Questions (and Answers) from the Unicode webinar “Documenting and Preserving Languages with Unicode”

Topics

The following questions were posed as part of the May 16, 2023, Unicode webinar presentation, “Documenting and Preserving Languages with Unicode,” by Debbie Anderson and Andrew Glass, with answers by Debbie Anderson and other Unicode experts.

Importance of Unicode for language preservation, documentation, and archiving

QUESTION: How is Unicode related to language preservation and why is it important?

ANSWER: For those who wish to preserve languages with language documentation (in text) or who want to preserve written materials of a language, using Unicode (with a Unicode-based font) means the text materials will be accessible to others electronically and archivable for the long-term.

Using a non-Unicode font to represent materials, however, will mean the texts may be difficult to exchange electronically and may not be recoverable in the future. In a time when many languages are in danger of disappearing and text materials may be lost, using the stable standard Unicode to represent texts is critical.

QUESTION: I am doing language documentation and would like to know more about how Unicode could be of use.

ANSWER: Language documentation involves use of various characters, including the International Phonetic Alphabet and other notational systems, as well as other letters and symbols. Most of these characters are in the international standard, Unicode. By using a font that is Unicode-based, your text will be able to be typed, sent, and received by others as you intend. The text should also allow reliable copying and pasting of text between applications.

Also, because Unicode is an international standard that is widely supported, your language documentation materials should be stable.

If you have a letter or symbol that you believe is not in Unicode, you should review the FAQs https://www.unicode.org/faq/prop_new_characters.html and https://www.unicode.org/faq/char_proposal.html. (Note that many letters or symbols are representable in Unicode with a sequence of characters, such as q̀ which can be represented by <0071, 0300>.)

If the letter or symbol is not representable in Unicode, please use the following template for proposing new characters: https://www.unicode.org/L2/L2023/23104-addl-script-template-april2023.pdf.

QUESTION: Is Unicode a sustainable solution for the long-term archiving of humanities research data in different languages?

ANSWER: Yes! Because Unicode is an international standard, it is a stable format that will be supported into the future. It is also widely supported on operating systems and browsers today.

Making a script in Unicode usable

QUESTION: What is the process for making a script in Unicode be usable?

ANSWER: Most operating systems update to each Unicode release within a few months to a year after that release, allowing programs that run on them to use the script. Once a script has been released (published) in the Unicode Standard, a Unicode-based font and keyboard are needed to be able to use them — but those might not be included in the operating systems. In addition to the keyboard, some complex scripts need input method engines.

For a more detailed description of the general process to bring a language and its script online, see the Zero to Digital guides on the Translation Commons page: https://translationcommons.org/programs/resources/. Other handy resources and materials are available from SIL: https://scripts.sil.org/default.

Fixing implementation problems for a script

QUESTION: The combining diacritics are not working reliably in implementations, rendering text unreadable for my language e.g. Kildin Sámi. How can this situation be rectified?

ANSWER: Unicode does what it can to provide documentation and other resources for implementations, but it doesn't impose any QA assessment on implementations. You should contact vendors for the fonts you use to report shortcomings for display of particular languages, such as Sámi.

Unicode projects rely heavily on volunteer contributions. Such work could certainly benefit smaller language communities, and it could be helpful if there were volunteers to drive the work. One resource related to this issue is documentation of what graphemes are used for different languages. The Unicode CLDR project collects many different kinds of information for supporting particular languages, one of these is lists of characters used. You can see an example for Norwegian here: https://www.unicode.org/cldr/charts/42/summary/no.html. CLDR doesn't appear to have information for Sámi, though; it would be great to have someone contribute data for Sámi.

How to gain the skills needed to support a script in software

QUESTION: I am from a community with no software that supports its script. How may I further develop the necessary skills to support my script and community?

ANSWER: There are many online resources available to help develop the skills needed to develop a Unicode proposal (if needed), keyboards, and fonts.

- For a general overview of the process of bringing a language online, the "Zero to Digital" online guidelines from Translation Commons may be useful: https://translationcommons.org/programs/resources/ . Other handy resources are available from SIL: https://scripts.sil.org/default

- When creating a new orthography for a community, you can refer to UTN #19 Recommendations for Creating New Orthographies https://www.unicode.org/notes/tn19/ and to the SIL paper https://www.sil.org/sites/default/files/best_practice_for_non-alphabetic_characters_v2_0.pdf

- For help on creating a Unicode proposal for a character not in Unicode, refer to the proposal templates (for a new script) https://www.unicode.org/L2/L2023/23105-new-script-template-april2023.pdf and (to add a character to an existing script) https://www.unicode.org/L2/L2023/23104-addl-script-template-april2023.pdf.

