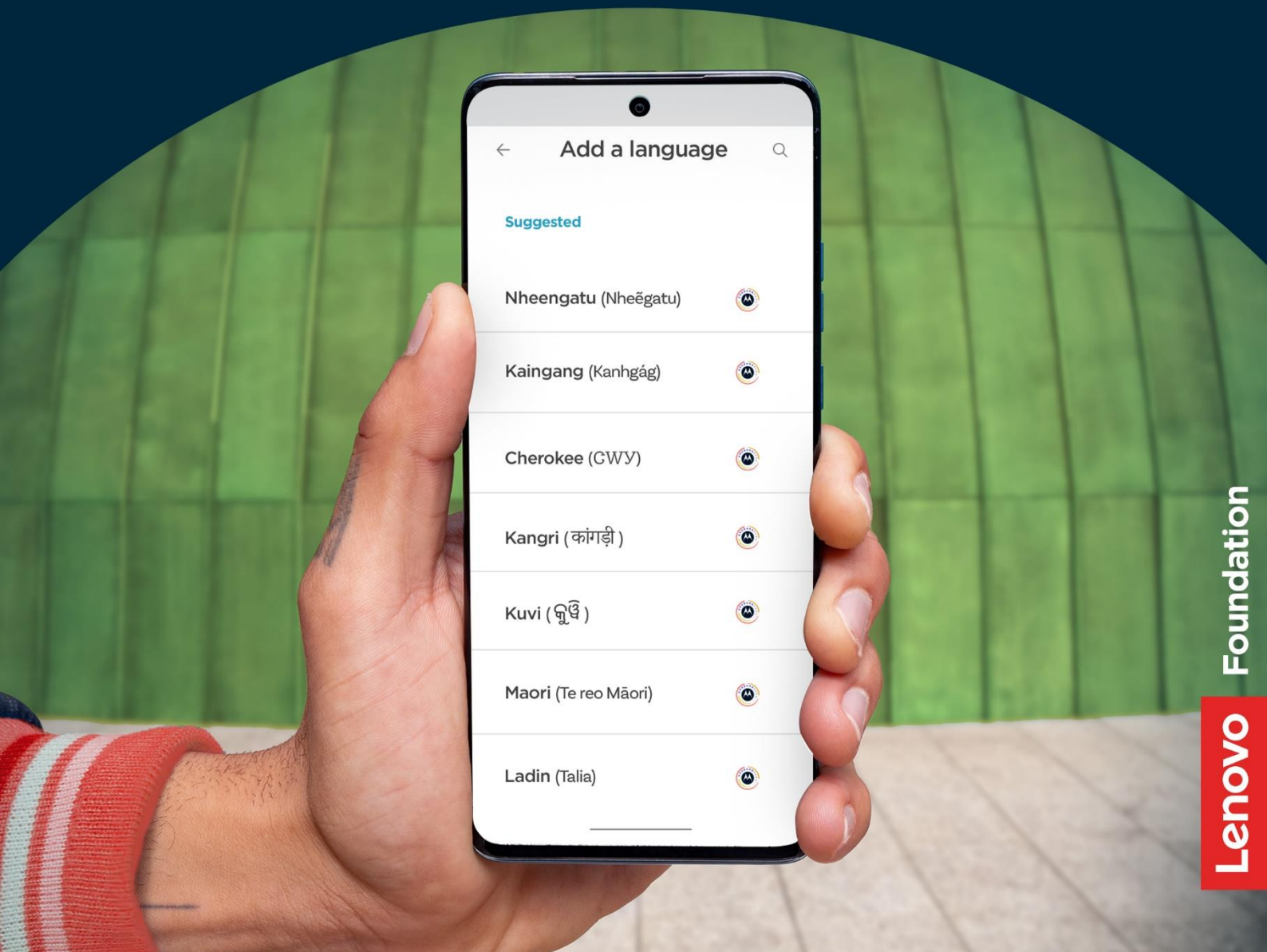
motorola

hello indigenous

A Blueprint on the Preservation of Endangered Indigenous Languages Through Digital Inclusion

PUBLISHED IN 2024



Lenovo Foundation

This publication has the cooperation of UNESCO regarding the “Indigenous Languages on Mobile” project, which aims to produce a study and an executive summary on the technological development that made possible the insertion of Indigenous Languages present in Brazilian, North American, European and Asia-Pacific territories into the donor’s devices. The designations employed in this publication and its presentation do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area of its authorities, or concerning the delimitation of its frontiers or boundaries. The publication expresses the opinions of the authors and do not necessarily represent the opinions of UNESCO and do not commit the Organization.

Authors

Janine Oliveira (Author, Supervisor)

Marison Ranieri Rodrigues de Freitas (Author, Compiler)

Delaney Gomez-Jackson (Author, Researcher)

Juliana Peres Rebelatto Pereira

Natalia Sarmento Tenório Falcão

Roy Yokoyama

Sushil Garg

Yukitomi Fujinaga

Reviewed by

Guilherme Saba (Reviewer, Diagrammer)

Livia Teixeira

Marcelo Mazzini

Natàlia Borja Gutés

Contribution by

Adauto Candido Soares

Jaco Du Toit

Hezekiel Dlamini

Maria Luzia de Cerqueira Gomes

Luciana Vedovato

Manuella Foz

Mahmoud Ebrahim

Monica Hauser

Alice Damasceno Saiki

Pratima Harite

Santiago Mendez Galvis

Sydni Behm

Cover design by

Giovanna Paganini

ISBN 979-8-218-37915-5

Table of contents

1. Digital Inclusion: endangered Indigenous language initiative goals and non-goals

1.1 Background on Indigenous languages in the world

An Indigenous language is a language that is native to a region and spoken by Indigenous peoples – they are frequently reduced to the status of minority languages in their respective regions because they are spoken by small-scale groups. Thus, these languages are not generally national languages (although they can be, such as Aymara, an official language in Bolivia). Indigenous languages that are not recognized by their governments often do not receive support for language education and revitalization. In fact, there is a correlation between Indigenous languages and endangered languages, insofar as Indigenous communities usually assimilate to speak the majority language of their region.

According to the Department of Economic and Social Affairs of the United Nations (UN DESA) report, 40% of the 6,700 languages spoken worldwide are in danger of disappearing.[1] Many of these are Indigenous languages, which play an essential role in various facets of Indigenous cultures, such as defining Indigenous relationships with the Earth, preserving Indigenous territory, and transmitting Indigenous history, science, and general worldviews. Although indigenous people make up 6% of the global population, they speak more than 4,000 of the world's languages.[2] Many Indigenous languages are facing the danger of disappearing in the near future due to various factors such as colonization, globalization, assimilation and discrimination. In "The World's Languages in Crisis," it is predicted that half of the world's languages will be lost during this century.[3] Consequently, efforts to engage in language revitalization projects are essential to the preservation of Indigenous languages and cultures.

1.2 Initiative goals and non-goals

This document is a detailed report of best practices for the Digital Inclusion of speakers of endangered Indigenous languages and is targeted to civil society, encompassing, as an example, private individuals, non-governmental organizations, and companies. The civil society plays a pivotal role in bringing unique perspectives to the table, driving social change, and advocating beyond profit-making, through ethical business practices. Together, this collective force fosters collaboration, inclusivity, and empowerment to build a more inclusive and prosperous future for all.

The proclamation of the period between 2022 and 2032 as the International Decade of Indigenous Languages (IDIL 2022-2032), as per Resolution A/RES/74/135 from United Nations General Assembly, shines a spotlight on the issue and invites the world to pay more attention to the critical and fragile situation of many Indigenous languages. This document aims to provide a blueprint for the addition of Indigenous languages into software in order to aid Digital Inclusion efforts and to provide a set of recommended steps to follow in order to successfully achieve it.

The Digital Inclusion initiative presented in this document is primarily, but not exclusively, focused on Android on a technical level. For other operating systems, there may be additional requirements and standards, therefore it is recommended to refer to related documentation.

Goals

As new generations of Indigenous people increase their literacy and use of technology, it is crucial that they are able to use their native language in digital formats to avoid the endangerment and loss of the language.[4] UNESCO estimates that we lose one Indigenous language every two weeks, resulting in around 3,000 unique languages being lost by the end of the century.[5] To help preserve our human heritage and the unique histories of Indigenous cultures, as well as empower the next generation, this Digital Inclusion initiative aims to integrate endangered Indigenous languages into smartphones.

As the Digital Inclusion initiative continues over the next decade, its main goal is to serve the

communities through raising awareness of endangered Indigenous languages. The Digital Inclusion initiative delivery may also serve and address the needs of Indigenous peoples since the work allows easier access to technology and also brings action toward the survival of endangered languages. Finally, it may help to empower future generations of Indigenous communities to use technology in their native language.

Kaingang (spoken in Southern Brazil), Nheengatu (spoken in the Amazon), Cherokee (spoken in the United States), Kangri (spoken in India), Te reo Māori (spoken in New Zealand and Australia), and Ladin (spoken in Italy) are now part of the more than 90 languages offered in Motorola mobile user interface. Additionally, a virtual keyboard in Kuvi (spoken in India) was developed and made available for download in smartphones.

The main goals of this initiative involve promoting the written form of a language in a natural vehicle such as technology, bringing awareness to the endangered Indigenous languages, and working towards their revitalization. Motorola has open-sourced over 800,000 translated Indigenous words as of April 2023 through its official website, enabling other OEMs (Original Design Manufacturers) and companies to promote the languages through their interfaces and paving the way for broader use and revitalization efforts. As a start, in 2022, Motorola's parent company Lenovo has deployed the integration of Latin American languages Nhengatu and Kaingang on its PCs; additionally, Gboard, Android's native virtual keyboard, now supports the two languages for broader usage.

Non-goals

Knowing there are over 3,000 Indigenous languages spoken in the world, in order to prioritize languages that are at higher risk of endangerment while working towards achieving the goal, the Digital Inclusion initiative driven by Motorola primarily focuses on the UNESCO categories "Definitely endangered," "Critically endangered," or "Severely endangered."

2. Language selection criteria

2.1 UNESCO categorization

The UNESCO levels of language endangerment provide classifications regarding how threatened a language is based on intergenerational transfer. There are five levels[6] (excluding a non-endangered status) and a brief summary of each level is as follows:[7]

- Vulnerable: Most children speak the language, but it may be restricted to certain domains, e.g. at home.
- Definitely endangered: Children no longer learn the language as their mother tongue in their home.
- Severely endangered: The language is spoken by grandparents and older generations; while the parent generation may understand it, they do not speak it to children or among themselves.
- Critically endangered: The youngest speakers are grandparents and older, and they speak the language partially and infrequently.
- Extinct: There are no speakers left.

Motorola's Digital Inclusion strategy focuses primarily on three categories: Definitely endangered, Severely endangered and Critically endangered (although other endangered languages are also considered).

2.2 How languages are selected

The criteria for selecting endangered languages for Motorola's Digital inclusion initiative is buttressed by four main pillars, to be expanded below: Language endangerment status, Digital Inclusion status, Community input and feeling towards the initiative, and Availability of subject matter experts (SMEs).

Language endangerment status is the first criterion for determining the inclusion and can be assessed using hard data, primarily using the vast dataset provided by UNESCO activities, and acts as a "thermometer" to evaluate the risk of a certain language and thus the eligibility into the digital inclusion initiative.

Digital Inclusion status is a measure of how well a language is represented and supported in the digital world. It depends on various factors, such as the availability of digital resources, tools, and services for the language, the level of access and participation of the speakers, and the degree of recognition and protection of the language rights and diversity.

One of the factors is whether the alphabet of the language is supported by Unicode, which is a standard for character representation that covers most of the world's writing systems.[8] Unicode supports not only the Latin alphabet but also the Greek, Cyrillic, Arabic, Hebrew, and Thai alphabets, as well as Japanese (Katakana, Hiragana), Chinese, and Korean (Hangul) writing systems, for example. In addition, there are also mathematical, commercial, and technical characters, and historical control characters for teleprinters. However, some languages may not have their alphabets encoded in Unicode yet, which is explored in further detail in §5.

Another factor is whether there is locale data available for the language, which is a set of parameters that defines the user's language, region, and cultural preferences. Locale data can affect how dates, times, numbers, currencies, and other formats are displayed, as well as how text is sorted and searched. Locale data can be provided by the operating system, the programming language, or the application. Depending on the peculiarities of a language and its endangerment status, an effort may be made to create a keyboard and layout and implement it in the platform (Android is the focus in this document in terms of Operating System, as detailed in §1). An example is also found in §5.

Community input and feeling towards the Digital Inclusion initiative is an important part of the selection method, as some communities may be more supportive of the revitalization initiative than others. This can be decisive in moving a project ahead or halting it since, due to the very nature of

the initiative, there is a small quorum of subjects, so input and cooperation are paramount. That can only be achieved by a positive feeling towards the Digital Inclusion initiative from the part of the community, which facilitates the communication flow.

The availability of subject matter experts (SMEs) is also of great importance, as the need for researchers, professors, translators, and linguists are all part of the multi-pronged revitalization strategy. SMEs help bridge the gap between the community and the institution in terms of communication, understanding ideas, providing existing research, and academic work, finding common ground and leveraging talent and efficiency, which all help the initiative. It can also present quite the challenge, as some languages may have more researchers, professors, translators, and linguists than others, so procurement of SMEs can sometimes prove to be difficult.

As stated earlier, Motorola's efforts have primarily been focused on three categories laid out by UNESCO (Definitely endangered, Severely endangered, and Critically endangered) since we have seen a gap in terms of players and efforts in these categories. Thus, examples present in this document reflect this strategic decision.

3. Digital Inclusion initiative process setup

3.1 Indigenous peoples' Digital Inclusion acceptance and receptiveness

It is important to respect and acknowledge cultural protocols while collaborating with Indigenous communities and groups. Thus, the recommendation is that interactions happen through scholars, organizations, or institutions that already have previous meaningful and long-lasting relationships with the Indigenous peoples. Initiating the relationship with the Indigenous peoples through someone of their trust might bridge the gap on knowing the cultural protocols that are the etiquette, customs, codes, and other behaviors of a particular cultural community or group, as well as the appropriate processes for collaborating on an initiative with that group.

Although protocols share some common themes and practices, it is important to recognize that there is great diversity among Indigenous peoples and communities. Each community has their own culture, heritage and language, which influence proper protocols. Indigenous traditions and ethical conduct are important to maintaining networks and establishing respectful relationships. It is recommended that one consults the community representatives about appropriate protocol in the community.

The requester should share, beforehand, high-level details of the Digital Inclusion initiative and its objectives, for example, "the goal is to digitize a written language by enabling it in the user interface of a smartphone" or "the goal is to promote workshops that aim to document a spoken language for the first time in order to develop a keyboard for smartphones that support it, allowing its speakers to communicate in their mother tongue."

When presenting the Digital Inclusion initiative, it is vital to be mindful of the fact that the matter in discussion is their own native language and that, unless the requester is from the particular community, this initiative is an idea coming from an outsider. Dialogue between the company sponsoring the initiative, linguists, and the speaking community is encouraged so that the initiative fits the needs of the community. For example, when deciding on which dialect is chosen for the writing system, discussions among the speaking community are essential. Thus, active listening is advised while someone is speaking in order to ask questions related to the real benefits the initiative would bring to their community, the level of engagement of their community with internet and smartphones, and if their children learn and speak the language at school and at home. Such questions might better define the initiative's impact as well as show respect to the community's culture and needs.

Once all the community language needs are defined and agreed upon both by the community and the requester, a proposal is made to identify speakers (ideally people with writing or translating experience or professional translators) willing and available to work on the localization of the language data of a specific content into their language during a predetermined period of time.

As long as the community is respectfully involved from the beginning, understands and agrees on the benefits of the initiative for the revitalization of their language as well as the compensation for the parties involved, acceptance and receptiveness should come as a consequence.

3.2 Vendor partner selection criteria and non-profit partnerships

The vendor selection process is a series of procurement-related steps that determine service and product requirements and match them with vendor capabilities and pricing. For Digital Inclusion initiatives, it is common to consider engaging with customary partners which the company already works with (normally a language service provider), similarly to what one would do with any language inclusion project in a technology company. However, the selection of language services providers for endangered Indigenous languages may have additional requirements as the partnership with Indigenous peoples can be a very new endeavor for most suppliers.

Almost as important as assessing suppliers' financial stability is to ensure – through open conversation – that vendors and partners share the same ethical pillars and values. In many cases,

because the selected language has not yet been a part of any well-structured pool of resources of global languages in any vendor, it is required that one is selected for a customized service delivery. That demands the definition of new quality and productivity agreements to ensure all parties involved will be respectful and thoughtful of the Indigenous community's culture. While there is overlap in understanding milestones and deliveries for widely spoken languages, the milestones and deliveries cannot be the same as the ones that need to be defined for endangered languages. Thus, the goal during vendor partner selection should be to ensure the work done during the language revitalization and digitization initiative demonstrates continuing commitment to Indigenous peoples' rights while meeting the corporate limitations in terms of resources and predefined milestones, such as product release date.

An aspect that must be considered during the vendor partner selection is to ensure that all personnel involved in the process (translators, reviewers and project managers) are aware of and acknowledge that the localized content produced is the intellectual property of the client. In Motorola and Lenovo's revitalization initiative, such an agreement and awareness among all parties (e.g., Indigenous community, translators, reviewers and project managers) are key to the strategy of open-sourcing the corpus, making the content available to others who are equally committed to language revitalization work.

Corporate-Indigenous relations can easily rely on a top-down commitment from business leaders, and that is why the criteria for a language consultant and/or translator must be differentiated from other language consultants and/or translators who are selected primarily based on their availability (in hours or days) to work on the project. To make this distinction, it is recommended that the question is made directly to the linguist in terms of capacity at early stages of the negotiation process with vendor partners so that the data is taken into consideration as a criterion for the language selection process.

Normally, the recommendation is that the company shortlists two or three potential vendors and applies the vendor selection criteria checklist. For vendor due diligence, different team members participate in the evaluation process. Items to negotiate in a contract include:

- Number of speakers for the selected Indigenous language;
- Number of professional translators for the selected Indigenous language;
- Pricing per new word translated;
- Milestones in terms of daily capacity and availability per speaker/translator;
- Final acceptance of the completed language data;
- Initial service delivery date and word count;
- Total word count and leverage discount to be covered per future purchase order release.

Alternatively, a company may consider partnering up with a nonprofit organization (NPO). Similarly to the benefits illustrated in §3.1 about initiating the relationship with the Indigenous peoples through someone of their trust, working with a nonprofit that already supports an Indigenous community might facilitate the flow and ensure an appropriate process for conducting business and interacting with that community.

The same criteria applied to a language service provider company is valid for the NPO selection criteria. There are a couple of advantages of pursuing an NPO, such as cost efficiency – for example, NPOs will not add a markup on top of the translators' cost per word, unlike language service providers – as well as the ability to have payments made through the philanthropic funds, assuming the sponsor organization has such a division. Additionally, certain specialized NPOs already connected with the relevant Indigenous community may provide further local academic and/or data from research.

3.3 Linguistic and scholar partnerships

As mentioned in §3.1, the recommendation is that engagement with communities happens through scholars, organizations or institutes that already have previous meaningful and long-lasting relationships with the Indigenous peoples. They will lead the formation of the team to work on the

initiative, with the goal of guaranteeing that the deliverable is meaningful for the Indigenous peoples and their language.

It is recommended that the linguistic and engineering teams, together with the selected language service providers engage and train scholars, native speakers, and/or professional translators if applicable, on Computer Assisted Translation (CAT) tools, processes, software language particularities, and all that pertains to the digitization of a language, including but not limited to the choice of dialect to be used and methodologies to ensure terminology consistency throughout the product database.

Besides being the expert-matter of linguistics, the role of scholars is one of a consultant that ensures the process of digitization begins and ends in a respectful manner to the Indigenous community and focuses exclusively on language revitalization. Linguists also aim to promote technical discussion among the community and the onboarded translators regarding language specifications, such as the tone of voice to be applied in technology, dialect differences, and choice of style of writing. At times, this is not an easy task given that it is at this phase where technology-related terminologies arise for the first time and, although the concept might not be new, a word or an expression to convey it has not yet been determined in a given language.

If the chosen language is spoken in communities that have good penetration of technology, including but not limited to access to internet and computers, the first steps would be to:

a. have a clear understanding of people's availability and willingness to dedicate hours of their lives to the initiative,
b. agree on milestones in terms of words to be translated in a given period of time, and
c. agree on the monetizing compensation to be received per translated word.

That is recommended to be done either through the intermediation of the selected language service provider or the NPO responsible for the partnership between the sponsor company and scholars and linguists.

If the chosen language is spoken in communities that do not have access to the internet at the target speeds, it is recommended that the donation of laptops/computers and the providing of support for internet connection is considered. This will be yet another step towards closing the gap in Digital Inclusion, referred to as the digital divide, commonly seen in many Indigenous communities, which prevents them from having their routines enhanced by the use of technology both socially and economically through access to information.

The process of exposing speakers and translators to localization-specific matters follows the initial logistics-related phase. That includes training on tools and workshops on localization processes and best practices, such as:

- The usage of clear, simple and consistent language;
- Accountability for linguistic, cultural, and technical differences (such as idiomatic expressions, significance of specific colors, units of measurement, and currencies);
- Maintenance of the original intent and functionality of the original application; and
- Assurance that the localized product will look and function as if it were native to that specific market and peoples.

This phase is normally coordinated by the linguistic and engineering teams of the company ideating and sponsoring the initiative, while the logistics and people management happen through the NPO or language service provider.

For the digitization process described in this paper, the company has worked closely with universities (Unicamp in Brazil, the University of North Carolina - Chapel Hill in the United States, the Free University of Bozen-Bolzano), their linguists, and KISS (Kalinga Institute of Social Science) in India to engage with citizens and linguists.

4. Localization process and linguistic considerations

4.1 Localization Process

Localization is the systematic process of adapting a product or content to a specific location or market, including translation, associated imagery, and cultural elements that influence how the content will be perceived.[9] Details of such adaptations can be found in §6. This process involves modifying, restructuring, and adapting content for the target audience and requires a solid strategy and a clear roadmap, backed up by efficient communication. Besides translation, successful localization may involve, and is not limited to, planning, content preparation, post-editing and proofreading, quality assurance and in-context review.[10] This process applies to Indigenous and non-Indigenous languages.

TMS and CAT Tools considerations

A Translation Management System (TMS), also commonly known as translation management software, is a software platform that facilitates the lifecycle of a localization process. Translation management software eliminates repetitive and laborious manual tasks through built-in, automated machine applications while simultaneously enabling workflow control, increasing collaboration and efficiency. A TMS is a tool for efficiency and is not limited purely to translations.[11] Top-quality systems also automate workflows for improved project management, as well as offer other services such as integration with content management systems, financial tracking, analytics, resource allocation, and vendor neutrality (allowing companies to engage with multiple suppliers). Additionally, some other features to consider when selecting a TMS are:

- Extensive support for different file formats and type;
- In-context translation (the possibility of associating images for visual reference to be visible in the TMS editor to a specific segment);
- Capacity to have multiple translators working on the same project simultaneously (allowing segments from one project to be assigned to specific resources).

Several factors need to be considered when choosing the right TMS. Translation management systems are designed to support complex tasks in order to make translation and localization processes manageable and efficient. Many TMS tools offer Computer Assisted Translation (CAT) and built-in machine translation (MT). These all-in-one applications allow users to manage and plan projects through a single platform.[11] Translation management systems help manage translation workflows but do not perform the translations. CAT tools can work within Translation management systems to support a given localization workflow. Workflows are customizable, allowing for multiple steps to be added and assigned to different linguists as needed, based on type of content and project requirements.

CAT tools are an important part of the pipeline, improving translators and reviewers' efficiency and productivity in a localization process and are commonly known and used by Language Service Providers (LSPs). They are also frequently used by individual translators and bilingual employees who work for organizations with global audiences or extensive localization needs.

CAT tools automate tasks that a linguist would otherwise have to conduct manually, including the management of translated content. It breaks content in a specific Source language down into translation units, often referred to as segments (usually phrases or paragraphs) for localization into the Target languages. For enlightenment purposes, this document will refer to "Source language" when stating about the default language used for software development - popularly English for this intent; while "Target language" will be the term used to refer to the language in which translation is needed.

These tools often consolidate MT integration, terminology management, Translation Memory (TM) leveraging, string alignment, scope analysis, lookup capabilities, quality assurance, spell and grammar checkers, and other functions. There are many sophisticated CAT tools in the market designed for diverse types of tasks. A typical CAT tool consists of at least three major components:

A TM, where previous translations are stored; a Terminology Database containing a list of approved terminologies required to be referenced during localization; and an Editor, a page through which translation takes place.

Translation Memory

A Translation Memory is a linguistic database that enables the reutilization of translations for future work, a major functionality in a CAT tool. It stores content from the Source language and the corresponding translations in segments, also known as translation units (TU), that can be as long as a paragraph or as short as a word. The program searches for matches with the text in the Source language among ongoing and former translation projects, providing outcome suggestions accordingly. A match can be exact or fuzzy, determined in accordance to the percentage of content similarity – 100%, 99%, 95%, 80% and so on.

