LENGUAS INDÍGENAS:

DE CERO A DIGITAL

UNA GUÍA PARA LLEVAR SU IDIOMA A INTERNET

TRANSLATION COMMONS

2022-2032 | INTERNATIONAL DECADE OF
Indigenous Languages

Lenguas indígenas: “De cero a digital”

Autores: Deborah W. Anderson, Lee Collins, Craig Cornelius, Craig Cummings.
Revisores y colaboradores: Andrew Owen, Julia Nee, Lawrence Wolf-Sonkin, Anna Luisa Daigneault, Julie Anderson, Daniel Bogre Udell.
Diseño y Marketing: Mette Attar, Johanna Behm, Leonidas Pappas.
Coordinación de proyectos: Ester Perez, Jeannette Stewart

1. INTRODUCCIÓN

Translation Commons es una comunidad de voluntarios sin fines de lucro que apoya la digitalización de los idiomas, asesora a los profesionales de la lengua, brinda cursos y recursos para el sector lingüístico.

Uno de los principales programas de Translation Commons es la Iniciativa de digitalización de idiomas (LDI, por sus siglas en inglés), que busca acercar las capacidades digitales a las comunidades lingüísticas que las deseen. Cerca de 6000 idiomas en todo el mundo tienen una presencia digital pequeña o inexistente. La LDI proporciona una guía que una comunidad puede seguir con el fin de digitalizar su idioma.

Translation Commons se asoció con el 2019 - Año Internacional de las Lenguas Indígenas; una iniciativa de la UNESCO para centrar más la atención en las comunidades indígenas y la digitalización de sus idiomas. Como parte de su misión, la LDI apoya el acceso digital equitativo a los idiomas indígenas y de otras minorías con el fin de garantizar que estas comunidades lingüísticas puedan participar en actividades globales en línea y, a su vez, beneficiarse de todas las aplicaciones informáticas modernas en su idioma nativo. La creación de pautas que equipen a las comunidades con las herramientas y el entendimiento para digitalizar los escritos y publicarlos en la web, les otorga el conocimiento para facilitar el proceso sin perder la autonomía. Además de las pautas, Translation Commons ofrece tutoriales, talleres y ayuda a las comunidades con la digitalización del idioma, poniéndolos en contacto con expertos de la industria que los guiarán a través del proceso de normalización.

1.1 Serie “De cero a digital”

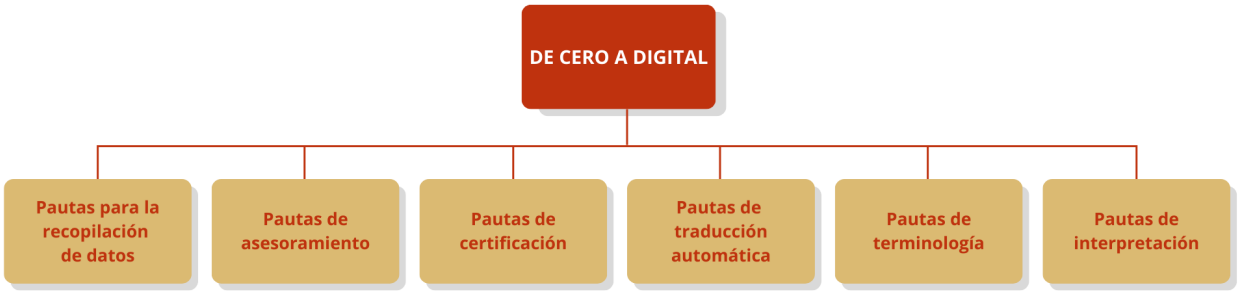
Este documento forma parte de una serie de pautas titulada “De cero a digital”, que aborda de manera integral las prácticas de digitalización del lenguaje. Los autores de las pautas son expertos en lingüística y tecnología del lenguaje. El público objetivo es cualquier comunidad lingüística que desee tener la capacidad de usar su idioma en sistemas digitales.

La digitalización amplifica las vías de comunicación de una comunidad lingüística. Para obtener más información sobre cómo la digitalización de un idioma beneficia a las comunidades indígenas y al mundo en general, consulte el *Apéndice sobre los Beneficios de la digitalización del idioma*.

Para obtener más información sobre el proceso de digitalización del idioma, consulte la página de Recursos del sitio web de Translation Commons, la cual provee información adicional relacionada con la LDI, como guías, presentaciones, videos y otros documentos.

En la Figura 1, a continuación, encontrará todas las pautas que se crearon para ayudar a las comunidades indígenas.

Figura 1: Serie “De cero a digital”



Este documento, en particular, describe cómo habilitar el software para dispositivos móviles y de escritorio con el fin de admitir un idioma escrito. La implementación que se recomienda permite a los hablantes nativos comunicarse en línea, compartir conocimientos, intercambiar documentos y utilizar software y dispositivos que de otro modo serían inaccesibles para ellos.

Los destinatarios de este documento son:

- Comunidades indígenas que desean que su idioma sea accesible en dispositivos móviles y computadoras.
- Tecnólogos que apoyan la digitalización de uno o más idiomas.
- Organizaciones que deseen habilitar comunidades lingüísticas.

Este documento tiene como objetivo ayudarlo a determinar qué herramientas necesita y cómo usarlas. También puede ayudarlo a descubrir las herramientas disponibles para usar su idioma en línea.

Cuando las personas preguntan cómo usar su idioma en internet, existen varias respuestas posibles. Es útil comprender que hay varios niveles de tecnología, que se deben tomar en cuenta al comenzar a usar un idioma en línea, tanto para la tecnología basada en la web como para la tecnología móvil.

Este documento trata sobre el uso de los idiomas escritos en línea por sus hablantes, lectores y creadores. Esto incluye conversaciones cotidianas, mensajes de texto, correo electrónico (email), redes sociales y blogs. El objetivo es ayudar a las personas a desarrollar sitios web y una variedad de contenido, y fomentar la comunicación con las comunidades locales y la diáspora lingüística en cualquier parte del mundo. Pueden existir varias convenciones de escritura o de script para un idioma determinado. La intención no es establecer una norma sobre cómo se deben usar las tecnologías digitales, sino hacer posible que las personas usen las tecnologías aquí descritas con el fin de aumentar el prestigio, la percepción pública y la usabilidad de las lenguas indígenas en todo el mundo. Si bien los métodos de documentación, las gramáticas y

los diccionarios formales son útiles en los estudios lingüísticos y la normalización, estos no son un requisito para utilizar su idioma en línea.

