TRANSLATION COMMONS

2019 | INTERNATIONAL YEAR OF Indigenous Languages

INDIGENOUS LANGUAGES:

ZERO TO DIGITAL

A GUIDE TO BRING YOUR LANGUAGE ONLINE

DECEMBER 2019

Indigenous Languages: Zero to Digital

Authors: Deborah W. Anderson, Lee Collins, Craig Cornelius, Craig Cummings
Reviewers and contributors: Andrew Owen, Julia Nee, Lawrence Wolf-Sonkin, Anna Luisa Daigneault, Julie Anderson, Daniel Bogre Udell.
Design and Marketing: Mette Attar, Johanna Behm,
Project Coordination: Ester Perez, Jeannette Stewart

1. INTRODUCTION

Translation Commons is a nonprofit volunteer community that supports the digitization of languages, mentors language professionals, and provides courses and resources for the language industries.

One of the principal programs at Translation Commons is the Language Digitization Initiative (LDI), which seeks to bring digital capabilities to the language communities that desire them. Nearly 6,000 languages throughout the world have a small or nonexistent digital presence. The LDI provides a roadmap that a community can follow to achieve digitization of their language.

Translation Commons partnered with the 2019 International Year of Indigenous Languages, a UNESCO initiative, to focus more attention on Indigenous communities and digitization of their languages. Supporting equitable digital access to Indigenous and other minority languages is part of the LDI's mission to ensure that such language communities are able to participate in global online activities and have all the benefits of modern computer applications in their native language. Creating guidelines to equip communities with the tools and understanding to digitize their scripts and bring their languages to the internet gives them the knowledge to facilitate the process while maintaining their autonomy. In addition to guidelines, Translation Commons provides tutorials, workshops and assists communities with language digitization by introducing them to industry experts that guide them through the standardization process.

1.1 Zero to Digital Series

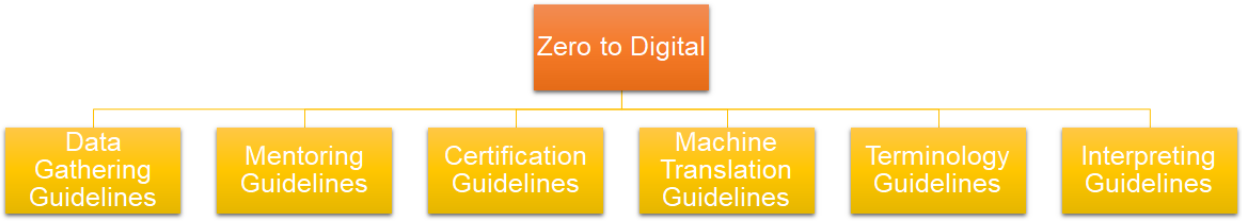
This document is one in a series of guidelines entitled *Zero to Digital*, which holistically addresses language digitization practices. The authors of the guidelines are experts in language technology and linguistics. The intended audience is any language community that wants the capability to use their language on digital systems.

Digitization expands a language community's avenues for communication. See the *Benefits of Language Digitization Appendix* for more detail on how digitizing a language benefits both Indigenous communities and the world at large.

To learn more about the language digitization process, see the Translation Commons' Resources web page provides additional LDI and related information, including guidelines, presentations, videos, and other documents.

Below, in Figure 1, you will find all guidelines created to assist Indigenous Communities.

Figure 1: Zero to Digital Series



This document in particular describes how to enable mobile and desktop software to support a written language. The recommended implementation allows native speakers to communicate online, share knowledge and documents, and to use software and devices that would otherwise be inaccessible to them.

The intended audiences of this document are:

- Indigenous communities wanting to make their language accessible on mobile devices and computers
- Technologists supporting the digitization of one or more languages
- Organizations wanting to enable language communities

This document aims to help you determine what tools you need and how to use them. It may also assist you in discovering the available tools for using your language online.

When people ask how to use their language on the Internet, there are many possible answers. It is helpful to understand that there are several levels of technology that must be considered when getting started with using a language online, both for web-based and mobile technology.

This document is about the use of written languages online by their speakers, readers, and creators. This includes ordinary conversation, text messages, email, social media, and blogs. The aim is to help people develop websites and a variety of content, fostering communication with local communities and language diaspora anywhere in the world. There may be multiple scripts or writing conventions for a given language. The intention is not to prescribe how people should use digital technologies, but to enable them to use the technologies described here in ways that enhance the prestige, public perception, and usability of indigenous languages worldwide. While formal documentation methods, grammars, and dictionaries are useful in linguistic studies and standardization, they are not a requirement to use your language online.

This multi-stakeholder partnership is made up of a Steering Committee to oversee implementation, ad hoc groups to provide relevant advice, and contributing partners.

