

Proposal to encode WHITE IDEOGRAPHIC COMMA

Gen Kojitani(genkjtn@gmail.com)

2026 June 9

This proposal requests the encoding of one Japanese punctuation character, WHITE IDEOGRAPHIC COMMA, used historically and occasionally in modern scholarly or literary contexts as an inline punctuation mark corresponding especially to the European semicolon.

Code Point	Glyph	Character Name
U+16FE5	ㄥ	WHITE IDEOGRAPHIC COMMA

Background

Japanese printed texts have historically used several punctuation marks in addition to the ordinary ideographic comma and ideographic full stop. One such mark is “*shiroten*,” “*shiro-goma ten*” or “*goma-shiro ten*”, literally “white dot” or “white sesame dot”(Despite this name, it is not U+FE46 WHITE SESAME DOT), a hollow form resembling an ideographic comma (U+3001).

Shiroten emerged in the context of Meiji-period experiments with Japanese punctuation and the development of modern written Japanese. Its introduction is often attributed to *Yamada Bimyo*, who was active in the reform of Japanese prose style and in the use of punctuation in modern Japanese writing. However, the mark did not become part of ordinary modern Japanese punctuation. It was tried by several Meiji-period writers and translators, but it eventually fell out of general use.

The mark appears in discussions of Japanese punctuation and in printed literary, lexicographic, governmental, and translation contexts. In these sources, *shiroten* is not used as a decorative or emphatic mark beside the text. Rather, it functions as punctuation within the text line. Its use is particularly associated with a punctuation value intermediate between the Japanese comma and full stop, and in translation contexts it may correspond to the European semicolon.

A particularly important example is *Futabatei Shimei*’s translation “*Meguriai*,” where the use of *shiroten* has been analyzed in relation to the punctuation of the Russian source text. In this context, the mark was used to preserve distinctions such as those expressed by the semicolon in the source language, rather than merely replacing the ordinary Japanese comma.

Encoding this character is needed for accurate representation of Japanese texts, especially historical Japanese documents, scholarly editions, digitized literary works, dictionaries, and translations that preserve punctuation distinctions found in the source material.

Period and geographical area

The attested use of WHITE IDEOGRAPHIC COMMA is in Japanese texts, primarily from the Meiji period, with later appearances in scholarly or literary editions and translations. The known examples are from Japan. No independent tradition of use in Chinese, Korean, or Vietnamese text has been identified.

User community and need

The user community consists primarily of scholars, editors, librarians, digital humanists, lexicographers, historians of Japanese writing, and publishers preparing accurate transcriptions or editions of Japanese texts.

The character is needed for plain-text representation of Japanese works in which shiroten is distinguished from the ordinary ideographic comma, the semicolon, and emphasis marks. Without a separate encoded character, users must substitute visually similar but semantically incorrect characters, use images, markup, private-use characters, or ad hoc font-dependent solutions.

Such substitutions are not adequate for searchable text, scholarly quotation, text encoding, collation, corpus analysis, or preservation of original punctuation practices.

Character Name

The proposed name is WHITE IDEOGRAPHIC COMMA. This name follows established Unicode naming patterns for CJK punctuation. It is structurally parallel to U+3001 IDEOGRAPHIC COMMA and uses WHITE in the same descriptive sense found in other CJK punctuation names, such as WHITE CORNER BRACKET characters.

Distinction from U+3001 IDEOGRAPHIC COMMA

WHITE IDEOGRAPHIC COMMA is not a glyphic variant of U+3001 IDEOGRAPHIC COMMA. U+3001 represents the ordinary Japanese ideographic comma. The proposed character has a distinct punctuation value in sources that contrast it with the ordinary comma. In particular, shiroten has been used as a semicolon-like mark, where the ordinary ideographic comma would not preserve the punctuation distinction intended by the source.

The distinction is therefore not merely typographic. Replacing WHITE IDEOGRAPHIC COMMA with U+3001 IDEOGRAPHIC COMMA would erase a meaningful punctuation contrast in plain text.