Esta asociación de múltiples partes interesadas está compuesta por un Comité Directivo, encargado de supervisar la implementación; grupos ad hoc, que brindan el asesoramiento relevante y socios contribuyentes.

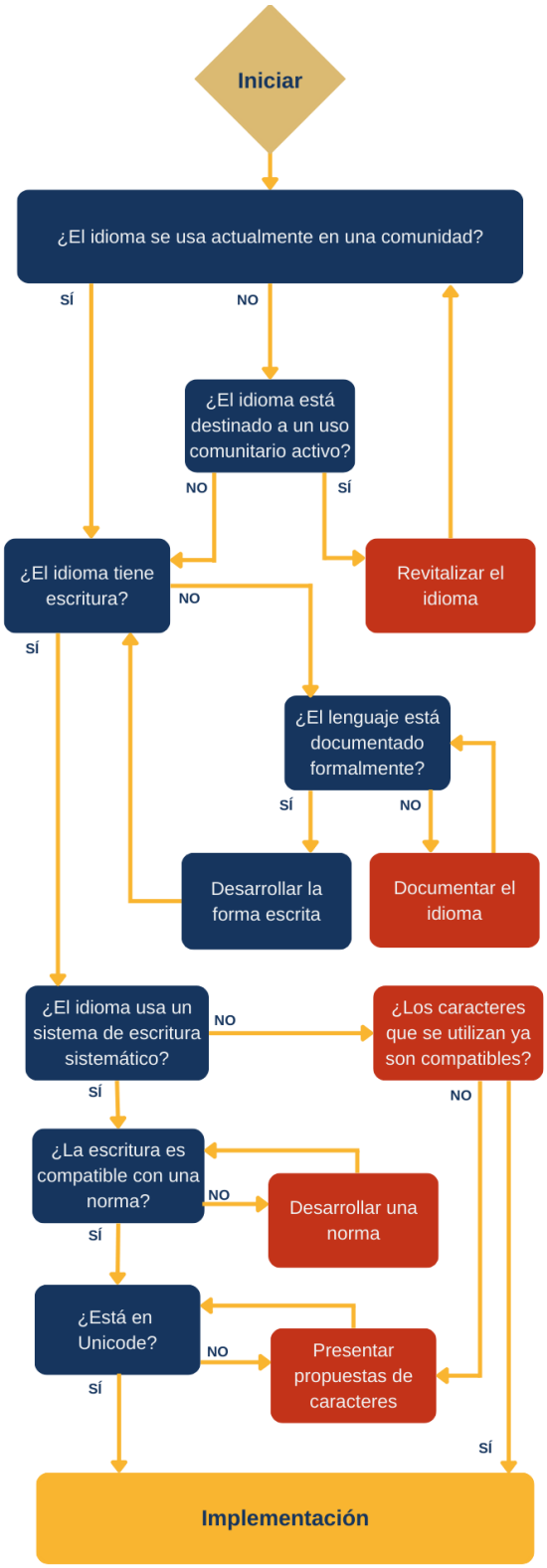
2. RESUMEN DEL PROCESO

Este documento proporciona dos flujos de trabajo. El primero se utiliza para determinar el estado actual del idioma. El segundo se utiliza para desarrollar soluciones tecnológicas que permitan el uso digital del idioma. Los pasos de los flujogramas se refieren a las secciones del documento donde encontrará información más detallada. Estos pasos son únicamente para fines consultivos y algunos de ellos pueden realizarse simultáneamente.

2.1. Flujo de trabajo sobre el estado del idioma

Este flujo de trabajo (Figura 2) describe los pasos para determinar el estado actual de un idioma en preparación para su uso en línea.

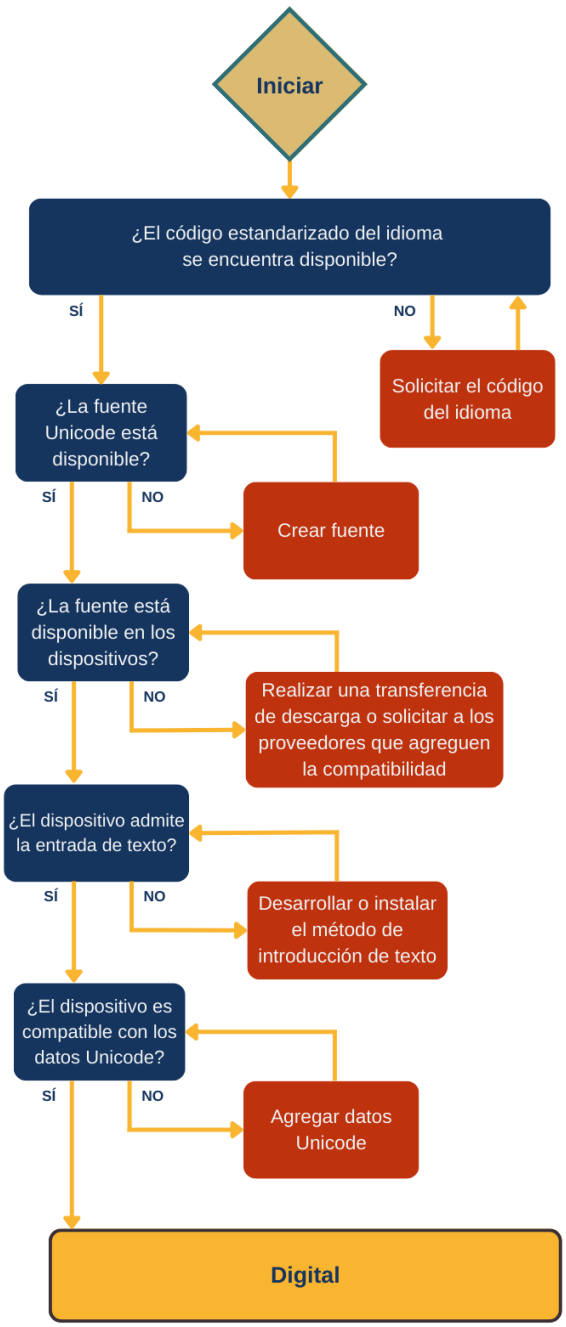
Figure 2: Determinar el estado de un idioma



2.2. Flujo de trabajo para la implementación de tecnología

Una vez determinado el estado del idioma, utilice este flujo de trabajo (Figura 3) para dar inicio a la tecnología disponible que llevará el idioma a la web.