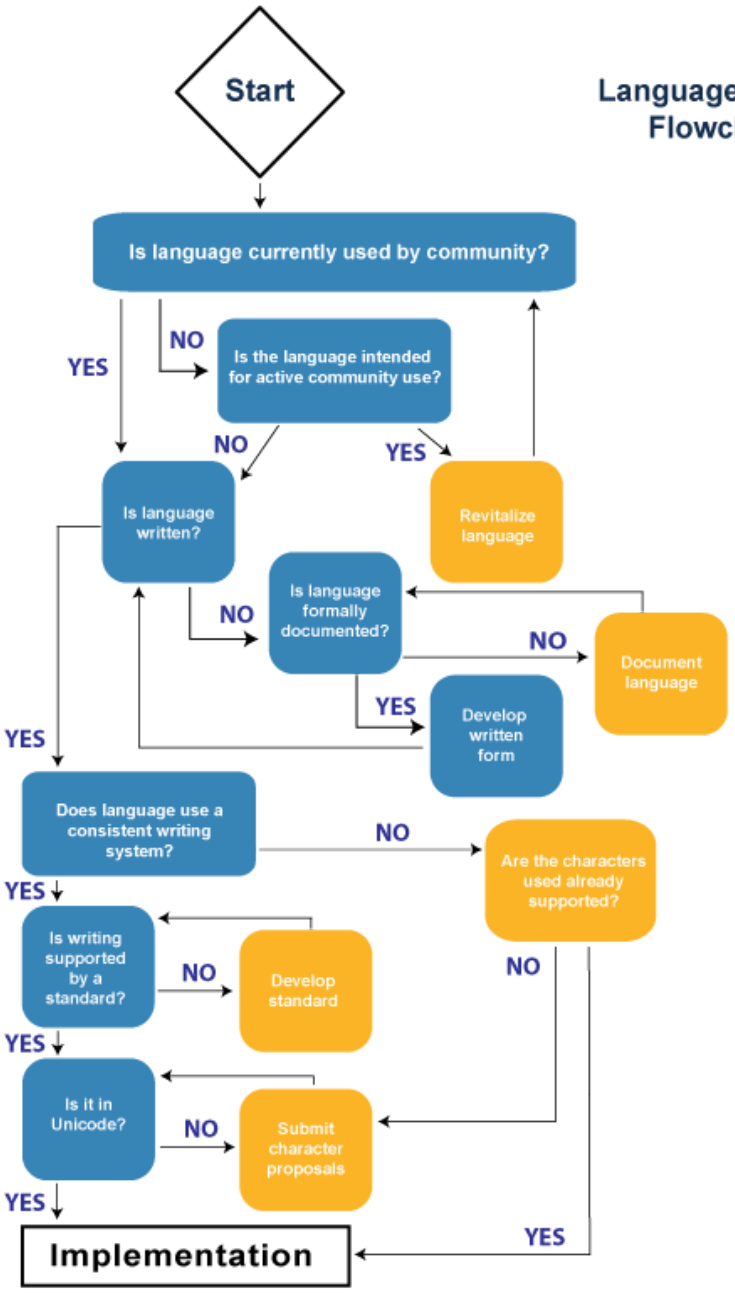
2. PROCESS OVERVIEW

This document provides two workflows. The first is used to determine the current status of the language. The second is used to develop technology solutions to enable digital use of the language. Steps in the flowcharts refer to sections in the document where you will find more detailed information. Steps are advisory only and some steps can be carried out concurrently.

2.1. Language Status Workflow

This workflow (Figure 2) describes the steps to determine the current status of a language in preparation for using it online.