Terminology Database

A Terminology Database, or termbase, is an integrated module within a CAT tool that serves as a compilation of terminology and may also contain associated information related to the defined words or phrases. Information available in a termbase for entries may include a term and its version in the Source and Target language, its definition, usage examples, and metadata. Termbase entries are useful to elucidate words or phrases that are potentially ambiguous in meaning, technical and product-specific information (marketing/branding), or expressions that can possibly be translated into a forbidden or taboo expression. Pre-organized termbases and glossaries may be imported into CAT tools but may also be organized and edited by a user with set permissions. Utilizing a pre-organized termbase can enable a translation project to maintain consistent standards as well as a consistent message, since suggested or mandated Target language terminology for specified words or phrases are clearly defined for the end-user. Projects containing multiple assigned translators will have a uniform degree of quality if adherence to data contained in the termbase is followed.[13]

CAT Tool Editor

The Editor of the CAT tool is the means through which translation and review takes place. Robust editors have a side-by-side view of content in Source and Target languages with TM search capabilities and Terminology Database results in display to make translation and review as efficient and smooth as possible. Some editors also have a pop-up window for preview of the translated document, which can be useful to determine the desired layout of translated text. Translatable texts in the Source language are displayed in segments; each segment may contain either sentences or paragraphs according to the defined setting. Once a translation is complete for a segment, it can be finalized or locked to prevent unintentional editing and to keep track of the translation progress. Also, if a TM match exists for a segment, the CAT tool will display that result denoting the match type (e.g., 100%, Fuzzy Match, No Match, Repetition, In-Context Match) which often reduces translation efforts and costs.

When working with an Indigenous endangered language, determining which tool to use is an important decision and this decision must consider the needs of the community. If the tool is only available online, it may not be available in areas with limited connectivity to the Internet and might prevent underserved communities from taking advantage of its features during the localization process. Also, it is not common to have Indigenous languages available by default in Translation Memory Systems, and these languages may need to be added as a supported language, which is a process that takes time. In such cases, another available language can be used instead as a placeholder until such a request is completed.

There are several factors to consider when deciding which TMS and CAT solution suits the project needs and budget involved. There is no one-size-fits-all solution, and companies of all sizes use them differently. Nevertheless, when working with translation or localization of content, using a TMS with an integrated CAT tool is necessary. A successful TMS will deliver accurate and relevant results quickly, all while maintaining accurate timelines and project management efficiency.[12]

4.2 Linguistic considerations when creating orthographies

It is imperative to consider both the linguistic structures and sociolinguistic aspects of a language when creating an orthographic system. At its core, an orthography represents a combination of linguistic considerations and the needs of a speaking community. Crucially, orthographies should be agreed upon by native speakers and be able to be reformed if needed.

An effective orthography is morphophonemic, such that it should generally capture phonemic contrasts while also maintaining consistent spelling for sounds which have different pronunciations in specific contexts (e.g., the plural *-s* in English is pronounced as [z] in certain contexts but is consistently transcribed as *-s*). Additionally, phonology, paired with native speaker intuitions, should be considered when transcribing word breaks within compound words and sentences. Other linguistic facts to consider are how features – such as consonant/vowel length, nasalization, and tone – should be represented (e.g., as a single letter, diacritic, etc.).

The orthography should be as widely accepted as possible by the Indigenous community for which it is made. In order for the orthography to be used by speakers, it needs to be intuitive enough to learn and teach. For example, the orthographies for many Indigenous Mesoamerican languages have similar conventions to that of the Spanish orthography, since such communities are already familiar with this orthography and can transfer their knowledge of written Spanish to the written form of the Indigenous language. A new orthography for a particular Indigenous language should be as similar as possible to related Indigenous languages, if the languages do not have substantial phonological differences. With respect to dialectical differences, the speaking community should decide on whether a standard orthography would best fit the needs of a community and, if so, how to approach standardizing any differences. In general, discussions about creating or adapting an orthography should involve members from each community and can involve collaboration with linguists and local or regional institutions such as language commissions and universities. For example, in Latin America, institutions such as the Museu do Índio and Instituto Emílio Goeldi can be consulted for collaborating on an orthography with indigenous communities in the region. Another example of collaboration between Indigenous communities and institutions involves the University of California campuses with communities from North and Central America. The University of California, Santa Cruz has worked with a Sierra Norte Zapotec variety to adapt the existing Zapotec alphabet to capture phonological differences in the variety which are represented by new characters. Ultimately, whether a certain dialect is chosen over another dialect, or whether a standard form is created, is dependent on the community's particular initiative.

An orthographic system of an Indigenous language must be approved by the speakers who will be using it. If an established orthography is created, the speakers have the authority to make any changes they deem necessary. Thus, dialogue between linguists and Indigenous communities is always at the forefront of this collaborative endeavor.

4.3 Software Integration, Leveraging, Source & Target, and more

Similarly to how software development can be conducted in several different ways, software localization can follow a particular methodology as well. Traditionally, it follows a waterfall process whereby the localization is done all at once, usually after the default language version of the software has finished development or has already been released. This can be a good option if more control is needed over the process and if there is a limited budget. However, this approach has some significant drawbacks, the most important being that it can take a long time to get the localized software or product in the market. Since all the content has to be translated at once, it can be very time-consuming. The other main drawback is that it can be more difficult to catch errors and mistakes, as there are fewer opportunities for feedback and iterations. Once the content has been translated, it can be very difficult and expensive to make changes.[14]

In contrast, continuous localization is an approach in which translations are done constantly as new content is available. Continuous localization ensures that localized content is ready for simultaneous release at all times. This approach also has the advantage of being more flexible and iterative. However, continuous localization can be harder to manage and coordinate since it requires a tech-

savvy localization team that is comfortable enough to work according to an agile environment. Along with that, having a stack of tech tools that can complete the task is also important. Ultimately, continuous localization has proven to be a more efficient and cost-effective way to localize a software product.[13]

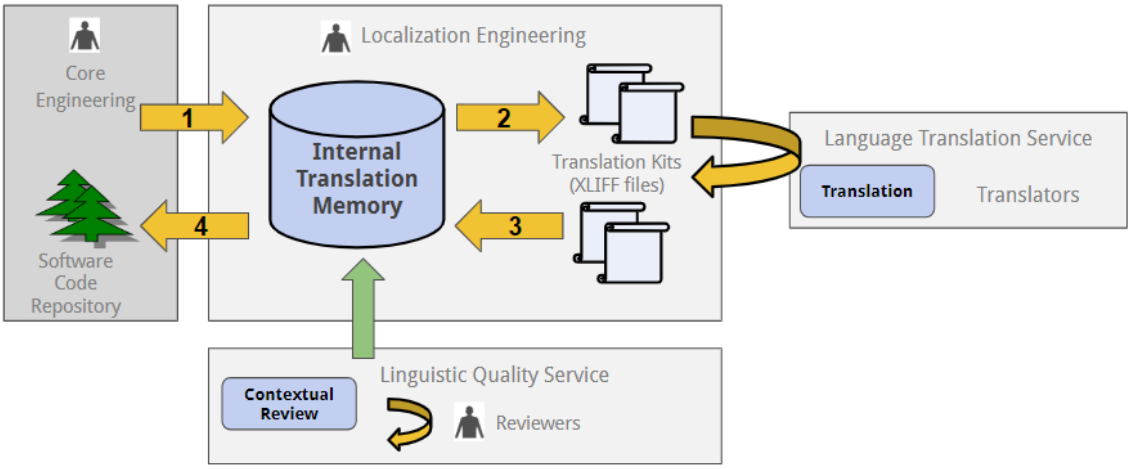


Figure 4.3-1. Continuous Localization Workflow

Figure 4.3-1 shows an example of how a continuous workflow for software localization can be set up with the help of a toolset/system for software resource management. Once the core engineering or development team pushes code with new strings to the software repository, the Localization platform/tool, as a resource management system, would:

1. Automatically detect new or modified strings and search in an internal translation memory for existing translations to leverage from;
2. Push new content that needs to be localized to the TMS.

When translation is completed, the Localization platform would:

3. Automatically pull the translated content from the translation management system and add it to internal translation memory;
4. Merge the translated content back into the software code repository.

The translation step between 2 and 3 would usually occur in a TMS (§4.1). The steps 2 and 3 can be automated if the TMS provides the necessary Application Programming Interface (API); otherwise, it can be performed manually as well by a Localization/Language Project Manager.

The implementation of steps 1 and 4 depends on the software platform or content management system used for authoring the content. For the initiative described in this paper, content localized by Motorola was for Android™ operating system, and strings were present in hundreds of different "git" repositories. Git is a very popular distributed version control system used in software development. The localization tool platform would crawl through all the code repositories, detect strings present inside the Android resource folders in various XML files, and check if the translation already exists in the internal translation memory. All the strings without a respective existing translation are extracted for translation.

The above workflow would usually work in a language pair of Source and Target languages. Source language strings are sent out to be translated into the Target language using a .xliff file. XML Localization Interchange File Format (XLIFF) is an XML-based bitext format commonly used to

exchange data between and among tools during a localization process and is a popular format used in CAT tools.

When working with an Indigenous language, it is possible that a Source language is not useful for direct translation to a Target Indigenous language. As Indigenous communities may live concentrated in a single region and are only familiar with the dominant language in that specific region, more translation steps might be needed. For instance, this process would first require translation from the Source language to the dominant language of that region, then translation from the dominant language to the Indigenous language. When working on Indigenous languages in the Amazon region, Motorola had to translate first from English (United States) to Brazilian Portuguese and then from Brazilian Portuguese to the impacted Indigenous languages.

This multi-step process can be achieved by using the same localization workflow shown above (Figure 4.3-1) twice. First, the development language is used as the Source and the regional dominant language is used as the Target. Once this step is completed, the regional development is used as the Source and the Indigenous language is used as the Target. Some Translation Management Systems have the capability to add multi-step translations within the set workflow which can be an effective option to work without making further changes in the localization workflow.

While the first approach can work with any TMS, it can cause some delays because of additional back-and-forth of files. Also, as the number of intermediary translation steps increases (e.g. Chinese > English > Odia > Kuvi), this approach can become very costly and time-consuming. The second approach can scale well but needs to have a TMS in place in which this option can be enabled.

4.4 Linguistic Quality Assurance considerations and sensitivity to cultures

Linguistic Quality Assurance (LQA) is the process during which linguists use a particular methodology to review translations and determine whether they include errors. Such errors might involve, but are not limited to, the following: irregularities in glosses, inconsistent spelling or grammar, and contextual inadequacies. Additionally, it is important to collaborate with native speakers of a language to ensure that the translations accurately represent the language and the community's needs.

When working with an Indigenous community, researchers should familiarize themselves with the culture of the speakers of the language, as cultural factors may directly correlate with aspects of the language (for example, kinship terms).[15] Researchers should establish consistent communication with a community to build personal relationships and mutual trust with members of the community. For example, as aforementioned, researchers must come to agreements with a community regarding payment for their language expertise. While some communities are open to receiving appropriate monetary compensation, other communities prefer to be compensated with practical supplies (e.g., flashlights, umbrellas). Cameron et al. (1992) outline three types of linguistic research methods[16] (1-3), and Czaykowska-Higgins (2009) provides an expansion[17] to this model (4), defined below:

1. **Ethical research**: Model in which the research is *on* subjects; in other words, there are ethical considerations driving the methodology, but the model does not go beyond data collection.
2. **Advocacy research**: Model in which the research is *on* and *for* subjects; the work directly aims to benefit the community.
3. **Empowering research**: Model in which the research is *on*, *for*, and *with* subjects; the work benefits the community and explicitly involves the community in decision-making and leading the research project.
4. **Community-based language research**: Model in which research is *"... on a language, and that is conducted for, with, and by the language-speaking community within which the research takes place and which it affects. This kind of research involves a collaborative relationship, a partnership, between researchers and (members of) the community..."* [16]

Linguistic endeavors should strive to emulate a Community-based language research model, which ultimately regards native speakers as the experts of their language, and that mutual communication and collaboration are key aspects of a linguistic project.

5. Internationalization language support levels

Internationalization provides the ability to make products and services available globally, while meeting regional and legal requirements of a targeted country. In the given scenario, it is the process of adapting the software code into a behavior in which inputs and displays of various languages are enabled independently of the interface language. This process also includes the compliance to locale standards, such as adhering to currencies, measurement units, and date and time formats. The following is a list of the five levels of support that are needed to cover the basic requirements for the Digital Inclusion of a language, and is applicable either to Indigenous and non-Indigenous languages. Refer to Appendix for a summarized flowchart that illustrates the support levels.

5.1 Writing System

Although writing systems are a recent form of language in comparison to spoken language, they are crucial in the transmission of information in modern times.[18] Since writing systems came into existence within the past few thousand years, many modern-day systems can trace their foundations back to only a few scripts. For example, the Roman alphabet used for English script originates from the second millennium BCE Phoenician alphabet.[18] However, it is difficult to determine the origins of certain scripts, such as Chinese script or Mayan script, which seem to have developed independently from other scripts.[18]

Researchers have recently recognized the study of writing systems as its own branch under linguistics. In the study of writing systems, it is crucial to classify them by type, similar to how spoken languages are classified in families.[18] One particular type of writing system is a semasiographic system, which utilizes symbols and imagery. For example, road signs and signs for garment care are systems which must be learned and represent ideas without explicitly utilizing segments of language. These systems have been attested in Indigenous cultures, and also have applications for people who have disabilities which interfere with reading or writing.

Another type of system is a glottographic system, which expresses ideas in the form of segments (e.g., words, phonemes) of spoken language. Within the glottographic systems, there are different levels of representation which range from logographic to phonographic.[18] A logographic script generally uses one symbol, which does not reference pronunciation, to represent a word or concept. An example of a logographic system is the Chinese script. In contrast, phonographic systems explicitly correspond with phonetic units of a language. Units that can be represented include larger, syllabic units; Japanese *hiragana* and *katakana*, for example, have a script in which syllabaries sharing the same symbol for a consonant or vowel do not necessarily look similar. Another phonetic unit that can be represented in a written system is features, as in the Korean alphabet called *hangul*, in which the place of articulation of the particular sound is represented by a graphic form. Finally, phonographic systems can be alphabetic (i.e., phonemic) and intended to represent distinct sounds (i.e., phonemes) in a language, such as English.

Aside from the type of writing system being used, another aspect used to classify written languages is by their completeness.[18] For example, some systems represent vowel length, as in Finnish: *kaatua* 'to fall' and *katua* 'to regret.'[18] Scripts can also represent tone; for example, in tonal languages, the tonal difference between two words with the same orthography can be represented illustrated by tonal accent marks or by a tonal number system.

When creating a writing system for an unwritten language, the choice of system depends on many factors, including cultural, historical, and technological aspects. Writing systems help to transmit and document language, so it is important to collaborate with the speaking community of the language to devise a system that allows speakers to communicate with each other in an effective and standardized way.

A language can be digitized in various ways, including audio and video formats. Recently, language revitalization efforts have utilized modern technologies such as social media and video games to promote language learning. However, these efforts should be supplemented by a complete writing system which can be (i) learned by speakers of a language and (ii) taught by speakers of the language

to pass down to future generations. Thus, writing systems should be agreed upon by a speaking community, as well as amenable to future change as the community deems appropriate.

An example of the process of creating a writing system can be seen in the development of the Kuvi writing system, which was a partnership effort between the Kalinga Institute of Social Sciences (KISS) and the Lenovo Foundation. Kuvi is classified by UNESCO as a potentially endangered language and is spoken in the Odisha and Andhra Pradesh states of India. In addition to involving the speakers of this language in creating the Kuvi character sets, the process also involved Kuvi language experts, linguists, and IT experts. Firstly, the following questions were considered by those participating in this project:

- Will a new script be made for the Kuvi character sets, or will an existing script be adopted for them?
- If the script is adopted, which script should be chosen? Will the community accept the adopted script and does it adequately represent Kuvi?
- Does the script phonemically represent Kuvi? If so, how are the phonemes represented with the graphemes of the script?
- How are the community members involved with the decision-making process of the development of the script?

During the development phase of this project, community members collaborated with language experts and linguists to describe the phonological features of Kuvi. Community members were asked to pronounce the sounds in the context of vocabulary. They were then presented with scripts for Odia, Telugu, and Hindi and were asked to choose a script which best represented the phonology and phonetics of Kuvi. A parallel table with all of the scripts was developed. Words were elicited from members of the speaking community to ensure that the character sets accurately represented Kuvi. Finally, the community provided feedback during testing stages and necessary changes were made to best meet the needs of Kuvi speakers. This process ultimately highlighted the importance of community involvement, the need for the Kuvi character sets to be harmonious with other languages spoken in the regions where Kuvi is spoken, and how linguistic considerations should inform the process of creating writing systems.

For character(s) of writing systems that are not yet supported by Unicode – an information technology that represents text – proposals listing the properties of such characters must be submitted to the Unicode Standard. After submitting the proposal, it is reviewed by the Unicode Consortium and Unicode Technical Committee; this process is highly involved, and sponsors of character(s) should be prepared to answer questions from the committee as well as organize discussions about the proposal. The following section provides an in-depth discussion of the technicalities of Unicode.

5.2 Unicode

Unicode is an information technology standard for encoding, representing, and consistently managing text in a given writing system. It is maintained by the Unicode Consortium and contains an array of characters, coding methodology, standard character encodings, character properties (such as uppercase and lowercase), and rules for normalization, decomposition, sorting, and rendering. It is widely used in localization and internalization of software and is updated on a yearly basis, with new languages, characters, and emoji.[19]

Unicode publishes and maintains a chart of supported characters.[20] Most or all characters of a newly supported Indigenous language may have already been defined in the Unicode standard. A newly supported Indigenous language might require characters from multiple scripts. For instance, the digits 0-9 are in widespread use; also, the Devanagari *danda* is used across many Indic scripts.[21] The example below illustrates a character U+203B (※) under the "General Punctuation" section in Unicode. It is neither defined in Japanese Hiragana script nor in Urdu Arabic script but is used in both languages.

U+203B ※ REFERENCE MARK

= Japanese kome

= Urdu paragraph separator

Figure 5.2-1: Unicode character U+203B

In the event of Unicode not yet supporting characters from a language being digitized, there are two paths to be taken, each with its pros and cons, and they require equal attention.

a. Private Use Areas (PUA)[22] in Unicode Standard

There are a total of 137,468 possible code points and these code points are designated for private use. PUA code points are in the ranges U+E000..U+F8FF in Basic Multilingual Plane (BMP), U+F0000..U+FFFFD, and U+100000..U+10FFFD in Planes 15 and 16. Any PUA code point can be used to support the Indigenous characters; however, external systems will not be able to support non-Standard Unicode code points.

Additionally, character properties like line breaking standards and case conversions can be defined through PUA characters, though some operating systems may not support those.

PROS	CONS
● Easy to define custom private code points inside Unicode standard ● Does not cause character corruption with other Unicode scripts ● Works within the internal system	● Does not work with external systems ● The use of PUA must be implemented in each system component ● Requires the addition of private-use character glyphs in the system font (refer to §5.4 - PUA for more details) ● Requires the creation of a custom IME (refer to §5.5 for details)

Table 5.2-1: Unicode PUA Pros and Cons

b. Submitting a Character Proposal[19] to Unicode

Proposals for inclusion of new characters and scripts can be submitted to the Unicode Consortium. A proposal should first consider if a particular script or character has already been proposed. Once the proposal has been identified as new, then it should include the following relevant information for submission:

Unicode Character Properties

[Basic Information]

- Code point - *Unicode code point (optional)*
- Glyph - *graphical representation of the character*
- Name - *name of the character*

The most basic information required about characters includes name, code point and other identifying information. A sample listing of code point, glyph and name for the Telugu Letter La is the following:

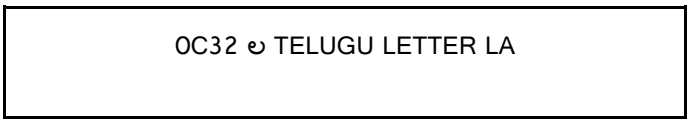


Figure 5.2-2. Telugu Letter La

[General Category and other properties]

- General Category - *most general classification of the code point*
- Canonical Combining Class - *Canonical Ordering Algorithm*
- Bidirectional Class - *Bidirectional Algorithm*
- Decomposition Type/Mapping - *Decomposition string-value property*
- Numeric Type/Value - *Numeric type property*
- Bidi Mirrored - *"Y" if mirrored text in bidirectional text*
- Unicode 1 Name - *Old name published in Unicode 1.0*
- ISO Comment - *ISO comment field*
- Simple Uppercase Mapping - *If uppercase equivalent exists*
- Simple Lowercase Mapping - *If lowercase equivalent exists*
- Simple Title Case Mapping - *If title-case exists*

The properties are documented in the Unicode Character Database and following line is the entry for the Telugu Letter La:

Figure 5.2-3. Properties of Telugu Letter La

If the proposed characters exhibit shaping behavior (contextual shaping, ligatures, conjuncts or stacking), one must provide a description of that behavior with glyph examples.[18] Information about the sorting order should also be provided. If the proposed characters are symbols, one should consult "Criteria for Encoding Symbols" in the Unicode Standard.

Once all the data is defined and collected, then the proposal can be submitted to the Unicode Standard. A series of meetings with the Unicode Technical Committee will then be held and discussed for acceptance.

In summary, the submission of new characters and scripts encompasses many challenges, time and effort; but the reward of being included in the Unicode Standard is substantial.

PROS	CONS
● Characters used in Indigenous language will be included in Unicode Standard ● Characters will be supported in most computers and digital devices in the world	● A submission takes a long time to be accepted and included in Unicode Standard

Table 5.2-2: Submission to Unicode Standard Pros and Cons

5.3 ICU and CLDR

International Components for Unicode (ICU)[23] is a library for modern programming languages, such as C, C++, Java, and JavaScript, and provides Unicode and Globalization support for software applications. It is portable and modular, and gives applications the same results on multiple platforms. It is released under an open source, nonrestrictive license for commercial use or to be adapted to other open-source software. ICU provides, but is not limited to:

- Code Page conversion - between Unicode and Code Page;
- Collation - rule for comparing characters;
- Formatting - numbers, dates, times, currency, gender and plural;
- Calendar - weekday, month names with full, medium and short abbreviation;
- Timezone - timezone names;
- Unicode character properties - name, upper/lower case, bi-directionality, more;
- Regular expression - regular expression matching to Unicode character string;
- Bidirectional text handling - algorithm for handling the bidirectional text;
- Text Boundaries - character boundary, word boundary, line-break boundary, sentence boundary.

In case the system uses Unicode PUA code points (§5.2-a), the code point converter between PUA and charsets should be developed and integrated into the ICU library.

The Unicode Common Locale Data Repository (CLDR)[24] is the largest and most extensive standard repository of locale data available, and its data provides companies with the constituent elements for software to support a wide array of languages. CLDR uses the Unicode Locale Data Markup Language (LDML) format to store and interchange the locale data. ICU compiles and builds the CLDR data into libraries so that it can provide the consistent locale conventions and standards across the multiple programming languages and platforms.

CLDR includes information such as numbering systems, date, time, time zones, measurement units, country names, cities, scripts, grammar rules, and much more. CLDR provides different levels of necessary base support, which is especially important for endangered languages as information can be more difficult to acquire.

```
<ldml>
  <identity>
    <version number="$Revision$"/>
    <language type="te"/>
```

```
</identity>
<localeDisplayNames>
 <localeDisplayPattern>
  <localePattern>{0} ({1})</localePattern>
  <localeSeparator>{0}, {1}</localeSeparator>
  <localeKeyTypePattern>{0}: {1}</localeKeyTypePattern>
 </localeDisplayPattern>
 <languages>
   <language type="elx">ఎలామైట్</language>
   <language type="en">ఇంగ్లీష్</language>
   <language type="en_AU">ఆస్ట్రేలియన్ ఇంగ్లీష్</language>
   <language type="en_CA">కెనడియన్ ఇంగ్లీష్</language>
   <language type="en_GB">బ్రిటిష్ ఇంగ్లీష్</language>
   <language type="en_GB" alt="short">యు.కె. ఇంగ్లీష్</language>
   <language type="en_US">అమెరికన్ ఇంగ్లీష్</language>
   <language type="en_US" alt="short">యు.ఎస్. ఇంగ్లీష్</language>
   <language type="enm">మధ్యమ ఆంగ్లం</language>
   <language type="eo">ఎస్పెరాంటో</language>
   <language type="es">స్పానిష్</language>
   <language type="es_419">లాటిన్ అమెరికన్ స్పానిష్</language>
   <language type="es_ES">యూరోపియన్ స్పానిష్</language>
   <language type="es_MX">మెక్సికన్ స్పానిష్</language>
   <language type="et">ఎస్టోనియన్</language>
   <language type="tcy">తుళు</language>
   <language type="te">తెలుగు</language>
   <language type="tem">టిమ్నె</language>
   <language type="teo">టెసో</language>
   <language type="ter">టెరెనో</language>
 </languages>
<localeDisplayNames>

<dates>
 <calendars>
  <calendar type="gregorian">
   <months>
    <monthContext type="format">
     <monthWidth type="abbreviated">
        <month type="1">జన</month>
        <month type="2">ఫిబ్ర</month>
        <month type="3">మార్చి</month>
        <month type="4">ఏప్రి</month>
        <month type="5">మే</month>
```

```
            <month type="6">జూన్</month>
            <month type="7">జులై</month>
            <month type="8">ఆగ</month>
            <month type="9">సెప్టెం</month>
            <month type="10">అక్టో</month>
            <month type="11">నవం</month>
            <month type="12">డిసెం</month>
        </monthWidth>
        <monthWidth type="narrow">
            <month type="1">జ</month>
            <month type="2">ఫి</month>
            <month type="3">మా</month>
            <month type="4">ఏ</month>
            <month type="5">మే</month>
            <month type="6">జూ</month>
            <month type="7">జు</month>
            <month type="8">ఆ</month>
            <month type="9">సె</month>
            <month type="10">అ</month>
            <month type="11">న</month>
            <month type="12">డి</month>
        </monthWidth>
      </monthContext>
    </months>
  </calendar>
 </calendars>
</dates>
</ldml>
```

Table 5.3.-1: A sample CLDR for Telugu language in LDML format

One must check if the locale ID for Indigenous language is defined in CLDR. The list of supported languages is available at the CLDR website. For example, Telugu Language is defined and classified as a Modern Coverage level:

```
Locale ID ; Coverage Level ; Name

te ;  modern ;   Telugu
```

Figure 5.3-1. Telugu Language Coverage

The definition of CLDR Coverage Levels is as follows:

a. Core Data

This level has minimal data about the language and writing system that is required before other information can be added using the CLDR survey tool.