Figura 3: Implementación tecnológica



3. ESTADO DEL IDIOMA

Esta sección describe los pasos presentados en el flujo de trabajo sobre el estado del idioma. El objetivo es determinar el nivel actual de soporte existente para el idioma en computadoras y dispositivos móviles. Incluye los siguientes pasos sugeridos para ayudarlo a poner el idioma en línea. No se necesitan conocimientos técnicos para responder a estas preguntas.

3.1. ¿El idioma se usa actualmente en una comunidad?

Sí: ¿El idioma tiene escritura?
No: ¿El idioma está destinado a un uso comunitario activo?

3.2. ¿El idioma está destinado a un uso comunitario activo?

Sí: Revitalizar el idioma
No: ¿El idioma tiene escritura?

3.2.1. Revitalizar el idioma

Con el compromiso y la organización de la comunidad, se pueden revitalizar idiomas, incluso aquellos que no tengan hablantes actualmente. Esto comprende la utilización de muchos de los recursos para la documentación del idioma, pero también son necesarios materiales de enseñanza, maestros que los usen y un gran esfuerzo por parte de la comunidad para mantener el trabajo durante muchos años, quizás décadas.

En Nueva Zelanda se adoptó un enfoque para revitalizar el idioma maorí, también conocido como Te Reo, basado, en parte, en el enfoque adoptado en Gales para revitalizar el galés. Medios de difusión en lenguas indígenas, escuelas nido, etc. También se ha adoptado un enfoque similar con el irlandés. Otros ejemplos de los muchos proyectos en todo el mundo incluyen:

- **Hebreo**: un ejemplo de revitalización de un idioma tras muchos siglos sin que nadie lo hablara.
- **Cheroqui o Cherokee**: las comunidades y los gobiernos están dedicando esfuerzos que comprenden desde escuelas de inmersión para niños y educación continua hasta la escuela primaria, secundaria y la formación de adultos.
- **Mutsun**: la tribu Amah Mutsun en California tiene un programa de revitalización para su idioma.
- **Chakma**: un idioma de Bangladés y el este de India. La comunidad está comenzando a utilizar su *script* histórico en las actividades de educación y alfabetización.
- **Tunica:** el idioma ancestral de la tribu Tunica-Biloxi quedó inactivo en 1948, y desde entonces ha sido revitalizado por personas que adquirieron la lengua por

herencia, quienes dan formación a nuevos hablantes fluidos y animan a otros a inscribirse en cursos de inmersión.

- **Córnico:** el córnico se revitalizó por primera vez a principios de la década de 1900, pero el movimiento para desarrollar el idioma ganó impulso en la década de 2000, cuando los hablantes de córnico aprovecharon los foros en línea para encontrarse y utilizar el idioma a diario.

3.3. ¿El idioma se encuentra en un registro público?

Si el idioma se encuentra en un registro público o en una lista de idiomas.

Sí: ¿El idioma tiene escritura?
No: ¿El código estandarizado del idioma se encuentra disponible?

3.4. ¿El idioma tiene escritura?

¿Su idioma tiene forma escrita?

Sí: Documentar el idioma
No: Desarrollar la forma escrita

3.4.1. Desarrollar la forma escrita

Los idiomas sin escritura se encuentran fuera del alcance de este documento. Sin embargo, aún se puede utilizar el idioma en línea con recursos de audio y video. Existen organizaciones que ofrecen orientación y herramientas para idiomas que son principalmente orales. Los departamentos académicos de lingüística también son buenos recursos.

Los recursos de audio y video son más útiles cuando los usuarios e investigadores pueden encontrarlos. Todas las grabaciones deben incluir una etiqueta o código de idioma para poder encontrar estos datos mediante indexación automática. Utilice códigos estandarizados del idioma, como IETF BCP-47.

3.4.2. Documentar el idioma

¿Su idioma está documentado formalmente, por ejemplo, con una gramática, un diccionario o un estudio lingüístico?

Sí: El idioma está documentado
No: El idioma no está documentado

3.4.2.1. El idioma está documentado

Si existen diccionarios, gramáticas y otra información lingüística, ¿son accesibles para los usuarios del idioma? Considere opciones para hacer que esta información esté disponible y sea útil para las comunidades. Entre estas, se puede incluir la creación de recursos en línea y materiales educativos adecuados para escuelas de inmersión y educación primaria. Trabaje en la asignación de los derechos de autor de dicho material a las comunidades lingüísticas. Obtenga o cree formatos digitales para poder convertir la información en un formato en línea, siempre que las comunidades lo consideren oportuno.

Si el acceso a estos recursos no es posible, tome medidas para intercambiar libros, repositorios en línea y otra información lingüística con las comunidades que usan el idioma.

3.4.2.2. El idioma no está documentado

Si el idioma aún no está documentado, puede ser posible utilizarlo en forma oral o escrita. Sin embargo, la compatibilidad en forma de sugerencias ortográficas, búsqueda y texto predictivo será limitada.

3.5. ¿El idioma usa un sistema de escritura sistemático?

Un sistema de escritura es un conjunto de reglas que tiene el propósito de utilizar uno o más *scripts* para escribir un idioma en particular. Muchos idiomas se escriben con más de un *script*. Por ejemplo, el serbocroata se escribe en caracteres cirílicos o latinos según las diferentes comunidades. Además, la mayoría de los sistemas de escritura se utilizan en más de un idioma. Por ejemplo, el birmano, el shan, el mon y otros idiomas se pueden escribir con el *script* de Myanmar (Birmania).

¿El idioma utiliza por lo menos un sistema de escritura de manera sistemática, incluida la ortografía?

Sí: ¿La escritura es compatible con una norma?
No: ¿Los caracteres que se utilizan ya son compatibles?

3.5.1. ¿Los caracteres que se utilizan ya son compatibles?

Si no existe un sistema de escritura constante, la escritura puede ser informal, sin usar una ortografía o gramática sistemática (o incluso puede usar más de un tipo de escritura). Aunque dicho texto puede crearse y visualizarse con las herramientas existentes, las herramientas de ortografía y gramática tendrán una utilidad limitada. En situaciones en las que existen múltiples sistemas de escritura contrapuestos, es importante que las personas ajenas a la comunidad perciban que esas ortografías pueden representar diferentes intereses dentro de la comunidad lingüística.

