Title: Proposal to Encode 9 Additional Kanbun Marks

Author: Wang Yifan

Date: 2020-09-17 (Updated: 2020-09-18)

1. Summary

Since when sixteen Kanbun marks were included in Unicode at version 1.0, no addition has ever been made. We have found nine more marks in actual use to be encoded, among which seven are natural expansion of two existing series, and two are novel addition.

2. Background

The method of 漢文訓読 *kanbun kundoku*, practiced over a millennium in Japan to read Classical Chinese text using the Japanese grammar, features various annotation marks (返り点 *kaeriten*) which are employed to manipulate the reading order of characters into comprehensible Japanese word order.

The origin of *kaeriten* is thought to be old proofreading markups, which has undergone gradual sophistication into a system akin to that in modern days by the end of the 14th century (Kobayashi, 1974). There were still varying details in the convention among schools, until it was finally standardized when a formal report was published in 1912 (Ministry of Education, 1912), which has been deemed the source of norms up to date.

Kaeritens function as step-by-step instruction to signify the reading order of a specific character, put beside every character that deviates from the written order. The reader is expected to (1) read every unmarked characters in order until seeing a character with any starting sign, (2) jump from a character with a sign of the same series to another following the priority, and (3) if no character which is unmarked or with a same- or lower-leveled mark is left unread prior to where the starting sign is, continue to read from the next eligible character. Multiple series of signs are provided to enable nested jumping, during which you will completely fly over those from inner tiers. Each series contains highly uneven number of items due to unsystematized conventional origin, which is shown below:

Level	Maximal Repertoire and Priority
0	レ (for single swap)
1	一→二→三→四→五→六→七→八→九(→十?) [Chinese numerals]
2	上→中→下 [top, middle, and bottom]
3	甲→乙→丙→丁→戊→己→庚→辛→壬→癸 [the heavenly stems]
4	天→地→人 [heaven, earth, and human]
(higher)	元→亨→利→貞 [an Yijing set phrase] 乾→坤 [an Yijing set phrase] 春→夏→秋→冬 [the four seasons] etc.

In the table above, items in **boldface sans-serif** are encoded in the Kanbun block as of Unicode 13.0, and **boldface serif** are manifested in the 1912 report (Fig. 4), including the nesting hierarchy. The rest are mostly theoretical and hypothetical entries; those of undetermined levels are especially obscure, seeming to have little modern real-world usage despite literatures that cite their existence.

3. Newly Attested Marks

3.1. Identities and sources

We have totally found nine new marks:

- Level 1 addition
  - **五・六・七・八・九** (fifth through ninth)
- Level 3 addition
  - **戊・己** (fifth and sixth)
- Of undetermined level
  - **乾・坤** (first and last; all for this tier)

Appearing in two documents:

a) *Taishō Tripiṭaka* 大正新脩大藏經 (vol. 12, p. 803; vol. 9, p. 39) [Fig. 1, 2]

b) *Book of Jin* 晉書, annotated by 志村楨幹 (fasc. 40, folio 3 recto) [Fig. 3]

Source a) is a letterpress printing book series sequentially published during 1924–1934 in Tokyo, and b) is a wood engraving book series published in 1702 in Kyoto, with image taken from a facsimile reprint in 1971.

3.2. Detailed Information

Fully (and redundantly) annotated reproduction of each passage that contains new elements in respective figures is given to the right in vertical writing.

(1) It is intended to be read in Japanese "當に一心に是の涅槃經を受持・讀誦・書寫・解說・供養・恭敬・尊重・讚歎すべし (*You must wholeheartedly remember, recite, copy, preach, devote to, revere, and praise the (Mahāpari)nirvāṇa Sūtra*)". *Kaeritens* in **black** are the only option according to the standard method. The cause of extended *kaeritens* is the consecutively coordinated verbs that share their object. In principle, *kaeriten* is supposed to progress backward unnesting layers of the VO/VC structure, but when multiple coordinated verbs shares an object, each verb has to be marked individually in sequence, since their object is already read out of order (Kotajima and Yuki, 2011; p. 53). Such