Distinction from U+FE46 WHITE SESAME DOT

WHITE IDEOGRAPHIC COMMA is not unifiable with U+FE46 WHITE SESAME DOT. U+FE46 WHITE SESAME DOT is an emphasis mark. The Unicode Standard describes U+FE45 SESAME DOT and U+FE46 WHITE SESAME DOT as marks used in vertical text, where a series of sesame dots may appear beside the main text as sidelining to provide visual emphasis. The Standard further notes that the function of an ideographic comma in connected text is distinct from that of a sesame dot.

The proposed character has a different function. It is used inline in running Japanese text as punctuation. It is not placed beside vertical text as an emphasis mark, and it does not form a series of sidelining marks applied to adjacent base text.

The resemblance between WHITE IDEOGRAPHIC COMMA and U+FE46 WHITE SESAME DOT is therefore glyphic only. Unifying the two would conflate an inline punctuation mark with a sidelining emphasis mark, leading to incorrect text semantics, line breaking, vertical layout behavior, searching, and scholarly transcription.

Distinction from U+FF1B FULLWIDTH SEMICOLON

Although WHITE IDEOGRAPHIC COMMA may correspond to a European semicolon in translation contexts, it is not a semicolon character. It is a Japanese punctuation mark with an ideographic-comma-shaped glyph and Japanese layout behavior. Replacing it with U+003B or U+FF1B would preserve only an approximate punctuation value, while losing the actual character used in the source text.

Properties

U+16FE5; WHITE IDEOGRAPHIC COMMA;Po;0;ON;;;;;N;;;;;

Script=Common
East_Asian_Width=W
Line_Break=CL
Vertical_Orientation=Tu
Terminal_Punctuation=Yes
Sentence_Terminal=No

The proposed properties are modeled on U+3001 IDEOGRAPHIC COMMA, because the character is comma-like inline CJK punctuation.

Line breaking and vertical layout

WHITE IDEOGRAPHIC COMMA is an inline Japanese punctuation mark. It should participate in Japanese line-breaking behavior as U+3001.

The Line_Break property should be CL. The character should not appear at the start of a line, and it should be kept with the preceding character, as is expected for Japanese comma-like punctuation. This follows the behavior of U+3001 IDEOGRAPHIC COMMA and U+3002 IDEOGRAPHIC FULL STOP.

The known attestations are in vertical Japanese text. In such text, the glyph is placed in the upper-right area of the ideographic character cell, following the Japanese convention for comma-like punctuation in vertical layout.

No horizontal Japanese examples are currently known. However, because the character is an inline Japanese punctuation mark corresponding in form and behavior to comma-like punctuation, its horizontal presentation should follow the ordinary Japanese convention for such punctuation. In horizontal Japanese text, the glyph should be placed in the lower-left area of the ideographic character cell.

Horizontal	Vertical
	

The representative code chart glyph should follow the Japanese convention, since the attested use of the character is in Japanese texts. No attested use is currently known in ROC or Hong Kong typographic contexts where ideographic punctuation may be centered in the character cell. However, as with U+3001 IDEOGRAPHIC COMMA and U+3002 IDEOGRAPHIC FULL STOP, centered presentation may be appropriate in fonts or layout environments following Traditional-Chinese regional conventions. Such regional glyph variation should be treated as font-level or standardized-variation behavior if necessary, not as a distinct character.

Collation, searching, and normalization

No canonical or compatibility decomposition is proposed.

WHITE IDEOGRAPHIC COMMA should not normalize to U+3001 IDEOGRAPHIC COMMA, U+FE46 WHITE SESAME DOT, U+003B SEMICOLON, or U+FF1B FULLWIDTH SEMICOLON. Such normalization would erase a distinction that is meaningful in source texts.

Search systems may choose to provide loose matching between WHITE IDEOGRAPHIC COMMA and semicolon-like punctuation for user convenience, but this should be a higher-level tailoring. At the character encoding level, the distinction should be preserved.