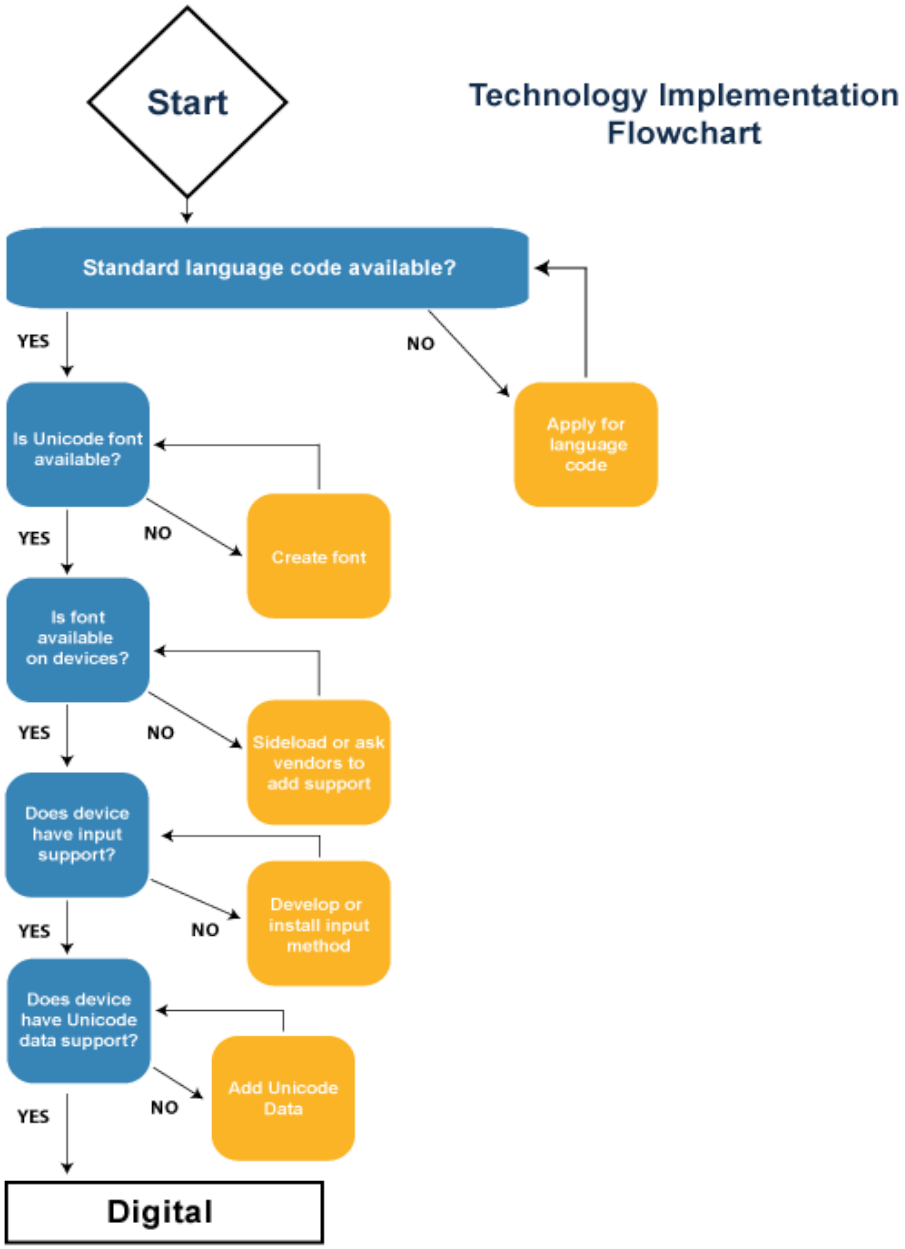
Figure 2: Determining a language status



2.2. Technology Implementation Workflow

After determining the status of your language, use this workflow (Figure 3) to get started with the technology available to take your language online.

Figure 3: Technology implementation



3. LANGUAGE STATUS

This section describes the steps in the language status workflow. The aim is to determine the current level of support for your language on computers and mobile devices. It includes suggested next steps to help you take your language online. You do not need a technical background to answer these questions.

3.1. Is the Language Currently Used by a Community?

Yes: Is language written?
No: Is language intended for active community use?

3.2. Is Language Intended for Active Community Use?

Yes: Revitalize language
No: Is language written?

3.2.1. Revitalize Language

With community commitment and organization, even languages with no current speakers can be revived. This involves many of the resources for language documentation, but also needs teaching materials, teachers to use them, and a strong effort by the community to maintain the effort over many years, perhaps decades.

The approach taken in New Zealand to revitalizing the Māori language, also known as Te Reo, was based in part on the approach taken in Wales to revitalizing Welsh. Indigenous language broadcast media, nest schools, and so on. A similar approach has also been taken with Irish. Other examples of the many projects worldwide include:

- **Hebrew**: An example of reviving a language after many centuries without anyone speaking it.
- **Cherokee**: Communities and governments are dedicating efforts from immersion schools for children, continuing education through elementary and high school, and adult training.
- **Mutsun**: The Amah Mutsun tribe in California has a language revitalization program for their language.
- **Chakma**: A language of Bangladesh and eastern India. Community is starting to use its historic script for education and literacy efforts.

- **Tunica:** the ancestral language of the Tunica-Biloxi Tribe became dormant in 1948 and has since been reawakened by heritage speakers training new fluent speakers and encouraging others to enroll in immersion courses.
- **Cornish:** Cornish was first revitalized in the early 1900s, but the movement to grow the language gained momentum in the 2000s, as Cornish speakers leveraged online forums to find one another and use the language on a daily basis.

3.3. Is Language in a Public Registry?

If the language is in a public registry or list of languages.

Yes: Is language written?
No: Standard language code available

3.4. Is Language Written?

Does your language have a written form?

Yes: Document language
No: Develop written form

3.4.1. Develop Written Form

Non-written languages are beyond the scope of this document. However, you can still use your language online with audio and video resources. There are organizations that offer guidance and tools for languages that are primarily oral. Academic linguistics departments are also good resources.

Audio and video resources are more useful when they can be found by users and researchers. Any recordings should include a language tag or code so that automatic indexing can find these data. Use standard language codes such as IETF BCP-47.

3.4.2. Document Language

Is your language formally documented, for example, with a grammar, a dictionary, or linguistic study?