(3) (2) (1)

或人有坤背レ充ニ以要スルニ權一貴ヲ者乾

若聽己我一等於ニ佛一滅一後一在ニ此娑一婆世一界一勤加ニ精一進一護乙持・讀丙誦・書丁寫・供戊養是經一典甲者。

應下當一心一受ニ持・讀三誦・書四寫・解五說・供六養・恭七敬・尊八重・讚九歎是涅一槃一經一。

long enumeration is arguably rare in classics (which were willingly studied and annotated) where the rhetorical beauty is in consideration, but of course not uncommon in legal and practical documents.

The *kaeriten* on the top in **red** is wrong, presumably a confusion in typesetting, as the next line contains an identical markup pattern (which is correct; see Fig. 1). Were there not for it, it possibly had the tenth mark of this series as well.

(2) Read "若し我等の仏滅後に於て此の娑婆世界に在りて勤めて精進を加へ是の經典を護持・讀誦・書寫・供養するを聞かば (*If you hear us earnestly pursuing the faith, and maintaining, reciting, copying, and devoting to this scripture at this Sahā world after Shakyamuni's death…*)". The marking is the only valid way to annotate it, with the situation alike to the previous. The level 3 marks directly wrap the level 1, because the level 2 has an insufficient number of marks to cover.

(3) Read "或るひと充に背きて權貴を要する者あり (*When there was one who turned his back on Chong seeking advancement…*)". The 乾 and 坤 marks appear as level 2 marking as shown in Section 2. Prior to the 1912 report, there was no agreement in mark hierarchy. The author might have chosen them for the sake of clarity that the tier has no more than two elements. *Okurigana* (phonetic aid) in orange shows minor nonconformities to current practice, which are irrelevant to this problem.

4. Proposal

4.1. Names and Arrangement

We propose to add the said nine characters into the Unicode Standard. Our suggested arrangement in a 16-code column is as below:

Code Point	Glyph	Name
U+XXXX0	五	IDEOGRAPHIC ANNOTATION FIVE MARK
U+XXXX1	六	IDEOGRAPHIC ANNOTATION SIX MARK
U+XXXX2	七	IDEOGRAPHIC ANNOTATION SEVEN MARK
U+XXXX3	八	IDEOGRAPHIC ANNOTATION EIGHT MARK
U+XXXX4	九	IDEOGRAPHIC ANNOTATION NINE MARK
U+XXXX5		(reserved; *for the possible TEN MARK to come*)

U+XXXX6	乾	IDEOGRAPHIC ANNOTATION CREATIVE HEAVEN MARK
U+XXXX7	坤	IDEOGRAPHIC ANNOTATION RECEPTIVE EARTH MARK
U+XXXX8	戊	IDEOGRAPHIC ANNOTATION FIFTH MARK
U+XXXX9	己	IDEOGRAPHIC ANNOTATION SIXTH MARK
U+XXXXA		(reserved; *for the possible SEVENTH MARK to come*)
U+XXXXB		(reserved; *for the possible EIGHTH MARK to come*)
U+XXXXC		(reserved; *for the possible NINTH MARK to come*)
U+XXXXD		(reserved; *for the possible TENTH MARK to come*)
U+XXXXE		(reserved)
U+XXXXF		(reserved)

Reserved code points are left for possible future addition to each series of marks, probably up to ten for each, but 乾 and 坤 are a closed set on their own that has no expansion expected. Their naming practice follows the existing ones in the Kanbun block, with 乾 and 坤 kept in harmony with Yijing Hexagram Symbols counterparts.

If possible, opening up room to accommodate 32 code points in total would be desirable if it consists of a standalone block, as there is a considerable possibility that other marks such as those listed in Section 2 will join.