Es posible que necesite ayuda de lingüistas y de la comunidad técnica:

- Muchas comunidades utilizan diversas ortografías y conjuntos de caracteres para escribir sus idiomas, incluidas las variaciones de los dialectos. Los desarrolladores de herramientas lingüísticas, tales como teclados, deberían ser conscientes de esto y proporcionar más de un conjunto de caracteres, signos diacríticos y sugerencias de ortografía para intentar satisfacer las necesidades de todos los grupos comunitarios.
- La comunidad técnica debería considerar métodos para identificar dichas variaciones, hacerlas detectables y posiblemente desarrollar formas de facilitar la conversión entre tales variantes, según lo requieran los miembros de la comunidad.
- En el caso de que los caracteres utilizados en la escritura ya tengan algún tipo de compatibilidad informática, la comunidad técnica podrá trabajar con la comunidad lingüística para identificar y desarrollar fuentes y teclados.
- Si la falta de sistematización es una barrera para utilizar el idioma, los educadores locales, los responsables de la formulación de políticas, los líderes de la comunidad lingüística, los lingüistas, etc., pueden ayudar a adoptar una forma escrita más sistemática para su uso en línea.

3.6. ¿La escritura es compatible con una norma?

¿El sistema de escritura está respaldado por una ortografía comúnmente aceptada y un conjunto de reglas gramaticales, ya sea formal o informal?

Incluso sin tener una norma, la escritura y las comunicaciones informales son posibles. Sin embargo, la ortografía, la gramática y el vocabulario no normalizados, así como las grandes variaciones regionales, pueden dificultar el uso más avanzado de las redes sociales, el intercambio de documentos, la búsqueda en línea de sitios web e información y otras herramientas.

Sí: Presentar propuestas de caracteres
No: Desarrollar un estándar

3.6.1. Presentar propuestas de caracteres

La ortografía, la gramática, la puntuación y otros aspectos compartidos de la lengua escrita pueden mejorar en gran medida la capacidad de los miembros de la comunidad para utilizar servicios y herramientas disponibles en línea, tales como sugerencias de entrada de texto, texto predictivo, la búsqueda en línea y las herramientas de escritura.

El estándar Unicode especifica *scripts* en lugar de los idiomas. Un *script*, como el latín, se puede utilizar para más de un idioma, por ejemplo, inglés, swahili, indonesio, etc. Si

los caracteres de su idioma son compatibles, puede continuar con el flujo de trabajo de tecnología del lenguaje.

Cuando el sistema de escritura está disponible y en uso, pero sus caracteres aún no están incluidos en el estándar Unicode, será útil normalizar los caracteres para poder utilizar el idioma en dispositivos conectados, incluyendo equipos móviles, computadoras portátiles y de escritorio. Aún sin la estandarización, es posible intercambiar archivos con las mismas fuentes y métodos de entrada de textos. Sin embargo, las herramientas y los servicios en línea serán limitados.

Para cumplir el estándar Unicode:

1. Verificar si los caracteres están incluidos o son compatibles con Unicode.
2. Si no lo están, prepare y envíe una propuesta.

3.6.2. Desarrollar un estándar

Si bien no es un paso necesario y los idiomas puedan tener diferentes formas de escritura, cuando la formalización se ajusta al objetivo actual de la comunidad, se deben desarrollar pautas de ortografía, puntuación y gramática que permitan a las personas comunicar sus ideas digitalmente. Esto también mejora la capacidad de los teclados para proponer opciones de palabras y corrección ortográfica. Además, permite que los motores de búsqueda y otros servicios y herramientas en línea presenten resultados más útiles y relevantes.

3.7. Continuar a la implementación

La compatibilidad de caracteres está disponible en Unicode. Los caracteres del sistema de escritura se encuentran en los *scripts* compatibles. Ahora puede continuar con el flujo de trabajo de tecnología del idioma.

4. FLUJO DE TRABAJO DE IMPLEMENTACIÓN DE TECNOLOGÍA DEL IDIOMA

4.1 Nota sobre Tecnología para texto en sistemas digitales

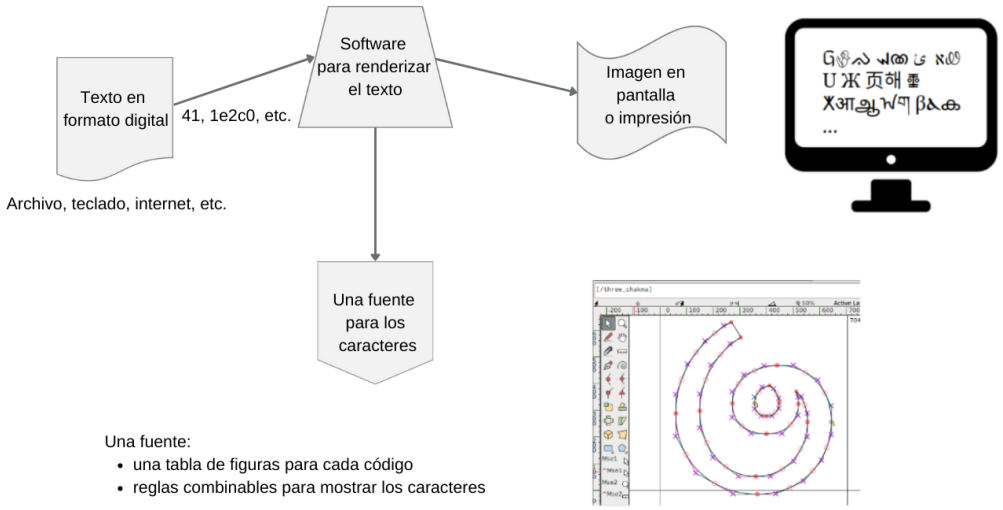
El texto en dispositivos digitales se almacena como patrones de bits que se utilizan como *punto de código*. Por ejemplo, el punto de código U+0041 podría representar “A”. Este código se puede crear por un teclado, almacenarse en un archivo o transmitirse a otra aplicación, y se visualiza o se *renderiza* a través de sistemas que entienden el sistema de codificación utilizado para los caracteres.

Un *estándar de codificación* asigna un significado a cada uno de los códigos posibles. Por ejemplo, ISO-8859-1 define caracteres para el intervalo de 0 a 255, y proporciona un significado para cada uno de estos puntos de código. Unicode es un estándar para

caracteres que incluye muchos sistemas de escritura, asignando un valor único a cada caracter entre más de 150 *scripts* diferentes, que incluyen latín, cirílico, chino, árabe, hebreo, devanagari, tamil, birmano y muchas otras.