Yes: Language is documented
No: Language is not documented

3.4.2.1. Language is Documented

Where dictionaries, grammars and other linguistic information exist, are they accessible to the users of the language? Consider options to make this information more widely available and useful to communities. This could include building online resources, educational materials suitable for immersion schools and primary education. Work to assign copyrights of such material to the language communities. Obtain or create digital formats of such information so it can be converted to online form as the communities see fit.

Where these resources are not accessible, take steps to share books, online repositories, and other language information with the communities that use the language.

3.4.2.2. Language is Not Documented

If the language is not yet documented, it may still be possible to use the language in oral or written form. However, support in the form of spelling suggestions, search, and predictive text will be limited.

3.5. Does Language Use a Consistent Writing System?

A writing system is a set of rules for using one or more scripts to write a particular language. Many languages are written with more than one script, for example Serbo-Croatian is written in both Cyrillic and Latin characters by different communities. Also, most writing systems are used by more than one language. For example, Burmese, Shan, Mon, and other languages can all be written in the Myanmar script.

Does the language use at least one writing system in a consistent way, including spelling?

Yes: Is writing supported by a standard?
No: Are the characters used already supported?

3.5.1. Are the Characters Used Already Supported?

If there is no constant writing system, writing may be informal, not using consistent spelling or grammar (or even using more than one script). Although such text can be created and displayed by existing tools, spelling and grammar tools will have limited utility. In situations where there are multiple competing writing systems, it is important

for outsiders to realize that those orthographies may represent different interests in the language community.

You may require help from linguists and the technical community:

- Many communities use various orthographies and character sets for writing languages, including variations in dialects. Developers of language tools such as keyboards should be aware of this, and provide more than one set of characters, diacritic marks and spelling suggestions to attempt to meet the needs of all community groups.
- The technical community should consider methods to identify such variations, make them discoverable, and possibly develop ways to facilitate conversion between such variants, as needed by community members.
- Where the characters used in writing are already supported on computers in some form, the technical community can work with the language community to identify and develop fonts and keyboards.
- Where the inconsistency is a barrier to using the language, local educators, policy makers, language community leaders, linguists, etc., can help to adopt a more consistent written form for online use.

3.6. Is Writing Supported by a Standard?

Is the writing system supported by commonly accepted orthography and set of grammar rules, either formal or informal?

Even without a standard, informal writing and communications are possible. However, non-standard spelling, grammar, vocabulary, or large regional variations may make more advanced usage difficult for social media, sharing documents, and finding websites and information with online search and other tools.

Yes: Submit character proposals
No: Develop standard

3.6.1. Submit Character Proposals

Shared spelling, grammar, punctuation, and other aspects of written language can greatly improve the ability of community members to utilize available online services and tools such as typing suggestions, predictive text, web search, and writing tools.

The Unicode Standard specifies scripts rather than languages. A script, such as Latin, can be used for more than one language, for example English, Swahili, Indonesian and so on. If the characters of your language are supported you can proceed to the Language Technology workflow.

When the writing system is available and in use, but its characters are not yet in the Unicode Standard, it is useful to standardize the characters so that the language can be used on connected devices, including mobile, laptop, and desktop computers. Without standardization, it is still possible to share files among users with the same fonts and input methods. However, online tools and services will be limited.

To achieve Unicode standardization:

1. Check if the characters are included or supported in Unicode.
2. If the characters are missing, prepare and submit a proposal.

3.6.2. Develop Standard

Although it is not a necessary step and languages may have different forms of writing, when formalization fits the community's current goal, you should develop guidelines for spelling, punctuation, and grammar that enable people to communicate their ideas digitally. This also improves the ability of keyboards to propose word choices and spell correction. In addition, it enables search engines and other online services and tools to return more useful and relevant results.

3.7. Proceed to Implementation

Character support is available in Unicode. The characters of the writing system are found in the supported scripts. You can now proceed to the Language Technology workflow.

