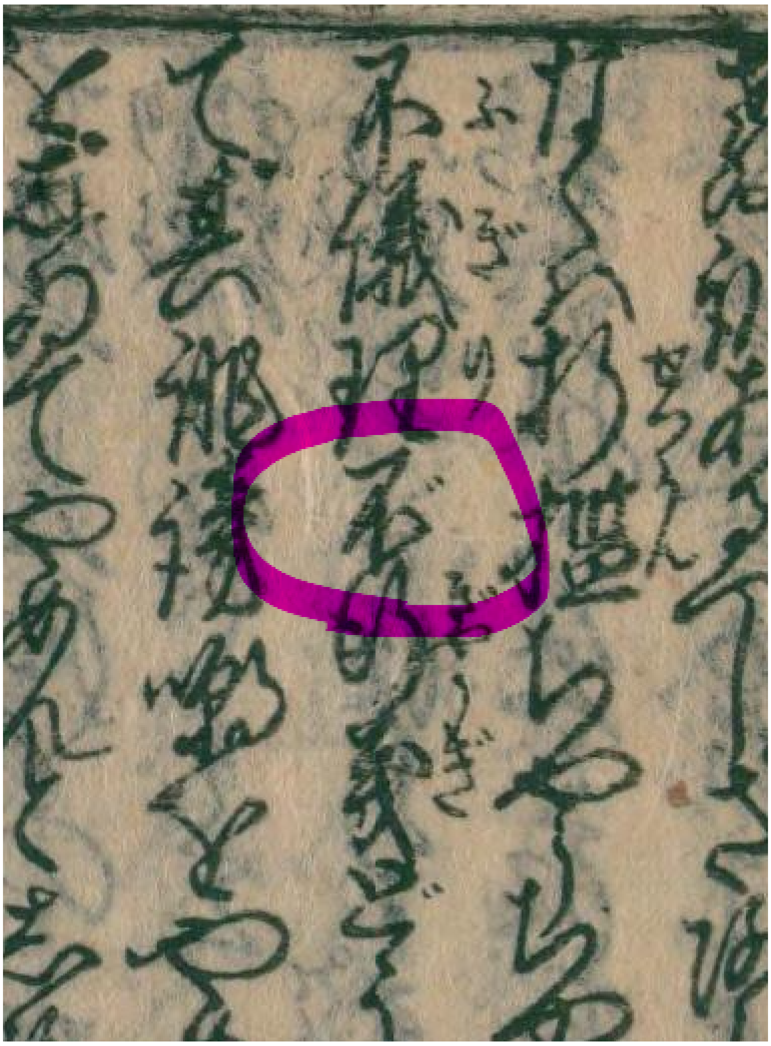


Fig.6 仮名草子『可笑記』

In Fig.6, the kana voicing mark is used with the Han Ideograph 不 to indicate that its pronunciation should be ぶ instead of a normal ふ. The context is 不儀理不行義, so it must be a Han Ideograph 不 but not a Hiragana ふ.

From these two examples, we may consider to encode the characters without the kana voicing mark. Also for U+1AFA8 (⿰言ギ), U+1AFB8 (⿱ド寸) and U+1AFB9 (⿱⺍ド).

About the script properties

For such kind of Hybrid−script Ideographs, we may need to discuss about the script properties in Unicode if to be encoded. My suggestion is that, all of them should have Script=“Common”, and Script_Extension=“Hani Latn” or “Hani Hira” or “Hani Kana” or “Hang Hani”.

(End of document)