4.2. Rationale for Inclusion

Their usefulness is as much as existing marks in the Kanbun block. As there is no *de jure* standard governs them, it is not immediately clear why the current repertoire is extracted. For the level 1 and 3 characters, the 1912 government report contains up to 五 and 戊 respectively, with a commercial font support the same range[1]. *Taishō Tripitaka* embraces tens of occurrences up to 七 and 戊 for each besides aforementioned examples, which is among issues that prevent completion of the text encoding in SAT Text Database (https://21dzk.l.u-tokyo.ac.jp/SAT/index_en.html).

[1] https://www.iwatafont.co.jp/font/kanbunsp.html

4.3. Properties

Suggested properties for U+XXXX0 to U+XXXX4:

```
<XXXXY>;IDEOGRAPHIC ANNOTATION <name> MARK;No;0;L;;;;<n>;N;;;;;
East_Asian_Width=W
Script=Common
Script_Extensions=Hani
```

Suggested properties for U+XXXX6 to U+XXXX9:

```
<XXXXY>;IDEOGRAPHIC ANNOTATION <name> MARK;So;0;L;;;;;N;;;;;
East_Asian_Width=W
Script=Common
Script_Extensions=Hani
```

Whereas existing ideograph-like symbols in the Kanbun block are assigned compatibility decomposition mapping for historical reasons, the designation is hardly justifiable either in semantics or in typography, inasmuch as virtually considered deprecated, as described in the Unicode 13.0 standard (The Unicode Consortium, 2020). New characters need not adhere to the convention.

Acknowledgements

Our gratitude to Dr. Kiyonori Nagasaki for providing the valuable text data, @JUMANJIKYO and @KAN0U for the inspiring suggestion and reference, Dr. Ken Lunde, Eiso Chan, Alexander Zapryagaev, and Unihan contributors for the kindest advice and support in content and format.

References

Fang, Xuanling 房玄齡 et al. (eds.) (1971[1702]). *Book of Jin* 晉書. 3 volumes. In: *Wakokubon Seishi* 和刻本正史 series. Tokyo: Kyūko Shoin.

Kobayashi, Yoshinori 小林芳規 (1974). "Kaeriten no enkaku" 返點の沿革. *Diacritical language and diacritical materials* 訓点語と訓点資料 54: 86–111. http://ir.lib.hiroshima-u.ac.jp/00025038.

Kotajima, Yosuke 古田島洋介 and Yoshinobu Yuki 湯城吉信 (2011). *Kanbun kundoku nyūmon* 漢文訓読入門. Tokyo: Meiji Shoin.

Ministry of Education, Japan 文部省 (1912). "Kanbun kyōju ni kansuru chōsa hōkoku" 漢文教授ニ關スル調査報告. *Official Gazette* 官報 (Meiji) 8630: 703–706. https://dl.ndl.go.jp/info:ndljp/pid/2951987.

Takakusu, Junjirō 高楠順次郎 (ed.) (1925). *Taishō Tripiṭaka* 大正新脩大藏經 (vols. 9, 12). Tokyo: Taisho Issaikyo Kankokai 大正一切經刊行會.

Unicode Consortium, the (2020). *The Unicode Standard, Version 13.0.0*. Mountain View: The Unicode Consortium. http://www.unicode.org/versions/Unicode13.0.0/.