4. LANGUAGE TECHNOLOGY IMPLEMENTATION WORKFLOW

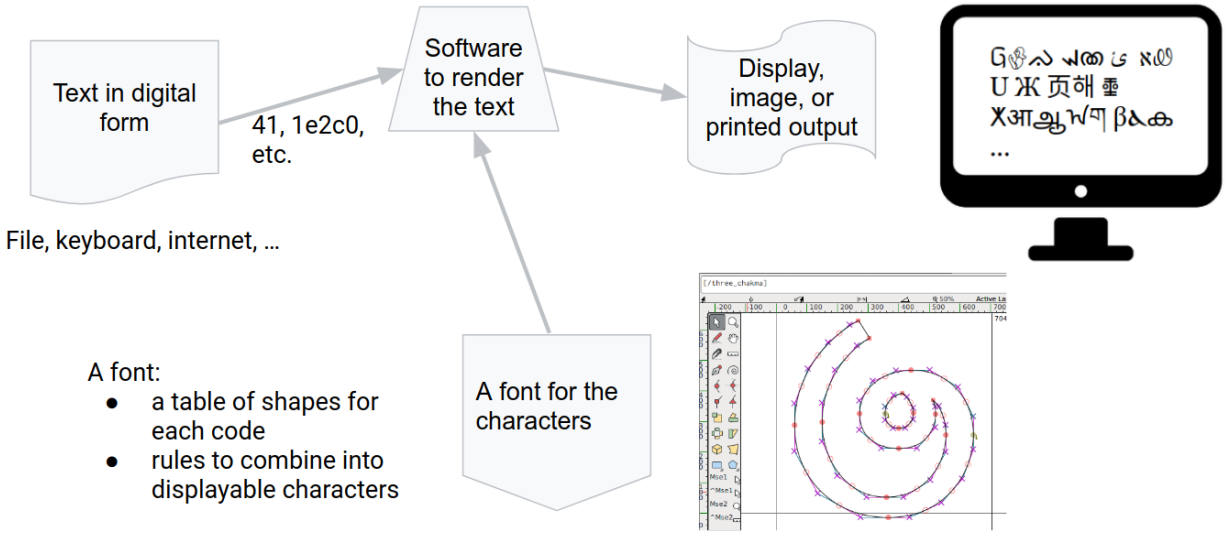
4.1 Note on Technology for Text in Digital Systems

Text in digital devices is stored as patterns of bits used as *code point*. For example, the code point U+0041 could represent "A". This code can be created by a keyboard, stored in a file or transmitted to another application, and is displayed or *rendered* by systems that understand the coding system used for characters.

An *encoding standard* assigns a meaning to each of the possible codes. For example, ISO-8859-1 defines characters for the range 0 to 255, providing a meaning for each of these code points. Unicode is a standard for characters that includes many writing systems, assigning a unique value for each character of more than 150 different scripts, including Latin, Cyrillic, Chinese, Arabic, Hebrew, Devanagari, Tamil, Myanmar, and many others.

When used together, text created with a defined encoding can be created with a compatible input method and displayed using a compatible font. This process is shown in the diagram below (Figure 4), illustrating that a scheme such as Unicode may display characters of many languages.

Figure 4: Unicode displays characters of many languages



Many writing systems use base characters with optional diacritic or modifier marks, such as “e” with an acute accent, resulting in the output “é”. An encoding may define that base characters can be followed by combining diacritic codes to produce the combined character. In some cases, an encoding may include precomposed characters, using single code points that incorporate both the base and combining code points.

Any written language used in digital devices must therefore include:

- A standard encoding system for characters, for example Unicode,
- Rendering systems plus fonts for characters used in the language, and
- Methods or applications to render the code points to desired media.

In this document, we emphasize the Unicode Standard because it is used in all modern digital devices. It is also the default encoding for many sources of text. Unicode simplifies much of the effort required to bring a language online.

4.2. Definitions for implementing digital support

This section aims to help you implement digital support for your written language. Many languages use non-standardized fonts and scripts. This does not prevent the use of the

language online, but it can limit the degree of support that is available. The following steps describe aspects of digital support for written languages on the Internet.

The basic requirements for a digital writing system include:

- **Characters**: Identify the set of letters, diacritics, ligatures, punctuation, numbers, ideographs, and other symbols that are used in the language's orthography.
- **ISO identifier**: For language and script. This can include optional regional identifiers, for example *es-MX* for Spanish as used in Mexico.
- **Encoding Standard**: Either a formal standard (Unicode with collation, sorting, and character combinations) or an informal standard (font encoding).
- **Font support**: The information required to create a typeface that includes all the required character shapes:
    - Details of text presentation such as joining, ligatures, combining groups (this may already be provided in Unicode documents).
    - Sample text for font designers and developers to use for testing. Updates may be required to rendering engines or other software.
- **Input method**: A way of entering the characters for all devices. Typically a physical or on-screen keyboard.

Additional language resources are available:

- Common Locale Data Repository (CLDR) which includes additional information about a language such as calendars.
- Language code such as IETF BCP-47.
- Software application support for word segmentation:
    - Dictionaries of words written without explicit breaks.
    - Rules, for example spaces, punctuation.

4.3. Standard language code available?

Is a standard language code available and agreed upon?

Yes: Is Unicode font available?
No: Apply for language code

4.3.1. Apply for language code

If no language code is available you will need to establish a standard language code with optional regional variant, script or both. Non-standard language codes and tags can lead to confusion and mistaken identification of language resources. Follow the guidelines of ISO-639 and IETF BCP-47.

4.4. Is Unicode font available?

Is a Unicode-compatible font available to community members for desktop or laptop use?

Yes: Is font available on devices?
No: Create font

4.4.1. Create font

Initially after a script is added to Unicode, there may not yet be any Unicode fonts to support the characters. In this case, you may need to engage with font designers and the technical community to create such fonts to support text in the script.

4.5. Is font Available on Devices?

Is the font available on mobile and other devices available to the community? If so, text on websites, social media, and other applications will be readable.