Plain-text requirement

WHITE IDEOGRAPHIC COMMA is required for plain-text representation. It cannot be represented accurately by U+3001 IDEOGRAPHIC COMMA, because that would erase the distinction between ordinary comma punctuation and shiroten. It cannot be represented by U+FE46 WHITE SESAME DOT, because that character is a sideling emphasis mark rather than inline punctuation. It cannot be represented by U+003B SEMICOLON or U+FF1B FULLWIDTH SEMICOLON, because those characters do not preserve the Japanese typographic form used in the sources.

Private-use characters, images, or markup are not adequate for interchange, searching, indexing, quotation, or scholarly transcription.

Samples

二、終止の形をとつてゐても、その文意が
続く場合にはテンをうつ（例23）。
たゞし、他のテンとのつり合ひ上、
この場合にマルをうつこともある。（例
4）。

〔附記〕この項のテンは、言はゞ、半終止
符ともいふべきものであるから、将来、
特別の符号（例へば「ゝ」^{シロテン}）のこ
きもの（が）が広く行はれるやうになること
は望ましい。

用例の〔参照二〕は本則によるもの。
また〔参照二〕は「ゝ」^{シロテン}を使つて
みたもの。

Fig. 1 文部省『くぎり符号の使ひ方〔句読法〕(案)』, 1946

○おもただし（第一上）及（第四上）又（第四上）
 形（面正し）（一）オモテブセデナイ。||
 面目ヲホドコスヤウデアル。||メイヨデ
 アル。―枕ノ草紙、内侍ノスケナド云ヒテ、
 ヲリヲリ内ヘマ井リ、祭リノ使ヒナドニ
 出デタルモおもただしカラズヤアル。―
 源氏、年ゴロウレシク、おもただしキツイ
 デニノミ打チヨリタマヒシモノヲ。

Fig. 2 山田美妙『日本大辞書』第3巻 う，え，お，p.345，日本大辞書発行所，1892

山田武太郎編輯
 緑山散史
 花の都は程遠し。
 頼みし親の膝を去り。
 岩をも徹を桑の弓。
 東の空は遙々と、
 いつか飾る綾錦。

Fig. 3 山田美妙 編『新体詞選』，p.1，香雲書屋，1886

い、今も夜だ。彼の時のやうな、燈火の射した見知ぬ
小坐舗から不意な聲が響わたった。彼の時のやうな、
自分一人だ。胸が荐りに躍りだした。「夢ぢやな
いか？」と思つた。ソレまた聞えた最後の Vienn が

Fig. 4 ツルゲーネフ『めぐりあひ』, translated by 二葉亭四迷, 『都の花』 第一卷第三号, p.12, 1888

婦人はジロリと自分の顔を見て、急に起上りて、其
 場を去らうとした。
 是れは餘り残酷な！ 自分は婦人の腕を捉へた。
 「何卒、といひだした」——マア、下に坐て下さい、
 私のいふ所を聞て下さい。
 婦人は考へてゐたが、また腰を掛けた。

Fig. 5 『二葉亭四迷全集』第一巻『めぐりあひ』, p.203, 岩波書店, 1964

彼は余と故郷を全じふし、
 余と手を携へて都へ上りにき——
 京都に出で、琵琶を後にし
 三州の沃野を過りて、濱名に若き、
 富士の麓に出て、函根を越し、
 遂に花の都へは着たりき。
 愛といひ戀といふには科あれど、
 吾等雙個の愛は精神にあり、
 花の美しくしきは美しくしけれど、
 吾が花嫁の美は、其處にあり、
 梅が枝にさへずる鳥は多情あれ、
 吾が情はたい赤き心にあり。

Fig. 6 北村透谷 著『楚囚之詩』, p.5, 春祥堂, 1889

— 句読点は慣行のもの以外に、現代欧文の例にならない、すでに『饗宴』においても採用したもの、すなわちコーロンの代りに「」を、またセミコーロンの代りに『』を用いた。文章の構造を一目瞭然たらしむるに役立つと信じたからである。