Fig. 1 Example with 五 to 九

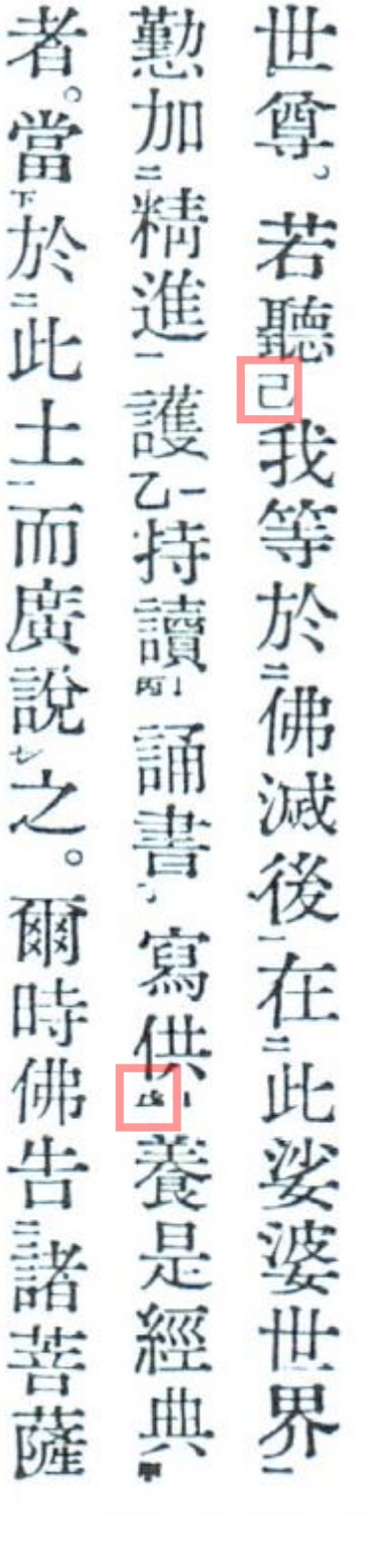


Fig. 2 Example with 戊 and 己