- For help on creating a keyboard, assistance is available from Translation Commons at: https://translationcommons.org/keyboard-resources/

- For additional information on keyboards, see the following:
  MSKLC: https://www.microsoft.com/en-us/download/details.aspx?id=102134
  SIL Keyman: https://keyman.com/developer/
  IME authoring: https://learn.microsoft.com/en-us/windows/apps/design/input/input-method-editors
  About Windows keyboards: https://kbdlayout.info/

Finding Unicode characters

QUESTION: How do I find the Unicode characters I need to write my language?

ANSWER: There are several different ways to find the characters you need. You can start with third-party applications, like Apple's Character Viewer and Window's Character Map, or check WIkipedia or Omniglot for the characters used by a given language.

Another approach is to look through the Unicode code chart for the specific script. (Code charts are accessible from: https://www.unicode.org/charts/.) However, if the characters are in a script like Latin and Arabic, which have many "blocks" of characters, this can be unwieldy. For Latin or Arabic, using a general search engine may be the most expedient way to find a character. Alternatively, some third-party apps can be useful, such as https://r12a.github.io/pickers/index.html.

For Chinese, Japanese, and Korean ideographs, it is best to use the Unihan Database Lookup: https://www.unicode.org/charts/unihan.html. For emoji, you can refer to resources on the Emoji page https://www.unicode.org/emoji/techindex.html or Emojipedia.

Shortcut for typing a Unicode character by code point

QUESTION: Can you show what shortcut we can use to type any Unicode character by using the code point, for example if we type in Word on a Mac?

ANSWER: Search "macros Unicode hex input" or, for Mac-specific input, "Unicode hex input on Mac" Another resource is: https://en.wikipedia.org/wiki/Unicode_input (For Unicode hex input on Mac, hold down option key and, while still holding down the option key, type the hex code point)

To type more than a few characters, a more practical approach is to use a character picker (such as Insert special characters in Google docs, Apple's Show Emoji & Symbols, or Insert Advanced Symbol in MW Word).

Proposing characters

QUESTION: What is the encoding process for new characters and scripts?

ANSWER: If a character or script is not in Unicode, and you believe it should be, a Unicode proposal is needed. The templates for the new character additions or new scripts will give you an idea of what is required for a proposal. It is often advisable to work with someone who participates on the development of the Unicode Standard or ask for assistance from the Script Encoding Initiative. Getting such feedback can save you considerable time. In some cases, the proposed characters can already be represented by Unicode characters or, for a new script, if not enough evidence is available, it may not be ready for encoding.

Once a proposal has been prepared with the required information, it is submitted to docsubmit@unicode.org and routed to the Unicode Script Ad Hoc (SAH). (Note: The Script Ad Hoc does not review requests for CJK ideographs or emoji; those proposals go to either the Unicode CJK & Unihan Group or the Emoji Subcommittee.)

If the character or script is deemed to be likely eligible for encoding, the proposal is reviewed by the Script Ad Hoc. The SAH group often has questions or comments requiring the proposal to be revised, typically several times, before all the required information is provided.

Once the SAH group agrees that all the necessary information has been provided, the proposal is passed on to a second group, the Properties and Algorithms Group (PAG) which makes sure the data is complete, accurate, and in alignment with other Unicode data.

Once the proposal has been reviewed and considered complete and accurate, a recommendation is made to the Unicode Technical Committee, which is the group that officially approves the character or script. The UTC will decide to approve the character (or not). The UTC will also eventually identify which version the character or script will be published in (visible on the Pipeline page).

Once a character (or script) has been published in the Unicode Standard, then Unicode-based fonts, keyboards, and software can be made available.

The process from the first Unicode proposal until publication is **at least two years,** often many more. The incorporation of newly approved characters and scripts into readily accessible fonts, keyboards and software can take additional time.

In order for modern scripts to be incorporated in software, such as showing the names of the days of the week, months, etc., in a given language on your device or computer, locale data is required. For further information, please see the Unicode CLDR project for information on how to submit such locale data.