Al usarse en conjunto, el texto creado con una codificación definida puede crearse con un método de entrada de textos compatible y visualizarse utilizando una fuente compatible. Este proceso se presenta en el diagrama a continuación (Figura 4), que ilustra cómo un estándar como Unicode puede mostrar caracteres de muchos idiomas.

Figura 4: Unicode muestra caracteres de varios idiomas



Muchos sistemas de escritura utilizan caracteres base con signos diacríticos o marcas modificadoras opcionales, como una “e” que, con un acento agudo, da como resultado “é”. Una codificación puede definir que los caracteres base puedan estar seguidos por códigos diacríticos combinados para producir el carácter combinado. En algunos casos, una codificación puede incluir caracteres precompuestos, y utilizar puntos de código únicos que incorporan tanto la base como los puntos de código combinados.

Por lo tanto, cualquier idioma escrito que se utilice en dispositivos digitales debe incluir:

- Un sistema de codificación estandarizado para caracteres, como por ejemplo, Unicode;
- Sistemas de renderizado y fuentes para los caracteres utilizados en el idioma; y
- Métodos o aplicaciones para renderizar los puntos de código en los medios deseados.

En este documento, hacemos hincapié en el estándar Unicode porque se utiliza en todos los dispositivos digitales modernos. También es la codificación predeterminada para muchas fuentes de texto. Unicode simplifica gran parte del esfuerzo necesario para poner un idioma en línea.

4.2. Definiciones para implementar la compatibilidad digital

Esta sección pretende implementar la compatibilidad digital del idioma escrito. Muchos idiomas utilizan fuentes y *scripts* no estandarizados. Esto no impide usar el idioma en línea, pero puede limitar el grado de compatibilidad disponible. Los siguientes pasos describen aspectos de la compatibilidad digital para idiomas escritos en internet.

Los requisitos básicos para un sistema de escritura digital incluyen:

- **Caracteres**: identificar el conjunto de letras, signos diacríticos, ligaduras, puntuación, números, ideogramas y otros símbolos que se utilizan en la ortografía del idioma.
- **Identificador ISO**: para el idioma y la escritura. Esto puede incluir identificadores regionales opcionales: por ejemplo, *es-MX* para español de México.
- **Estándar de codificación**: puede ser un estándar formal (Unicode con recopilación, clasificación y combinaciones de caracteres) o informal (codificación de fuentes).
- **Compatibilidad de fuentes**: la información necesaria para crear una clasificación tipográfica que incluya todas las formas necesarias de los caracteres:
    - Detalles de la presentación del texto, tales como uniones, ligaduras, grupos de combinación (puede ya estar proporcionado en documentos Unicode).
    - Texto de muestra para que los diseñadores y desarrolladores de fuentes lo utilicen en las pruebas.
      Podrá ser necesario realizar actualizaciones para los motores de renderizado u otro software.
- **Método de entrada de texto**: una manera de introducir los caracteres en todos los dispositivos. Normalmente, un teclado físico o en pantalla.

Se encuentran disponibles recursos lingüísticos adicionales:

- Repositorio de datos de configuración regional común (Common Locale Data Repository, CLDR) que incluye información adicional sobre un idioma, como calendarios.
- Códigos de idioma como IETF BCP-47.
- Software de soporte de aplicaciones para la segmentación de palabras:
    - Diccionarios de palabras escritas sin separaciones explícitas.
    - Reglas: por ejemplo, espacios, puntuación.

4.3. ¿El código estandarizado del idioma se encuentra disponible?

¿Hay un código estandarizado del idioma disponible y aprobado?

Sí: ¿La fuente Unicode está disponible?
No: Solicitar el código del idioma

4.3.1. Solicitar el código del idioma

Si no existe un código de idioma disponible, deberá establecer un código estandarizado del idioma con una variante regional opcional, un *script* o ambos. Los códigos y etiquetas de idioma no estandarizado pueden generar confusión e identificación errónea de los recursos del idioma. Siga las normas de ISO-639 y IETF BCP-47.

4.4. ¿La fuente Unicode está disponible?

¿Existe una fuente compatible con Unicode disponible para que los miembros de la comunidad puedan usarla en computadoras de escritorio o portátiles?

Sí: ¿La fuente está disponible en los dispositivos?
No: Crear fuente

4.4.1. Crear fuente

Inicialmente, después de agregar un *script* a Unicode, es posible que aún no existan fuentes Unicode que admitan los caracteres. En este caso, es posible que deba colaborar con los diseñadores de fuentes y la comunidad técnica con el fin de crear dichas fuentes para compatibilizar el texto en el *script*.

4.5. ¿La fuente está disponible en los dispositivos?

¿La fuente está disponible en dispositivos móviles y otros dispositivos utilizables por la comunidad? De ser así, los textos en sitios web, redes sociales y otras aplicaciones serán legibles.

Es importante tener en cuenta que la mayoría de las computadoras y dispositivos móviles funcionarán mejor con texto Unicode, fuentes compatibles y métodos de introducción que producen textos Unicode.

Verifique si la fuente se proporciona como parte de las fuentes Noto u otros sitios web para descargar y usar fuentes Unicode. Cabe mencionar que muchos dispositivos móviles pueden no ser directamente compatibles con la instalación o descarga de fuentes. Las computadoras de escritorio y las portátiles sí permiten la instalación de fuentes descargables.

En algunas aplicaciones, como procesadores de texto y navegadores, podría ser necesario un paso de configuración para que la aplicación presente la fuente en su documento o página web.

Los sitios web pueden utilizar fuentes web para presentar textos y utilizar una fuente proporcionada por el sitio web. Esto permite a los creadores de contenido especificar la fuente utilizada y permite que el sitio sea legible en los dispositivos, incluso si la fuente no está instalada en el mismo. Tenga en cuenta que una fuente web solo funciona con páginas que la definen específicamente, y dicho uso no instala una fuente en un dispositivo de forma permanente. Las últimas fuentes Noto están disponibles como fuentes web, lo cual es útil si el dispositivo aún no tiene la versión relevante de la fuente o no se puede instalar.