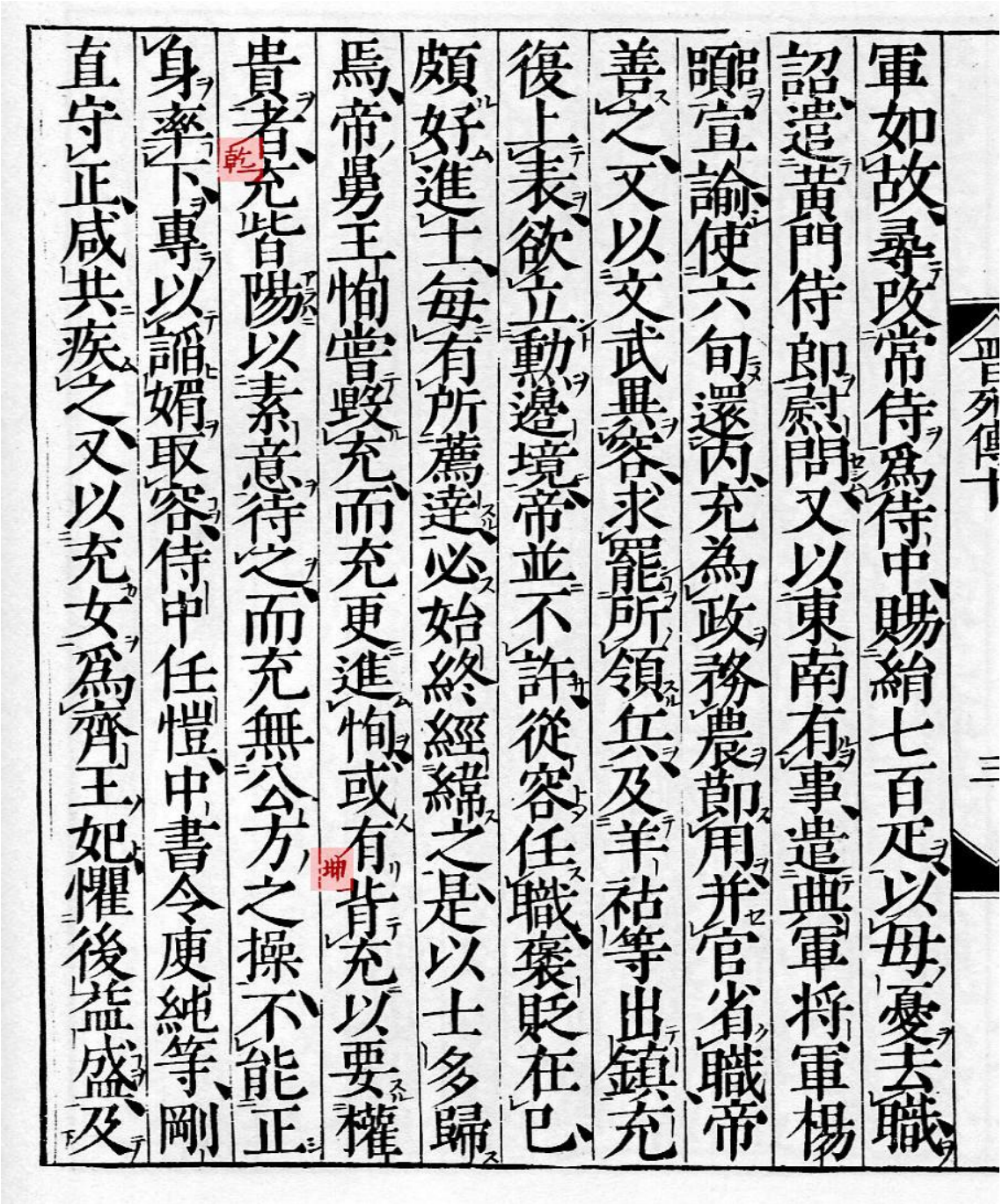
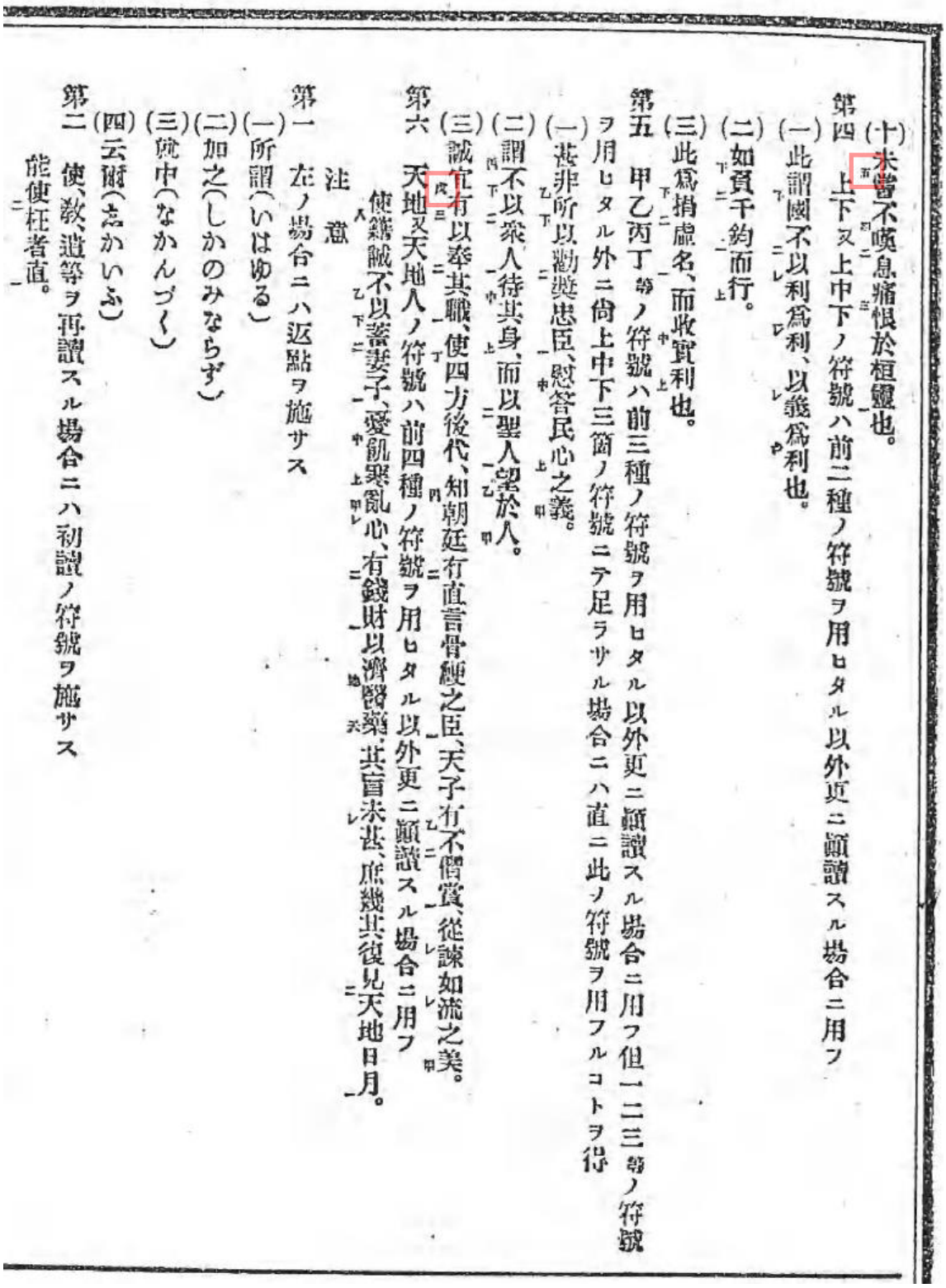


