

TO: UTC
FROM: Debbie Anderson, SEI, UC Berkeley
DATE: April 17 2023 -- Revised July 2, 2023
SUBJECT: Draft Template for new script proposals

Attached is a draft template for proposal authors who wish to propose a new script.

Coversheet for New Proposal Submissions

This coversheet is to provide information used by the review committees to help in their evaluation of proposals. Information provided in the coversheet is treated as confidential by the review committees and will not be published as a part of the proposal document, in the event that the proposal is deemed appropriate for posting.

**Required field*

*Your name (and any other proposal authors):

*Email (for main proposal author):

Your address:

*Date:

*Describe your relationship to the user community for the script/characters. *Examples:*

- *I have been contracted to prepare the proposal by the user community by the Language Board of [].*
- *I have been in email contact with [names of users] on [dates].*
- *I am the creator of the script/characters.*
- *I am a member of the user community, as speaker of the [] language using the [] script.*
- *I discovered the script or characters during research as a student of [field] at [institution].*

*How do you define the user community (for this script or character[s])?

Examples:

- *Modern users of the Bamum script who live in Fumban, Cameroon, and want to be able write the script on their mobile devices.*
- *Researchers in the Middle Eastern Languages and Cultures Department at UC Berkeley who are studying cuneiform tablets from Iraq and want to put their text online.*

TEMPLATE FOR NEW SCRIPT PROPOSAL AUTHORS

To: Script Ad Hoc /Unicode Technical Committee

From: [your name and affiliation]

Subject: [name of script]

Date:

Include a short paragraph summarizing what is being asked the UTC to do (such as, “We propose the addition of the xxx script into the Unicode Standard”).

I Background

Provide background on the proposed script, addressing the questions listed below. This information is needed in order to understand the use of the script today. For modern scripts, widespread usage is looked upon favorably.

Information to include:

- Who is the user community (for example, scholars, government, school children, or elders of a community) for whom the script is needed?
- What is the approximate size of the user community?
- What is the time period and geographical area (i.e., countries or region) when/where the script has been used?
- Give a brief history of the script. Is it related to other scripts? Has the script changed significantly over time?
- For a modern script
 - Is the script being taught? If so, is it taught in public schools or private schools or informally? About how people are learning the script today? Are there any graduates of the schools yet and, if so, how many?
 - Is the script included in primers?
 - Does the script appear in newspapers or other publications and/or signage? If you can, provide an estimate on the number of publications and type (e.g., 5 book titles and a monthly magazine). Note the number of written works written by people who are not the script’s creator.
 - How many people are using the script daily to write the language?
 - Is there government support for the script?
 - Is there a font currently being used to print the script?
 - Does the script appear on any keyboards today? If so, can you provide information on the keyboard(s) that use the characters for the script?
 - Are there other scripts in the community that might be competing with the proposed script?

Tip: A criterion for adding a new, modern-use script is that it has spread across a community (beyond the creator), with evidence to support this claim.

- For a historical script
 - Was the script used in certain contexts (such as, in religious materials or to document commercial transactions)?
 - What is the importance of getting this historical script into Unicode?
- Is there an ongoing digitization project that requires the script?
- What language(s) is it used for? Give the [ISO 639](#) code (e.g., “gym” is the ISO 639 code for Ngäbere). If only one language could be specified for the script, what language would it be?
- Give any history on whether the script has been proposed before (with proposal document numbers). *Tip:* Do a search on the Unicode website to see if any reference is made to the script.

II Script Name

Discuss the proposed name of the script. If a script is accepted for encoding, the Unicode Standard will require a script name that uses only Latin letters A – Z, with no diacritics (hence Ngäbere would not be permitted). This may require a variation on script names preferred by the user community. Please provide a name that uses only A – Z and that would be acceptable to the user community. Script names can be a sticking point, so it is best to find one that the user community does not object to.