Sí: ¿El dispositivo admite la entrada de texto?
No: Instalar manualmente o pedir a los proveedores que se agregue compatibilidad

4.5.1. Instalar manualmente o pedir a los proveedores que se agregue compatibilidad

Es posible que los sistemas de escritura estandarizados recientemente aún no estén incluidos en los dispositivos móviles, aunque exista una fuente disponible. Inste a los proveedores de dispositivos a que incluyan las fuentes necesarias para el idioma.

Una alternativa es instalar manualmente una fuente Unicode en el dispositivo móvil. Esto puede ser eficaz, pero requiere conocimientos especializados, generalmente.

Advertencia: la instalación de una fuente de la web en un dispositivo móvil puede presentar riesgos de seguridad y privacidad. Además, si se realiza dicha instalación, la garantía del dispositivo o los acuerdos de compatibilidad pueden perder su validez.

4.6. ¿El dispositivo admite la entrada de texto?

Los usuarios de un idioma necesitarán un método de entrada de texto para crear mensajes, correos electrónicos, publicaciones de blogs u otro contenido en ese idioma de manera eficaz. ¿El dispositivo (móvil o de otro tipo) admite de forma propia algún modo de introducción del idioma?

Sí: ¿La entrada de texto es compatible con aplicaciones o dispositivos de terceros?
No: Desarrollar método de entrada de texto

4.7. ¿La entrada de texto es compatible con aplicaciones o dispositivos de terceros?

¿La compatibilidad de entrada de texto es compatible con los teclados estándar del sistema? Si es así, aprenda a habilitar o instalar la compatibilidad con el teclado. Esto es específico para cada dispositivo, pero los recursos en línea pueden ayudar con este procedimiento.

Los proveedores de dispositivos móviles ya tienen muchos teclados existentes en iOS y Android. Por ejemplo:

- Gboard, el teclado de Google para dispositivos móviles (https://support.google.com/gboard/answer/6380730?hl=en&co=GENIE.Platform=Android), es compatible con varios idiomas, tanto en iOS como en Android.
- Google Input Tools (https://www.google.com/inputtools/) cuenta con teclados virtuales para utilizar en las páginas web del navegador Chrome en computadoras (no en dispositivos móviles).
- Existen muchas aplicaciones de terceros para la entrada de texto en la App Store de Apple y en Google Play.

Tenga en cuenta que muchos idiomas ya tienen publicadas disposiciones de teclado, por ejemplo, en el repositorio de teclados CLDR.

4.7.1. Desarrollar método de entrada de texto

Existen numerosas herramientas para instalar métodos de entrada de texto para varios idiomas y también para desarrollar nuevos teclados. Estas incluyen:

- Aplicaciones de teclado disponibles en la App Store, Google Play u otras fuentes.
- Keyman: una herramienta para teclados de idiomas (https://keyman.com/).
- Microsoft Keyboard Layout Creator (MSKLC) Versión 1.4 (https://www.microsoft.com/en-us/download/details.aspx?id=102134).
- SIL Ukelele (https://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=ukelele).

Advertencia: la instalación de un teclado u otro método de entrada de texto desde la web puede presentar riesgos de seguridad y privacidad.

4.8. ¿El dispositivo es compatible con la base de datos Unicode?

¿Hay disponible una base de datos Unicode?

Sí: continúe usando su idioma.

No: si fuera necesario, considere agregar datos esenciales del idioma a CLDR. Estos incluyen el nombre y la etiqueta del idioma. Obtenga más información sobre CLDR en Ayuda adicional de idiomas (Additional Language Support).

5. AYUDA ADICIONAL DE IDIOMAS (ADDITIONAL LANGUAGE SUPPORT)

La estandarización Unicode cubre muchos de los aspectos de procesamiento de texto de un sistema de escritura y un idioma. Con compatibilidad para fuentes y entrada de texto, muchas cosas funcionarán sin problema, como la mayoría de los aspectos básicos del procesamiento de textos, hojas de cálculo y correo electrónico.

Los pasos aquí descritos son solo el comienzo. El Repositorio de datos de configuración regional común (CLDR) proporciona “elementos fundamentales para que el software sea compatible con los idiomas del mundo” mediante la recopilación de información útil para diferentes regiones (idioma y país). Estos datos pueden proporcionar nombres de idiomas, países, meses, días de la semana y otras informaciones. También permite el formato de fecha, hora, números y otra información formateada comúnmente en función de la configuración regional.

Si bien los datos del CLDR no son necesarios para la comunicación de texto básico en idiomas indígenas, esta información mejora la funcionalidad del idioma. Casi todas las herramientas, como el correo electrónico, los mensajes de texto, las redes sociales, etc., funcionarán bien cuando las fuentes y el teclado estén presentes.

Los programadores utilizan esta información para crear soluciones regionales para aplicaciones en línea en contextos específicos, como calendarios, hojas de cálculo, datos numéricos y selecciones de menú, como también otros contextos de la interfaz del usuario.

CLDR también almacena información adicional sobre el idioma, como los caracteres utilizados en la escritura y la distribución del teclado para la entrada de texto. Consulte los teclados CLDR para obtener más información.

Sin embargo, pueden ser necesarias herramientas de procesamiento de texto y datos adicionales para una compatibilidad lingüística más completa en aplicaciones en línea y móviles. Los siguientes son algunos de los elementos necesarios para lograr la paridad de las lenguas indígenas con las lenguas totalmente compatibles:

- **Segmentación y separación de palabras**: la separación del texto es necesaria para su correcta disposición, como también la selección de grupos de grafemas, palabras y oraciones por parte del usuario. En muchos idiomas no se proporciona un signo explícito para indicar los límites de las palabras: por ejemplo, espacios o puntuación. En tales casos, serán necesarios datos de diccionarios o algoritmos para proporcionar dicha información. Para obtener más detalles, consulte: https://unicode-org.github.io/icu/userguide/boundaryanalysis/.
- **Salto de línea**: los idiomas tienen una variedad de reglas sobre las posiciones en las que el texto puede interrumpirse para pasar a una nueva línea. Por

ejemplo, las citas deben ir entre diversos caracteres específicos del idioma y seguir diferentes reglas para separar palabras y oraciones. Además, las posiciones en las que la línea de texto puede interrumpirse dependen de las propiedades de los caracteres Unicode en el sistema de escritura. Adicionalmente, puede ser necesario juntar números con divisas, como $10, para evitar malentendidos sobre el significado. Las reglas de puntuación varían tanto por el idioma, como por la región. El salto de línea es un caso específico del análisis de límites Unicode (Unicode Boundary Analysis): https://unicode-org.github.io/icu/userguide/boundaryanalysis/#line-break-boundary).