Fig. 3 Example with 乾 and 坤 (Fang et al., 1971)


(十)未嘗不嘆息痛恨於桓靈也。

第四　上下又上中下ノ符號ハ前二種ノ符號ヲ用ヒタル以外更ニ顚讀スル場合ニ用フ

(一)此謂國不以利爲利、以義爲利也。

(二)如貢千鈞而行。

(三)此爲捐虛名、而收實利也。

第五　甲乙丙丁等ノ符號ハ前三種ノ符號ヲ用ヒタル以外更ニ顚讀スル場合ニ用フ但一二三等ノ符號ヲ用ヒタル外ニ尙上中下三箇ノ符號ニテ足ラサル場合ニハ直ニ此ノ符號ヲ用フルコトヲ得

(一)甚非所以勸奬忠臣、慰答民心之義。

(二)謂不以衆人待其身、而以聖人望於人。

(三)誠宜有以奉其職、使四方後代、知朝廷有直言骨鯁之臣、天子有不僭賞、從諫如流之美。

第六　天地又天地人ノ符號ハ前四種ノ符號ヲ用ヒタル以外更ニ顚讀スル場合ニ用フ

使饑饉不以害妻子、愛飢寒亂心、有錢財以濟醫藥、其盲未甚、庶幾其復見天地日月。

注意

第一　左ノ場合ニハ返點ヲ施サス

(一)所謂(いはゆる)

(二)加之(しかのみならず)

(三)就中(なかんづく)