1. Language code (eg. te for Telugu language)
2. Four exemplar sets: main, auxiliary, numbers and punctuation
3. Verified country data (i.e. population of speakers in the regions or countries)
4. Default content script and region
5. Time cycle used with the language in the default content region

b. Basic Data

This level includes a small set of data for supporting the language.

1. Delimiter Data - quotation start/end, including alternates
2. Numbering system - default numbering system plus native numbering system
3. Locale Pattern Info - locale pattern and separator as well as code pattern
4. Language Names - in the native language for the native language and for English
5. Script Name(s) - scripts customarily used to write the language
6. Country Name(s) - for countries where commonly used
7. Measurement System - metric vs. imperial
8. Fully defined Month and Day of Week names
9. AM/PM period names
10. Date and Time formats
11. Date and Time interval patterns
12. Timezone baseline formats - region, GMT, GMT-zero, hour
13. Number symbols - decimal and grouping separators; plus, minus, percent sign
14. Number patterns - decimal, currency, percent, scientific pattern

c. Moderate Data

This level includes the following additional locale attributes.

1. Plural and Ordinal rules
2. Casing information
3. Collation rules

d. Modern Data

This level is considered as “full” CLDR support level and it includes the following additional data.

1. Grammatical Features
2. Romanization table (non-Latin scripts only)

In case a newly introduced language is not supported by CLDR, there are two options and both require the Core Data Coverage Level at minimum.

1. Modify the existing ICU & CLDR libraries in the system

In Android, ICU and CLDR are integrated in the Android SDK (Software Development Kit). The Android SDK is a set of development tools that are used to develop applications for the Android platform. This SDK provides a selection of tools that are required to build Android applications and ensures the process goes as smoothly as possible.[25] To modify the underlying CLDR, the following steps need to be taken (note that the steps are based on Android 13 and they may change in future releases):

a. Identify the locale ID[26] for the Indigenous language;
b. Create LDML xml file for the locale ID and populate the data;
c. Rebuild the ICU and CLDR for Android SDK;
d. Rebuild the entire Android system.

PROS	CONS
● CLDR data (eg. locale translations, date time formats, etc) are available immediately	● The need to rebuild the Android System

Table 5.3-2: Pros and Cons of Customizing CLDR

2. Submit data via Survey tool[27]

Data for CLDR (Common Locale Data Repository) is gathered and processed via Survey tool. The Survey tool is a web-based tool for collecting CLDR data and includes various attributes that need to be specified before submission. The tool provides a way to propose new localized data, check what others have proposed, and communicate with them to resolve differences. During each submission period, contributors from Unicode Consortium members, other organizations, and the public at large are invited to review the data for their languages and countries and propose new translations of terms or modification, including language translations entirely new to the repository.[26]

There are four data collection stages in the Survey tool:

a. Shakedown

   This stage is the start of the tool. One needs to make sure that the Coverage level is set correctly and look for any issues as a vetter.

b. General submission

   This phase is where the CLDR data is being entered into the Survey tool. The Core Data section needs to be filled. It is highly encouraged to add the Basic Data section as well.

c. Vetting

   To resolve all of the errors and review open requests and discussions.

d. Resolution

   The vetting is complete and any further work will be conducted by the CLDR committee.

CLDR has two types of releases: Full release and Limited release. Typically, the Full release is open for contributions for all languages and data areas. The Limited release is open for contributions for selected locales and selected fields for all locales. For more information, consult the Survey Tool website.

PROS	CONS
● Data used in the newly supported language will be included in CLDR ● Data will be supported in most computers and digital devices in the world	● A submission takes a long time to be accepted and included in CLDR

Table 5.3-3: Submitting data to CLDR Committee Pros and Cons

5.4 Fonts

A font is a collection of glyphs to represent abstract characters. In other words, fonts are the bedrock for language support and need to be available in the operating system so characters and writing systems can be displayed to the end-user. It is a pivotal element in order to properly display information from Unicode, CLDR, and more to the user.

There are 2 types of font format: raster fonts and vector fonts. In raster fonts, a glyph is a bitmap that the system uses to draw the character or symbol in the font, whereas in vector fonts, a glyph is a collection of line endpoints that the system uses to draw the character.[28] The advantage of using the raster fonts is that they are very fast to render and can be optimized for 2D image processing. In modern computers, however, virtually all fonts are vector-based and can be used at any size without loss of sharpness.

Typeface and fonts are often used interchangeably. The difference is relevant to typographic designers. A typeface is the underlying visual design and a font is the one that implements the typeface. For example, "Times New Roman" is a typeface and "Times New Roman Italic" is a font. Examples of typefaces can be seen below:

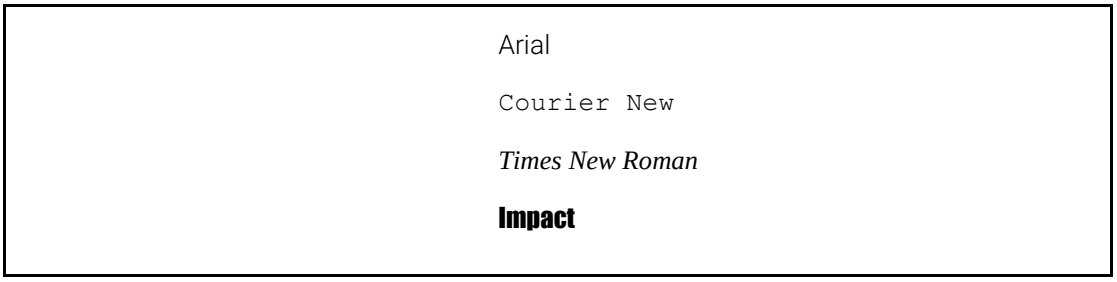


Figure 5.4-1: Font typefaces

It is important to note that fonts are highly copyrightable material and not to be distributed freely without the font designer's permission.

To determine if the system font has the character glyphs, the font (e.g. OpenType font) has an embedded table used to map Unicode character codes to glyphs indices in the font: The 'cmap' table. An example of 'cmap' table is available below:

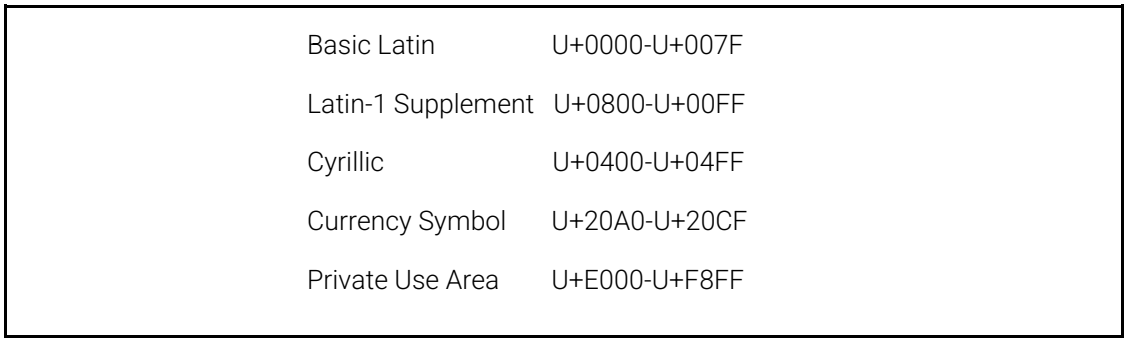


Figure 5.4-2: "cmap" in Font file

Incidentally, there are no fonts that include all the Unicode characters. It is, thus, important to determine if the system font supports the Indigenous language characters. Indigenous language characters may already be defined in the Unicode standard, but there is no guarantee that the system font includes such character glyphs. If the system font does not contain the Indigenous character glyphs, then a custom font needs to be created by a font designer and installed and distributed throughout the system. It is recommended to contact the operating system vendor (and font vendor)

to include the missing glyphs in the system so that the custom font does not need to be installed in the future operating system releases.

Fonts can support the Unicode Private Use Area (PUA) code point range (see 5.2-a). A font designer should add the glyphs and define the range in the 'cmap' table in the font. When the system needs to render the Unicode-PUA characters to the users, the system must have the custom font.

5.5 IME

An Input Method Editor (IME) is as important as the presence of fonts since it enables users to input information and data, broadly separated into three categories: keyboard entry, handwriting recognition, and voice recognition. For a new language to be properly supported in software, one of the categories of input methods has to be presented to the user. For mobile devices, the least complex implementation is the keyboard entry.

IME sits between the physical keyboard and the operating system. It interprets a physical key stroke to a character code point. By interpreting the key strokes, the operating system can support different characters with the same physical keyboard layout.

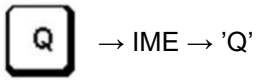


Figure 5.5-1: a keystroke via IME

Operating System vendors often provide many keyboard layouts for languages. However, not all languages are supported. Developing a keyboard for the Indigenous language is possible, but faces many challenges. Examples of questions to understand these challenges are: is the Indigenous community familiar with computers and digital devices?; where should the characters be placed on the keyboard?; how usable is the new keyboard layout?

Android SDK includes extensive documentation for creating an on-screen keyboard IME. To add an IME to the Android system, create an Android application containing a class that extends the InputMethodService. Each key should then be defined in the XML file. An example of XML file is below:

```
<Row>
  <Key android:keyOutputText="\u0C4D"
    android:keyWidth="@dimen/key_telugu_width_normal_row1"
    android:keyLabel="\u0C4D" android:keyEdgeFlags="left"/>
  <Key android:keyOutputText="\u0C3E"
    android:keyWidth="@dimen/key_telugu_width_normal_row1"
    android:keyLabel="\u0C3E"/>
  <Key android:keyOutputText="\u0C3F"
    android:keyWidth="@dimen/key_telugu_width_normal_row1"
    android:keyLabel="\u0C3F"/>
  <Key android:keyOutputText="\u0C40"
    android:keyWidth="@dimen/key_telugu_width_normal_row1"
    android:keyLabel="\u0C40"/>
  <Key android:keyOutputText="\u0C41"
    android:keyWidth="@dimen/key_telugu_width_normal_row1"
```

```
            android:keyLabel="\u0C41"/>
        <Key android:keyOutputText="\u0C42"
            android:keyWidth="@dimen/key_telugu_width_normal_row1"
            android:keyLabel="\u0C42"/>
        <Key android:keyOutputText="\u0C46"
            android:keyWidth="@dimen/key_telugu_width_normal_row1"
            android:keyLabel="\u0C46"/>
        <Key android:keyOutputText="\u0C47"
            android:keyWidth="@dimen/key_telugu_width_normal_row1"
            android:keyLabel="\u0C47"/>
        <Key android:keyOutputText="\u0C4A"
            android:keyWidth="@dimen/key_telugu_width_normal_row1"
            android:keyLabel="\u0C4A"/>
        <Key android:codes="-1"
            android:keyWidth="@dimen/key_telugu_width_normal_row1"
            android:keyIcon="@drawable/ic_zero_width_outlined"
android:keyEdgeFlags="right"/>
    </Row>
```

Table 5.5-1: A sample XML definition for Android Keyboard Layout

It is important to note that it is possible to define the keyboard layout using the Unicode PUA code points. It is simply a matter of specifying the PUA code points in the XML definition file.

6. Quality assurance

Quality Assurance is an important process responsible for guaranteeing that all products and services are compliant with some specific requirements, which includes the CLDR/ICU database that contains the locale standards such as adhering to currencies, measurement units, date and time formats, and more (refer to §5 for more information about CLDR/ICU). There are two main types of validations when dealing with Internationalization Quality Assurance: Functional validations and Internationalization domain validations.

6.1 Functional validations

Functional validations are a set of validations used to guarantee that the product works in the same way, independently of the language selected. During this testing phase, the following verifications are performed: the ability to change the device language and how the User Interface reacts and refreshes after the switch; the non-existence of residual effects after the modification of the language (i.e., whether some strings remain in the previously set language); the language is correctly preserved after a reboot; and the input method applicable to the newly selected language loads automatically in the operating system.

6.2 Internationalization domain validations

Internationalization domain validations are a set of verifications to ensure that regional conventions and requirements are followed and that the apps are Unicode-compliant. This refers to, but is not limited to, measurement units, text alignment, date and time formats, addresses, phone numbers, and alphabetical ordering. This type of validation checks if these all load automatically in the operating system in accordance with applicable conventions.

It is worth mentioning, for example, that different Input Method Editors (e.g. QWERTY and QWERTZ) apply to different languages. Due to these differences, the keyboard layout must be validated by checking its layout compared with internal reference documents or based on the Gboard keyboard. Together with the layout validation, there is a need to check its ability to input and display non-ASCII characters. ASCII stands for American Standard Code for Information Interchange and is a set that contains control characters, punctuation marks, digits, and uppercase and lowercase versions of letters in the English alphabet. Accent characters like “ã”, “á”, “é” and the cedilla “ç” are examples of non-ASCII characters.

Another scenario for validation is whether text (entered via any of the input methods provided on the device) is received in a support device without the omission of any characters when sent via email or messaging apps. A validation to check if characters from a received message or accessed via web are rendered without corruption is also part of the scope, as well as a verification if the installed font integrated to the software build (version) supports the glyphs – otherwise missing glyphs will be shown as the replacement characters “□” or “🞋”.

When it comes to time and date, Quality Assurance is also responsible for checking if the format of default system settings and information under various applications are in accordance with the latest official release version of regional conventions defined by ICU/CLDR as described in §3.3. For instance, “*Donnerstag, 27. April 2023*” is an example of an acceptable date format when the system language is changed to German (Germany) while “*jueves, 27 de abril de 2023*” is the acceptable format applied for Spanish (Mexico). If the locale convention is not followed, it might have a significant impact on the user’s comprehension. For example, for Portuguese (Brazil), the date format convention is DD/MM/YYYY, while for English (United States), it should be MM/DD/YYYY. The date 03/04/2023 is understood as April 3, 2023 by Brazilians whereas Americans would read it as March 4, 2023. Locale is the combination of language plus country and covers the region and cultural elements.

Other relevant locale-specific conventions are:

- The position of the % sign and decimal separators (e.g. “%5,5” for Turkish (Turkey) and “5,5 %” for French (France);
- Measurement units (e.g. °F or °C for temperatures; miles or km for distances; and lb or kg for weight);
- Address formats;
- Layout orientation (i.e. while most of the languages are written from left to right, some languages like Arabic are written from right to left, and the user interface should mirror the language conventions with the alignment on the right side).

7. Localization language support levels

There are several different localization support levels to which a company may decide to commit and engage in. In the case of the Digital Inclusion, an initiative primarily involving software, it is especially important to make that decision clear to all stakeholders, given the potential for added complexities such as scarcity of linguistic experts or native speakers. An additional complexity is that there may be a low or non-existing internationalization support level (which is a dependency for localization) by the targeted operating system or by Unicode. Both areas of consideration are described in §4 and §5. Localization support levels may include categorizations such as Basic, Partial or Full. When working on software localization, regardless of which operating system the initiative is targeted for, there are various content types to consider, including the operating system itself (core elements such as system settings, error messages and other types of user notifications), preloaded applications, server content, etc. There are also supporting documents that complement information to users, which may need to be localized to improve usability information of the software product/project in the Target languages and regions (like printed users guides, online help content and legal documents). For the purpose of this document, the following is assumed for content pertaining to software:

- A **Basic** support level refers to only some elements of a smartphone user interface being localized into a particular language. Such elements could be specific to format types for the Target language region (e.g., month names, weekdays names or time zones in the Settings app). The entire smartphone operating system, as well as preloaded applications, would remain in a default language (e.g., English or other available language) other than the desired Indigenous or non-Indigenous language intended to be localized.

- A **Partial** localization support level would entail that a larger portion of the smartphone user interface is translated into the desired language. Partial localization support could mean that the operating system, including apps like Settings, Dialer and Calendar for example, is localized, but preloaded applications (like company proprietary applications or other bundled apps) are not, or the other way around.

- A **Full** localization support level indicates the operating system plus a company's developed applications are localized, giving users a broader exposure to their native language. Third party applications are not considered here in the localization support level categorization, given that companies do not usually own or have direct control or access to their software content (which can be downloaded independently, from an app storefront). This means that a company can provide a Full localization support for a given language, while downloadable or other third-party apps may not support it and will be displayed in alternative languages.

When planning for digitizing an endangered Indigenous language, it is recommended that, before deciding and communicating on the intended localization support level, the considerations noted in the §2 and §3 are well-researched, given their impact on scope, budget, complexity and overall feasibility of the digitization process.

8. Indigenous peoples' feedback and continuity

Motorola and Lenovo Foundation introduced the Digital Inclusion initiative of endangered Indigenous languages in 2021, and as it moves forward over the next decade, it coincides with the period from 2022 to 2032 declared by the United Nations as the International Decade of Indigenous Languages (IDIL 2022-2032). Motorola and Lenovo Foundation expect to continue to raise awareness to the cause, acting for the survival of endangered languages, and encouraging the next generations of Indigenous communities to use the technology in their native languages. For that, it is important to look back on what has worked out well and what must be improved, based on the Indigenous peoples' candid feedback.

8.1 Feedback from UNESCO

As stated in §1, UNESCO estimates that we lose one Indigenous language every two weeks, resulting in around 3,000 unique languages being lost by the end of the century.[2] With the global initiative releasing first in Latin America with Indigenous languages spoken in Brazil, Colombia and Venezuela, partnering with UNESCO in Brazil to further promote the integration of endangered Indigenous languages into technology became essential for the continuity of this initiative. UNESCO Brasília's actions occur through technical cooperation projects in partnership with various government levels and different sectors of civil society whenever their purposes contribute to public policies for sustainable development related to themes of expertise that UNESCO works on.[29] Therefore, collaborating with their Communication and Information Unit has allowed us to discuss further impactful actions that could benefit the Indigenous peoples and languages. Moreover, this collaboration indicated to Motorola that the Digital Inclusion initiative was on the right path of broadly benefiting Indigenous communities.

For UNESCO, "language is a primary means for communicating information and knowledge, thus the ability to access content on the Internet in a language which one can use is a key determinant for the extent to which one can participate in the knowledge societies."[30] Consequently, Multilingualism and Accessibility are two of the six priorities of the UNESCO Information for All Program (IFAP). Therefore, Motorola's initiative to include Indigenous languages in smartphones enables the Digital Inclusion of underserved groups and maintains synergy with UNESCO's objectives regarding social inclusion by generating a sense of belonging and recognition of Indigenous cultures in the digital world.

In addition, UNESCO considers the transparency strategy of open-source a good practice, insofar as technical processes and language data are provided to the general public so that other companies producing smartphones can make their functionalities available to communities that speak endangered languages. The effort in adapting the keyboard to meet the needs of endangered Indigenous languages is a good practice when it comes to Digital Inclusion.

Motorola and Lenovo Foundation's initiatives to support the Digital Inclusion of endangered Indigenous languages through the writing systems in smartphones is significantly aligned with the Result 3 of the Global Action Plan for the International Decade of Indigenous Languages (IDIL 2022 - 2032): favorable conditions established for digital empowerment, freedom of expression, media development, access to information and language technology, alongside artistic creation in Indigenous languages.[31]

8.2 Feedback from partners

Important feedback regarding the initiative came from Wilmar da Rocha D'Angelis, a linguist at the State University of Campinas (Unicamp) who led the first phase of the initiative for Latin American Indigenous languages Nheengatu and Kaingang in 2020. He believes that an impact of the initiative that is difficult to measure is perhaps the most significant one: since Motorola is a company that is highly associated with cutting-edge technologies and has a good reputation and penetration in the Brazilian market, its initiative to support Indigenous communities, their cultures, and their languages is a powerful means of increasing the value of cultural diversity in Brazil. The initiative is, according

to Wilmar, a way of bringing awareness to the need of the industry to pay more attention to the earliest known inhabitants of an area.

According to Wilmar, "for the Indigenous people themselves – especially for a very large group, in each linguistic community of Kaingang or Nheengatu, of teachers and Indigenous intellectuals who are incessantly dedicated to strengthening their ancestral language – the Digital Inclusion initiative pointed out new spaces for action. The development and modernization of their languages and awakened new vocations, such as that of translator. There are already surveys of Indigenous teachers and young people coming to us, asking about the existence or possibility of developing translator courses, especially from the perspective of the relationship of their Indigenous languages with Portuguese and English. Some translation research groups at Brazilian universities also began to focus on the topic of translation to and from Indigenous languages."

He also added that "the specific keyboard aroused a great deal of interest in various communities that speak other Indigenous languages, and we have already been consulted several times about it; for some, it was enough to indicate that they adopt a smartphone equipped with Android 11 and select Nheengatu (in the case of the Guarani, for example) or Kaingang (in the case of the Apãniekrá, for example) in order to be able to take advantage of a useful keyboard for communication in their language. Obviously, it doesn't work for understanding the commands and instructions."

Wilmar points out to the fact that the collaborative work between speakers of varied dialects from three different regions in the Amazon to localize Nheengatu on Motorola smartphones generated the need for greater unification of the language.[32] That resulted in the creation of the Academy of Languages Nheengatu, a pioneering initiative among Indigenous languages in Brazil which came as a consequence of the Digital Inclusion initiative by Motorola and Lenovo.[33] The Academy of Languages Nheengatu aims to regularize Nheengatu, standardize it in three aspects and thus regain part of the lost space and in search of the place that was taken from it, the mother tongue that is the face of the identity of the Amazon.

In line with the feedback obtained from the first phase of the digitization initiative, Principal Chief Richard G. Sneed of the Eastern Band of Cherokee Indians, who, together with other Cherokee leaders who have spent several months consulting with Motorola, said that "having smartphone user interfaces localized in Cherokee is a tool that can be utilized to help the generation that is coming up to become more familiar with it, so it is going to be imperative that the Cherokee peoples incorporate that into the curriculum and into their everyday teaching methodology." He believes "Motorola has made a concerted effort to work with Indigenous people groups to incorporate the traditional languages into the technologies so that the languages don't disappear forever and that this is just one more piece of a very large puzzle of trying to preserve and proliferate the language."[34]

Dr. Benjamin Frey, assistant professor at UNC-Chapel Hill, who was the Eastern dialect expert on the Cherokee language team for this project, brought important enlightenment towards critical thinking regarding cultural differences. This feedback throughout the initiative deployment, especially when balancing between the corporate culture (e.g. product deadlines and technical requirements) and the Indigenous peoples' culture (their ways of living, needs, and communication style) were key for the success of the initiative's delivery to the community. Additionally, the scarcity of linguistic experts available to participate and the large scope of the project brought hurdles in the planning stages towards the end of the project. The whole involved team assimilated those difficulties and learned from that experience to further adapt the methods, styles, and channels of communication, resulting in a successful delivery of over 200,000 words of an open-sourced dataset in Cherokee language and the first smartphones with a fully localized user interface to promote it in 2022. The engagement with Chief Sneed from EBCI, Dr. Frey from UNC-Chapel Hill, and the Cherokee linguists was extremely valuable and necessary in order to shorten the learning curve that took place towards understanding the culture divide.

As for the Cherokee Indigenous peoples' feedback, Chief Sneed stated in 2022 that the initiative brought forth "a useful tool, especially adding it to Cherokee language classes in schools on the Qualla boundary." Additionally, Chief Sneed said that "what is happening with Lenovo and Motorola is that they recognize that there is a responsibility that goes with technology. There is a lot of power

that goes with having technology… in being the keepers of that technology. To see firms that are not just interested in profit but instead see their creation as a tool to enrich humanity and then to put the resources behind that to make it happen… I applaud that!" Noticeably, Cherokee Nation Principal Chief Chuck Hoskin Jr. said that "Anytime a business can incorporate the Cherokee language into its product for mass citizens to learn, it's a win not just for Cherokee Language preservation, but for the perpetuation of all Native languages."

While the feedback obtained from professors from North and Latin America pertains mostly to the impact of the open-sourced data and the inspirational aspect of the representation of Indigenous languages in technology, Professor Sharma Suhn, from Hamirpur- Kangra Valley, believes the ability to choose Kangri from the list of languages supported in Motorola smartphones has brought pride and meaningfulness to the population of his town.

The Digital Inclusion initiative released in 2023 involving the Kuvi language was in Asia Pacific and that brought positive feedback from Dr. Sushree Sangita Mohanty, Assistant Professor of Anthropology, who worked on the initiative. She believes the initiative "is also creating opportunities for the non Kuvi speakers to get engaged and learn this language through the four different mediums (Kuvi-Odia, Kuvi-Telugu, Kuvi-Devanagiri, Kuvi-Latin). Hence, this model is expanding its user numbers, which can help to increase the usage of Kuvi language."