Often the script name used by a user community includes a word that means ‘script’ or ‘writing’. In Unicode, these words are not normally included in the script name. For example, in South Asia, scripts are often known by a name that includes the Sanskrit word ‘lipi’, but in Unicode that would normally be omitted. For instance, a script name in Unicode would be ‘Modi’ rather than ‘Modi Lipi’.

III Structure

Identify the type of script (see [Section 6.1 of the Unicode Standard](#)): Is the script an alphabet, a syllabary, or another type of script? Does the script run left to right, or right to left? Are there combining diacritics?

IV Character Repertoire

List the proposed characters. For each character, provide a glyph (image), a code point, a proposed name for each and refer to at least one example in the figures.

Example (from [L2/20-115](#)):

	U+A7F0 LATIN SMALL LETTER ESH WITH DOUBLE BAR. Figures 1–9.
---	---

Glyph: Pick a representative glyph that most users would recognize.

Code point: You can suggest proposed code points for the characters, but this is not necessary, since the standards committees will assign code points. If you do suggest code points, verify they are available by comparing the latest [script code chart](#) and the [Pipeline of proposed characters](#). If you don’t suggest code points, use XXXX0...XXXX9, XXXXA, ...XXXXF, XXX10, XXX11, etc.

Name: It is best to suggest a name that fits with the same pattern as others in the same script and similar characters in other scripts. Check the [names list](#) or the [code charts](#) for the naming patterns.

Figure: At least one example of each character is required. Examples help to establish that the characters need to be interchanged digitally. In many cases, this can be demonstrated by providing two examples from different authors and different publishers. In some cases, one figure may suffice. The character should be circled in the figure with a caption that identifies the character and provides a reference to the source. (See further guidelines in section VIII.)

Chart

Although not strictly required, providing a chart of the proposed characters and a names list is helpful for those reviewing the proposal. Examples can be found in most script proposals, such as the chart and names list on page 10 of the Sidetic script proposal, [L2/23-019](#).

For advanced script proposal authors, using the Unibook software will create a chart and names list in the Unicode Standard format. It can be found here: <https://unicode.org/unibook/>
If using Unibook to create a chart, please also provide the names list file used to create the chart.

Additional information on characters

Phonetic value

Provide the phonetic value of the characters (preferably with IPA or a phonetic description), if applicable.

Example (from [L2/19-118](#)):

The first of these characters is commonly used when writing Wolof using the Arabic script to represent a few phonemes ñ, č, nǝ, nč. JEEM WITH THREE DOTS ABOVE typically functions to create nasalised sounds that are common in Wolof.

Joining information

Do any characters join or touch with one another? If so, describe how they join and provide examples in the figures. The more detail, the better. This is especially helpful for font designers.

Punctuation

If a proposed character is a mark of punctuation, describe its behavior (for example, if a given character is used as a paragraph marker). Provide examples that demonstrate the character's use.

Also, are Latin punctuation symbols or those of another encoded script also used? If so, list them with their names and code points.

Numbers

How are numbers represented in the script? Does the script use a decimal (radix 10) system or another system? Does the script have its own set of script-specific numbers or does it use numbers from another script?

V Properties

General Category and other properties

Give the General Category of the characters and other properties. *(You can skip this step initially)*

Tip: Find characters that you think are used in a similar way, and use the properties contained in <https://www.unicode.org/Public/UCD/latest/ucd/UnicodeData.txt>, modifying as needed.