It is important to note that most computers and mobile devices will work best with Unicode text, compatible-fonts, and input methods that produce Unicode text.
Check if the font is provided as part of Noto fonts or other websites for downloading and using Unicode fonts. Note that many mobile devices may not directly support font installation or download. Desktops and laptops allow installation of downloadable fonts.

Some applications such as word-processors and browsers may require a configuration step for the application to present the font in your document or web page.

Web fonts can be used by websites to present text using a font supplied by the web site. This enables content creators to specify the particular font used, and also makes the site readable on devices, even if the font is not already installed on the device. Note that a web font works only with pages that specifically define it, and such use does not permanently install a font on a device. The latest Noto fonts are available as web fonts, which is useful if the device does not already have the relevant version of the font or it cannot be installed.

Yes: Does device have input support?
No: Manual install or ask vendors to add support

4.5.1. Manual Install or Ask Vendors for Support

Recently standardized writing systems may not yet be included in mobile devices, even when a font is otherwise available. Urge device vendors to include the required fonts for the language.

An alternative is to manually install a Unicode font on the mobile device. This can be effective, but often requires specialized knowledge.

Caution: Installing a font from the web on a mobile device may introduce security and privacy risks. Also, device warranty or support agreements may become invalid if such an installation is performed.

4.6. Does the Device Have Input Support?

Users of a language will need an input method to effectively create messages, emails, blog posts, or other content in that language. Does the device (mobile or otherwise) natively support a way to enter the language?

Yes: Is input supported by third party apps or devices?
No: Develop input method

4.7. Is input Supported by Third Party Apps or Devices?

Is input support built into the system's standard keyboards? If so, learn how to enable or install keyboard support. This is device specific, but online resources can help with this procedure.

Vendors of mobile devices offer many existing keyboards on iOS and Android. For example:

- Gboard, the Google Keyboard for mobile (https://support.google.com/gboard/answer/6380730?hl=en&co=GENIE.Platform=Android) has support for many languages on both iOS and Android.
- Google Input Tools (https://www.google.com/inputtools/) has virtual keyboards for use in web pages in the Chrome browser on computers (not mobile).
- Many third party applications for text input are available in Apple's App Store and Google Play.

Note that many languages already have keyboard layouts that are publicly described, for example in the CLDR Keyboards repository.

4.7.1. Develop Input Method

There are numerous tools to install input methods for many languages, and also to develop new keyboards. These include:

- Keyboard applications available from the App Store, Google Play, or other sources.
- Keyman: a tool for language keyboards (https://keyman.com/).
- Microsoft Keyboard Builder (https://www.microsoft.com/en-us/download/details.aspx?id=22339).
- SIL Ukelele (https://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=ukelele).

Caution: Installing a keyboard or other input method from the web may introduce security and privacy risks.

4.8. Does the Device Have Unicode Data Support?

Is Unicode data available?

Yes: Proceed with using your language.
No: If needed, consider adding essential language data to CLDR. These include the language tag and the name of the language. See more about CLDR under Additional Language Support.

5. ADDITIONAL LANGUAGE SUPPORT

Unicode standardization covers many of the text processing aspects of a writing system and language. With font and input support, many things will just work. This includes most aspects of basic word processing, spreadsheets, and email.

The steps described here are only the beginning. The Common Locale Data Repository (CLDR) provides “key building blocks for software to support the world's languages” by collecting useful information for different locales (language and country). This data can provide the names of languages, countries, months, weekdays, and other information. It also enables locale-aware formatting of date, time, numbers, and other commonly formatted information.

Although CLDR data is not required for basic text communication in indigenous languages, this information enhances language functionality. Almost all tools such as email, texting, social media, and so on will work well when fonts and keyboard are present.

This information is used by programmers to create output for online applications in specific contexts such as localized calendars, spreadsheets, numeric output, menu selections, and other user interface contexts.

CLDR also stores additional language information such as the characters used in writing and keyboard layouts for text entry. See CLDR Keyboards for more information.

However, additional data and text processing tools may be needed for more complete language support in online and mobile applications. The following are some things needed for parity of indigenous languages with fully supported languages:

- **Segmentation and word breaking**: Separation of text is needed for correct layout of text, and also user selection of grapheme clusters, words, and sentences. For many languages, no explicit signal is provided to indicate word boundaries, for example spaces or punctuation. In such cases, dictionary data or algorithms are needed to provide such information. For more information, see: userguide.icu-project.org/boundaryanalysis.
- **Line breaking**: Languages have a variety of rules for positions where text can be interrupted to move to a new line. For example, quoted text is surrounded with various language-specific characters that follow different rules for breaking between words and sentences. In addition, positions where a line of text can be broken depends on the properties of the Unicode characters in the writing system. Also, numbers and currencies such as $10 may need to be glued together to prevent misunderstanding of the meaning. The rules of punctuation vary both by the language and the region. Line breaking is a specific case of Unicode Boundary Analysis (userguide.icu-project.org/boundaryanalysis#TOC-Line-break-Boundary).
- **Identifying language of textual content:** Documents should be explicitly labeled with a language code or other identifier that describes the human language of the document. When such information is available in a document, tools may use this to find appropriate information for users more effectively. In a multi-language document, individual sections or even paragraphs may be tagged with the language of the text. Use of standard tags such as IETF BCP-47 (https://tools.ietf.org/html/bcp47) is especially important:
    - The mechanism for such identification varies among applications, and the user may need to learn if and how this is done. Note that such identification may provide identifiers from a list rather than supporting any possible tag
    - For online documents and websites, HTML provides the *lang* attribute (https://www.w3schools.com/tags/att_global_lang.asp) to explicitly label the language of any HTML component. The value of this attribute should be taken from standard sets of language identifiers rather than an arbitrary user-defined string. For example, use *rs* for Serbian, *de* for German, *zh-Hans* or *zh-CN* (or simply *zh*) for Chinese with simplified script.

- **Language detection**: Online services and other applications can often give more relevant and useful results when the language of text is known. For text that is not explicitly tagged, language detectors such as cld2 (https://github.com/optimaize/language-detector) have been developed. Such tools typically perform a statistical analysis on characters of the text, providing a probable identification of the human language. This is required because most writing systems are used for multiple languages, for example Latin letters for Swahili, Lakota, Warlpiri, Finnish; Cyrllic for Russian, Ukranian, Khazah; Myanmar for Burmese, Shan, Mon, and so on.
- **Dictionaries for word processing**: Most word processing applications support basic text creation, editing, sharing, and printing. Predictive text, spell correction, grammar suggestions, and other tools employ word lists with occurrence frequency data, dictionaries, and other linguistic data. Synonyms and commonly used idioms are also useful in tools such as online search.
- **Non-ASCII digits**: Many scripts have digits distinct from the western digits. Examples include Myanmar, Adlam, Arabic, and Farsi. However, many applications such as spreadsheets do not interpret these digits as numeric values but rather as textual values. Implementers of such applications may consider Unicode properties of such characters in order to process them as numerals, but this support is not implemented consistently (https://en.wikipedia.org/wiki/Numerals_in_Unicode).
- **Translated user interfaces**: In some applications, especially for education or information specifically in the user's language, it may be useful to translate text that appears in the user interface (UI). For example, the menu items for the operating system functions may be translated, such as "Start" or "Open file". In many situations, however, it is not feasible for the owner of the application to provide translations for small language communities. When user interface is available in at least one of the languages that a user can understand, a translated interface is of less immediate value.
- **Optical Character Recognition (OCR):** Many languages have a substantial written literature in books and other documents. OCR can be used to convert the text in such documents into digital form. Open source OCR projects are available (https://pdf.iskysoft.com/ocr-pdf/open-source-ocr.html), and can be trained on new writing systems. It is important to note that a language model with lists of common words greatly improves the accuracy of OCR methods.

5.1. Public Language Resources

There are many available resources, free and on subscription that communities can use:

5.1.1. Language Resources

- Panlex (https://panlex.org/)

- Wikitongues (https://wikitongues.org/)

5.1.2. Tool Resources

- SIL - Keyman
- Font tools
- Open source tools for dictionaries

5.2. Advanced Language Technology

The following capabilities require very large amounts of data for training machine learning systems. Open source, public software is starting to become available, but most work in this field is in academic research or corporate products. Most of these features are unlikely to be available to most languages in the near future.

- **Speech To Text**: Recognizing words spoken by a person, and converting the sounds to text of the spoken message. Such text can be transcribed into documents or may be used for control of applications and devices.
- **Text to Speech**: Producing natural speech output from text. This is useful for hands free interfaces, and for machine reading to humans from textual sources.
- **Transcription of Academic Audio Media**: Important for language documentation, especially in academic linguistic studies. Some open source projects are beginning to address this need, including:
- **Accelerated Transcription for Linguists**: https://github.com/CoEDL/elpis
- **Machine Translation**: Converting text in one human language to another is one of the hardest tasks for computers. Current systems can translate among a limited set of supported languages. However, such systems generally do not understand context and are not at the level of human translation. With new machine learning techniques, however, the quality and reliability of machine translation is increasing rapidly for those with large corpora. Machine translation support for indigenous languages is not widely available, but open source and university efforts are starting to appear. For example, www.apertium.org is an online tool supporting machine translation efforts for non-dominant languages.

6. GLOSSARY

ASCII: American Standard Code for Information Interchange: A character encoding for electronic communication.

BCP-47: An *IETF* tag to identify languages.

CLDR: Common Locale Data Repository. Contains additional language information.

Character: The smallest component of written language that has semantic value; refers to the abstract meaning and/or shape, rather than a specific shape. ref: Unicode Glossary

Code Point: a number that represents a specific character or formatting.

Diacritic: a character added to a letter or basic glyph, typically to modify its sound or semantic meaning.