Finally, the latest phase of the Digital Inclusion initiative, which was released in 2024, involved Ladin, a language spoken in the Dolomite region of Italy. This phase was well-received by the community, including our collaborator Professor Paul Videsott, Professor of Romance Philology at the Free University of Bozen-Bolzano (UNIBZ). Professor Videsott believes this initiative "definitively will help Ladin and other minority languages to be more visible… mobile phones are like the pencil of the 21st century, and having minority languages, and having Ladin in [them], has the same importance of having a language in a book… in the centuries before. Having Ladin in these mobile phones shows, not only for the Ladins, but to all users of smaller languages in the world, that their language was not only useful for the twenty centuries before our time, but will be useful also in the future." Additionally, Johann Gamper, Professor of Computer Science and Vice-Rector of Research at UNIBZ, remarks that this initiative also impacts younger generations of speakers: "Young people use these mobile phones much more than we do. My feeling is that they would definitely use this Ladin interface… I'm sure they will use their local language, in this case, Ladin."

8.3 Final thoughts and continuity

At its core, the Indigenous language project represents the social impact that inclusive technology can make. The project aligns with Lenovo Foundation's mission to empower diverse populations with access to technology, as well as with Lenovo's vision to provide smarter technology for all. There were key factors present that distinguished the opportunity for its quality social impact:

- Linguistic expertise: The foundation of this project is the passionate linguistic expertise of Motorola's Globalization team. Their passion for the project and commitment to its quality and the communities it has impacted was a positive sign for its effectiveness. Lenovo Foundation's support was focused on the empowerment of Indigenous peoples and promoting endangered language preservation, aligned with Motorola's Globalization team's initiatives. The project's impact and credibility were made possible because of Motorola's subject matter experts and leadership.

- Respect for Indigenous cultures: A notable conclusion from Motorola's technologists through the partnership with scholarly experts and educational institutions is that the project would not be possible without peers and scholars in the field as well as. a significant amount of trust from the Indigenous communities invited for participation. That trust could only be earned by engaging scholarly experts who helped the team translate, localize, and interact with Indigenous communities with sensitivity and care. Thus, those peers, scholars, and peoples and cultures empowered through the project deserve the utmost respect.

- Commitment to quality: Powered by linguistic expertise, academic experts, and collaboration with Indigenous communities, Motorola's experts ensured digitization through evaluation of the current existence (or not) of languages on the Unicode platform. The leaders carefully selected languages in alignment to UNESCO's guidelines and the assessments of what could be possible as employee volunteers. Their focus on attempting what was feasible to achieve at the highest standard of localization has led to a quality contribution to the digitization of Indigenous languages, recognized by UNESCO and with plans to continue each year.

- Impact beyond Lenovo and Motorola: Most importantly, the digitization was not conducted just for Motorola users. The resources created by the Motorola team through support from Lenovo Foundation can be openly-sourced and leveraged so that other technology OEMs can localize their devices and embrace Lenovo Foundation's mission to empower underrepresented populations with access to technology. For the Lenovo Foundation, it is critical that this initiative is trumpeted as an act of inclusion to build awareness for Indigenous communities and is not trumpeted as an act of having a competitive advantage. The team maintains a vision of more OEMs joining this initiative to preserve language and ensure the preservation of human heritage and diversity of language on our planet.

With the core tenants of passionate employee experts, a commitment to respect for diverse and underrepresented cultures, commitment to quality through collaboration of scholars and institutions, and a vision for impact broader than Motorola devices, the Indigenous language project will continue. The team anticipates further contributions during the International Decade of Indigenous Languages (IDIL 2022-2032) by continuing the project to include endangered Indigenous languages.

Endnotes

1 UN Department of Social and Economic Affairs. (February 10, 2023). Why indigenous languages matter: The International Decade on Indigenous Languages 2022–2032. Retrieved from https://www.un.org/development/desa/dpad/publication/un-desa-policy-brief-no-151-why-indigenous-languages-matter-the-international-decade-on-indigenous-languages-2022-2032/

More information about Indigenous languages can be found at UNESCO, via https://www.unesco.org/en/articles/motorola-and-lenovo-foundation-announce-next-phase-initiative-revitalize-endangered-indigenous

2 United Nations. (April 19, 2018). The United Nations Permanent Forum on Indigenous Issues. Retrieved from https://www.un.org/development/desa/indigenouspeoples/wp-content/uploads/sites/19/2018/04/Indigenous-Languages.pdf

3 Krauss, M. (1992). The World's Languages in Crisis. ResearchGate. Retrieved from https://www.researchgate.net/publication/300468999_The_world's_languages_in_crisis

4 UNESCO. Motorola and Lenovo Foundation announce next phase of Endangered Indigenous Languages Revitalization Initiative at UNESCO HQ. (December 19, 2022). Retrieved from https://www.unesco.org/en/articles/motorola-and-lenovo-foundation-announce-next-phase-initiative-revitalize-endangered-indigenous

5 UNESCO IESALC. (February 21, 2022). A decade to prevent the disappearance of 3,000 languages. Retrieved from https://www.iesalc.unesco.org/en/2022/02/21/a-decade-to-prevent-the-disappearance-of-3000-languages/

6 UNESCO World Atlas of Languages. (n.d.). Retrieved from https://en.wal.unesco.org/

7 University of Connecticut. (November 15, 2023). Endangered languages: UNESCO classification. Retrieved from https://guides.lib.uconn.edu/c.php?g=1232158&p=9415488

8 IONOS. (June 12, 2023). What is Unicode?. IONOS Digital Guide. Retrieved from https://www.ionos.com/digitalguide/websites/website-creation/unicode/

9 Everything you need to know about localization. Smartling. (n.d.). Retrieved from https://www.smartling.com/resources/101/localization-101/

10 8 key steps in the localization process. Redokun Blog. (n.d.). Retrieved from https://redokun.com/blog/localization-process

11 MotionPoint. (March 30, 2022). Translation Management Systems (TMS): A comprehensive guide. MotionPoint. Retrieved from https://ru.motionpoint.com/blog/translation-management-systems-tms-a-comprehensive-guide/

12 Prevaly, S. (September 1, 2022). Computer-Assisted Translation (CAT): A complete guide. MotionPoint. Retrieved from https://www.motionpoint.com/blog/computer-assisted-translation-cat-a-complete-guide/

13 Dalibor. (November 28, 2022). Translation memory: What it is, and how to use it. Phrase. Retrieved from https://phrase.com/blog/posts/translation-memory/

14 Sokolov, I. (May 16, 2023). Software localization: Getting your product ready for the global market. Translation & Localization Blog. Retrieved from https://www.smartcat.com/blog/software-localization/

15 Dimmendaal, G. (2001). "Places and People: Field Sites and Informants" in Newman, P. & Ratliff, M. (eds.) Linguistic Fieldwork. Cambridge University Press. 55-75.

16 Cameron, D., Frazer, E., Harvey, P., Rampton, M.B.H., and Richardson, K. (1992). Researching language: Issues of power and method. London and New York: Routledge.

17 Czaykowska-Higgins, E. (2009). Research Models, Community Engagement, and Linguistic Fieldwork: Reflections on Working within Canadian Indigenous Communities. Language Documentation & Conservation, 3, 15-50. Retrieved from http://hdl.handle.net/10125/4423

18 Sampson, G. (2016). "Writing systems: methods for recording language" in Allan, K. (ed.) The Routledge Handbook of Linguistics. Routledge. 47-61.

19 Unicode. Submitting Character Proposals. (April 1, 2016). Retrieved from https://www.unicode.org/pending/proposals.html

20 Unicode 15.0 character code charts. Unicode (n.d.). Retrieved from https://www.unicode.org/charts

21 Unicode: Where is my character? (September 28, 2018). Retrieved from https://unicode.org/standard/where/

22 Unicode. Private-use characters, Noncharacters & Sentinels FAQ. Unicode. (n.d.-a). Retrieved from https://www.unicode.org/faq/private_use.html

23 International Components for Unicode. ICU. (n.d.). Retrieved from https://icu.unicode.org/

24 Unicode. Common Locale Data Repository. Unicode CLDR. (n.d.). Retrieved from https://cldr.unicode.org/

25 Rouse, M. (October 6, 2020). Android SDK. Retrieved from https://www.techopedia.com/definition/4220/android-sdk

26 ISO 639 code tables: ISO 639. SIL International. (n.d.). Retrieved from https://iso639-3.sil.org/code_tables/639/data

27 CLDR survey tool. Unicode CLDR. (n.d.). Retrieved from https://cldr.unicode.org/index/survey-tool

28 Raster, Vector, TrueType, and OpenType fonts. Microsoft Learn. (January 7, 2021). Retrieved from https://learn.microsoft.com/en-us/windows/win32/gdi/raster--vector--truetype--and-opentype-fonts

29 UNESCO Brasília. UNESCO.org. (n.d.). Retrieved from https://www.unesco.org/en/fieldoffice/brasilia.

30 UNESCO. Multilingualism. UNESCO.org. (April 20, 2023). Retrieved from https://www.unesco.org/en/ifap/multilingualism

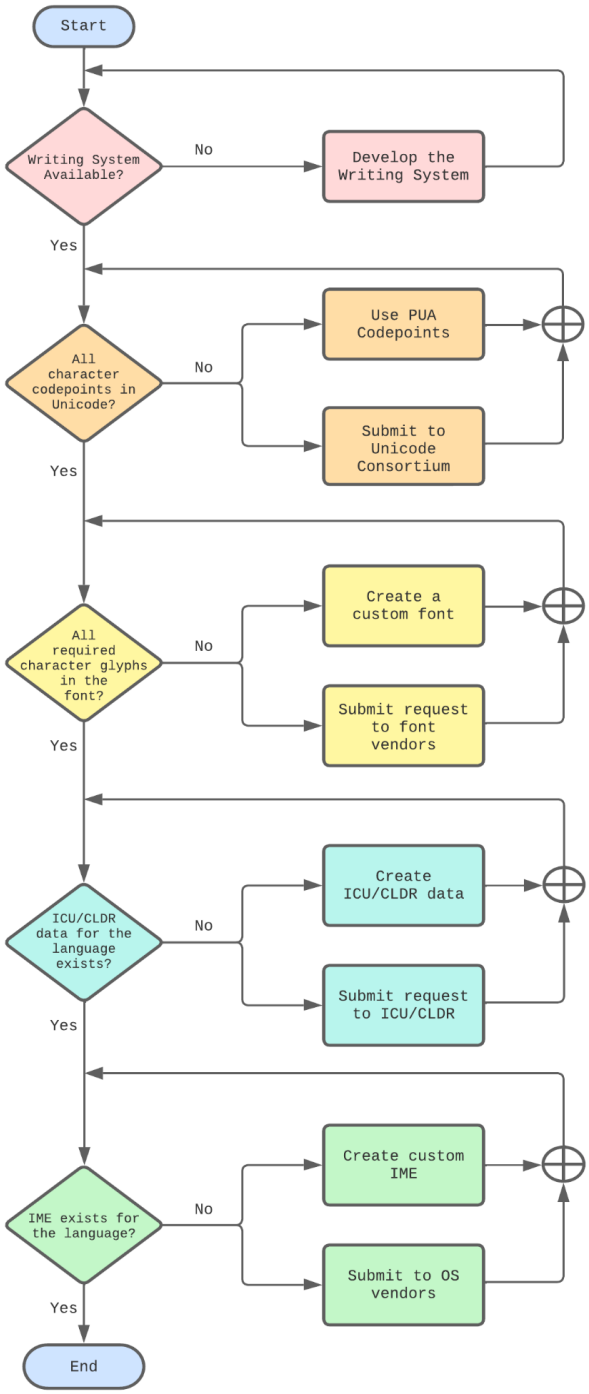
31 UNESCO. Global action plan of the International Decade of Indigenous Languages (IDIL 2022-2032). UNESDOC Digital Library. (2021). Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000379851

32 da Rocha D'Angelis, W. (February, 2023). A língua nheengatu e Suas Ortografias: Questões técnicas e de política linguística. ResearchGate. Retrieved from https://www.researchgate.net/publication/369577326_A_lingua_Nheengatu_e_suas_ortografias_questoes_tecnicas_e_de_politica_linguistica

33 Ternes, P. (2021). Projeto inédito cria configuração de smartphone em Kaingang e Nheengatu. Kamuri. Retrieved from https://kamuri.org.br/kamuri/projeto-inedito-cria-configuracao-de-smartphone-em-kaingang-e-nheengatu/

34 Hodge, R. (March 2, 2022). Cell phone technology keeps endangered Cherokee language a tap away. WLOS. Retrieved from https://wlos.com/news/local/cherokee-motorola-endangered-language-android-12-eastern-band-of-cherokee-indians-principal-chief-richard-sneed-north-carolina

Appendix



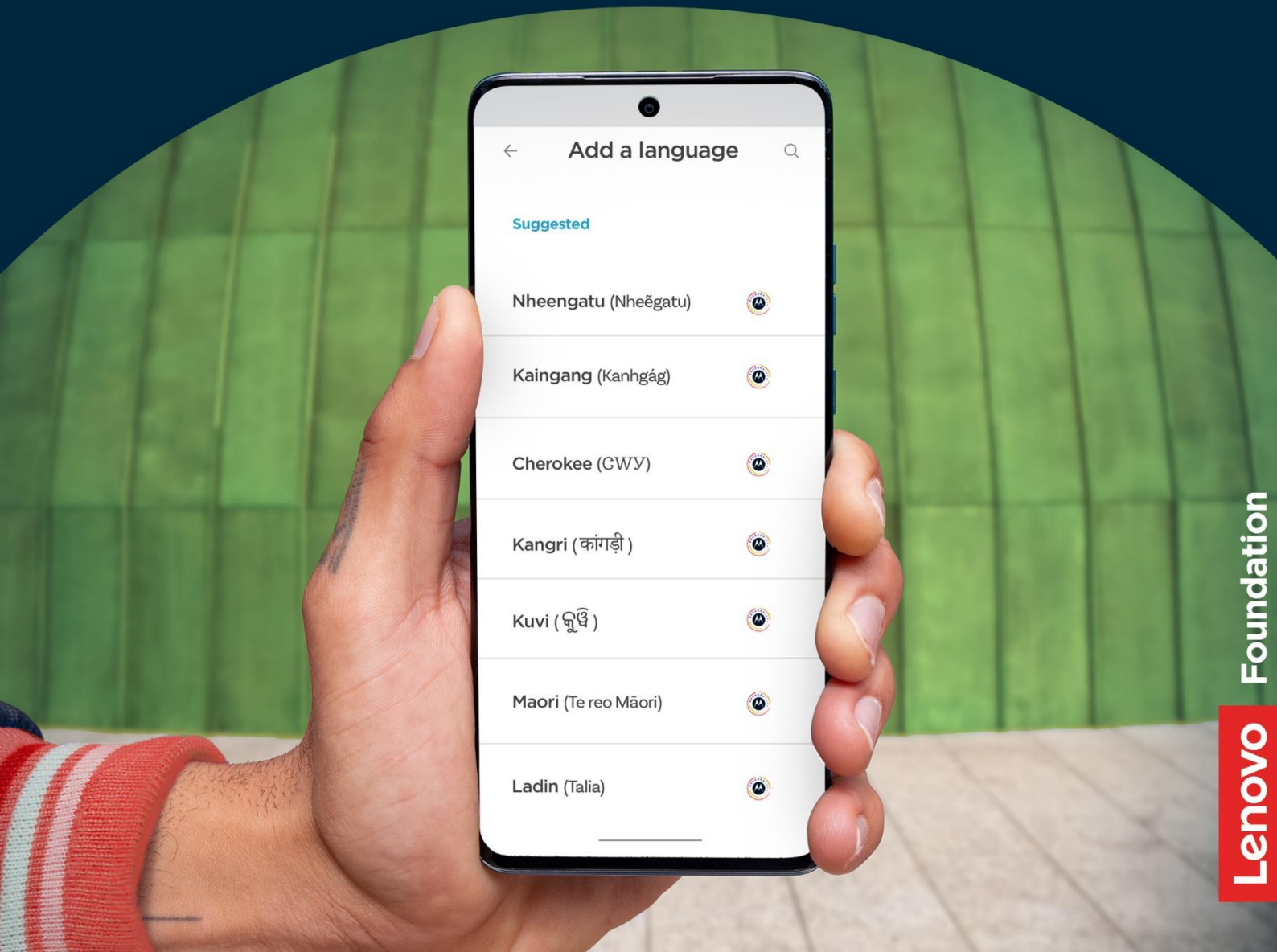
Flowchart illustrating Internationalization support levels by color.

motorola

hello indigenous

Um modelo de projeto de inclusão digital para preservação de línguas indígenas ameaçadas

PUBLICADO EM 2024



Lenovo Foundation

Esta publicação conta com a cooperação da UNESCO na parceria Indigenous Languages on Mobile. Ela visa produzir um estudo e um resumo executivo sobre o desenvolvimento tecnológico que possibilitou a inserção de línguas Indígenas presentes no Brasil, na América do Norte, na Europa e nos territórios da Ásia-Pacífico nos dispositivos da doadora. As indicações de nomes e a apresentação desta publicação não implicam a manifestação de qualquer opinião por parte da UNESCO a respeito da condição jurídica de qualquer país, território, cidade, região ou de suas autoridades, tampouco da delimitação de suas fronteiras ou limites. As ideias e opiniões expressas nesta publicação são as dos autores e não refletem obrigatoriamente as da UNESCO nem comprometem a Organização.

Autoria

Janine Oliveira (Autora, Supervisora)
Marison Ranieri Rodrigues de Freitas (Autor, Compilador)
Delaney Gomez-Jackson (Autora, Pesquisadora)
Juliana Peres Rebelatto Pereira
Natalia Sarmento Tenório Falcão
Roy Yokoyama
Sushil Garg
Yukitomi Fujinaga

Revisão

Guilherme Saba (Revisor, Diagramador)
Livia Teixeira
Marcelo Mazzini
Natàlia Borja Gutès

Colaboração

Adauto Candido Soares
Jaco Du Toit
Hezekiel Dlamini
Maria Luzia de Cerqueira Gomes
Luciana Vedovato
Manuella Foz
Mahmoud Ebrahim
Monica Hauser
Alice Damasceno Saiki
Pratima Harite
Santiago Mendez Galvis
Sydni Behm

Capa

Giovanna Paganini

ISBN 979-8-218-37915-5

1. Inclusão Digital: Metas e não metas desta iniciativa para preservação de línguas Indígenas ameaçadas de extinção

1.1 Antecedentes das línguas Indígenas no mundo

Uma língua Indígena é uma língua nativa de uma região e falada por povos Indígenas. Essas línguas são frequentemente reduzidas ao status de línguas minoritárias em suas respectivas regiões porque são faladas por grupos de pequena escala. Assim, essas línguas geralmente não são línguas nacionais (embora possam ser, como o aimará, uma língua oficial na Bolívia). As línguas Indígenas não reconhecidas por seus governos muitas vezes não recebem apoio para a educação e revitalização. De fato, existe uma correlação entre as línguas Indígenas e as línguas ameaçadas, na medida em que as comunidades Indígenas geralmente assimilam a língua majoritária de sua região.

De acordo com o informe do Departamento das Nações Unidas para os Assuntos Econômicos e Sociais (UN DESA, na sigla em inglês), 40% das 6.700 línguas faladas no mundo estão em risco de desaparecer[1]. Muitas delas são línguas Indígenas, que desempenham um papel essencial em várias facetas das culturas Indígenas, tais como definir as relações Indígenas com a Terra, preservar o território Indígena e transmitir a história, a ciência e suas visões gerais de mundo. Embora os povos Indígenas representem 6% da população global, eles falam mais de 4.000 das línguas do mundo[2]. Muitas línguas Indígenas enfrentam o perigo de desaparecer em um futuro próximo devido a vários fatores, como colonização, globalização, assimilação e discriminação. Em *The World's Languages in Crisis*, prevê-se que metade das línguas do mundo serão perdidas durante este século[3]. Consequentemente, os esforços que promovam o engajamento em projetos de revitalização linguística são essenciais para a preservação das línguas e culturas Indígenas.

1.2 Metas e não metas da iniciativa

Este documento é um relatório detalhado de melhores práticas para a Inclusão Digital de falantes de línguas Indígenas ameaçadas de extinção e é direcionado à sociedade civil, abrangendo, como um exemplo, pessoas físicas, organizações não governamentais e empresas. A sociedade civil desempenha um papel fundamental em trazer perspectivas únicas para a mesa, impulsionando a mudança social e defendendo práticas comerciais éticas, além daquelas apenas lucrativas. Juntos, essa força coletiva promove a colaboração, a inclusão e o empoderamento para construir um futuro mais abrangente e próspero para todos.

A proclamação do período entre 2022 e 2032 como a Década Internacional das Línguas Indígenas (IDIL 2022-2032), conforme a Resolução A/RES/74/135 da Assembleia Geral das Nações Unidas, ilumina a questão e convida o mundo a prestar mais atenção à situação crítica e frágil de muitas línguas Indígenas. Este documento visa fornecer um modelo para a adição de idiomas Indígenas ao software, a fim de auxiliar outras ações de Inclusão Digital, e fornecer um conjunto de etapas recomendadas a serem seguidas para alcançá-lo com sucesso.

A iniciativa de Inclusão Digital apresentada neste documento é focada principalmente, mas não exclusivamente, no Android em um nível técnico. Para outros sistemas operacionais, pode haver requisitos e padrões adicionais, portanto, recomenda-se consultar a documentação relacionada.

Metas

À medida que as novas gerações de povos Indígenas aumentam sua alfabetização e uso da tecnologia, é fundamental que eles sejam capazes de usar sua língua nativa em formatos digitais para evitar que ela entre em risco de extinção e seja perdida[4]. A UNESCO estima que perdemos uma língua Indígena a cada duas semanas, resultando na perda de cerca de 3.000 línguas únicas até o

final do século[5]. Para ajudar a preservar nossa herança humana, as histórias únicas das culturas Indígenas e a empoderar a próxima geração, esta iniciativa de Inclusão Digital visa integrar em smartphones línguas Indígenas ameaçadas de extinção.

À medida que esta iniciativa de Inclusão Digital prossegue no decorrer da próxima década, seu objetivo principal é servir às comunidades por meio da conscientização sobre as línguas Indígenas ameaçadas de extinção. A execução da iniciativa de Inclusão Digital também pode atender a outras e diversas necessidades dos povos Indígenas, pois permite um acesso mais fácil à tecnologia, traz ações para a sobrevivência de línguas ameaçadas e capacita as futuras gerações de comunidades Indígenas a usar a tecnologia em sua língua nativa.

Kaingang (falado no sul do Brasil), nheengatu (falado na Amazônia), cherokee (falado nos Estados Unidos), kangri (falado na Índia), te reo māori (falado na Nova Zelândia e Austrália) e o ladino (falado na Itália) agora fazem parte dos mais de 90 idiomas oferecidos na interface de usuário de smartphones da Motorola. Além disso, um teclado virtual em kuvi (falado na Índia) foi desenvolvido e disponibilizado para download em smartphones.

Com os objetivos específicos de inserir a forma escrita de uma língua oral em um meio usual de comunicação, como a tecnologia digital, conscientizar a sociedade sobre as línguas Indígenas ameaçadas e trabalhar para sua revitalização, a Motorola disponibilizou em seu website o código-fonte de mais de 800.000 palavras Indígenas traduzidas em abril de 2023, permitindo que outras Fabricantes de Equipamento Original (OEMs, na sigla em inglês) e empresas promovam os idiomas por meio de suas interfaces proporcionando, dessa forma, oportunidades para aplicações e esforços mais amplos de revitalização. Para começar, em 2022, a Lenovo, empresa controladora da Motorola, implantou a integração dos idiomas latino-americanos nheengatu e kaingang em seus PCs; e o Gboard, teclado virtual nativo do Android, agora suporta essas duas línguas para uso mais extenso.

Não metas

Sabendo que existem mais de 3.000 línguas Indígenas faladas no mundo, e a fim de priorizar as que correm maior risco de extinção enquanto trabalha-se para alcançar o objetivo, a iniciativa de Inclusão Digital impulsionada pela Motorola se concentra principalmente nas categorias da UNESCO de "Definitivamente em perigo", "Severamente em perigo" ou "Criticamente em perigo".

2. Critérios de seleção de idioma

2.1 Categorização da UNESCO

Os níveis de risco linguístico da UNESCO fornecem classificações sobre o quão ameaçada uma língua é com base na transferência intergeracional. Existem cinco níveis[6], e um breve resumo de cada nível é o seguinte[7]:

- **Vulnerável:** a maioria das crianças fala o idioma, mas pode ser restrito a certos domínios, por exemplo, em casa.
- **Definitivamente em perigo:** as crianças não aprendem mais, em sua casa, o idioma como língua materna.
- **Severamente em perigo:** a língua é falada pelos avós e pelas gerações mais velhas; embora a geração dos pais possa entendê-la, eles não a falam com os filhos ou entre si.
- **Criticamente em perigo:** os falantes mais jovens são avós e mais velhos, e falam a língua parcialmente e com pouca frequência.
- **Extinta:** não há mais falantes.

A estratégia de Inclusão Digital da Motorola se concentra principalmente em três categorias: Definitivamente em perigo, Severamente em perigo e Criticamente em perigo (embora outras línguas em perigo também sejam consideradas).