Fig. 7 クリトン『ソクラテスの弁明』, translated by 久保勉, p.65, 岩波文庫, 1964

References

1. 吉村明子「[明治の文体：特に欧文体を中心として](#)」『日本文學』第 21 巻、1963 年 10 月 20 日、44-57 頁。
2. 岡崎晃一「[巖谷小波・幸田露伴の白ゴマ点](#)」『神戸親和女子大学言語文化研究』第 2 巻、2008 年、1-9 頁。
3. 岡田和子「[二葉亭四迷の『めぐりあひ』といロシア語原文における句読点の比較 — 明治時代の洋語学習と《白抜き点》](#)」『文学研究論集』第 12 巻、筑波大学比較・理論学会、1995 年 3 月 20 日、17-42 頁。

**ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646¹**

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 WHITE IDEOGRAPHIC COMMA		
2. Requester's name:	Gen Kojitani		
3. Requester type (Member body/Liaison/Individual contribution):	Individual		
4. Submission date:	2026/6/9		
5. Requester's reference (if applicable):	genkjtn@gmail.com		
6. Choose one of the following:			
This is a complete proposal:			yes
(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:			yes
Name of the existing block:	Ideographic Symbols and Punctuation		
2. Number of characters in proposal:			1
3. Proposed category (select one from below - see section 2.2 of P&P document):			
A-Contemporary	B.1-Specialized (small collection)	x	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?			
a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document?			yes
b. Are the character shapes attached in a legible form suitable for review?			yes
5. Fonts related:			
a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?	Gen Kojitani(requester)		
b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):	Gen Kojitani, genkjtn@gmail.com. The font is licensed under the SIL Open Font License 1.1.		
6. References:			
a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?			yes
b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?			yes
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)?			yes
See sections on properties, line breaking and vertical layout, collation, searching, and normalization.			
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?	CJK & Unihan Working Group; experts in Japan consulted through UTC review.
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:	Section "User community and need".
4. The context of use for the proposed characters (type of use; common or rare)	Rare
Reference:	Historical and scholarly Japanese text representation; see Background and Evidence of use.
5. Are the proposed characters in current use by the user community?	Yes
If YES, where? Reference:	Rarely, in scholarly and editorial contexts.
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)?	N/A
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?	
If YES, reference:	
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:	

(Detach the following pages from the filled summary form before submitting).

1. Information accompanying submissions

The process of deciding which characters should be included in the repertoire of the standard by WG 2 depends on the availability of accurate and comprehensive information about any proposed additions. WG 2, at its San Francisco meeting 26, designed a form (template) that will assist the submitters in gathering and providing the relevant information, and will assist WG 2 in making more informed decisions.

This form is part of the [WG 2 Principles and Procedures document](http://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html) and is available on line at <http://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html>.

A duly completed proposal summary form must accompany each new submission. The latest version of this form must be used in submissions. Such a form will assist WG 2 to better evaluate the proposal, and progress the proposal towards a speedier acceptance and inclusion in the standard. Submitters are also requested to ensure that a proposed character does not already exist in the standard.

Submitters are encouraged to visit the “[Where is my Character](#)” page on the Unicode web site for more information on checking if their proposed character or script is already encoded in the standard, or a similar proposal has already been made by someone else. There are also several electronic discussion lists maintained by the Unicode consortium that one could use to discuss with other experts internationally on various subjects related to the standard. Submitters are also encouraged to familiarize themselves with ISO/IEC TR15285 – Character Glyph Model. The latest version of ISO/IEC 10646 as well as ISO/IEC TR15285 are available on line under 'Freely Available Standards' at: <http://standards.iso.org/ittf/PubliclyAvailableStandards/>.

National bodies should take note of the guidelines in section *C.4 Some guidelines on proposing new material as ballot comments* in the [P&P document](#).

In addition to text extracted from the standard in *Annex L: Character-naming guidelines* in the [P&P document](#), the following definitions from the standard are also referenced in the proposal summary form:

Clause 4.12 Combining character:

A member of an identified subset of the coded character set of ISO/IEC 10646 intended for combination with the preceding non-combining graphic character, or with a sequence of combining characters preceded by a non-combining character (see also 4.14).

NOTE – ISO/IEC 10646 specifies several subset collections, which include combining characters.

Clause 4.14 Composite sequence:

A sequence of graphic characters consisting of a non-combining character followed by one or more combining characters (see also 4.12).