Example (from L2/19-111):

```
08C5;ARABIC LETTER LAM WITH SMALL ARABIC LETTER TAH  
ABOVE;Lo;0;AL;;;;;N;;;;;
```

The fields include the following (with links to descriptions and range of values):

Code_point

NAME

[General Category](#) [range of [gc values](#)]

[Canonical Combining Class](#) [range of [CCC values](#)]

[Bidi Class](#) [range of [Bidi Class values](#)]

[Decomposition Type/Decomposition Mapping](#)

[Numeric Type](#)

[Numeric Value](#)

[Bidi Mirrored](#)

Unicode_1_Name (obsolete as of Unicode 6.2.0)

ISO_Comment (obsolete as of Unicode 5.2.0)

[Simple Uppercase Mapping](#)

[Simple Lowercase Mapping](#)

[Simple Titlecase Mapping](#)

Line breaking information

Answering the following questions will help the Unicode experts determine the appropriate line break properties:

- Do the major languages using the script separate words using U+0020 SPACE or some other character? If some other character, which one?
- Can line breaks only occur at word boundaries, or can they occur before any spacing character, or (for Brahmic scripts) before any orthographic syllable?
- If line breaks are usually only allowed at word boundaries, is there a special mode (“hyphenation”) that allows line breaks within words at select positions with the insertion of a marker that this mode was used? Is the marker U+002D HYPHEN-MINUS, or which other character?
- Can line breaks occur within numbers?
- Are there restrictions on line breaking before or after certain punctuation characters? E.g., no line breaks before full stop or before closing parenthesis.
- Are there any other special considerations for line breaking in this script?

Properties for Arabic and Indic characters

Additional information is required for characters that join, such as Arabic, and for various Indic scripts. (See details in <https://www.unicode.org/Public/UCD/latest/ucd/ArabicShaping.txt> and, for Indic scripts, <https://www.unicode.org/Public/UCD/latest/ucd/IndicPositionalCategory.txt> and <https://www.unicode.org/Public/UCD/latest/ucd/IndicSyllabicCategory.txt>.)

VI Collation

What is the expected sorting order of the characters in the script? In other words, where would you expect them to occur in a dictionary or list of words?

Give the order characters are expected to occur. A prose description is acceptable, or you can use the following format (which shows ALEPH occurs before GENERIC ALEPH-NUN, etc.):

ⴰ ALEPH < ⴢ GENERIC ALEPH-NUN < ⴣ ALEPH WITH LEFT TAIL < ⴥ BETH <
⴦ GENERIC BETH-YODH < ⴧ GIMEL-HETH < ⴨ WAW < ⴩ ZAYIN < ⴫ FINAL HETH <
⴬ YODH < ⴭ KAPH < ⴮ LAMEDH < ⴯ MEM < ⴰ NUN < ⴱ SAMEKH < ⴲ PE <

(Note: The characters in a code chart do not necessarily reflect the expected order in a dictionary.)

VI References

Provide reference works that you used and standard reference works that can be consulted, with full bibliographic information.

Example (from [L2/19-111](#)):

Muhammad Yar. Circa 1792. *Afarinish Nama*. Courtesy British Library.

VII Acknowledgements

(optional)

VIII Examples

Provide examples of each proposed character in a line of printed text, circling the characters. In the caption, name the circled character and provide the reference.

Printed materials are generally preferred, if possible.

It is highly advisable to include texts that show the character in running text, that is, not just a chart or a keyboard, but sentences that use the character. Such examples help demonstrate spacing and where line breaks can occur.

Provide a transliteration of the text of a few examples.

Note: Proposals without any examples will not likely be reviewed, stalling any progress on a proposal. Only those characters with evidence will be considered.

Example (from L2/19-111):