2.2 Como os idiomas são selecionados

Os critérios para a seleção de idiomas ameaçados pela iniciativa de Inclusão Digital da Motorola são sustentados por quatro pilares principais, a serem expandidos abaixo: categoria de ameaça ao idioma, categoria de Inclusão Digital, contribuição e sentimento da comunidade em relação à iniciativa e disponibilidade de especialistas no assunto (SMEs, na sigla em inglês).

A categoria de ameaça ao idioma é o primeiro critério para determinar a inclusão e pode ser avaliado usando dados concretos, principalmente usando o vasto conjunto de dados fornecido pelas atividades da UNESCO, e atua como um "termômetro" para avaliar o risco de uma determinada língua e, portanto, a elegibilidade para a iniciativa de Inclusão Digital.

A categoria de Inclusão Digital é uma medida de quão bem um idioma é representado e suportado no mundo digital. Ele depende de vários fatores, como a disponibilidade de recursos, ferramentas e serviços digitais para a língua, o nível de acesso e participação dos falantes e o grau de reconhecimento e proteção dos direitos e diversidade linguística.

Um dos fatores é se o alfabeto da língua é suportado pelo Unicode, que é um padrão para representação de caracteres que abrange a maioria dos sistemas de escrita do mundo[8]. O Unicode suporta não apenas o alfabeto latino, mas também os alfabetos grego, cirílico, árabe, hebraico e tailandês, bem como os sistemas de escrita japonês (katakana, hiragana), chinês e coreano (hangul), por exemplo. Além disso, há também caracteres matemáticos, comerciais e técnicos, e caracteres de controle histórico para teleimpressoras. No entanto, alguns idiomas podem ainda não ter seus alfabetos codificados em Unicode, o que é explorado em mais detalhes na seção 5.

Outro fator é se há dados de localidade disponíveis para o idioma, que é um conjunto de parâmetros que define a língua, a região e as preferências culturais do usuário. Os dados de localidade podem afetar a forma como as datas, horas, números, moedas e outros formatos são exibidos, bem como a forma como o texto é ordenado e pesquisado. Esses dados de localidade podem ser fornecidos

pelo sistema operacional, pela linguagem de programação ou pelo aplicativo. Dependendo das peculiaridades de um idioma e de sua categoria de perigo, pode ser feito um esforço para criar um teclado e um layout e implementá-lo na plataforma (o Android é o foco deste documento em termos de sistema operacional, conforme detalhado na seção 1). Um exemplo também é encontrado na seção 5.

A contribuição e o sentimento da comunidade em relação à iniciativa de Inclusão Digital é uma parte importante do método de seleção, pois algumas comunidades podem apoiar mais a iniciativa de revitalização do que outras. Isso pode ser decisivo para levar um projeto adiante ou interrompê-lo, já que devido à própria natureza da iniciativa, há um pequeno quórum de sujeitos e, portanto, a contribuição e a cooperação são primordiais. Isso só pode ser alcançado por um sentimento positivo em relação à iniciativa de Inclusão Digital por parte da comunidade, o que facilita o fluxo de comunicação.

A disponibilidade de especialistas no assunto (SMEs) também é de grande importância, pois a necessidade de pesquisadores, professores, tradutores e linguistas faz parte da estratégia de revitalização multifacetada. Os SMEs ajudam a preencher a lacuna entre a comunidade e a instituição em termos de comunicação, compreensão de ideias, fornecimento de pesquisas e trabalhos acadêmicos existentes, além de encontrar um terreno comum e alavancar talentos e eficiência, o que facilita a iniciativa. Isso também pode representar um grande desafio, pois alguns idiomas têm mais pesquisadores, professores, tradutores e linguistas do que outros, portanto, a aquisição de SMEs pode ser difícil.

Como afirmado antes, os esforços da Motorola concentraram-se principalmente em três das categorias estabelecidas pela UNESCO: Definitivamente em perigo, Severamente em perigo e Criticamente em perigo, pois vimos uma lacuna em termos de players e esforços nessas categorias. Assim, os exemplos presentes neste documento refletem essa decisão estratégica.

3. Configuração do processo de iniciativa de Inclusão Digital

3.1 Aceitação e receptividade da Inclusão Digital dos povos Indígenas

É importante respeitar e reconhecer os protocolos culturais ao colaborar com comunidades e grupos Indígenas. Portanto, a recomendação é que as interações aconteçam por meio de acadêmicos, organizações ou instituições que já tenham relações anteriores significativas e duradouras com os povos Indígenas. Iniciar o relacionamento com os povos Indígenas por meio de alguém de sua confiança pode preencher a lacuna que se pode ter em conhecer os protocolos culturais que são a etiqueta, os costumes, os códigos e outros comportamentos de uma determinada comunidade ou grupo cultural e os processos apropriados para colaborar em uma iniciativa com esse grupo.

Embora os protocolos compartilhem alguns temas e práticas comuns, é importante reconhecer que há muita diversidade entre os povos e comunidades Indígenas. Cada comunidade tem sua própria cultura, patrimônio e língua, o que influencia o protocolo adequado. As tradições Indígenas e as condutas éticas são importantes para manter redes e estabelecer relacionamentos respeitosos. É recomendável consultar os representantes da comunidade sobre o protocolo apropriado nela.

O solicitante deve compartilhar previamente os detalhes de alto nível da iniciativa de Inclusão Digital e seus objetivos, por exemplo, "a meta é digitalizar um idioma escrito, habilitando-o na interface do usuário de um smartphone" ou "a finalidade é promover workshops que visem documentar um idioma falado pela primeira vez para desenvolver um teclado para smartphones que o suporte, permitindo que seus falantes se comuniquem em sua língua materna".

Ao apresentar a iniciativa de Inclusão Digital, é vital estar atento ao fato de que o assunto em discussão é seu próprio idioma nativo e que, a menos que o solicitante seja da comunidade em questão, essa é uma ideia vinda de alguém de fora. O diálogo entre a empresa patrocinadora da iniciativa, os linguistas e a comunidade falante é incentivado, para que a iniciativa se adeque às necessidades da comunidade. Por exemplo, ao decidir qual dialeto é escolhido para o sistema de escrita, as discussões entre a própria comunidade falante são essenciais. Assim, a escuta ativa é aconselhada enquanto alguém está falando, a fim de fazer perguntas relacionadas aos benefícios reais que a iniciativa traria para sua comunidade, o nível de engajamento da comunidade com internet e smartphones e se as crianças aprendem e falam a língua na escola e em casa. Essas questões podem definir melhor o impacto da iniciativa, bem como demonstrar respeito à cultura e às necessidades da comunidade. .

Uma vez que todas as necessidades linguísticas da comunidade são definidas e acordadas tanto pela comunidade quanto pelo solicitante, uma proposta é feita para identificar falantes (idealmente com experiência em escrita ou tradução) dispostos e disponíveis para trabalhar na localização dos dados de idioma de um conteúdo específico em sua língua durante um período de tempo predeterminado.

Contanto que a comunidade esteja respeitosamente envolvida desde o início, entenda e concorde com os benefícios da iniciativa para a revitalização do sua língua e com a compensação que as partes envolvidas terão pelo trabalho realizado, a aceitação e a receptividade devem vir como consequência.

3.2 Critérios de seleção de parceiros fornecedores e parcerias com organizações sem fins lucrativos

O processo de seleção de fornecedores é uma série de etapas relacionadas a aquisições que determinam os requisitos de serviço e produto e os combinam com os recursos e preços do fornecedor. Para iniciativas de Inclusão Digital, é comum considerar o envolvimento com parceiros habituais com os quais a empresa já trabalha (normalmente, um provedor de serviços linguísticos), da mesma forma que se faria com qualquer projeto de inclusão linguística em uma empresa de tecnologia. No entanto, a seleção de prestadores de serviços linguísticos para línguas Indígenas

ameaçadas de extinção pode ter requisitos adicionais, pois a parceria com comunidades Indígenas pode ser uma novidade para a maioria dos fornecedores.

Quase tão importante quanto avaliar a estabilidade financeira dos fornecedores é garantir, por meio de conversas abertas, que fornecedores e parceiros compartilhem os mesmos pilares e valores éticos. Em muitos casos, como nenhum fornecedor conta com um pool global bem estruturado de staff linguístico no idioma escolhido ainda, é necessário selecionar essa equipe para uma prestação de serviços personalizada. Isso exige a definição de novos acordos de qualidade e produtividade para garantir que todas as partes envolvidas sejam respeitosas e atenciosas com a cultura da comunidade Indígena. Ainda que existam semelhanças, os marcos e entregas que precisam ser definidos para as línguas ameaçadas não podem ser os mesmos dos idiomas amplamente falados. Por isso, o objetivo durante a seleção de parceiros fornecedores deve ser garantir que o trabalho realizado durante a iniciativa de revitalização e digitalização linguística demonstre um compromisso contínuo com os direitos dos povos Indígenas, atendendo às limitações corporativas em termos de recursos e marcos predefinidos, como a data de lançamento do produto.

Um aspecto que deve ser considerado durante a seleção do parceiro fornecedor é garantir que todo o pessoal envolvido no processo (tradutores, revisores e gerentes de projeto) esteja ciente e reconheça que o conteúdo localizado produzido é Propriedade Intelectual do cliente. Na iniciativa de revitalização da Motorola e da Lenovo, esse acordo e conscientização entre todas as partes (por exemplo, comunidade Indígena, tradutores, revisores e gerentes de projeto) são fundamentais para a estratégia de disponibilização via código aberto (open sourcing) do corpus: tornar o conteúdo disponível para outros que estão igualmente comprometidos com o trabalho de revitalização de idiomas.

As relações corporativo-Indígenas podem facilmente contar com um compromisso de cima para baixo dos líderes empresariais, e é por isso que os critérios para consultores linguísticos e/ou tradutores devem ser diferenciados daqueles aplicados a recursos menos experientes selecionados com base na disponibilidade (em horas ou dias) para trabalhar no projeto. Para fazer essa diferenciação, nas fases iniciais do processo de negociação com parceiros fornecedores, recomenda-se que o questionamento seja feito diretamente ao linguista em termos de capacidade, para que os dados sejam levados em consideração como critério para o processo de seleção do idioma.

Normalmente, a recomendação é que a empresa pré-selecione dois ou três fornecedores em potencial e aplique a lista de verificação de critérios de seleção de fornecedores. Para a *due diligence* do fornecedor, diferentes membros da equipe participam do processo de avaliação. Os itens a serem negociados em um contrato incluem:

- Número de falantes para o idioma Indígena selecionado;
- Número de tradutores profissionais para o idioma Indígena selecionado;
- Preços por nova palavra traduzida;
- Marcos em termos de capacidade diária e disponibilidade por falante/tradutor;
- Aceitação final dos dados de idioma preenchidos;
- Data de entrega inicial do serviço e contagem de palavras;
- Contagem total de palavras e desconto de alavancagem a serem cobertos pela liberação futura do pedido de compra.

Alternativamente, uma empresa pode considerar a parceria com uma organização sem fins lucrativos (NPO, na sigla em inglês). Da mesma forma que os benefícios ilustrados na subseção 3.1 sobre iniciar o relacionamento com os povos Indígenas por meio de alguém de sua confiança, trabalhar com uma organização sem fins lucrativos que já apoia a comunidade Indígena pode facilitar o fluxo e garantir um processo apropriado para conduzir negócios e interagir com essa comunidade.

Para a seleção de NPOs, aplicam-se os mesmos critérios usados para uma empresa prestadora de serviços linguísticos. Existem algumas vantagens de buscar uma NPO, como a eficiência de custos

(já que as NPOs não adicionam uma margem sobre o custo por palavra dos tradutores, diferentemente dos provedores de serviços linguísticos), bem como a capacidade de ter pagamentos feitos através dos fundos filantrópicos, supondo que a organização patrocinadora tenha tal divisão. Além disso, certas NPOs especializadas já conectadas com a comunidade Indígena relevante podem fornecer mais dados acadêmicos e/ou de pesquisa locais.

3.3 Parcerias linguísticas e acadêmicas

Conforme mencionado na subseção 3.1, a recomendação é que o engajamento com as comunidades aconteça por meio de estudiosos, organizações ou institutos que já tenham relações anteriores significativas e duradouras com os povos Indígenas. Elas liderarão a formação da equipe para trabalhar na iniciativa, com o objetivo de garantir que a execução seja significativa para os povos Indígenas e sua língua.

Recomenda-se que as equipes linguísticas e de engenharia, juntamente com os provedores de serviços linguísticos selecionados, envolvam e treinem acadêmicos, falantes nativos e/ou tradutores profissionais, se aplicável, em ferramentas de Tradução Assistida por Computador (CAT, na sigla em inglês), processos, particularidades da linguagem de software e tudo o que diz respeito à digitalização de um idioma, incluindo, mas não se limitando, à escolha do dialeto a ser usado e metodologias para garantir a consistência da terminologia em todo o banco de dados do produto.

Além de especialistas em linguística, os acadêmicos têm o papel de um consultor que garante que o processo de digitalização comece e termine de forma respeitosa com a comunidade Indígena e se concentre exclusivamente na revitalização do idioma. Os linguistas também visam promover a discussão técnica entre a comunidade e os tradutores integrados em relação às especificações linguísticas, como tom de voz a ser aplicado na tecnologia, diferenças de dialeto e escolha do estilo de escrita. Às vezes, essa não é uma tarefa fácil, dado que é nessa fase que as terminologias relacionadas à tecnologia surgem pela primeira vez e, embora o conceito possa não ser novo, talvez ainda não haja uma palavra ou expressão definida para o transmitir em um determinado idioma.

Se o idioma escolhido for falado em comunidades com boa penetração de tecnologia, incluindo, mas não se limitando, ao acesso à internet e a computadores, os primeiros passos seriam:

a. ter uma compreensão clara da disponibilidade e disposição das pessoas para dedicar horas de suas vidas à iniciativa,
b. concordar com marcos em termos de palavras a serem traduzidas em um determinado período de tempo e
c. concordar com a compensação de monetização a ser recebida por palavra traduzida.

Recomenda-se que isso seja feito por meio da intermediação do provedor de serviços linguísticos selecionado ou da NPO responsável pela parceria entre a empresa patrocinadora e acadêmicos e linguistas.

Caso o idioma escolhido seja falado em comunidades que não possuem acesso à internet nas velocidades adequadas, recomenda-se que seja considerada a doação de laptops/computadores e o fornecimento de suporte para conexão à internet. Este será mais um passo para fechar a lacuna na Inclusão Digital, conhecida como a exclusão digital, comumente vista em muitas comunidades Indígenas, o que as impede de ter suas rotinas aprimoradas social e economicamente pelo uso da tecnologia, através do acesso à informação.

O processo de exposição de falantes e tradutores a assuntos específicos de localização segue a fase inicial relacionada à logística. Isso inclui treinamento em ferramentas e workshops sobre processos de localização e melhores práticas, tais como.

- O uso de linguagem clara, simples e consistente;
- Responsabilização por diferenças linguísticas, culturais e técnicas (como expressões idiomáticas, significado de cores específicas, unidades de medida e moedas);

- Manutenção da intenção e funcionalidade originais do aplicativo original; e
- Garantia de que o produto localizado parecerá e funcionará como se fosse nativo daquele mercado e povos específicos.

Esta fase é normalmente coordenada pelas equipes de linguística e engenharia da empresa idealizando e patrocinando a iniciativa, enquanto as de logística e gestão de pessoas acontecem por meio de gestão feita pela NPO ou prestadora de serviços linguísticos.

Para o processo de digitalização descrito neste relatório, a empresa trabalhou em estreita colaboração com universidades (Unicamp no Brasil, University of North Carolina – Chapel Hill nos Estados Unidos, e a Universidade Livre de Bozen-Bolzano na Itália), através de seus linguistas, e com o KISS (Kalinga Institute of Social Science) na Índia, para se envolver com cidadãos e linguistas.

4. Processo de localização e considerações linguísticas

4.1 Processo de Localização

A localização é o processo sistemático de adaptação de um produto ou conteúdo a um local ou mercado específico, incluindo tradução, imagens associadas e elementos culturais que influenciam a forma como o conteúdo será percebido[9]. Detalhes de tais adaptações podem ser encontrados na seção 6. Esse processo envolve modificar, reestruturar e adaptar o conteúdo para o público-alvo e requer uma estratégia sólida e um roteiro claro, respaldado por uma comunicação eficiente. Além da tradução, a localização bem-sucedida pode envolver, e não se limita a, planejamento, preparação de conteúdo, pós-edição e revisão, garantia de qualidade e revisão no contexto[10]. Este processo se aplica a línguas Indígenas e não Indígenas.

Considerações sobre TMS e CAT tools

Um Sistema de Gerenciamento de Tradução (TMS, na sigla em inglês), também conhecido como software de gerenciamento de tradução, é uma plataforma de software que facilita o ciclo de vida de um processo de localização. O software de gerenciamento de tradução elimina tarefas manuais repetitivas e trabalhosas por meio de aplicativos de máquina automatizados integrados, ao mesmo tempo em que permite o controle do fluxo de trabalho, aumentando a colaboração e a eficiência. Um TMS é uma ferramenta de eficiência e não se limita puramente a traduções[11]. Os sistemas de alta qualidade também automatizam fluxos de trabalho para melhorar o gerenciamento de projetos, além de oferecer outros serviços, como integração com sistemas de gerenciamento de conteúdo, rastreamento financeiro, análise, alocação de recursos e neutralidade de fornecedores (permitindo que as empresas se envolvam com vários fornecedores). Além disso, alguns outros recursos a serem considerados ao selecionar um TMS são:

- Amplo suporte para diferentes formatos e tipos de arquivo;
- Tradução no contexto (a possibilidade de associar imagens de referência visual para serem visíveis, em um segmento específico, no editor do TMS);
- Capacidade de ter vários tradutores trabalhando no mesmo projeto simultaneamente (permitindo que segmentos de um projeto sejam atribuídos a recursos específicos).

Vários fatores precisam ser considerados ao escolher o TMS certo. Os sistemas de gerenciamento de tradução são projetados para suportar tarefas complexas, a fim de tornar os processos de tradução e localização gerenciáveis e eficientes. Muitas ferramentas de TMS oferecem Tradução Assistida por Computador (CAT, na sigla em inglês) e tradução automática integrada (MT, na sigla em inglês). Esses aplicativos tudo-em-um permitem que os usuários gerenciem e planejem projetos por meio de uma única plataforma[11]. Os sistemas de gerenciamento de tradução ajudam a gerenciar os fluxos de trabalho de tradução, mas não realizam as traduções propriamente ditas. As CAT tools podem trabalhar dentro dos sistemas de gerenciamento de tradução para suportar um determinado fluxo de trabalho de localização. Os fluxos de trabalho são personalizáveis, permitindo que várias etapas sejam adicionadas e atribuídas a diferentes linguistas, conforme necessário, com base no tipo de conteúdo e nos requisitos do projeto.

As CAT tools são uma parte importante do passo a passo do processo, melhorando a eficiência e a produtividade dos tradutores e revisores em um processo de localização e são comumente conhecidas e usadas pelos Provedores de Serviços Linguísticos (LSPs, na sigla em inglês). Elas também são frequentemente usadas por tradutores individuais e funcionários bilíngues que trabalham para organizações com audiências globais ou extensas necessidades de localização.

As CAT tools automatizam tarefas que um linguista teria que realizar manualmente, incluindo o gerenciamento de conteúdo traduzido. Ela divide o conteúdo em um idioma de origem específico em unidades de tradução, muitas vezes referidas como segmentos (geralmente frases ou parágrafos) para localização nos idiomas de destino. Para fins de esclarecimento, este documento se referirá ao "idioma de origem" ao iniciar pelo idioma padrão usado para o desenvolvimento de

software – geralmente para essa finalidade, o idioma é o inglês; enquanto o "idioma de destino" será o termo usado para se referir ao idioma no qual a tradução é necessária.

Essas ferramentas geralmente consolidam a integração de MT, gerenciamento de terminologia, aproveitamento de Memória de Tradução (TM, na sigla em inglês), alinhamento de linhas, análise de escopo, recursos de pesquisa, garantia de qualidade, verificadores ortográficos e gramaticais e outras funções. Existem muitas CAT tools sofisticadas no mercado projetadas para diversos tipos de tarefas. Uma CAT tool típica consiste em pelo menos três componentes principais: uma TM, onde as traduções anteriores são armazenadas; um banco de dados de terminologia contendo uma lista de terminologias aprovadas que devem ser referenciadas durante a localização; e um editor, uma página através da qual a tradução ocorre.

Memória de Tradução

Uma Memória de Tradução é um banco de dados linguístico que permite a reutilização de traduções para trabalhos futuros, uma das principais funcionalidades em uma CAT tool. Ela armazena conteúdo do idioma de origem e as traduções correspondentes em segmentos, também conhecidos como Unidades de Tradução (TU, na sigla em inglês), que podem ser tão longos quanto um parágrafo ou tão curtos quanto uma palavra. O programa procura correspondências com o texto no idioma de origem entre projetos de tradução em andamento e antigos, fornecendo sugestões de resultados de acordo. Uma correspondência pode ser exata ou difusa, determinada de acordo com a porcentagem de similaridade de conteúdo – 100%, 99%, 95%, 80% e assim por diante.

Banco de Dados de Terminologia

Um banco de dados de terminologia, ou termbase, é um módulo integrado dentro de uma CAT tool que serve como uma compilação de terminologia e também pode conter informações associadas relacionadas às palavras ou frases definidas. As informações disponíveis para entradas em um termbase podem incluir um termo e sua versão no idioma de origem e de destino, sua definição, exemplos de uso e metadados. As entradas de termbase são úteis para elucidar palavras ou frases que são potencialmente ambíguas em significado, informações técnicas e específicas do produto (marketing/marca) ou expressões que podem ser traduzidas em uma expressão proibida ou tabu. Termbases e glossários pré-organizados podem ser importados para as CAT tools, mas também podem ser organizados e editados por um usuário com permissões definidas.

A utilização de um termbase pré-organizado pode permitir que um projeto de tradução mantenha padrões consistentes, bem como uma mensagem consistente, uma vez que a terminologia da língua-alvo sugerida ou obrigatória para palavras ou frases especificadas é claramente definida para o usuário final. Os projetos que contenham vários tradutores designados terão um grau uniforme de qualidade se a adesão aos dados contidos no termbase for seguida[13].

Editor de CAT Tool

O editor de CAT tool é o meio pelo qual a tradução e a revisão ocorrem. Editores robustos têm uma visão lado a lado do conteúdo nos idiomas de origem e de destino com recursos de pesquisa de TM e resultados do banco de dados de terminologia em exibição para tornar a tradução e a revisão as mais eficientes e suaves possíveis. Alguns editores também têm uma janela pop-up para visualização do documento traduzido, o que pode ser útil para determinar o layout desejado do texto traduzido. Textos traduzíveis no idioma de origem são exibidos em segmentos, e cada segmento pode conter frases ou parágrafos de acordo com a configuração definida. Depois que uma tradução é concluída para um segmento, ela pode ser finalizada ou bloqueada para evitar a edição não intencional e acompanhar o progresso da tradução. Além disso, se existir uma correspondência de TM para um segmento, a CAT tool exibirá esse resultado denotando o tipo de correspondência (por exemplo, 100%, Correspondência Difusa, Sem Correspondência, Repetição, Correspondência no Contexto), o que geralmente reduz os esforços e custos de tradução.

Ao trabalhar com uma língua Indígena ameaçada, determinar qual ferramenta usar é uma decisão importante e deve considerar as necessidades da comunidade. Se a ferramenta estiver disponível apenas on-line, ela pode não estar disponível em áreas com conectividade limitada à Internet e pode impedir que comunidades carentes aproveitem seus recursos durante o processo de localização. Além disso, não é comum ter idiomas Indígenas disponíveis por padrão nos sistemas de Memória de Tradução, e esses idiomas podem precisar ser adicionados como um idioma suportado, um processo que leva tempo. Nesses casos, outro idioma disponível pode ser usado em vez disso, como um espaço reservado, até que tal solicitação seja concluída.

Existem vários fatores a serem considerados ao decidir qual solução de TMS e CAT se adapta às necessidades e ao orçamento do projeto envolvido. Não existe uma solução única para todos, e empresas de todos os tamanhos as usam de forma diferente. No entanto, ao trabalhar com tradução ou localização de conteúdo, é necessário usar um TMS com uma CAT tool integrada. Um TMS bem-sucedido fornece resultados precisos e relevantes rapidamente, ao mesmo tempo em que mantém cronogramas precisos e eficiência no gerenciamento de projetos[12].

4.2 Considerações linguísticas ao criar ortografias

É imperativo considerar tanto as estruturas linguísticas quanto os aspectos sociolinguísticos de uma língua ao criar um sistema ortográfico. Em sua essência, uma ortografia representa uma combinação de considerações linguísticas e as necessidades de uma comunidade falante. Fundamentalmente, as ortografias devem ser acordadas por falantes nativos e podem ser reformadas, se necessário.

Uma ortografia eficaz é morfofonêmica, de modo que geralmente deve capturar contrastes fonêmicos, ao mesmo tempo em que mantém uma ortografia consistente para sons que têm pronúncias diferentes em contextos específicos (por exemplo, o plural *-s* em inglês é pronunciado como [z] em certos contextos, mas é consistentemente transcrito como *-s*). Além disso, a fonologia, emparelhada com as intuições do falante nativo, deve ser considerada ao transcrever quebras de palavras dentro de palavras e frases compostas. Outros fatos linguísticos a serem considerados são como as características – como comprimento da consoante/vogal, nasalização e tom – devem ser representadas (ou seja, como uma única letra, diacrítico, etc.).