NOTE 1 – A graphic symbol for a composite sequence generally consists of the combination of the graphic symbols of each character in the sequence.

NOTE 2 – A composite sequence is not a character and therefore is not a member of the repertoire of ISO/IEC 10646.

2. Submitter's responsibilities

The national body or liaison organization (or any other organization or an individual) proposing new character(s) or a new script shall provide:

3. Proposed category for the script or character(s), character name(s), and description of usage.
4. Justification for the category and name(s).
5. A representative glyph(s) image on paper:

If the proposed glyph image is similar to a glyph image of a previously encoded ISO/IEC 10646 character, then additional justification for encoding the new character shall be provided.

Note: Any proposal that suggests that one or more of such variant forms is actually a distinct character requiring separate encoding should provide detailed, printed evidence that there is actual, contrastive use of the variant form(s). It is insufficient for a proposal to claim a requirement to encode as characters in the Standard, glyphic forms which happen to occur in another character encoding that did not follow the Character-Glyph Model that guides the choice of appropriate characters for encoding in ISO/IEC 10646.

Note: WG 2 has resolved in Resolution [M38.12](#) not to add any more Arabic presentation forms to the standard and suggests users to employ appropriate input methods, rendering and font technologies to meet the user requirements.
6. Mappings to accepted sources, for example, other standards, dictionaries, accessible published materials.
7. Font resource:

A font must be provided to the Editor promptly after the characters have been approved by WG2 for use in printing the charts. If a font is not provided, the Editor cannot include the glyphs in the charts and, as a result, the repertoire of characters corresponding to these glyphs will not be included in draft amendments. Fonts must be in one of the following formats (in preferential order): OpenType, TrueType, Postscript Type 1.

Fonts submitted must have no license restrictions that prevent embedding into PDF documents. Because of synchronization between ISO/IEC 10646 and The Unicode Standard, any grant of license must cover use in publishing both standards and related documents (see resolution [M45.30](#)). For technical reasons, the editors must be able under that license to freely modify or replace glyph outlines in their copies of the fonts. The submitter should be prepared to provide fonts of suitable quality and license conditions, unless equivalent fonts are already available to the editors.
8. List of all the parties consulted:

Towards ensuring that all significant stakeholders have been consulted, WG 2 requests that submitters solicit feedback from other groups who use the proposed characters, and list these contacts. Submitters are encouraged to provide the email id-s of the submitters as well as other experts who have been consulted to facilitate any clarification queries.
9. Equivalent glyph images:

If the submission intends using composite sequences of proposed or existing combining and non-combining characters, a list consisting of each composite sequence and its corresponding glyph image shall be provided to better understand the intended use.
10. Compatibility equivalents:

If the submission includes compatibility ideographic characters, identify (per resolution [M45.29](#)):

 - the source, which contains two distinct code positions that correspond to a single unified CJK Ideograph character of ISO/IEC 10646
 - the ISO 10646 unified CJK Ideograph
 - the code position in the source for the unified CJK ideograph
 - the code position in the source for the proposed compatibility ideograph
11. Properties that may affect the BiDi processing

Any BiDirectional algorithm related properties associated with the characters should be spelled out (see UAX#9 - <http://www.unicode.org/reports/tr9/>).
12. Any additional information that will assist in correct understanding of the different characteristics and linguistic processing of the proposed character(s) or script.
13. If any of the proposed characters are suitable as *syntax* characters please take note of guideline D.2.5 (on page 24 and of Unicode Standard Annex 31 at <http://www.unicode.org/reports/tr31/>).
14. If the proposal is for a new script, identify the standardized characters that are commonly used directly in the script, or proposed to be unified with the characters of the script, in particular standardized characters allocated in different blocks. Examples include punctuation marks and combining marks. Such information will assist in assigning properties for characters shared across multiple scripts or in identifying character repertoires needed to support particular languages.
15. If you are aware of already standardized characters that are visually close to any of the proposed characters, you are invited to list them in the proposal. This will assist in the analysis of the script for 'visually confusables', towards providing additional guidance on use of the standard from a security perspective (see UTR#36 - Unicode Security Considerations – at <http://www.unicode.org/reports/tr36/>).