Dormant Language: a language with no current speakers.

Font: A collection of glyphs used for the visual depiction of character data

Glyph: one of a set of symbols representing a readable character for the purpose of writing.

Grammar: the rules governing the composition of a *natural language*.

IETF: Internet Engineering Task Force.

Indigenous Language: a language that is native to a specific region.

IYIL2019: 2019 International Year of Indigenous Languages.

Language Revitalization: reversing the decline of a language or reviving a *dormant language*.

Ligature: a combination of two or more glyphs into a single glyph.

Natural Language: a language that has evolved naturally in humans as distinct from formal languages such as those used in computing. Includes oral, visual, visual-manual (signed) and written languages.

Noto: A font family comprising over 100 individual fonts, which together are intended to cover all the scripts encoded in Unicode (currently covering all scripts existing in Unicode 6.0 and earlier).

Orthography: a set of conventions for writing a language.

PUA: Private Use Area: a range of *code points* in *Unicode* that will never have characters assigned to them.

Punctuation: spacing and marks that are not sounded that aid the understanding of a text.

Script: A collection of letters and other written signs used to represent textual information in one or more writing systems

Translation Commons: an online community and platform for freely sharing linguistic knowledge.

UNESCO: United Nations Educational, Scientific and Cultural Organization.

Unicode: the most commonly used standard for digitally encoding the characters of the world's writing systems.

Writing system: A set of rules for using one or more scripts to write a particular language.

7. REFERENCES

7.1. Language Revitalization

Routledge Handbook of Language Revitalization (Hinton, Huss, & Roche 2018)
The Green Book of Language Revitalization in Practice (Hinton & Hale 2001)
Language Documentation and Revitalization in Latin American Contexts (Perez-Baez, Rogers, & Roses Labrada 2016)
Developing Orthographies for Unwritten Languages (Cahill & Rice 2014)
http://cherokeepreservation.org/what-we-do/cultural-preservation/cherokee-language/
https://language.cherokee.org/
http://amahmutsun.org/language
https://rising.globalvoices.org/blog/2011/11/29/languages-online-activism-to-save-chakma-language/
https://www.languageconservancy.org/programs/indigenous-language-program-support/

7.2. Language Registries

https://www.ethnologue.com/
https://glottolog.org/

7.3. Unicode and Font Encoding

https://unicode.org/main.html
https://unicode.org/standard/supported.html
https://unicode.org/standard/where/
https://unicode.org/pending/proposals.htm
https://unicode.org/glossary/
https://linguistics.berkeley.edu/sei/

Specialized fonts can use non-standardized approaches for characters, such as Private Use Areas (PUAs) or other character ranges such as ASCII or Arabic with customized glyphs for code points. This is called font encoding. Such non-standard approaches enable users to see the characters they type, though others will not, unless they use the

same fonts. Online services and tools will also not be able to correctly interpret text in such font encodings, because the underlying encoding does not contain information on the actual intended meaning of the character.

While the text that is based on Unicode has advantages over font-encoded text, it may be necessary to use specialized fonts until the characters in a writing system are included in the Unicode Standard. In this case, a font should use only PUAs of the Unicode range instead of reusing code values that are already reserved for other scripts. This prevents overlapping code values and enables easier use of existing scripts. Converting such a font to Unicode, after a script is standardized, is relatively easy for PUA codes, given that the PUA codes themselves are used consistently.

7.4. Language Codes

https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes

7.5. Fonts

https://www.google.com/get/noto/

Tools for developing fonts include:
- FontForge
- FontLab
- Glyphs
- BirdFont (https://birdfont.org/)

The University of Reading has a masters program in font design (typefacedesign.net/) whose students may be able to assist in creating a new Unicode font.

Commercial font developers can create new Unicode fonts.

8. NOTES

The scope of this document is limited to written languages. Other forms of communication include:

- Emojis
- Oral languages
- Visual-manual (signed) languages
- Visual languages

Languages with their own script may also be written with other scripts. For example:

- Turkish was originally written with Arabic script but is now written with Latin script.
- Chinese can be written with Latin script (pinyin).

This document does not cover language dialects, but these are worth considering when proposing language standards.

High quality language search has its own specific requirements:

- Language identification (from text)
- Segmentation: breaking text into words
- Determining the *stems* of words, as in a language with various forms, for example: *housing* and *houses* to *house*.

Although oral languages are beyond the scope of this document, the following resources may be of use:

- Google Earth (https://docs.google.com/forms/d/e/1FAIpQLSdphaDaz33syPoUDyTOTwwkaLWZx90zopUklha4uadfkUKG8A/viewform)
- https://www.blog.google/products/earth/indigenous-speakers-share-their-languages-google-earth/
- https://www.gerlingo.com/
- XTrans (https://www.ldc.upenn.edu/language-resources/tools/xtrans) is a next generation multi-platform, multilingual, multi-channel transcription tool that supports manual transcription and annotation of audio recordings.