A ortografia deve ser tão amplamente aceita quanto possível pela comunidade Indígena para a qual é feita. Para que a ortografia seja usada pelos falantes, ela precisa ser intuitiva o suficiente para aprender e ensinar. Por exemplo, as ortografias de muitos idiomas Indígenas mesoamericanos têm convenções semelhantes às da ortografia espanhola, pois essas comunidades já estão familiarizadas com essa ortografia e podem transferir seu conhecimento do espanhol escrito para a forma escrita do idioma Indígena. Uma nova ortografia para uma determinada língua Indígena deve ser o mais semelhante possível às línguas Indígenas relacionadas, se as línguas não tiverem diferenças fonológicas substanciais. Com relação às diferenças de dialetos, a comunidade falante deve decidir se uma ortografia padrão se adequaria melhor às necessidades de uma comunidade e, em caso afirmativo, como abordar a padronização de quaisquer diferenças. Em geral, as discussões sobre a criação ou adaptação de uma ortografia devem envolver membros de cada comunidade e podem envolver a colaboração com linguistas e instituições locais ou regionais, como comissões de idiomas e universidades. Por exemplo, na América Latina, instituições como o Museu do Índio e o Instituto Emílio Goeldi podem ser consultadas para colaborar em uma ortografia com comunidades Indígenas da região. Outro exemplo de colaboração entre comunidades e instituições Indígenas envolve os campi da Universidade da Califórnia com comunidades da América do Norte e Central. A Universidade da Califórnia, Santa Cruz, tem trabalhado com uma variedade do zapoteca da Sierra Norte para adaptar o alfabeto zapoteca existente e capturar as diferenças fonológicas na multiplicidade que são representadas por novos caracteres. Em última análise, se um determinado dialeto é escolhido sobre outro dialeto, ou se uma forma padrão é criada, depende da iniciativa particular da comunidade.

Um sistema ortográfico de uma língua Indígena deve ser aprovado pelos falantes que o utilizarão. Se uma ortografia estabelecida for criada, os falantes têm autoridade para fazer as alterações que

julgarem necessárias. Assim, o diálogo entre linguistas e comunidades Indígenas está sempre na vanguarda desse esforço colaborativo.

4.3 Integração de software, alavancagem, origem e destino e mais

De maneira semelhante ao desenvolvimento de software, que pode ser conduzido de várias maneiras diferentes, a localização de software também pode seguir uma metodologia específica. Tradicionalmente, ela segue um processo em cascata pelo qual a localização é feita de uma só vez, geralmente após a versão padrão do idioma do software ter terminado o desenvolvimento ou já ter sido lançada. Esta pode ser uma boa opção se for necessário mais controle sobre o processo e se o orçamento for limitado. No entanto, essa abordagem tem algumas desvantagens significativas, a mais importante é que pode levar muito tempo para colocar o software ou produto localizado no mercado. Como todo o conteúdo precisa ser traduzido de uma só vez, pode ser muito demorado. A outra principal desvantagem é que pode ser mais difícil detectar erros e equívocos, pois há menos oportunidades de feedback e iterações. Uma vez que o conteúdo foi traduzido, pode ser muito difícil e caro fazer alterações[14].

Em contraste, a localização contínua é uma abordagem na qual as traduções são feitas constantemente à medida que novos conteúdos estão disponíveis. A localização contínua garante que o conteúdo localizado esteja sempre pronto para lançamento simultâneo. Essa abordagem também tem a vantagem de ser mais flexível e iterativa. No entanto, a localização contínua pode ser mais difícil de gerenciar e coordenar, pois requer uma equipe de localização experiente em tecnologia, que se sente confortável o suficiente para trabalhar de acordo com um ambiente ágil. Além disso, ter uma pilha de ferramentas tecnológicas à altura da tarefa também é importante. No fim das contas, a localização contínua provou ser uma maneira mais eficiente e econômica de localizar um produto de software[13].

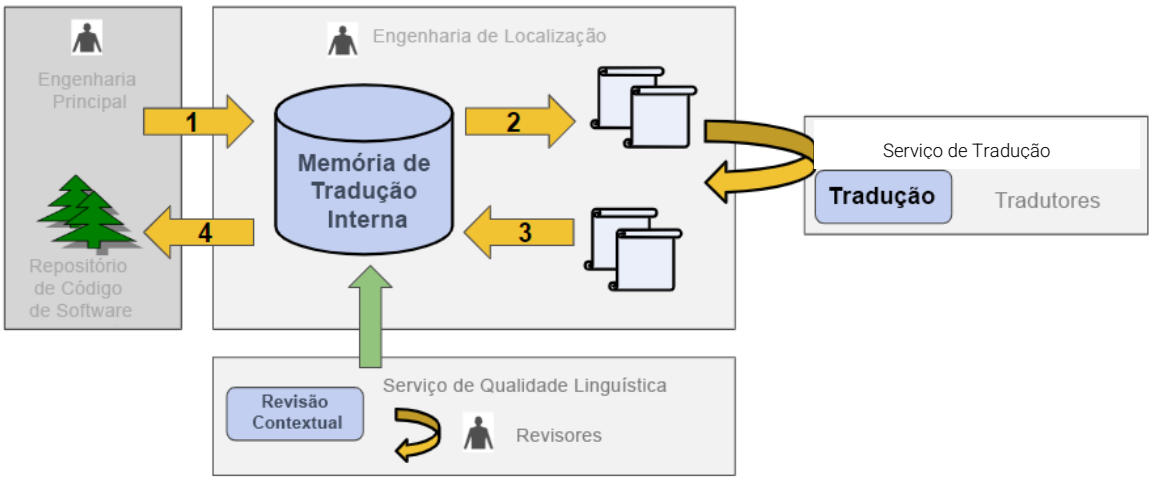


Figura 4.3-1. Fluxo de trabalho de localização contínua

A Figura 4.3-1 mostra um exemplo de como um fluxo de trabalho contínuo para localização de software pode ser configurado com a ajuda de um conjunto de ferramentas/sistema para gerenciamento de recursos de software. Uma vez que a equipe de engenharia ou desenvolvimento principal envia o código com novas linhas para o repositório de software, a plataforma/ferramenta de localização, como um sistema de gerenciamento de recursos, poderia:

1) Detectar automaticamente linhas novas ou modificadas e pesquisar em uma memória de tradução interna por traduções existentes para aproveitar;

2) Enviar o novo conteúdo que precisa ser localizado para o TMS.

Quando a tradução for concluída, a plataforma de localização:

3) Extrai automaticamente o conteúdo traduzido do sistema de gerenciamento de tradução e adiciona ele à memória interna de tradução;
4) Mescla o conteúdo traduzido de volta ao repositório de códigos do software.

A etapa de tradução entre 2 e 3 normalmente ocorreria em um TMS (subseção 4.1.1). As etapas 2 e 3 podem ser automatizadas se o TMS fornecer a Interface de Programação de Aplicativos (API) necessária; caso contrário, ela também pode ser executada manualmente por um Gerente de Projeto de Localização/Idioma.

A implementação das etapas 1 e 4 depende da plataforma de software ou do sistema de gerenciamento de conteúdo usado para a criação do conteúdo. Para a iniciativa descrita neste artigo, o conteúdo localizado pela Motorola era para o sistema operacional Android™, e as linhas estavam presentes em centenas de diferentes repositórios Git. O Git é um sistema de controle de versão distribuído, muito popular, usado no desenvolvimento de software. A plataforma da ferramenta de localização rastrearia todos os repositórios de código, detectaria linhas presentes dentro das pastas de recursos do Android em vários arquivos XML e verificaria se a tradução já existe na memória de tradução interna. Todas as linhas sem uma respectiva tradução existente são extraídas para tradução.

O fluxo de trabalho acima geralmente funcionaria em um par de idiomas de origem e destino. As cadeias de caracteres do idioma de origem são enviadas para serem traduzidas para o idioma de destino usando um arquivo.xliff. O XML Localization Interchange File Format (XLIFF) é um formato bitext baseado em XML comumente usado para trocar dados entre ferramentas durante um processo de localização e é um formato comum usado em CAT tools.

Ao trabalhar com um idioma Indígena, é possível que um idioma de origem não seja útil para a tradução direta para um idioma Indígena de destino. Como as comunidades Indígenas podem viver concentradas em uma única região e só estão familiarizadas com a língua dominante naquela região específica, mais etapas de tradução podem ser necessárias. Por exemplo, o processo exigiria primeiro a tradução do idioma de origem para o idioma dominante daquela região; e depois a tradução do idioma dominante para o idioma Indígena. Ao trabalhar com línguas Indígenas na região amazônica, a Motorola teve que traduzir primeiro do inglês (Estados Unidos) para o português brasileiro, e só então do português brasileiro para as línguas Indígenas impactadas.

Esse processo de múltiplas etapas pode ser alcançado usando o mesmo fluxo de trabalho de localização mostrado acima (Figura 4.3-1) duas vezes. Primeiro, usando a língua de desenvolvimento como origem, e a língua dominante regional como destino. Uma vez concluída essa etapa, o desenvolvimento regional é usado como origem, e a língua Indígena como destino. Alguns Sistemas de Gerenciamento de Tradução têm a capacidade de adicionar traduções de várias etapas dentro do fluxo de trabalho definido e podem ser uma opção eficaz para trabalhar sem fazer mais alterações no fluxo de trabalho de localização.

Embora a primeira abordagem possa funcionar com qualquer TMS, ela pode causar alguns atrasos devido a idas e vindas adicionais de arquivos. Além disso, à medida que o número de etapas intermediárias de tradução aumenta (por exemplo, chinês > inglês > odia > kuvi), essa abordagem pode se tornar muito cara e demorada. A segunda abordagem pode escalar bem, mas precisa ter um TMS em vigor no qual essa opção possa ser ativada.

4.4 Considerações sobre garantia de qualidade linguística e sensibilidade às culturas

A Garantia de Qualidade Linguística (LQA, na sigla em inglês) é o processo durante o qual os linguistas usam uma metodologia específica para revisar as traduções e determinar se elas incluem erros. Tais erros podem envolver, mas não estão limitados ao seguinte: Irregularidades nos glossários, ortografia ou gramática inconsistente e inadequações contextuais. Além disso, é importante colaborar com falantes nativos de um idioma para garantir que as traduções representem com precisão o idioma e as necessidades da comunidade.

Ao trabalhar com uma comunidade Indígena, os pesquisadores devem se familiarizar com a cultura dos falantes da língua, pois os fatores culturais podem se correlacionar diretamente com aspectos da língua (por exemplo, termos de parentesco)[15]. Os pesquisadores devem estabelecer uma comunicação consistente com uma comunidade para construir relacionamentos pessoais e confiança mútua com os membros da comunidade. Por exemplo, como mencionado acima, os pesquisadores devem chegar a acordos com uma comunidade em relação ao pagamento por sua experiência linguística. Enquanto algumas comunidades estão abertas a receber uma compensação monetária adequada, outras comunidades preferem ser compensadas com suprimentos práticos (por exemplo, lanternas, guarda-chuvas). Cameron et al. (1992) descrevem três tipos de métodos de pesquisa linguística[16] (1-3), e Czaykowska-Higgins (2009) fornece uma expansão[17] para este modelo (4), definido abaixo:

1. **Pesquisa ética:** modelo em que a pesquisa é *sobre* sujeitos; ou seja, há considerações éticas norteando a metodologia, mas o modelo não vai além da coleta de dados.
2. **Pesquisa de advocacia:** modelo em que a pesquisa é *sobre* e *para* os sujeitos; o trabalho visa diretamente beneficiar a comunidade.
3. **Pesquisa empoderadora:** modelo em que a pesquisa é *sobre*, *para* e *com* sujeitos; o trabalho beneficia a comunidade e envolve explicitamente a comunidade na tomada de decisões e na condução do projeto de pesquisa.
4. **Pesquisa linguística baseada na comunidade:** modelo em que a pesquisa é*"… em uma língua, e que é conduzida para, com e pela comunidade linguística dentro da qual a pesquisa ocorre e que afeta. Esse tipo de pesquisa envolve uma relação colaborativa, uma parceria, entre pesquisadores e (membros da) comunidade…"*[16]

Os esforços linguísticos devem empenhar-se para imitar um modelo de pesquisa linguística baseado na comunidade, que, em última análise, considera os falantes nativos como os especialistas de sua língua, e que a comunicação mútua e a colaboração são aspectos-chave de um projeto linguístico.

5. Níveis de suporte linguístico de internacionalização

A internacionalização fornece a capacidade de disponibilizar globalmente produtos e serviços, ao mesmo tempo em que atende aos requisitos regionais e legais de um país-alvo. No cenário dado, é o processo de adaptação do código de software em um comportamento no qual entradas e exibições de vários idiomas são ativadas independentemente do idioma da interface. Esse processo também inclui a conformidade com os padrões locais, como aderir a moedas, unidades de medida e formatos de data e hora. A seguir está uma lista dos cinco níveis de apoio necessários para cobrir os requisitos básicos para a Inclusão Digital de um idioma, e é aplicável a idiomas Indígenas e não Indígenas. Consulte o Apêndice para obter um fluxograma resumido que ilustra os níveis de suporte.

5.1 Sistema de escrita

Embora os sistemas de escrita sejam uma forma recente de linguagem em comparação com a língua falada, eles são primordiais na transmissão de informações nos tempos modernos[18]. Como os sistemas de escrita surgiram nos últimos milhares de anos, muitos sistemas modernos podem ter suas bases em apenas alguns escritos. Por exemplo, o alfabeto romano usado para a escrita em inglês é originário do alfabeto fenício do segundo milênio a.C.[18]. No entanto, é difícil determinar as origens de certos escritos, como a escrita chinesa ou a escrita maia, que parecem ter se desenvolvido independentemente de outros scripts[18].

Pesquisadores reconheceram recentemente o estudo dos sistemas de escrita como seu próprio ramo sob a linguística. No estudo dos sistemas de escrita, é necessário classificá-los por tipo, semelhante à forma como as línguas faladas são classificadas nas famílias[18]. Um tipo particular de sistema de escrita é um sistema semasiográfico, que utiliza símbolos e imagens. Por exemplo, sinais de trânsito e sinais para cuidados com roupas são sistemas que devem ser aprendidos e representam ideias sem utilizar explicitamente segmentos da linguagem. Esses sistemas, que interferem na leitura ou na escrita, foram atestados nas culturas Indígenas e também têm aplicações para pessoas com deficiências.

Outro tipo de sistema é um sistema glotográfico, que expressa ideias na forma de segmentos (por exemplo, palavras, fonemas) da língua falada. Dentro dos sistemas glotográficos, existem diferentes níveis de representação que variam do logográfico ao fonográfico[18]. Um escrito logográfico geralmente usa um símbolo, que não faz referência à pronúncia, para representar uma palavra ou conceito. Um exemplo de sistema logográfico é o alfabeto chinês. Em contraste, os sistemas fonográficos correspondem explicitamente às unidades fonéticas de uma língua. As unidades que podem ser representadas incluem unidades silábicas maiores; o hiragana e o katakana japoneses, por exemplo, têm uma escrita na qual os silabários que compartilham o mesmo símbolo para uma consoante ou vogal não são necessariamente semelhantes. Outra unidade fonética que pode ser representada em um sistema escrito são as características, como no alfabeto coreano chamado hangul, no qual o local de articulação do som particular é representado por uma forma gráfica. Finalmente, os sistemas fonográficos podem ser alfabéticos (ou seja, fonêmicas) e destinados a representar sons distintos (ou seja, fonemas) em uma língua, como o inglês.

Além do tipo de sistema de escrita utilizado, outro aspecto utilizado para classificar as línguas escritas é pela sua completude[18]. Por exemplo, alguns sistemas representam o comprimento da vogal, como em finlandês: *kaatua* "cair" e *katua* "se arrepender"[18]. Os escritos também podem representar tom; por exemplo, em línguas tonais, a diferença tonal entre duas palavras com a mesma ortografia pode ser representada ilustrada por marcas de acento tonal ou por um sistema de números tonais.

Ao criar um sistema de escrita para uma língua não escrita, a escolha do sistema depende de muitos fatores, incluindo aspectos culturais, históricos e tecnológicos. Os sistemas de escrita ajudam a transmitir e documentar a língua, por isso é importante colaborar com a comunidade falante do idioma para elaborar um sistema que permita que os falantes se comuniquem uns com os outros de maneira eficaz e padronizada.

Uma língua pode ser digitalizada de várias maneiras, incluindo formatos de áudio e vídeo. Recentemente, os esforços de revitalização linguística utilizaram tecnologias modernas, como mídias sociais e videogames, para promover o aprendizado de idiomas. No entanto, esses esforços devem ser complementados por um sistema de escrita completo que pode ser (i) aprendido por falantes de uma língua e (ii) ensinado por falantes da língua para transmitir às gerações futuras. Assim, os sistemas de escrita devem ser acordados por uma comunidade falante, bem como passíveis de mudanças futuras, conforme a comunidade julgar apropriado.

Um exemplo do processo de criação de um sistema de escrita pode ser visto no desenvolvimento do sistema de escrita kuvi, que foi um esforço de parceria entre o KISS (Kalinga Institute of Social Science) e a Fundação Lenovo. O kuvi é uma língua considerada potencialmente ameaçada pela UNESCO, falada nos estados indianos de Odisha e Andhra Pradesh. Além de envolver os falantes dessa língua na criação dos conjuntos de caracteres kuvi, o processo também envolveu especialistas em línguas kuvi, linguistas e especialistas em TI. Primeiramente, as seguintes questões foram consideradas pelos participantes deste projeto:

- Um novo escrito será feito para os conjuntos de caracteres kuvi, ou um escrito existente será adotado para eles?
- Se o escrito for adotado, qual escrito deve ser escolhido? A comunidade aceitará o escrito adotado e ele representa adequadamente a língua kuvi?
- O escrito representa foneticamente a língua kuvi? Em caso afirmativo, como os fonemas são representados com os grafemas do escrito?
- Como os membros da comunidade estão envolvidos com o processo de tomada de decisão do desenvolvimento do escrito?

Durante a fase de desenvolvimento deste projeto, os membros da comunidade colaboraram com especialistas em idiomas e linguistas para descrever as características fonológicas da língua kuvi. Os membros da comunidade foram solicitados a pronunciar os sons no contexto do vocabulário. A eles foram, apresentados os escritos para odia, telugu e hindi e, então, solicitados a escolher um escrito que melhor representasse a fonologia e a fonética da língua kuvi. Foi desenvolvida uma tabela paralela com todos os escritos. Palavras foram extraídas de membros da comunidade falante para garantir que os conjuntos de caracteres representassem com precisão o kuvi. Finalmente, a comunidade forneceu feedback durante as etapas de teste e as mudanças necessárias foram feitas para melhor atender às necessidades dos falantes de kuvi. Esse processo acabou destacando a importância do envolvimento da comunidade, a necessidade dos conjuntos de caracteres do kuvi serem harmoniosos com outras línguas faladas nas mesmas regiões onde se fala kuvi e como as considerações linguísticas devem informar o processo de criação de sistemas de escrita.

Para os caracteres de sistemas de escrita que ainda não são suportados pelo Unicode – um padrão da tecnologia da informação que representa texto – propostas listando as propriedades de tais caracteres devem ser submetidas ao Unicode. Após a apresentação da proposta, ela é revisada pelo Unicode Consortium e pelo comitê técnico do Unicode; esse processo é altamente envolvido e os patrocinadores de caracteres devem estar preparados para responder a perguntas do comitê, bem como organizar discussões sobre a proposta. A seção a seguir fornece uma discussão aprofundada dos aspectos técnicos do Unicode.

5.2 Unicode

O Unicode é um padrão de tecnologia da informação para codificar, representar e gerenciar texto em um determinado sistema de escrita de forma consistente. Ele é mantido pelo Unicode Consortium e contém uma matriz de caracteres, metodologia de codificação, codificações de caracteres padrão, propriedades de caracteres (como maiúsculas e minúsculas) e regras para normalização, decomposição, classificação e renderização. É amplamente utilizado na localização e internalização de software e é atualizado anualmente, com novos idiomas, caracteres e emojis[19].

O Unicode publica e mantém um gráfico de caracteres suportados[20]. A maioria ou todos os caracteres de uma língua Indígena recém-suportada já podem ter sido definidos no padrão Unicode. Um idioma Indígena recém-suportado pode exigir caracteres de vários escritos. Por exemplo, os dígitos 0-9 estão em uso generalizado; além disso, a devanagari *danda* é usada em muitos escritos índicos[21]. O exemplo abaixo ilustra um caractere U+203B (※) na seção "Pontuação Geral", em Unicode. Não é definido na escrita japonesa hiragana nem na escrita árabe urdu; mas é usado em ambas as línguas.

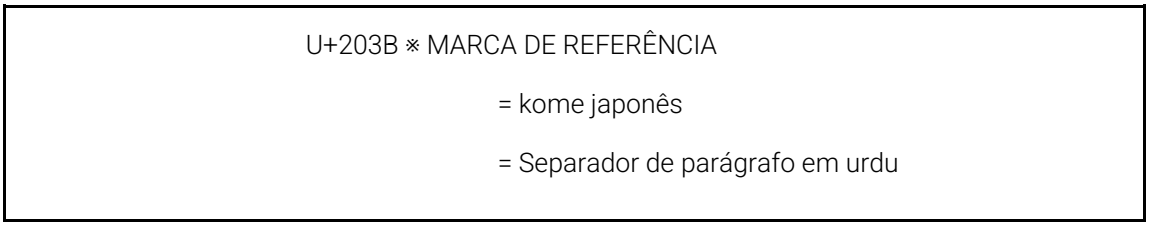


Figura 5.2-1: Caractere Unicode U+203B

No caso do Unicode ainda não suportar caracteres de um idioma que está sendo digitalizado, há dois caminhos a serem seguidos, cada um com seus prós e contras, e eles exigem atenção igual.

a. Áreas de Uso Privado (PUA, na sigla em inglês)[22] no Padrão Unicode

Há um total de 137.468 pontos de código possíveis e esses pontos de código são designados para uso privado. Os pontos de código PUA estão nas faixas U+E000..U+F8FF no Plano Multilíngue Básico (BMP), U+ F0000..U+ FFFFD e U+100000..U+10FFFD nos Planos 15 e 16. Qualquer ponto de código PUA pode ser usado para suportar os caracteres Indígenas; no entanto, talvez sistemas externos não suportem pontos de código Unicode não padrão.

Além disso, as propriedades dos caracteres, como padrões de quebra de linha e conversões de maiúsculas e minúsculas, podem ser definidas por meio de caracteres PUA; no entanto, alguns sistemas operacionais podem não ter suporte.

PRÓS	CONTRAS
● Fácil de definir pontos de código privados personalizados dentro do padrão Unicode ● Não causa corrupção de caracteres com outros scripts Unicode ● Trabalha dentro do sistema interno	● Não funciona com sistemas externos ● O uso de PUA deve ser implementado em cada componente do sistema ● Requer a adição de glifos de caracteres de uso privado na fonte do sistema (consulte a seção Fonte – PUA para obter mais detalhes) ● Requer a criação de um IME personalizado (consulte a seção IME para obter detalhes)

Tabela 5.2-1: Prós e contras do PUA Unicode

b. Enviar uma proposta de caractere[19] à Unicode

As propostas de inclusão de novos caracteres e escritos podem ser submetidas ao Unicode Consortium. Uma proposta deve primeiro considerar se um determinado escrito ou caracteres já foram propostos. Uma vez que a proposta tenha sido identificada como nova, ela deve incluir as seguintes informações relevantes para envio:

Propriedades de caracteres Unicode

[Informações Básicas]

- Ponto de código – *ponto de código Unicode (opcional)*
- Glifo – *representação gráfica dos caracteres*
- Nome – *nome dos caracteres*

As informações mais básicas necessárias sobre os caracteres incluem nome, ponto de código e outras informações de identificação. Uma listagem de exemplo de ponto de código, glifo e nome para a letra La em telugu é a seguinte:

```
0C32 ల TELUGU LETTER LA
```

Figura 5.2-2. Letra La em telugu

[Categoria Geral e outras propriedades]

- Categoria Geral – *classificação mais geral do ponto de código*
- Classe de combinação canônica – *algoritmo de ordenação canônica*
- Classe Bidirecional – *algoritmo bidirecional*
- Tipo de Decomposição/Mapeamento – *propriedade linha-valor de decomposição*
- Tipo/Valor Numérico – *propriedade do tipo numérico*
- Bidi Espelhado – *"Y" se texto espelhado em texto bidirecional*
- Nome Unicode 1 – *nome antigo publicado em Unicode 1.0*
- Comentário ISO – *campo de comentário ISO*
- Mapeamento Simples em Maiúsculas – *se houver equivalente em maiúsculas*
- Mapeamento simples em minúsculas – *se houver equivalente em minúsculas*
- Mapeamento simples de caixa de título – *se houver caixa de título*

As propriedades são documentadas no Banco de Dados de Caracteres Unicode e a seguinte linha é a entrada para a letra La em telugu:

```
0C32;TELUGU LETTER LA;Lo;0;L;;;;;N;;;;;
```

Figura 5.2-3. Propriedades da letra La em telugu

Se os caracteres propostos exibirem comportamento de modelagem (modelagem contextual, ligaduras, conjunções ou empilhamento), deve-se fornecer uma descrição desse comportamento com exemplos de glifos[18]. Informações sobre a ordem de classificação também devem ser fornecidas. Se os caracteres propostos forem símbolos, deve-se consultar "Critérios para Codificação de Símbolos" no Padrão Unicode.