QUESTION: My language is spoken by a community of Sikkim India, we used Tibetan script but there is one sign missing from Unicode. How can I propose it?

ANSWER: If you believe a Tibetan character used to represent the Bhutia language is missing in Unicode, please first review the FAQs on proposing new characters: https://www.unicode.org/faq/prop_new_characters.html and https://www.unicode.org/faq/char_proposal.html .

If you believe the character is still not representable in Unicode, please use the following proposal template: https://www.unicode.org/L2/L2023/23104-addl-script-template-april2023.pdf.

QUESTION: When submitting a proposal to add new characters to Unicode, do we need to define our own code points?

ANSWER: A proposal to add new characters does not need to provide code points, though a proposal can recommend code points. The proposal template offers a few comments regarding this topic: https://www.unicode.org/L2/L2023/23104-addl-script-template-april2023.pdf

QUESTION: I have created a font, keyboard, and instructions for my script. Can I submit my request to Unicode for the script, even if I don't yet have agreement from the communities in the countries who use this script?

ANSWER: Yes, any individual (or group) may submit a proposal for a script or request for characters. Proposal authors don't need to be native speakers of a language that uses that script. However, we strongly encourage people to work with such communities where possible.

Please be sure to review the script proposal template, which outlines information that is needed: https://www.unicode.org/L2/L2023/23105-new-script-template-april2023.pdf
Note that any script being proposed for the international standard must be free from any Intellectual Property protections.

Determining the correct characters for a language

QUESTION: How do I determine which are the correct Unicode characters to use for my language?

ANSWER: Ideally, having a guide that lists the Unicode characters for writing a given language would be very useful. For example, Unicode Technical Note #11 provides information on the Unicode characters used to write various languages that use the Myanmar script. Many languages, however, do not have such guidelines. Creating publicly accessible documentation on which Unicode characters to use for a given language would be a very practical and useful tool for others, though this may involve working with the language communities to arrive at consensus on which characters to use.

The following are other resources you can check:

- For many languages, the Unicode Common Locale Data Project lists exemplar characters, that is, the basic set of characters used for a language, based on feedback provided by contributors. (The exemplar characters listed are the Unicode characters.)

- A names list accompanies the Unicode code charts pages. The names list often provides brief annotations (which appear after a bullet). These annotations are meant to assist users in identifying some characters. If you believe an annotation would be useful to add – particularly in cases where there could be confusion regarding which character to use – you can suggest them on the Error Report: https://corp.unicode.org/reporting/error.html

- For those who are devising an orthography or who are trying to determine which Unicode character to use, the following article by SIL may be useful: Best practice when using non-alphabetic characters in orthographies. See also the Unicode Technical Note #14 "Recommendations for Creating New Orthographies"

Note that for the average user, having a Unicode-based keyboard and Unicode font available for one's language – which already provides the characters that should be used -- avoids having to worry about which Unicode characters to use.

Unicode's access, sharing and usage policies

QUESTION: What are the general data access, sharing, and usage policies for Unicode?

ANSWER: Unicode is a freely-accessible international standard. For details on Unicode's policies on usage and sharing, please refer to the following pages:

https://www.unicode.org/policies/licensing_policy.html

https://www.unicode.org/copyright.html

https://www.unicode.org/license.txt

https://www.unicode.org/policies/logo_policy.html

https://www.unicode.org/policies/

Where to submit questions

QUESTION: I have additional questions about Unicode. How can I submit them?

ANSWER: To ask a question relating to Unicode, please use the Unicode reporting form: https://www.unicode.org/reporting.html

How to get slides and video from the May 16 2023 webinar

QUESTION: Where can we get the slides and the recorded presentation from this Webinar?

ANSWER: The slides from the May 16, 2023 Unicode webinar presentation "Documenting and Preserving Languages with Unicode" are available at: Character Encoding and Fonts and Keyboards.

The video is now available at: https://www.youtube.com/watch?v=trDxfjgC5Ys&list=PLMc927ywQmTM2OCb6uhmMebeCl2xBK7Em&index=12

Do check out other webinars on the script and character encoding playlist, found on the Unicode YouTube channel: https://www.youtube.com/playlist?list=PLMc927ywQmTM2OCb6uhmMebeCl2xBK7Em