- **Identificar el idioma del contenido textual:** los documentos deben estar explícitamente etiquetados con un código del idioma u otro identificador que describa el lenguaje humano del documento. Cuando la información esté disponible en un documento, las herramientas podrán utilizarla para encontrar la información adecuada para los usuarios de manera más efectiva. En un documento en varios idiomas, las secciones individuales o incluso los párrafos pueden etiquetarse con el idioma del texto. Es particularmente importante utilizar etiquetas normalizadas como IETF BCP-47 (https://tools.ietf.org/html/bcp47):
  - El mecanismo para este tipo de identificación varía según las aplicaciones, y podrá ser necesario que el usuario tenga que aprender si se debe hacer y cómo se hace. Tenga en cuenta que dicha identificación puede proporcionar identificadores de una lista, en lugar de ser compatible con cualquier etiqueta posible.
  - Para documentos en línea y sitios web, HTML proporciona el atributo *lang* (https://www.w3schools.com/tags/att_global_lang.asp) para etiquetar explícitamente el idioma de cualquier componente HTML. El valor de este atributo debería tomarse a partir de conjuntos normalizados de identificadores de idioma, en lugar de una cadena arbitraria definida por el usuario. Por ejemplo, utiliza *rs* para serbio, *de* para alemán, *zh-Hans* o *zh-CN* (o simplemente *zh*) para chino con escritura simplificada.
- **Detección del idioma**: los servicios en línea y otras aplicaciones a menudo pueden presentar resultados más relevantes y útiles cuando se conoce el idioma del texto. Para un texto que no está explícitamente etiquetado, se han desarrollado detectores de idioma tales como cld2 (https://github.com/optimaize/language-detector). Estas herramientas suelen realizar un análisis estadístico de los caracteres del texto, lo que proporciona una identificación probable del lenguaje humano. Esto es necesario porque la mayoría de los sistemas de escritura se utilizan para diversos idiomas, por ejemplo, las letras latinas para swahili, lakota, warlpiri y finlandés; el cirílico para ruso, ucraniano y kazajo; el birmano para birmano, shan y mon, etc.
- **Diccionarios para procesamiento de texto**: la mayoría de las aplicaciones de procesamiento de texto admiten la creación, edición, uso compartido e impresión de texto básico. El texto predictivo, la corrección ortográfica, las sugerencias gramaticales y otras herramientas emplean listas de palabras con datos de frecuencia de aparición, diccionarios y otras referencias lingüísticas. Los sinónimos y modismos de uso común también son útiles en herramientas como la búsqueda en línea.

- **Dígitos no ASCII**: muchas escrituras tienen dígitos diferentes a los dígitos occidentales. Ejemplos: el ádlam, el árabe y el farsi. Sin embargo, muchas aplicaciones, como las hojas de cálculo, no interpretan estos dígitos como valores numéricos, sino como valores textuales. Los implementadores de tales aplicaciones pueden considerar las propiedades Unicode de dichos caracteres para procesarlos como números, pero esta compatibilidad no se implementa de manera consistente (https://en.wikipedia.org/wiki/Numerals_in_Unicode).
- **Interfaces de usuario traducidas**: en algunas aplicaciones, especialmente para educación o información específica en el idioma del usuario, puede ser útil traducir el texto que aparece en la interfaz de usuario (IU). Por ejemplo, se pueden traducir elementos del menú para funciones del sistema operativo, como "Inicio" o "Abrir archivo". Sin embargo, en muchas situaciones no es viable para el propietario de la aplicación proporcionar traducciones para comunidades lingüísticas pequeñas. Cuando la interfaz de usuario está disponible en por lo menos uno de los idiomas que el usuario puede entender, una interfaz traducida tiene un valor menos inmediato.
- **Reconocimiento óptico de caracteres (Optical Character Recognition, OCR):** muchos idiomas tienen una considerable cantidad de material publicado en libros y otros documentos. El OCR se puede utilizar para convertir el texto de tales documentos a formato digital. Existen proyectos de OCR de código abierto (*open source*) disponibles (https://pdf.iskysoft.com/ocr-pdf/open-source-ocr.html) y se pueden programar en nuevos sistemas de escritura. Es importante señalar que un modelo de lenguaje con listas de palabras comunes mejora en gran medida la precisión de los métodos de OCR.

5.1. Recursos lingüísticos públicos

Existen muchos recursos disponibles, tanto gratuitos como bajo suscripción, que las comunidades pueden utilizar:

5.1.1. Recursos lingüísticos

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

5.1.2. Recursos de herramientas

- SIL - Keyman
- Herramientas de fuentes
- Herramientas de código abierto para diccionarios

5.2. Tecnología lingüística avanzada

Las siguientes capacidades requieren de una gran cantidad de datos para programar sistemas de aprendizaje automático. El software público de código abierto estará

pronto disponible, pero la mayor parte del trabajo en este campo se realiza en investigación académica o productos corporativos. Es poco probable que la mayoría de estas funciones estén disponibles para la mayoría de los idiomas en un futuro cercano.

- **Voz a texto**: reconocer las palabras pronunciadas por una persona y convertir los sonidos en texto del mensaje hablado. Dicho texto se puede transcribir a documentos, o se puede utilizar para el control de aplicaciones y dispositivos.
- **Texto a voz**: producir una salida de voz natural a partir del texto. Esto es útil para interfaces de manos libres y para la lectura automática a humanos a partir de fuentes textuales.
- **Transcripción de medios de audio académicos**: importante para la documentación del idioma, especialmente en estudios académicos de lingüística. Algunos proyectos de código abierto están comenzando a abordar esta necesidad, entre ellos:
- **Transcripción acelerada para lingüistas (Accelerated Transcription for Linguists)**: https://github.com/CoEDL/elpis
- **Traducción automática (Machine Translation)**: convertir el texto de un idioma humano a otro es una de las tareas más difíciles para las computadoras. Los sistemas actuales pueden traducir entre un conjunto limitado de idiomas compatibles. Sin embargo, estos sistemas generalmente no comprenden el contexto y no están al nivel de la traducción humana. No obstante, con las nuevas técnicas de aprendizaje automático, la calidad y la confiabilidad de la traducción automática están aumentando rápidamente para los idiomas con grandes corpus. La compatibilidad de traducción automática para lenguas indígenas no está ampliamente disponible, pero los esfuerzos universitarios y de código abierto están apareciendo. Por ejemplo, www.apertium.org es una herramienta en línea que respalda los esfuerzos de traducción automática para idiomas no dominantes.