Uma vez que todos os dados são definidos e coletados, então a proposta pode ser submetida ao Padrão Unicode. Uma série de reuniões com o comitê técnico do Unicode será então realizada e discutida para aceitação.

Em resumo, a submissão de novos caracteres e escritos engloba muitos desafios, tempo e esforço; mas a recompensa de ser incluído no Padrão Unicode é substancial.

PRÓS	CONTRAS
• Os caracteres usados no idioma Indígena serão incluídos no Padrão Unicode • Os caracteres serão suportados na maioria dos computadores e dispositivos digitais do mundo	• Um envio leva muito tempo para ser aceito e incluído no Padrão Unicode

Tabela 5.2-2: Envio para Prós e Contras do Padrão Unicode

5.3 ICU e CLDR

A International Components for Unicode (ICU)[23] é uma biblioteca para linguagens de programação modernas, como C, C++, Java, JavaScript e fornece suporte Unicode e Globalização para aplicativos de software. É portátil, modular e dá aos aplicativos os mesmos resultados em várias plataformas. É lançado sob uma licença de código aberto e não restritiva para uso comercial ou para ser adaptado a outro software de código aberto. O ICU fornece, mas não está limitado a:

- Conversão de página de código – entre Unicode e página de código;
- Colação – regra para comparação de caracteres;
- Formatação – números, datas, horários, moeda, gênero e plural;
- Calendário – nomes de dias da semana, meses com abreviação completa, média e curta;
- Fuso horário – nomes dos fusos horários;
- Propriedades de caracteres Unicode – nome, maiúsculas/minúsculas, bidirecionalidade, mais;
- Expressão regular – expressão regular correspondente à cadeia de caracteres Unicode;
- Manipulação de texto bidirecional – algoritmo para manipulação do texto bidirecional;
- Limites de texto – limite de caracteres, limite de palavras, limite de quebra de linha, limite de frase.

Caso o sistema use pontos de código Unicode – PUA (subseção 5.2-a), o conversor de pontos de código entre PUA e conjuntos de caracteres deve ser desenvolvido e integrado à biblioteca do ICU.

O Unicode Common Locale Data Repository (CLDR) é o maior e mais extenso repositório padrão de dados de localidade disponível, e seus dados fornecem às empresas os elementos constituintes para o software suportar uma ampla gama de idiomas. O CLDR usa o formato Unicode Locale Data Markup Language (LMDL) para armazenar e trocar os dados de localidade. O ICU compila e constrói os dados CLDR em bibliotecas, para que possa fornecer as convenções e padrões de localidade consistentes nos vários idiomas e plataformas de programação.

O CLDR inclui informações como sistemas de numeração, data, hora, fusos horários, unidades de medida, nomes de países, cidades, escritos, regras gramaticais e muito mais. O CLDR fornece

diferentes níveis de suporte básico necessário, o que é especialmente importante para idiomas ameaçados, pois as informações podem ser mais difíceis de adquirir.

```
<ldml>
 <identity>
  <version number="$Revision$"/>
  <language type="te"/>
 </identity>
 <localeDisplayNames>
  <localeDisplayPattern>
   <localePattern>{0} ({1})</localePattern>
   <localeSeparator>{0}, {1}</localeSeparator>
   <localeKeyTypePattern>{0}: {1}</localeKeyTypePattern>
  </localeDisplayPattern>
  <languages>
   <language type="elx">ఎలామైట్</language>
   <language type="en">ఇంగ్లీష్</language>
   <language type="en_AU">ఆ స్ట్రేలియన్ ఇంగ్లీష్</language>
   <language type="en_CA">కెనడియన్ ఇంగ్లీష్</language>
   <language type="en_GB">బ్రిటిష్ ఇంగ్లీష్</language>
   <language type="en_GB" alt="short">యు.కె. ఇంగ్లీష్</language>
   <language type="en_US">అమెరికన్ ఇంగ్లీష్</language>
   <language type="en_US" alt="short">యు.ఎస్. ఇంగ్లీష్</language>
   <language type="enm">మధ్యమ ఆంగ్లం</language>
   <language type="eo">ఎస్పెరాంటో</language>
   <language type="es">స్పానిష్</language>
   <language type="es_419">లాటిన్ అమెరికన్ స్పానిష్</language>
   <language type="es_ES">యూరోపియన్ స్పానిష్</language>
   <language type="es_MX">మెక్సికన్ స్పానిష్</language>
   <language type="et">ఎస్టోనియన్</language>
   <language type="tcy">తుళు</language>
   <language type="te">తెలుగు</language>
   <language type="tem">టిమ్నె</language>
   <language type="teo">టెసో</language>
   <language type="ter">టెరెనో</language>
  </languages>
 <localeDisplayNames>

 <dates>
  <calendars>
   <calendar type="gregorian">
    <months>
     <monthContext type="format">
      <monthWidth type="abbreviated">
       <month type="1">జన</month>
       <month type="2">ఫిబ్ర</month>
       <month type="3">మార్చి</month>
       <month type="4">ఏప్రి</month>
```

```
            <month type="5">మే</month>
            <month type="6">జూన్</month>
            <month type="7">జులై</month>
            <month type="8">ఆగ</month>
            <month type="9">సెప్టెం</month>
            <month type="10">అక్టో</month>
            <month type="11">నవం</month>
            <month type="12">డిసెం</month>
          </monthWidth>
          <monthWidth type="narrow">
            <month type="1">జ</month>
            <month type="2">ఫి</month>
            <month type="3">మా</month>
            <month type="4">ఏ</month>
            <month type="5">మే</month>
            <month type="6">జూ</month>
            <month type="7">జు</month>
            <month type="8">ఆ</month>
            <month type="9">సె</month>
            <month type="10">అ</month>
            <month type="11">న</month>
            <month type="12">డి</month>
          </monthWidth>
        </monthContext>
      </months>
    </calendar>
  </calendars>
 </dates>
</ldml>
```

Tabela 5.3-1: Um exemplo de CLDR para a língua telugu no formato LDML

Deve-se verificar se o ID da localidade para o idioma Indígena está definido no CLDR. A lista de idiomas suportados está disponível no site do CLDR. Por exemplo, a língua telugu é definida e classificada como um nível de cobertura moderna:

```
ID do local; Nível de cobertura; Nome
te ; moderno ; telugu
```

Figura 5.3-1. Cobertura do idioma telugu

A definição dos níveis de cobertura CLDR é a seguinte:

a) Dados principais

Este nível tem dados mínimos sobre o idioma e o sistema de escrita que são necessários antes que outras informações possam ser adicionadas usando a ferramenta de pesquisa CLDR.

1. Código do idioma (por exemplo, “te” para o idioma telugu)
2. Quatro conjuntos exemplares: principal, auxiliar, números e pontuação
3. Dados verificados do país (ou seja, população de falantes nas regiões ou países)
4. Escritos e região de conteúdo padrão
5. Ciclo de tempo usado com o idioma na região de conteúdo padrão

b) Dados básicos

Este nível inclui um pequeno conjunto de dados para suportar o idioma.

1. Dados de delimitadores – início/fim de citações, incluindo alternativas
2. Sistema de numeração – sistema de numeração padrão mais sistema de numeração nativa
3. Informações do padrão de localidade – padrão e separador de localidade, além do padrão de código
4. Nomes de idiomas – no idioma nativo e em inglês
5. Nome(s) da(s) escrita(s) – escritas habitualmente usadas no idioma
6. Nome(s) do(s) país(es) – países onde é comumente usado
7. Sistema de medição – métrico vs. imperial
8. Nomes de mês e dia da semana totalmente definidos
9. Nomes dos períodos AM/PM
10. Formatos de data e hora
11. Padrões de data e intervalo de tempo
12. Formatos de linha de base do fuso horário – região, GMT, GMT-zero, hora
13. Símbolos numéricos – separadores decimais e de agrupamento; sinal de mais, menos, porcentagem
14. Padrões numéricos – decimal, moeda, porcentagem, padrão científico

c) Dados moderados

Este nível inclui os seguintes atributos de localidade adicionais.

1. Regras plurais e ordinais
2. Informações sobre maiúsculas e minúsculas
3. Regras de colação

d) Dados modernos

Este nível é considerado como nível de suporte CLDR "completo" e inclui os seguintes dados adicionais.

1. Características gramaticais
2. Tabela de romanização (apenas escritos não latinos)

Caso um idioma recém-introduzido não seja suportado pelo CLDR, existem duas opções e ambas exigem o Nível de Cobertura de Dados Básicos, no mínimo.

1. Modificar as bibliotecas ICU e CLDR existentes no sistema

No Android, ICU e CLDR estão integrados no Android SDK (*Software Development Kit*). O Android SDK é um conjunto de ferramentas de desenvolvimento que são usadas para desenvolver aplicativos para a plataforma Android. Este SDK fornece uma seleção de ferramentas necessárias para criar aplicativos Android e garante que o processo seja o mais tranquilo

possível[25]. Para modificar o CLDR subjacente, as seguintes etapas precisam ser seguidas. É importante lembrar que as etapas são baseadas no Android 13 e podem mudar em versões futuras:

a. Identificar o ID da localidade para o idioma Indígena;
b. Criar arquivo xml LDML para o ID de localidade[26] e preencher os dados;
c. Reconstruir o ICU e o CLDR para o SDK do Android;
d. Reconstruir todo o sistema Android.

PRÓS	CONTRAS
• Os dados CLDR (por exemplo, traduções locais, formatos de data e hora, etc.) estão disponíveis imediatamente	• Necessidade de reconstruir o sistema Android

Tabela 5.3-2: Prós e contras da personalização do CLDR

2. Enviar dados via *Survey Tool*[27]

Os dados para CLDR (*Common Locale Data Repository*) são coletados e processados por meio da *Survey Tool*. A *Survey Tool* é uma ferramenta baseada na web para coletar dados CLDR e inclui vários atributos que precisam ser especificados antes do envio. A ferramenta fornece uma maneira de propor novos dados localizados, verificar o que os outros propuseram e se comunicar com eles para resolver as diferenças. Durante cada período de envio, os contribuintes dos membros da Unicode Consortium, outras organizações e o público em geral são convidados a revisar os dados para seus idiomas e países e propor novas traduções de termos ou modificações, incluindo traduções de idiomas inteiramente novas para o repositório[26].

Existem quatro etapas de coleta de dados na *Survey Tool*:

a. Shakedown
   Esta etapa é o início da ferramenta. É preciso certificar-se de que o nível de cobertura está definido corretamente e procurar, como um examinador, quaisquer problemas.

b. Envio geral
   Esta fase é onde os dados CLDR estão sendo inseridos na *Survey Tool*. A seção *Core Data* (dados principais) precisa ser preenchida. É altamente recomendável adicionar a seção *Basic Data* (dados básicos) também.

c. Verificação
   Para resolver todos os erros e revisar solicitações e discussões abertas.

d. Resolução
   A verificação está concluída, e qualquer trabalho adicional será conduzido pelo comitê CLDR.

O CLDR tem dois tipos de versões: completa e limitada. Normalmente, a versão completa está aberta para contribuições para todos os idiomas e áreas de dados. A limitada está aberta a contribuições para localidades selecionadas e campos específicos para todas as localidades. Para mais informações, consulte o site da *Survey Tool*.

PRÓS	CONTRAS
• Os dados usados no idioma recém-suportado serão incluídos no CLDR • Os dados serão suportados na maioria dos computadores e dispositivos digitais do mundo	• Um envio leva muito tempo para ser aceito e incluído no CLDR

Tabela 5.3-3: Envio de dados aos prós e contras do Comitê CLDR

5.4 Fontes

Uma fonte é uma coleção de glifos para representar os caracteres abstratos. Em outras palavras, as fontes são a base para o suporte a idiomas e precisam estar disponíveis no sistema operacional para que os caracteres e os sistemas de escrita possam ser exibidos para o usuário final. É um elemento fundamental para exibir corretamente informações de Unicode, CLDR e outros para o usuário.

Existem 2 tipos de formato de fonte: fontes raster e fontes vetoriais. Em fontes raster, um glifo é um bitmap que o sistema usa para desenhar os caracteres ou símbolo na fonte; enquanto que em fontes vetoriais, um glifo é uma coleção de pontos finais de linha que o sistema usa para desenhar os caracteres28. A vantagem de usar as fontes raster é que elas são muito rápidas de renderizar e podem ser otimizadas para processamento de imagem 2D. Nos computadores modernos, no entanto, praticamente todas as fontes são baseadas em vetores e podem ser usadas em qualquer tamanho sem perda de nitidez.

Tipo de letra e fontes são frequentemente usados de forma intercambiável. A diferença é relevante para designers tipográficos. Um tipo de letra é o design visual subjacente e uma fonte é aquela que implementa o tipo de letra. Por exemplo, *Times New Roman* é um tipo de letra, e *Times New Roman Italic* é uma fonte. Exemplos de fontes podem ser vistos abaixo:

Arial

Courier New

Times New Roman

Impact

Figura 5.4-1: Tipos de fontes

É importante notar que as fontes são materiais altamente protegidos por direitos autorais e não devem ser distribuídas livremente sem a permissão do designer da fonte.

Para determinar se a fonte do sistema tem os glifos de caracteres, a fonte (por exemplo, fonte OpenType) tem uma tabela incorporada usada para mapear códigos de caracteres Unicode para índices de glifos na fonte: A tabela “cmap”. Um exemplo de tabela “cmap” está disponível abaixo:

Latim Básico	U+0000-U+007F
Latin-1 Suplemento	U+0800-U+00FF
Cirílico	U+0400-U+04FF
Símbolo de Moeda	U+20A0-U+20CF
Área de Uso Privada	U+E000-U+F8FF

Figura 5.4-2: “cmap” no arquivo de fonte

Aliás, não há fontes que incluam todos os caracteres Unicode. É, portanto, importante determinar se a fonte do sistema suporta os caracteres da língua Indígena. Os caracteres da língua Indígena já podem ser definidos no padrão Unicode, mas não há garantia de que a fonte do sistema inclua esses glifos de caracteres. Se a fonte do sistema não contiver os glifos de caracteres Indígenas, uma fonte personalizada precisará ser criada por um designer de fontes e instalada e distribuída por todo o sistema. Recomenda-se entrar em contato com o fornecedor do sistema operacional (e fornecedor da fonte) para incluir os glifos ausentes no sistema, para que a fonte personalizada não precise ser instalada nas versões futuras do sistema operacional.

As fontes podem suportar o intervalo de pontos de código da Área de Uso Privado Unicode (PUA) (consulte 5.2-a). Um designer de fontes deve adicionar os glifos e definir o intervalo na tabela “cmap” na fonte. Quando o sistema precisa renderizar os caracteres Unicode-PUA para os usuários, o sistema deve ter a fonte personalizada.

5.5 IME

Um Editor de Método de Entrada (IME, na sigla em inglês) é tão importante quanto a presença de fontes, pois permite que os usuários insiram informações e dados, amplamente separados em três categorias: entrada de teclado, reconhecimento de caligrafia e reconhecimento de voz. Para que um novo idioma seja devidamente suportado no software, uma das categorias de métodos de entrada deve ser apresentada ao usuário. Para dispositivos móveis, a implementação menos complexa é a entrada do teclado.

O IME fica entre o teclado físico e o sistema operacional. Ele interpreta um toque de tecla físico em um ponto de código de caractere. Ao interpretar os toques de tecla, o sistema operacional pode suportar diferentes caracteres com o mesmo layout de teclado físico.

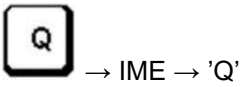


Figura 5.5-1: um toque de tecla via IME

Os provedores de sistemas operacionais geralmente fornecem muitos layouts de teclado para idiomas. No entanto, nem todos os idiomas são suportados. Desenvolver um teclado para a língua Indígena é possível, mas enfrenta muitos desafios. Exemplos de perguntas para entender esses

desafios são: A comunidade Indígena está familiarizada com computadores e dispositivos digitais? Onde os caracteres devem ser colocados no teclado? Quão utilizável é o novo layout do teclado?

O Android SDK inclui extensa documentação para criar um teclado IME na tela. Para adicionar um IME ao sistema Android, crie um aplicativo Android contendo uma classe que estenda o *InputMethodService*. Cada tecla deve então ser definida no arquivo XML. Um exemplo de arquivo XML está abaixo:

```
        <Row>
          <Key android:keyOutputText="\u0C4D"
            android:keyWidth="@dimen/key_telugu_width_normal_row1"
            android:keyLabel="\u0C4D" android:keyEdgeFlags="left"/>
          <Key android:keyOutputText="\u0C3E"
            android:keyWidth="@dimen/key_telugu_width_normal_row1"
            android:keyLabel="\u0C3E"/>
          <Key android:keyOutputText="\u0C3F"
            android:keyWidth="@dimen/key_telugu_width_normal_row1"
            android:keyLabel="\u0C3F"/>
          <Key android:keyOutputText="\u0C40"
            android:keyWidth="@dimen/key_telugu_width_normal_row1"
            android:keyLabel="\u0C40"/>
          <Key android:keyOutputText="\u0C41"
            android:keyWidth="@dimen/key_telugu_width_normal_row1"
            android:keyLabel="\u0C41"/>
          <Key android:keyOutputText="\u0C42"
            android:keyWidth="@dimen/key_telugu_width_normal_row1"
            android:keyLabel="\u0C42"/>
          <Key android:keyOutputText="\u0C46"
            android:keyWidth="@dimen/key_telugu_width_normal_row1"
            android:keyLabel="\u0C46"/>
          <Key android:keyOutputText="\u0C47"
            android:keyWidth="@dimen/key_telugu_width_normal_row1"
            android:keyLabel="\u0C47"/>
          <Key android:keyOutputText="\u0C4A"
            android:keyWidth="@dimen/key_telugu_width_normal_row1"
            android:keyLabel="\u0C4A"/>
          <Key android:codes="-1"
            android:keyWidth="@dimen/key_telugu_width_normal_row1"
            android:keyIcon="@drawable/ic_zero_width_outlined"
android:keyEdgeFlags="right"/>
        </Row>
```

Tabela 5.5-1: Uma definição XML de exemplo para o layout do teclado do Android

É importante notar que é possível definir o layout do teclado usando os pontos de código PUA Unicode. É simplesmente uma questão de especificar os pontos de código PUA no arquivo de definição XML.

6. Garantia de Qualidade

A Garantia de Qualidade é um processo importante responsável por garantir que todos os produtos e serviços estejam em conformidade com alguns requisitos específicos, que inclui o banco de dados CLDR/ICU que contém os padrões de localidade, como adesão a moedas, unidades de medida, formatos de data e hora e muito mais (consulte a seção 5 para obter mais informações sobre CLDR/ICU). Existem dois tipos principais de validações ao lidar com a Garantia da Qualidade de Internacionalização: Validações funcionais e Validações de domínio de internacionalização.

6.1 Validações funcionais

As validações funcionais são um conjunto de validações utilizadas para garantir que o produto funcione da mesma forma, independentemente do idioma selecionado. Durante esta fase de testes, as seguintes verificações são realizadas: A capacidade de alterar o idioma do dispositivo e como a Interface do Usuário reage e é atualizada após a troca; a inexistência de efeitos residuais após a modificação do idioma (ou seja, se algumas linhas permanecem no idioma definido anteriormente); o idioma é preservado corretamente após uma reinicialização; e o método de entrada aplicável ao idioma recém-selecionado carrega automaticamente no sistema operacional.

6.2 Validações de domínio de internacionalização

As validações de domínio de internacionalização são um conjunto de verificações para garantir que as convenções e requisitos regionais sejam seguidos e que os aplicativos sejam compatíveis com Unicode. Isso se refere, mas não se limita, a unidades de medida, alinhamento de texto, formatos de data e hora, endereços, números de telefone e ordenação alfabética. Esse tipo de validação verifica se todos esses itens são carregados automaticamente no sistema operacional, de acordo com as convenções aplicáveis.

Vale ressaltar, por exemplo, que diferentes editores de método de entrada (por exemplo, QWERTY e QWERTZ) se aplicam a diferentes idiomas. Devido a essas diferenças, o layout do teclado deve ser validado verificando seu layout em comparação com documentos de referência internos ou com base no teclado Gboard. Juntamente com a validação do layout, é necessário verificar sua capacidade de inserir e exibir caracteres não ASCII. ASCII significa *American Standard Code for Information Interchange* e é um conjunto que contém caracteres de controle, sinais de pontuação, dígitos e versões maiúsculas e minúsculas de letras do alfabeto inglês. Caracteres de sotaque como "ã", "á", "é" e a cedilha "ç" são exemplos de caracteres não ASCII.

Outro cenário para validação é se o texto inserido por meio de qualquer um dos métodos de entrada fornecidos no dispositivo é recebido em um dispositivo de suporte sem a omissão de quaisquer caracteres quando enviado por e-mail ou aplicativos de mensagens. Uma validação para verificar se os caracteres de uma mensagem recebida ou acessada via web são renderizados sem corrupção também faz parte do escopo, bem como uma verificação se a fonte instalada integrada à compilação do software (versão) suporta os glifos – caso contrário, os glifos ausentes serão mostrados como os caracteres de substituição "□" ou "".

Quando se trata de hora e data, a Garantia de Qualidade também é responsável por verificar se o formato das configurações e informações padrão do sistema em vários aplicativos está de acordo com a versão de lançamento oficial mais recente das convenções regionais definidas pela ICU/CLDR, conforme descrito na subseção 5.3. Por exemplo, "*Donnerstag, 27. April 2023*" é um exemplo de formato de data aceitável quando o idioma do sistema é alterado para alemão (Alemanha), enquanto *"jueves, 27 de abril de 2023*" é o aplicado para espanhol (México).

A falta de seguir a convenção local pode ter um impacto significativo na compreensão do usuário, considerando, por exemplo, que para o português (Brasil), a convenção de formato de data é DD/MM/AAAA, enquanto para o inglês (Estados Unidos) deve ser MM/DD/AAAA. A data 04/03/2023 é entendida como 3 de abril de 2023 pelos brasileiros, enquanto os americanos a leem como 4 de março de 2023. Localidade é a combinação de idioma mais país, que abrange a região e elementos culturais.

Outras convenções relevantes específicas do local são:

- A posição dos separadores de sinal % e decimal (por exemplo, "%5,5" para turco (Turquia) e "5,5 %" para francês (França);
- Unidades de medida (por exemplo, °F ou °C para temperaturas; milhas ou km para distâncias; e lb ou kg para peso);
- Formatos de endereço;
- Orientação da apresentação (ou seja, embora a maioria dos idiomas seja escrita da esquerda para a direita, alguns idiomas, como o árabe, são escritos da direita para a esquerda, e a interface do usuário deve espelhar as convenções do idioma com o alinhamento no lado direito).

7. Níveis de suporte linguístico de localização

Existem vários níveis diferentes de suporte de localização com os quais uma empresa pode decidir se comprometer e se envolver. No caso da Inclusão Digital, uma iniciativa que envolve principalmente software, é especialmente importante tornar essa decisão clara para todas as partes interessadas, dado o potencial de complexidades adicionais, como a escassez de especialistas linguísticos ou falantes nativos; nível de suporte à internacionalização baixo ou inexistente (que é uma dependência para localização) pelo sistema operacional alvo ou pelo Unicode, por exemplo. Ambas as áreas de consideração são descritas nas duas seções anteriores. Os níveis de suporte de localização podem incluir categorizações como Básico, Parcial ou Completo. Ao trabalhar na localização de software, independentemente do sistema operacional para o qual a iniciativa é direcionada, existem vários tipos de conteúdo a serem considerados, incluindo o próprio sistema operacional (elementos principais, como configurações do sistema, mensagens de erro e outros tipos de notificações do usuário), aplicativos pré-carregados, conteúdo do servidor, etc. Há também documentos de apoio que complementam as informações aos usuários que podem precisar ser localizadas para melhorar as informações de usabilidade do produto/projeto de software nos idiomas e regiões de destino (como guias de usuários impressos, conteúdo de ajuda online e documentos legais). Para os fins deste documento, o seguinte é presumido para o conteúdo referente ao software:

- Um nível de suporte **básico** refere-se a apenas alguns elementos de uma interface de usuário de smartphone sendo localizados em um idioma específico. Esses elementos podem ser específicos para tipos de formato para a região do idioma de destino (por exemplo, nomes de meses, nomes de dias da semana ou fusos horários no aplicativo de Configurações). Todo o sistema operacional do smartphone, bem como os aplicativos pré-carregados, permaneceriam em um idioma padrão (por exemplo, inglês ou outro idioma disponível) diferente do idioma Indígena ou não Indígena desejado destinado a ser localizado.

- Um nível de suporte de localização **parcial** implicaria uma porção maior da interface de usuário do smartphone traduzida para o idioma desejado. O suporte parcial à localização pode significar que o sistema operacional, incluindo aplicativos como Configurações, Discador e Calendário, por exemplo, está localizado, mas os aplicativos pré-carregados (como aplicativos proprietários da empresa ou outros aplicativos incluídos no pacote) não estão, ou o contrário. .