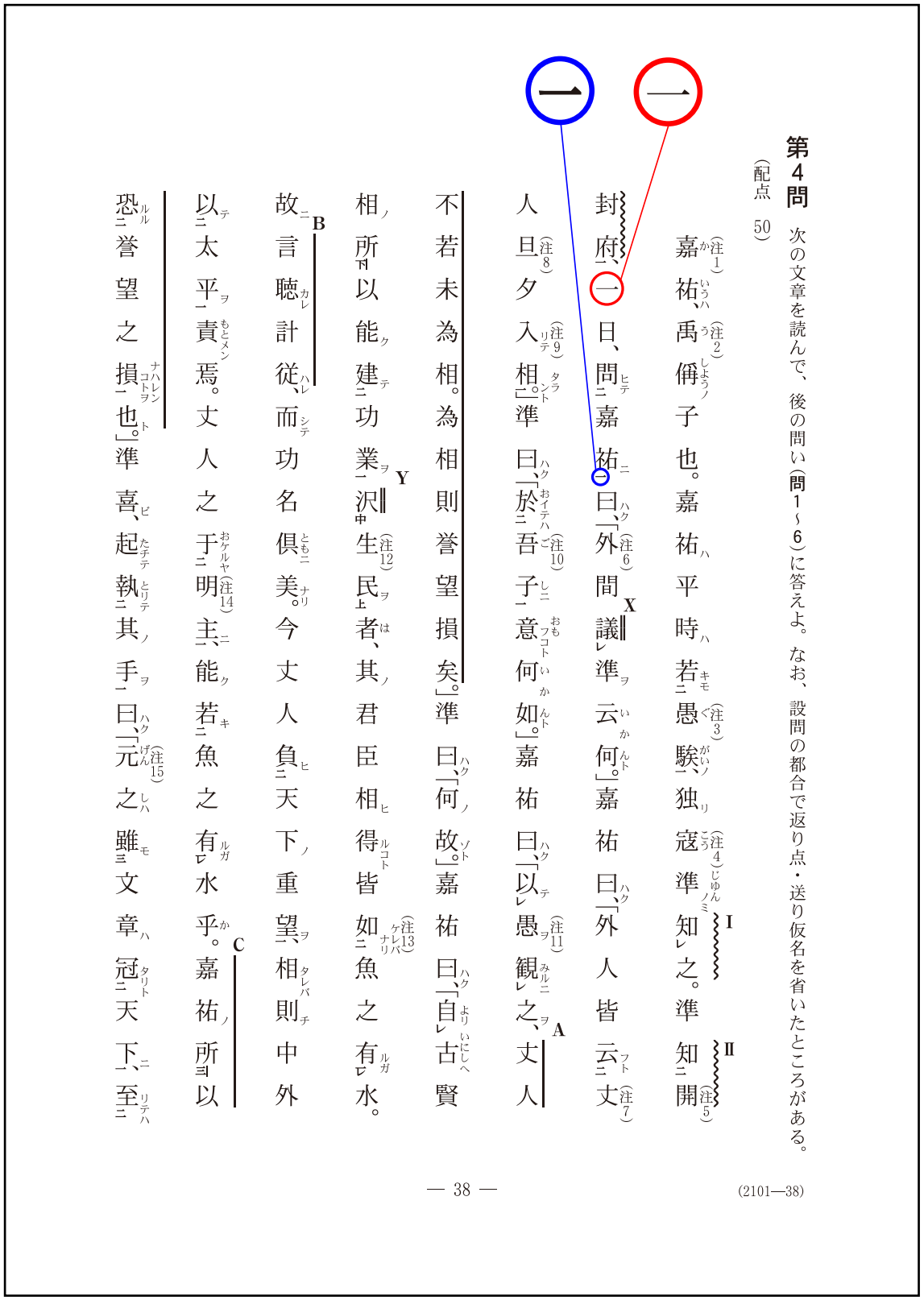
(四)云爾(ときいふ)

第二　使、敎、遣等ヲ再讀スル場合ニハ初讀ノ符號ヲ施サス

能使枉者直。

Fig. 4 A page from 1912 Official Gazette with 五 and 戊 (Ministry of Education, 1912)

Appendix: Current status and practice of Kanbun marks in Japan



第4問 次の文章を読んで、後の問い（問1～6）に答えよ。なお、設問の都合で返り点・送り仮名を省いたところがある。（配点 50）

嘉祐、禹偁子也。嘉祐平時若愚駭、独寇準知之。準知開封府、日、問嘉祐曰、「外間議準云何。」嘉祐曰、「外人皆云丈人旦夕入相。」準曰、「於吾子意何如。」嘉祐曰、「以愚観之、丈人不若未為相。為相則誉望損矣。」準曰、「何故。」嘉祐曰、「自古賢相之所以能建功業沢生民者、其君臣相得皆如魚之有水。故言聴計従、而功名倶美。今丈人負天下重望、相則中外以太平責焉。丈人之于明主、能若魚之有水乎。嘉祐所以恐誉望之損也。」準喜、起執其手曰、「元之雖文章冠天下、至

Above is a page from Literature (Japanese language) test papers in 2018 Center Test (National Center Test for University Admissions), which shows a typical instance of Kanbun typesetting. It illustrates

several facts on modern usage of Kanbun in Japan.