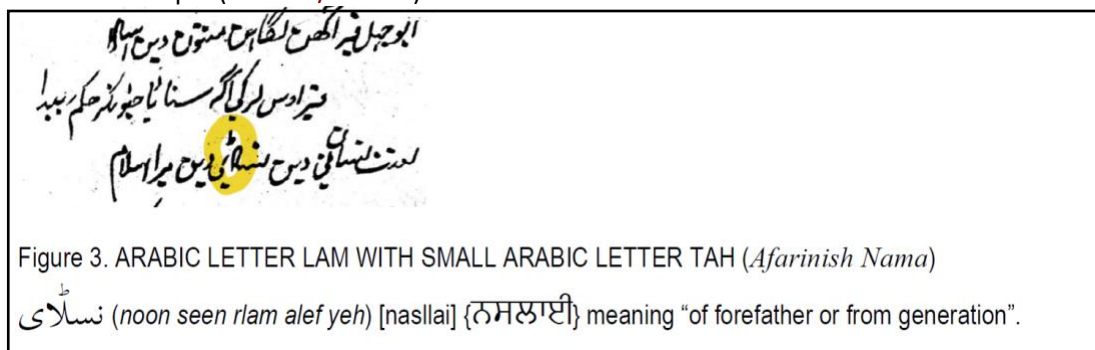


Figure 3. ARABIC LETTER LAM WITH SMALL ARABIC LETTER TAH (*Afarinish Nama*)

نسلای (noon seen lam alef yeh) [nasllai] {नमलायी} meaning “of forefather or from generation”.

IX ISO Proposal Summary Forms

Include the [ISO proposal summary forms](#) at the end. This is not necessary for the preliminary version of a proposal.

X. Next Steps

1. For submission to Unicode for review by Script Ad Hoc

- Number all the pages.
- Send a copy to Unicode (docsubmit@unicode.org) so it can be routed to the Script Ad Hoc to review.

The [Script Ad Hoc group](#) meets monthly and generally will review those proposals that have included information described in this template and have demonstrated the need for the characters to be carried in plain text. Typically proposals take several revisions until all the needed information is provided. Comments on proposals that are reviewed by the Script Ad Hoc appear in the [quarterly Script Ad Hoc recommendations](#).

NOTE: Not all proposed scripts and characters are approved. Some newly created scripts may not yet be ready for encoding, as the script generally needs to be widely accepted, used in various publications, and taught in schools.

Note: If the script is eventually accepted, a font will be required.

- If you do not have a font, you can create a font using [Glyphs](#) or other font-editing software. When creating glyphs that uses another font as its base, use a font that allows modification (such as Noto or a font with a SIL [Open Font License](#)).
- If the script includes combining marks (such as diacritics that appear above or below another character such as َ), do not include the dotted circle.
- In the font, map the glyphs to the proposed code points.

2. After the proposal is considered mature by the Script Ad Hoc

Once the Script Ad Hoc has decided the proposal has provided full information on the proposed characters and a strong case has been made to encode them, the document will be posted in the Unicode document register and then be reviewed by the Properties and Algorithms group, which may suggest changes. Once both the Script Ad Hoc and Properties and Algorithms have agreed the proposal is ready for encoding, the proposal goes to the Unicode Technical Committee (UTC).

The UTC will review the proposal and, if it considers the script ready to encode, will provisionally assign code points. The UTC will at a later point formally accept the characters and identify a Unicode version where the characters will be published. Before publication, the characters will go through an alpha and [beta](#) review phase.

Proposers should be aware that the entire approval process – if successful – can take at least two years. Proposals that are considered mature and advance to the UTC are required to have a font with the glyphs.

3. After publication

Once the characters are published in Unicode, having a font and keyboard ready – with the characters at the approved Unicode code points -- can speed up the use of the script. In addition, operating systems may need to be updated in order to support the new characters. Working with software, platform and font vendors may be useful in getting support for the new characters.

For support in locale data, see [CLDR project](#).

Additional Tips

Basically, the more information on the use and behavior of a character (with clear examples), the better. This information can be used by the committees to verify the character as proposed and to check if the specified properties are correct.

Consulting with linguists or specialists in the script is very helpful, since they can provide phonetic information and review the proposal. Also, if there are local or governmental language commissions or language institutes or societies, involving them can also be very beneficial.

The most successful proposal authors are those that work collaboratively with the standards committees. The committees review proposals to ensure the characters follow the Unicode Standard. Typically, proposal authors need to revise their proposals several times in order to present all the required information.