6. GLOSARIO

ASCII (acrónimo inglés de American Standard Code for Information Interchange): código normalizado estadounidense para el Intercambio de Información. Una codificación de caracteres para la comunicación electrónica.

BCP-47: una etiqueta *IETF* para identificar idiomas.

CLDR: repositorio de datos de configuración regional común (acrónimo inglés de Common Locale Data Repository). Contiene información adicional sobre el idioma.

Caracter: el componente más pequeño del lenguaje escrito que tiene un valor semántico; se refiere al significado y /o forma abstractos, en lugar de a una forma específica. Ref.: Glosario of Unicode Tems

Punto de código: un número que representa un caracter o formato específico.

Signo diacrítico: un caracter agregado a una letra o glifo básico, generalmente para modificar su significado sonoro o semántico.

Lengua durmiente: un idioma que no tiene hablantes actualmente.

Fuente: una colección de glifos utilizados para la representación visual de datos de un caracter.

Glifo: un conjunto de símbolos que representan un caracter legible empleado para escribir.

Gramática: reglas que rigen la composición de un *lenguaje natural*.

IETF: Grupo de Trabajo de Ingeniería de Internet (acrónimo inglés de Internet Engineering Task Force).

Lengua indígena: un idioma nativo de una región específica.

IYIL2019: 2019 - Año Internacional de las Lenguas Indígenas.

Revitalización del idioma: revertir el declive de un idioma o revitalizar una *lengua durmiente*.

Ligadura: una combinación de dos o más glifos en un único glifo.

Lenguaje natural: un lenguaje que ha evolucionado naturalmente en los seres humanos, a diferencia de los lenguajes formales como los que se utilizan en informática. Incluye lenguas orales, visuales, visuales-manuales (por señas) y escritas.

Noto: una familia de fuentes que comprende más de 100 fuentes individuales, que en conjunto están destinadas a cubrir todos los *scripts* codificados en Unicode (actualmente cubren todos los *scripts* existentes en Unicode 6.0 y versiones anteriores).

Ortografía: un conjunto de convenciones para escribir un idioma.

PUA: (acrónimo inglés de Private Use Area, Área de uso privado): un intervalo de *puntos de código* en *Unicode* que nunca tendrán caracteres asignados.

Puntuación: espaciado y marcas no sonoras que ayudan a entender un texto.

Script: un conjunto de letras y otros signos escritos utilizados para representar información textual en uno o más sistemas de escritura.

Translation Commons: una comunidad y plataforma en línea para el libre intercambio de conocimientos lingüísticos.

UNESCO: Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura.

Unicode: la norma más utilizada para la codificación digital de caracteres de los sistemas de escritura del mundo.

Sistema de escritura: un conjunto de reglas para usar uno o más *scripts* para escribir un idioma en particular.

7. REFERENCIAS

7.1. Revitalización del idioma

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

7.2. Catálogos de lenguas

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

7.3. Unicode y codificación de fuentes

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

Las fuentes especializadas pueden usar enfoques no normalizados para caracteres, tales como las Áreas de uso privado (PUA) u otros intervalos de caracteres como ASCII o árabe, con glifos personalizados para puntos de código. Esto se llama codificación de fuentes. Estos enfoques no normalizados permiten a los usuarios ver los caracteres que escriben, aunque otros no lo verán, a menos que utilicen las mismas fuentes. Los servicios y herramientas en línea tampoco podrán interpretar correctamente el texto en tales codificaciones de fuentes, porque la codificación subyacente no contiene información sobre el significado real del carácter.

Si bien el texto que se basa en Unicode tiene ventajas sobre el texto codificado con fuentes, podrá ser necesario utilizar fuentes especializadas hasta que los caracteres de un sistema de escritura se incluyan en la norma Unicode. En este caso, una fuente debe usar solo PUA de la gama Unicode en lugar de reutilizar valores de código reservados para otros *scripts*. Esto evita la superposición de valores de código y permite un uso más sencillo de los *scripts* existentes. Convertir una fuente de este tipo en Unicode después de normalizar un *script* es relativamente fácil para los códigos PUA, dado que los códigos PUA en sí mismos se utilizan de forma consistente.

7.4. Códigos de los idiomas

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

7.5. Fuentes

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

Las herramientas para desarrollar fuentes incluyen:
- FontForge
- FontLab
- Glyphs
- BirdFont (https://birdfont.org/)

La Universidad de Reading tiene un programa de maestría en diseño de fuentes (typefacedesign.net/) cuyos estudiantes pueden ayudar a crear una nueva fuente en Unicode.

Los desarrolladores de fuentes comerciales pueden crear nuevas fuentes en Unicode.

8. NOTAS

El alcance de este documento se limita a los idiomas escritos. Otras formas de comunicación incluyen:

- Emojis
- Lenguas orales
- Idiomas visual-manual (gestual)
- Idiomas visuales

Idiomas con escritura propia que también se pueden escribir con otros *scripts*. Por ejemplo:

- El turco se escribía originalmente con escritura árabe, pero ahora se escribe con escritura latina.
- El chino se puede escribir con *script* latino (pinyin).

Este documento no cubre los dialectos del idioma, pero vale la pena considerarlos al proponer normas para el idioma.

La búsqueda idiomática de alta calidad tiene sus propios requisitos específicos:

- Identificación del idioma (a partir de texto)
- Segmentación: separar el texto en palabras
- Determinar cuáles son las *raíces* de las palabras, como en un idioma con varias formas: por ejemplo, *caserío* y *casas* a *casa*.

Aunque las lenguas orales excedan el alcance de este documento, los siguientes recursos pueden ser útiles:

- Google Earth (https://docs.google.com/forms/d/e/1FAIpQLSdphaDaz33syPoUDyTOTwwkaLWZx90zopUklha4uadfkUKG8A/viewform)
- https://www.blog.google/products/earth/indigenous-speakers-share-their-languages-google-earth/
- https://www.gerlingo.com/
- XTrans (https://www.ldc.upenn.edu/language-resources/tools/xtrans) es una herramienta de transcripción multicanal, multilingüe y multiplataforma de próxima generación que admite la transcripción manual y la anotación de grabaciones de audio.