- Um nível de suporte de localização **completo** indica que o sistema operacional e os aplicativos desenvolvidos de uma empresa estão localizados, dando aos usuários uma exposição mais ampla ao seu idioma nativo. Aplicativos de terceiros não são considerados aqui na categorização do nível de suporte de localização, uma vez que as empresas geralmente não possuem ou têm controle direto ou acesso ao seu conteúdo de software (que pode ser baixado de forma independente, de uma loja de aplicativos). Isso significa que uma empresa pode fornecer um suporte de localização completo para um determinado idioma, enquanto aplicativos para download ou outros de terceiros podem não ser compatíveis e serão exibidos em idiomas alternativos.

Ao planejar a digitalização de um idioma Indígena ameaçado de extinção, recomenda-se que, antes de decidir e comunicar o nível de suporte de localização pretendido, as considerações observadas nas seções 2 e 5 sejam bem pesquisadas, dado seu impacto sobre o escopo, o orçamento, a complexidade e a viabilidade geral do processo de digitalização.

8. Feedback e continuidade

A Motorola e a Fundação Lenovo introduziram a iniciativa de Inclusão Digital de idiomas Indígenas ameaçados de extinção em 2021 e, à medida que avança pela próxima década, coincide com o período de 2022 a 2032 declarado pelas Nações Unidas como a Década Internacional das Línguas Indígenas (IDIL 2022-2032). A Motorola e a Fundação Lenovo esperam continuar a aumentar a conscientização sobre a causa, atuando para a sobrevivência de idiomas ameaçados e incentivando as próximas gerações de comunidades Indígenas a usar a tecnologia em seus idiomas nativos. Para isso, é importante olhar para o que funcionou bem e o que precisa melhorar, com base em feedbacks.

8.1 Feedback da UNESCO

A UNESCO estima que perdemos uma língua Indígena a cada duas semanas, resultando na perda de cerca de 3.000 línguas únicas até o final do século[2]. Com a iniciativa global de Inclusão Digital, lançada pela primeira vez na América Latina, que inclui idiomas Indígenas falados no Brasil, na Colômbia e na Venezuela, a parceria da UNESCO no Brasil com a Motorola, visa promover ainda mais, através da tecnologia, a integração de idiomas Indígenas ameaçados de extinção, torna-se essencial para a continuidade desse projeto. As ações da UNESCO Brasília ocorrem por meio de projetos de cooperação técnica em parceria com diversos níveis de governo e diferentes setores da sociedade civil sempre que seus propósitos contribuem para políticas públicas de desenvolvimento sustentável relacionadas a temas de *expertise* em que a UNESCO atua[29]. Portanto, colaborar com sua Unidade de Comunicação e Informação nos permitiu discutir outras ações de impacto que poderiam beneficiar os povos e línguas Indígenas. Além disso, essa colaboração indicou à Motorola que a iniciativa de Inclusão Digital estava no caminho certo para beneficiar comunidades Indígenas de forma mais ampla.

Para a UNESCO, "a língua é um meio primário para comunicar informação e conhecimento, portanto, a possibilidade de acessar conteúdo na Internet em uma língua que se pode usar é um determinante fundamental para a extensão em que se pode participar das sociedades do conhecimento"[30]. Assim, o Multilinguismo e a Acessibilidade são duas das seis prioridades do Programa Informação para Todos da UNESCO (IFAP, na sigla em inglês). Portanto, a iniciativa da Motorola de incluir línguas Indígenas nos smartphones permite a Inclusão Digital de grupos sub-representados e mantém sinergia com os objetivos da UNESCO em relação à inclusão social, gerando um sentimento de pertencimento e reconhecimento das culturas Indígenas no mundo digital.

Além disso, a estratégia de transparência para o código aberto é considerada pela UNESCO uma boa prática, com processos técnicos e dados linguísticos fornecidos ao público em geral para que outras empresas produtoras de smartphones possam disponibilizar suas funcionalidades para comunidades que falam idiomas ameaçados. O esforço em adaptar o teclado para atender às necessidades das línguas Indígenas ameaçadas de extinção é considerado pela organização uma boa prática quando se trata de Inclusão Digital.

Em suma, as iniciativas da Motorola e da Fundação Lenovo para apoiar a Inclusão Digital de línguas Indígenas ameaçadas no sistema de escrita por smartphone estão significativamente alinhadas com o Resultado 3 do Plano de Ação Global para a Década Internacional das Línguas Indígenas (IDIL 2022-2023): condições favoráveis estabelecidas para a capacitação digital, a liberdade de expressão, o desenvolvimento das mídias, o acesso a informação e às tecnologias da linguagem, junto com a criação artística nas línguas Indígenas.

8.2 Feedback dos parceiros

Um feedback importante sobre a iniciativa veio de Wilmar da Rocha D'Angelis, acadêmico da área de linguística da Unicamp, que liderou a primeira fase da iniciativa para as línguas Indígenas latino-americanas nheengatu e kaingang em 2020. Ele acredita que um impacto da iniciativa que é difícil

de medir é talvez o mais significativo: como a Motorola é uma empresa altamente associada a tecnologias de ponta e tem uma boa reputação e penetração no mercado brasileiro, sua iniciativa de apoiar as comunidades Indígenas, suas culturas e seus idiomas é um meio poderoso de aumentar o valor da diversidade cultural brasileira e uma forma de conscientizar a indústria sobre a necessidade de prestar mais atenção aos primeiros habitantes conhecidos de uma área.

Segundo Wilmar, "para os próprios povos Indígenas – especialmente para um grupo muito grande, em cada comunidade linguística kaingang ou nheengatu, de professores e intelectuais Indígenas que se dedicam incessantemente ao fortalecimento de sua língua ancestral – a iniciativa de Inclusão Digital apontou novos espaços de ação. O desenvolvimento e modernização de suas línguas e o despertar de novas vocações, como a de tradutor. Já existem pesquisas com professores e jovens Indígenas que nos procuram, questionando sobre a existência ou possibilidade de desenvolvimento de cursos de tradutores, especialmente sob a ótica da relação de suas línguas Indígenas com o português e o inglês. Alguns grupos de pesquisa em tradução em universidades brasileiras também começaram a se concentrar no tema da tradução de e para línguas Indígenas".

Ele também acrescentou que "o teclado específico despertou muito interesse em várias comunidades que falam outros idiomas Indígenas, e já fomos consultados várias vezes sobre isso; para alguns, bastava indicar que adotassem um smartphone equipado com Android 11 e selecionassem nheengatu (no caso dos guarani, por exemplo) ou kaingang (no caso dos apãniekrá, por exemplo) para poder aproveitar um teclado útil para a comunicação em seu idioma. Obviamente, isso não funciona para entender os comandos e instruções".

Wilmar aponta para o fato de que o trabalho colaborativo entre falantes de dialetos variados de três regiões diferentes da Amazônia para localizar o nheengatu em smartphones Motorola gerou a necessidade de maior unificação da língua[32]. Isso resultou na criação da Academia de Línguas Nheengatu, uma iniciativa pioneira entre as línguas Indígenas no Brasil. Isso veio como consequência da iniciativa de Inclusão Digital da Motorola e da Lenovo[33]. A Academia de Línguas Nheengatu tem como objetivo regularizar o nheengatu, padronizá-lo em três aspectos e assim recuperar parte do espaço perdido e em busca do lugar que lhe foi tirado, a língua materna que é a face da identidade da Amazônia.

De acordo com o feedback obtido na primeira fase da iniciativa de digitalização, o chefe principal Richard G. Sneed, da comunidade Eastern Band of Cherokee Indians (EBCI), que, junto com outros líderes cherokee, prestou consultoria à Motorola durante vários meses, disse que "ter interfaces de usuário de smartphones localizadas em cherokee é uma ferramenta que pode ser utilizada para ajudar a geração que está chegando a se familiarizar com elas, portanto, será imperativo que os povos cherokee incorporem isso ao currículo e à sua metodologia de ensino diária". Ele acredita que "a Motorola fez um esforço conjunto para trabalhar com grupos de povos Indígenas para incorporar as línguas tradicionais às tecnologias para que as línguas não desapareçam para sempre e que esta é apenas mais uma peça de um quebra-cabeça muito grande de tentar preservar e proliferar a língua"[34].

O Dr. Benjamin Frey, professor assistente da UNC-Chapel Hill, que foi o especialista em dialeto oriental da equipe do idioma cherokee para esse projeto, trouxe importantes esclarecimentos sobre o pensamento crítico em relação às diferenças culturais. Esse feedback ao longo da implantação da iniciativa, especialmente quando se trata de equilibrar a cultura corporativa (por exemplo, prazos de produtos e requisitos técnicos) e os povos Indígenas (seus modos de vida, necessidades e estilo de comunicação) foram fundamentais para o sucesso da entrega à comunidade. Além disso, a escassez de especialistas linguísticos disponíveis para participar e o grande escopo do projeto trouxeram obstáculos nas etapas de planejamento no final do projeto. Toda a equipe envolvida assimilou essas dificuldades e aprendeu com essa experiência para adaptar ainda mais os métodos, estilos e canais de comunicação, resultando em uma entrega bem-sucedida de mais de 200.000 palavras de conjunto de dados de código aberto no idioma cherokee e os primeiros smartphones com uma interface de usuário totalmente localizada para promovê-lo em 2022". O engajamento com o Chefe Sneed da EBCI, o Dr. Frey da UNC-Chapel Hill e os linguistas cherokee foi extremamente valioso e necessário para encurtar a curva de aprendizado que ocorreu para entender a divisão cultural.

Quanto ao feedback dos povos Indígenas cherokee, o chefe Sneed declarou em 2022 sobre a iniciativa que "essa é uma ferramenta útil, especialmente se for adicionada às aulas de língua cherokee nas escolas na fronteira de Qualla" e que "o que está acontecendo com a Lenovo e a Motorola é que elas reconhecem que há uma responsabilidade que acompanha a tecnologia". Há muito poder em ter tecnologia… em ser o guardião dessa tecnologia. Ver empresas que não estão interessadas apenas no lucro, mas que, em vez disso, veem sua criação como uma ferramenta para enriquecer a humanidade e, em seguida, colocam os recursos por trás disso para que aconteça… Eu aplaudo isso!" Notavelmente, o chefe principal da nação Cherokee, Chuck Hoskin Jr., disse que "sempre que uma empresa pode incorporar a língua cherokee em seu produto para os cidadãos em massa aprenderem, é uma vitória não apenas para a preservação da língua cherokee, mas para a perpetuação de todas as línguas nativas".

Embora o feedback obtido de professores da América do Norte e da América Latina se refira principalmente ao impacto dos dados de código aberto e ao aspecto inspirador da representação de idiomas Indígenas na tecnologia, o professor Sharma Suhn, do vale de Hamirpur-Kangra, acredita que a capacidade de escolher o kangri na lista de idiomas suportados pelos smartphones da Motorola trouxe orgulho e significado para a população de sua cidade.

Uma das mais recentes iniciativas de Inclusão Digital lançadas em 2023 envolvendo a língua kuvi foi na Ásia-Pacífico e que trouxe feedback positivo da Dra. Sushree Sangita Mohanty, professora assistente de antropologia, que trabalhou na iniciativa. Ela acredita que a iniciativa "também está criando oportunidades para os não falantes de kuvi se envolverem e aprenderem essa língua através dos quatro meios diferentes (kuvi-odia, kuvi-telegu, kuvi-devanagiri e kuvi-latino). Portanto, esse modelo está expandindo seus números de usuários, o que pode ajudar a aumentar o uso da língua kuvi".

Por último, a fase mais recente da iniciativa de Inclusão Digital, lançada em 2024, envolveu o idioma ladino, uma língua falada na região Dolomita, na Itália. Esta fase foi bem recebida pela comunidade e pelo Professor de Filologia Românica da Universidade Livre de Bozen-Bolzano (UNIBZ) Paul Videsott. Videsott, que também atuou como professor colaborador, acredita que a iniciativa "com certeza ajudará a dar visibilidade ao ladino e a outras línguas minoritárias", já que "os smartphones são como o lápis do século XXI". Para ele, "ver línguas minoritárias, assim como o ladino [neles] têm a mesma importância que, séculos atrás, elas teriam ao aparecer em um livro". Ele também pontua que "ter o idioma ladino nos smartphones mostra, não só para os ladinos, mas para todos os falantes de línguas menos faladas no mundo, que suas línguas não tiveram sua finalidade apenas durante os séculos anteriores ao nosso tempo, mas também terão sua utilidade no futuro". Além disso, Johann Gamper, Professor de Ciência da Computação e Vice-Reitor de Pesquisa da UNIBZ, pontua que esta iniciativa também impacta as gerações mais jovens de falantes do idioma: "Os jovens usam os telefones celulares muito mais do que nós. Eu tenho a sensação de que eles definitivamente utilizariam essa interface em ladino. Tenho certeza que usarão seu idioma local, neste caso, o ladino."

8.3 Considerações finais e continuidade

Em sua essência, o projeto de língua Indígena representa o impacto social que a tecnologia inclusiva pode causar. O projeto se alinha à missão da Fundação Lenovo de capacitar diversas populações com acesso à tecnologia, bem como à visão da Lenovo de fornecer tecnologia mais inteligente para todos. Houve fatores-chave que distinguiram a oportunidade por seu impacto social de qualidade:

- Experiência linguística: a base deste projeto é a experiência linguística da apaixonada equipe de Globalização da Motorola. A sua paixão pelo projeto e o compromisso com a sua qualidade e as comunidades impactadas, foi um sinal positivo para a sua eficácia. O apoio da Fundação Lenovo teve como foco o empoderamento dos povos Indígenas e a promoção da preservação de idiomas ameaçados, alinhado às iniciativas da equipe de Globalização

da Motorola. O impacto e a credibilidade do projeto foram possíveis graças aos especialistas no assunto e à liderança da Motorola.

- Respeito às culturas Indígenas: uma conclusão notável dos tecnólogos da Motorola por meio da parceria com especialistas acadêmicos e instituições educacionais é que o projeto não seria possível sem os colegas e estudiosos da área ou a confiança significativa das comunidades Indígenas convidadas à participação. Essa confiança só poderia ser conquistada envolvendo especialistas acadêmicos que ajudaram a equipe a traduzir, localizar e interagir, com sensibilidade e cuidado, com as comunidades Indígenas. Por isso, esses colegas e estudiosos merecem grande respeito, assim como as comunidades e culturas envolvidas no projeto.

- Compromisso com a qualidade: impulsionados por conhecimentos linguísticos, pesquisadores acadêmicos e pela colaboração das comunidades Indígenas, os especialistas da Motorola garantiram a digitalização por meio da avaliação da existência (ou não) atual de idiomas na plataforma Unicode. Os líderes selecionaram cuidadosamente os idiomas de acordo com as diretrizes da UNESCO e as avaliações do que poderia ser possível, com colaboradores voluntários. Seu foco em tentar o que era viável alcançar no mais alto padrão de localização levou a uma contribuição de qualidade, com planos de continuar a cada ano e reconhecida pela UNESCO, para a digitalização das línguas Indígenas.

- Impacto além da Lenovo e da Motorola: mais importante ainda, a digitalização não foi realizada apenas para usuários da Motorola. Os recursos criados pela equipe da Motorola através do apoio da Fundação Lenovo podem ser compartilhados em código aberto e aproveitados para que outros OEMs de tecnologia possam localizar seus dispositivos e abraçar a missão da Fundação Lenovo de capacitar populações sub-representadas com acesso à tecnologia. Para a Fundação Lenovo, é fundamental que essa iniciativa seja anunciada como um ato de Inclusão Digital para conscientizar as comunidades Indígenas, e não como uma vantagem competitiva. A equipe mantém uma visão de mais OEMs se unindo a esta iniciativa para preservar a língua e garantir a preservação do patrimônio humano e da diversidade de idiomas em nosso planeta.

Com os princípios básicos de funcionários especialistas dedicados, um compromisso com o respeito pelas culturas diversas e sub-representadas, um compromisso com a qualidade por meio da colaboração com acadêmicos e instituições e uma visão de impacto que vai além dos dispositivos da Motorola, o projeto das línguas Indígenas continuará. A equipe espera seguir contribuindo para a Década Internacional das Línguas Indígenas (IDIL 2022-2032).

Notas de fim do documento

1 UN Department of Social and Economic Affairs, Why indigenous languages matter: The International Decade on Indigenous Languages 2022–2032. Obtido em https://www.un.org/development/desa/dpad/publication/un-desa-policy-brief-no-151-why-indigenous-languages-matter-the-international-decade-on-indigenous-languages-2022-2032/

Mais informações sobre idiomas Indígenas estão disponíveis no site da UNESCO via https://www.unesco.org/en/articles/motorola-and-lenovo-foundation-announce-next-phase-initiative-revitalize-endangered-indigenous

2 United Nations. (19 de abril de 2018). The United Nations Permanent Forum on Indigenous Issues. Obtido em https://www.un.org/development/desa/indigenouspeoples/wp-content/uploads/sites/19/2018/04/Indigenous-Languages.pdf

3 Krauss, M. The World's Languages in Crisis. ResearchGate. (1992). Obtido em https://www.researchgate.net/publication/300468999_The_world's_languages_in_crisis

4 Motorola and Lenovo Foundation announce next phase of Endangered Indigenous Languages Revitalization Initiative at UNESCO HQ. (19 de dezembro de 2022). Obtido em https://www.unesco.org/en/articles/motorola-and-lenovo-foundation-announce-next-phase-initiative-revitalize-endangered-indigenous

5 UNESCO IESALC. (21 de fevereiro de 2022). A decade to prevent the disappearance of 3,000 languages. Obtido em https://www.iesalc.unesco.org/en/2022/02/21/a-decade-to-prevent-the-disappearance-of-3000-languages/

6 UNESCO World Atlas of Languages. (n.d.). Obtido em https://en.wal.unesco.org/

7 University of Connecticut. Endangered languages: UNESCO classification. Obtido em https://guides.lib.uconn.edu/c.php?g=1232158&p=9415488

8 IONOS. (12 de junho de 2023). What is Unicode?. IONOS Digital Guide. Obtido em https://www.ionos.com/digitalguide/websites/website-creation/unicode/

9 Everything you need to know about localization. Smartling. (n.d.). Obtido em https://www.smartling.com/resources/101/localization-101/

10 8 key steps in the localization process. Redokun Blog. (n.d.). Obtido em https://redokun.com/blog/localization-process

11 MotionPoint. (30 de março de 2022). Translation Management Systems (TMS): A comprehensive guide. MotionPoint. Obtido em https://ru.motionpoint.com/blog/translation-management-systems-tms-a-comprehensive-guide/

12 Prevaly, S. (1º de setembro de 2022). Computer-Assisted Translation (CAT): A complete guide. MotionPoint. Obtido em https://www.motionpoint.com/blog/computer-assisted-translation-cat-a-complete-guide/

13 Dalibor. (28 de novembro de 2022). Translation memory: What it is, and how to use it. Phrase. Obtido em https://phrase.com/blog/posts/translation-memory/

14 Sokolov, I. (16 de maio de 2023). Software localization: Getting your product ready for the global market. Translation & Localization Blog. Obtido em https://www.smartcat.com/blog/software-localization/

15 Dimmendaal, G. (2001). "Places and People: Field Sites and Informants" in Newman, P. & Ratliff, M. (eds.) Linguistic Fieldwork. Cambridge University Press. 55-75.

16 Cameron, D., Frazer, E., Harvey, P., Rampton, M.B.H., e Richardson, K. (1992). Researching language: Issues of power and method. Londres e Nova York: Routledge.

17 Czaykowska-Higgins, E. (2009). Research Models, Community Engagement, and Linguistic Fieldwork: Reflections on Working within Canadian Indigenous Communities. Language Documentation & Conservation, 3, 15-50. Obtido em http://hdl.handle.net/10125/4423

18 Sampson, G. (2016). "Writing systems: methods for recording language" in Allan, K. (ed.) The Routledge Handbook of Linguistics. Routledge. 47-61.

19 Submitting Character Proposals. Submitting character proposals. (1º de abril de 2016). Obtido em https://www.unicode.org/pending/proposals.html

20 Unicode 15.0 character code charts. Unicode (n.d.). Obtido em https://www.unicode.org/charts

21 Unicode: Where is my character? (28 de setembro de 2018). Obtido em https://unicode.org/standard/where/

22 Unicode. Private-use characters, Noncharacters & Sentinels FAQ. Unicode. (n.d.-a). Obtido em https://www.unicode.org/faq/private_use.html

23 International Components for Unicode. ICU. (n.d.). Obtido em https://icu.unicode.org/

24 Unicode. Common Locale Data Repository. Unicode CLDR. (n.d.). Obtido em https://cldr.unicode.org/

25 Rouse, M. (6 de outubro de 2020). Android SDK. Obtido em https://www.techopedia.com/definition/4220/android-sdk

26 ISO 639 code tables: ISO 639. SIL International. (n.d.). Obtido em https://iso639-3.sil.org/code_tables/639/data

27 CLDR survey tool. Unicode CLDR. (n.d.). Obtido em https://cldr.unicode.org/index/survey-tool

28 Raster, Vector, TrueType, and OpenType fonts. Microsoft Learn. (7 de janeiro de 2021). Obtido em https://learn.microsoft.com/en-us/windows/win32/gdi/raster--vector--truetype--and-opentype-fonts

29 UNESCO Brasília. UNESCO.org. (n.d.). Obtido em https://www.unesco.org/en/fieldoffice/brasilia.

30 UNESCO. Multilingualism. UNESCO.org. (20 de abril de 2023). Obtido em https://www.unesco.org/en/ifap/multilingualism

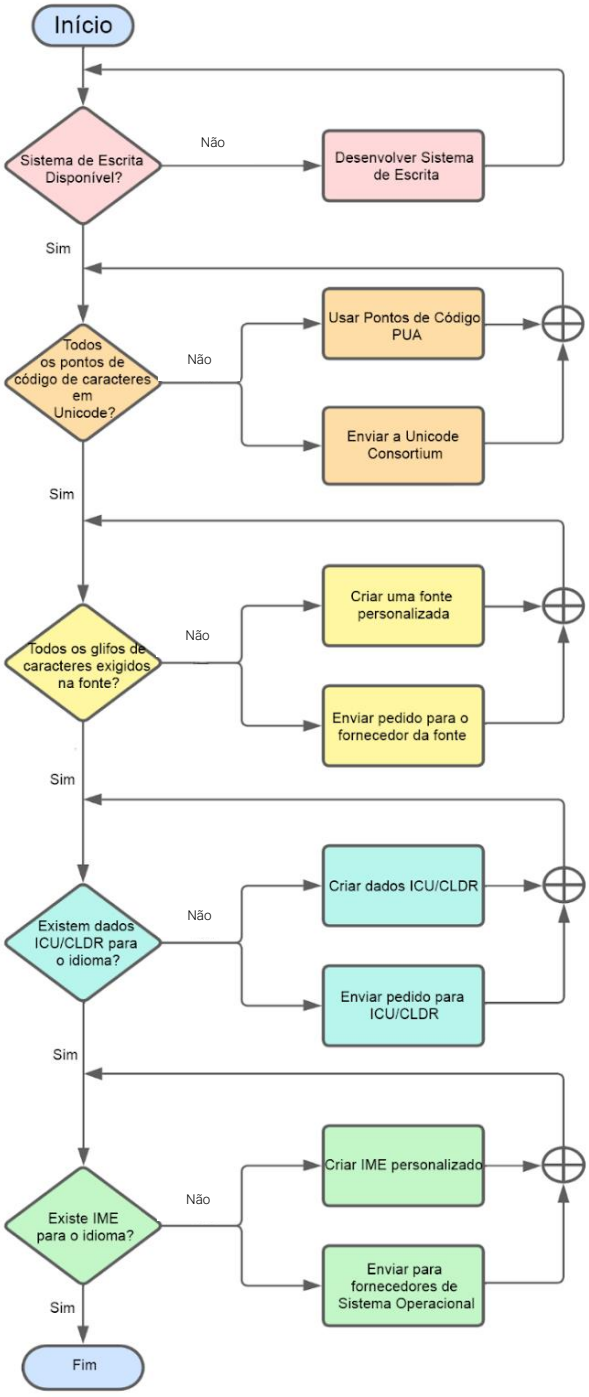
31 UNESCO. Global action plan of the International Decade of Indigenous Languages (IDIL 2022-2032). UNESDOC Digital Library. (2021). Obtido em https://unesdoc.unesco.org/ark:/48223/pf0000379851

32 da Rocha D'Angelis, W. A língua nheengatu e Suas Ortografias: Questões técnicas e de política linguística. ResearchGate. (fevereiro, 2023). Obtido em https://www.researchgate.net/publication/369577326_A_lingua_Nheengatu_e_suas_ortografias_questoes_tecnicas_e_de_politica_linguistica

33 Ternes, P. Projeto inédito cria configuração de smartphone em Kaingang e Nheengatu. Kamuri. (2021). Obtido em https://kamuri.org.br/kamuri/projeto-inedito-cria-configuracao-de-smartphone-em-kaingang-e-nheengatu/

34 Hodge, R. (2 de março de 2022). Cell phone technology keeps endangered Cherokee language a tap away. WLOS. Obtido em https://wlos.com/news/local/cherokee-motorola-endangered-language-android-12-eastern-band-of-cherokee-indians-principal-chief-richard-sneed-north-carolina

Apêndice



Fluxograma ilustrando os níveis de suporte à internacionalização por cor