1. Kanbun marks, distinct from CJK Ideographs characters, are widely in use. This is known from difference between Kanbun and non-Kanbun glyphs. Shown enlarged in the page above are Kanbun *kaeriten* 一 (U+3192) and CJK Ideograph 一 (U+4E00). Many Japanese typefaces, including the Yu family (游書体) that is bundled with Windows and MacOS, design Kanbun marks with such thicker strokes so that they harmonize with the bigger base character.
2. Kanbun marks are normally used inline with punctuation, unlike ruby (furigana) and other annotation elements that are put on the side of characters.
3. Kanbun marks usually appear in isolate, that means their stylistic distinction from CJK Ideographs is on the character basis, not the span basis like *italic* or **boldface**. Even when two marks appear in a row, they are conventionally transformed into composite ligatures in one-character height to clarify that they represent a single combined function (in line 5 and 7). Only the combination of [一 (U+3192) 上 (U+3196) 甲 (U+3199) 天 (U+319D)] + レ (U+3191) is possible.
4. In most cases, Kanbun marks are intermixed with texts that contains characters otherwise look identical except in styling, because they derive from those among the most basic and frequently used characters (numerals, adverbs, etc.) As most characters are semantically self-sustained in (Classical) Chinese, removing the distinction between Kanbun and CJK characters in the plain text would immediately result in ambiguity, as their occurrence is unpredictable. For example:
   a) 今丈人負二天下重望一、相則中外以二太平一責焉。 (line 6-7)
   b) 今丈人負二天下重望一、相則中外以二太平一責焉。

   a) is the actual sentence in the text above, and b) is the version replaced all its Kanbun marks with equivalent CJK Ideographs, that is still a valid sentence in Chinese with a different meaning. Restoring a) from b) is a non-evident and probabilistic task for both algorithm and human, because all underlined characters possibly have Kanbun usage, most of which can also be independently taken out from that sentence without breaking grammar.

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://std.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**
Please ensure you are using the latest Form from http://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html **.**
See also http://std.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** *Proposal to Encode 9 Additional Kanbun Marks*
2. Requester's name: *Wang Yifan*
3. Requester type (Member body/Liaison/Individual contribution): *Individual contribution*
4. Submission date: *2020-09-17*
5. Requester's reference (if applicable):
6. Choose one of the following:
   This is a complete proposal: *X*
   (or) More information will be provided later:

B. Technical – General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters):
      Proposed name of script:
   b. The proposal is for addition of character(s) to an existing block: *X*
      Name of the existing block: *Kanbun*
2. Number of characters in proposal: *9*
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary *X* B.1-Specialized (small collection) B.2-Specialized (large collection)
   C-Major extinct D-Attested extinct E-Minor extinct
   F-Archaic Hieroglyphic or Ideographic G-Obscure or questionable usage symbols
4. Is a repertoire including character names provided? *Yes*
   a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? *X*
   b. Are the character shapes attached in a legible form suitable for review?
5. Fonts related:
   a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?
   The requester
   b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
   747.neutron@gmail.com
6. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *X*
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *X*
7. Special encoding issues:
   Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? *No*

8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see Unicode Character Database (http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[1] Form number: N4502-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11, 2011-03, 2012-01)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? *No*
   If YES explain
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? *Yes*
   If YES, with whom? *Japanese font technical experts*
   If YES, available relevant documents:
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? *Yes*
   Reference: *See proposal*
4. The context of use for the proposed characters (type of use; common or rare) *Educational*
   Reference: *See proposal*
5. Are the proposed characters in current use by the user community? *Yes*
   If YES, where? Reference: *See proposal*
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *No*
   If YES, is a rationale provided?
   If YES, reference:
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *Yes*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character? *Yes*
    If YES, is a rationale for its inclusion provided? *Yes*
    If YES, reference: *See proposal appendix*
11. Does the proposal include use of combining characters and/or use of composite sequences? *No*
    If YES, is a rationale for such use provided?
    If YES, reference:
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?
    If YES, reference:
12. Does the proposal contain characters with any special properties such as control function or similar semantics? *No*
    If YES, describe in detail (include attachment if necessary)
13. Does the proposal contain any Ideographic compatibility characters? *No*
    If YES, are the equivalent corresponding unified ideographic characters identified?
    If YES, reference: