SERIE DE CERO A DIGITAL:

NORMAS PARA LA RECOPILACIÓN DE DATOS LINGÜÍSTICOS

UNA GUÍA PARA PONER SU IDIOMA EN LÍNEA

TRANSLATION COMMONS

2019 | INTERNATIONAL YEAR OF Indigenous Languages

Normas para la recopilación de datos lingüísticos

Autores: Julie Anderson y Tex Texin
Colaboradores: Grigory Sapunov, Jeannette Stewart, Kirti Vashee y Debbie Anderson
Editores: Andrew Owen, Akil Iyer, Shuto Kato y Paula Cirilo
Gráficas y mercadeo: Leonidas Pappas

Agradecemos el envío de comentarios que nos ayuden a mejorar nuestras normas.
Escríbanos a krista@translationcommons.org

Tabla de contenido

1. INTRODUCCIÓN

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

Uno de los principales programas de Translation Commons es la Iniciativa de digitalización de idiomas (LDI), que busca acercar las capacidades digitales a las comunidades lingüísticas que deseen beneficiarse de ellas. Existen cerca de seis mil idiomas en todo el mundo que 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, en asociación con la iniciativa de UNESCO 2019 - Año Internacional de las Lenguas Indígenas, busca centrar la atención en las comunidades indígenas y en la digitalización de sus idiomas. Como parte de su misión, la LDI apoya el acceso digital equitativo a los idiomas indígenas y 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 normas que brinden a las comunidades las herramientas y el entendimiento para digitalizar los escritos y publicarlos en la red, lo que les otorga el conocimiento para facilitar el proceso sin perder la autonomía. Además de las normas, 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.

Este documento forma parte de una serie de normas titulada *De cero a digital*, que aborda de manera integral las prácticas de digitalización del lenguaje. Los autores de las normas 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 para que

benefice tanto a las comunidades indígenas como 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 De cero a digital: una guía para poner su idioma en línea. En la página de Recursos del sitio web de Translation Commons, encontrará información adicional y relacionada con la LDI, entre las que se incluyen, guías, presentaciones, videos y otros documentos.

1.1 Acerca de este documento

Este documento tiene por objetivo:

- Registrar los tipos de datos lingüísticos que deben recopilarse con fines de digitalización.
- Describir un proceso para la recopilación de datos.
- Correlacionar los diversos tipos de datos lingüísticos con sus usos tecnológicos.

Para lograr la digitalización de un idioma es necesario recopilar una amplia variedad de ejemplos de cómo se usa. Los lingüistas y los expertos en tecnología utilizan estos ejemplos para el estudio y el análisis con el fin de diseñar reglas y componentes que permitan admitir el idioma en los sistemas digitales. Algunos ejemplos incluyen la lista de caracteres de un alfabeto, las diferentes formas en que se escriben dichos caracteres y las listas de palabras en la lengua y sus significados. Además de estos, se requieren muchos otros tipos de datos lingüísticos para lograr que un idioma tenga un soporte digital sólido.

Los sistemas digitales admiten muchos tipos de aplicaciones. Algunas aplicaciones requieren datos lingüísticos muy sencillos. Otros tienen requerimientos más avanzados. Por ejemplo, es posible que una aplicación de bloc de notas (*notepad*) solo necesite admitir la introducción y la visualización de un texto simple. Una aplicación de procesamiento de textos puede ofrecer opciones complejas de diseño y tipografía, revisión ortográfica y gramatical, clasificación lingüística, esquemas y otras funcionalidades que dependen del idioma. En este documento se describen los diversos requisitos de datos para una variedad de aplicaciones.

2. DESCRIPCIÓN GENERAL DEL PROCESO PARA LA RECOPILACIÓN DE DATOS LINGÜÍSTICOS

Esta guía describe los pasos para recopilar elementos de datos que son representativos e instructivos con el fin de digitalizar un idioma y ponerlo a disposición para su visualización en un repositorio digital. Estos pasos son los siguientes:

- Invitar fuentes confiables para que contribuyan con datos lingüísticos.
- Aceptar el aporte de datos lingüísticos.
- Documentar la fuente y los derechos de uso y la publicación de cada elemento.
- Convertir materiales no digitales a formato digital.
- Cargar los elementos de datos del lenguaje y las anotaciones en un repositorio de datos del idioma.
- Revisar (corregir, anotar, categorizar y autenticar).
- Publicar (hacer que los elementos revisados y aprobados estén disponibles para su visualización).

2.1. Identificación de las fuentes de datos lingüísticos

En primer lugar, las comunidades lingüísticas deben identificar a las posibles fuentes confiables de los datos lingüísticos. Estos pueden existir en una variedad de formas, tales como hablantes vivos, documentos históricos, archivos, obras de arte, etc. También es útil incluir ejemplos de conversaciones informales actuales (orales o escritas) y no solo de uso formal e histórico o de estilo literario.

2.2. Recopilación de materiales y formateo digital

Después se deberán reunir los materiales reales. Si los materiales aún no se encuentran en formato digital, deberán convertirse a un formato adecuado para su almacenamiento y transmisión digital (ya sea texto, imagen, audio, video, etc.). Por ejemplo, las historias de transmisión oral se pueden grabar como archivos de audio o video o se pueden

transcribir a texto. Las ilustraciones, libros, documentos e incluso las listas de palabras escritas a mano se pueden escanear y enviar como archivos de imagen.

- Identificación de materiales adecuados.
- Recopilación de material escrito.
- Recolección de conocimientos lingüísticos de los hablantes nativos.
- Grabación de audio o video.
- Transcripción de relatos orales.
- Escanéo de fotografía de material escrito.

Los miembros de la comunidad pueden comenzar a recopilar diversos tipos de datos lingüísticos en cualquier orden. Lo importante es comenzar el proceso de recopilar, anotar y almacenar los datos en el repositorio. Los materiales de ejemplo pueden incluir:

- Correspondencia escrita a mano y documentos impresos.
- Textos y libros.
- Diccionarios monolingües y de traducción y gramática.
- Sitios web.
- Redes sociales.
- Grabaciones de audio y video.
- Canciones, poesías y representaciones.
- Tradiciones orales.
- Obras de arte, dibujos y fotografías.
- El conocimiento de la comunidad y el uso del idioma por parte de los hablantes vivos.

2.3. Licencias, propiedad y procedencia

Al recopilar datos de un idioma, se debe asegurar obtener también el derecho legal para divulgar el contenido. Por ejemplo, si adquiere cartas o documentos personales o una grabación de un discurso, podrá necesitar el permiso del autor o del disertante (o de otras partes interesadas) antes de poner este material a disposición para su uso en el repositorio. Podrá ser necesario obtener un permiso por escrito para publicar los datos y también asegurarse de que quien otorgue el permiso tiene el derecho para hacerlo.

Probablemente, deberá consultar a los ancianos o a los abogados de la comunidad dentro de su jurisdicción para identificar los problemas que puedan surgir con la aceptación de los datos, su publicación en el repositorio y el uso de los mismos como base para la digitalización, así como la redacción adecuada para cualquier acuerdo de permiso.

Además, los enfoques tradicionales de propiedad intelectual, derechos de autor y soberanía de datos pueden requerir de una cuidadosa consideración y adaptación para proteger los derechos de la comunidad a su idioma, conocimiento, patrimonio y cultura. Existen numerosos sistemas y formas de abordar lo que respecta a propiedad intelectual de manera legítima. Es necesario asimilar los modelos de propiedad intelectual y soberanía de datos que incorporan los puntos de vista de la comunidad lingüística. Para leer más sobre el tema de la soberanía de los datos indígenas, especialmente en lo que se refiere al idioma, consulte:

- Battiste y Henderson, *"Cómo proteger el patrimonio y la sabiduría indígenas"*
- Lovett et al., *"Good Practices for Indigenous Data Sovereignty"* *(Buenas prácticas para la soberanía de los datos indígenas)*
- Te Taka Keegan, *"Māori Sovereignty over Māori Language Data"* *(Soberanía maorí sobre los datos de la lengua maorí)*
- Christopher Hutton, *"Who Owns Language? Mother Tongues as Intellectual Property and the Conceptualization of Human Linguistic Diversity"* *(¿Quién es dueño del idioma? Las lenguas maternas como propiedad intelectual y la conceptualización de la diversidad lingüística humana")*

Además de los aspectos legales de la recopilación de datos, es útil documentar los orígenes de cada elemento y cómo llegó a ser incorporado en el repositorio. Incluir esta información de procedencia es apropiado para certificar que los datos pertenecen a la lengua y será tomada en consideración durante el proceso de revisión.

2.4. Notas

Es útil registrar la información sobre cada uno de los datos recopilados en la medida de lo posible. Una lengua viva cambia de forma natural con el tiempo. Las lenguas también

desarrollan diferencias regionales y dialectales. Los factores sociales, como la edad, la fluidez, el género o el estatus de cada hablante y la circunstancia (ceremonial, formal, informal, etc.) pueden afectar los actos de habla (hablados o escritos).

El registro de información sobre los hablantes, las circunstancias, cuándo y dónde ocurrieron los actos de habla y otras anotaciones crea una imagen más precisa del idioma.

Los ejemplos de anotaciones incluyen:

- La fecha y lugar de origen del elemento.
- La información sobre el hablante o autor y el receptor o público previsto (edad, género, título y si el hablante es nativo o fluido).
- La relación o parentesco entre hablantes.
- La circunstancia.
- El estilo literario (prosa, poesía, letra, rito, etc.).

2.5. Almacenamiento del repositorio

Los archivos se pueden subir a un repositorio de almacenamiento en el que la comunidad lingüística, los lingüistas, los expertos en tecnología y otras terceras partes a los que la comunidad haya concedido acceso pueden visualizar y revisar la información.

Los pasos a seguir para subir el contenido pueden depender de la configuración e implementación específicas de su repositorio. Consulte la documentación y la administración de este, para obtener más información.

2.6. Control de acceso comunitario

Los miembros de la comunidad lingüística o sus representantes controlan el acceso al repositorio. La comunidad decide quién puede actuar como administrador del mismo. Los administradores manejan quién puede subir, editar, visualizar o trabajar con los contenidos.

Por lo general, el control de acceso a un repositorio se maneja mediante la definición de roles o perfiles de usuario. A cada rol se le otorga o niega cada uno de los diferentes tipos de derechos de acceso (llamados permisos). Luego, a medida que se van agregando usuarios y se les permite hacer uso el repositorio, se les asigna roles. La asignación de estos determina sus privilegios para realizar funciones en el repositorio, por ejemplo, derechos para ver, crear, editar o eliminar registros. Los registros del repositorio pueden tener muchos campos, incluidos aquellos que representan metadatos sobre cuándo se agrega un registro, quién lo hace, información de la ubicación, etc. Por lo tanto, podrá ser necesario especificar cuáles son los permisos para ver o editar cada uno de estos campos, especialmente si algunos datos representan información cuya privacidad debe controlarse.

También puede haber permisos administrativos para ocultar o hacer público un registro, crear o editar categorías para organizar registros y permisos para administrar usuarios (agregar, eliminar, cambiar rol de usuario, etc.) Los detalles de roles, permisos, etc. dependerán de cómo esté configurado su repositorio. Nota: se pueden crear permisos específicos para registros particulares o subconjuntos de usuarios. Por ejemplo, un especialista en lenguas índicas puede recibir permisos de edición para registros relacionados con lenguas índicas, pero no para registros en otras lenguas. El moderador de un grupo que trabaja con lenguas nigerocongolesas puede tener permisos para administrar a los usuarios que trabajan con esas lenguas, pero no a los usuarios que trabajan en otras lenguas.

Algunos permisos se pueden utilizar para crear una verificación u otro tipo de proceso para el contenido del repositorio. Por ejemplo, cuando el material se sube por primera vez al repositorio, es posible que solo lo pueda ver un grupo de revisores. Esto le da al grupo la oportunidad de hacer preguntas y validar los contenidos como apropiados y representativos de la lengua. Si los revisores lo aprueban, el contenido podrá publicarse o marcarse como visible para la comunidad en general. Aquellos contenidos que requieren un nivel de madurez para verlos, también pueden marcarse como tal. Del mismo modo, los contenidos que infrinjan las normas regionales pueden marcarse y ocultarse. Todos estos tipos de contenidos también se pueden ocultar. Estas funciones

dependen de las características particulares del repositorio. Para obtener más detalles, consulte la documentación de su repositorio.

Por lo general, la comunidad lingüística concede acceso a:

- Miembros de la comunidad.
- Expertos en lenguas.
- Expertos en informática.
- Partes interesadas.

El control de acceso puede complicarse y requerir un conjunto sofisticado de funciones de usuario y permisos de acceso. Sin embargo, la Tabla 1 ilustra el conjunto más básico de funciones y permisos. Cuando la celda en una tabla está vacía indica que un permiso o privilegio de acceso ha sido **negado** a las personas con el rol de asociado. Los registros nuevos o editados quedarán *ocultos* al público hasta que un revisor configure el registro para darle una configuración de *publicado*.

Tabla 1 Manejo de acceso al repositorio

	Permisos					
	Ver	**Comentar**	**Crear/ Editar**	**Ocultar/ Publicar**	**Eliminar**	**Categorizar**
Funciones						
Invitado	Concedido					
Usuario aprobado	Concedido	Concedido				
Miembro de la comunidad	Concedido	Concedido	Concedido			
Investigador o profesional de la lengua	Concedido	Concedido	Concedido			
Revisor	Concedido	Concedido	Concedido	Concedido		Concedido
Administrador de la comunidad	Concedido	Concedido	Concedido	Concedido	Concedido	Concedido

- Un **invitado** es un usuario anónimo que desea ver el contenido del repositorio.
- Un **usuario aprobado** es una persona que ha proporcionado credenciales que justifiquen su capacidad para contribuir con conocimiento y respeto.
- Un **miembro de la comunidad** es un miembro nativo o aceptado de la comunidad lingüística.
- Un **investigador** o **profesional de la lengua** es un experto que ha sido invitado a estudiar o contribuir con el repositorio.
- Un **revisor** es una persona con habilidades avanzadas para moderar debates de contenido y entiende los sentimientos y requisitos de la comunidad, así como cuestiones legales y de otro tipo.
- Un **administrador de la comunidad** tiene el máximo control sobre los usuarios, los privilegios y los datos del repositorio.

2.7. Manejo de errores

Incluso con un proceso de investigación exhaustivo, es inevitable que los datos lingüísticos presenten errores. Los errores tipográficos, de transcripción, la falible memoria humana, etc., son factores que pueden inducir fallos durante su recopilación de datos. Los datos se deben revisar antes de que estén disponibles y documentados como representativos de la lengua. Las revisiones periódicas continuas permiten corregir datos lingüísticos erróneos que pueden haber pasado a formar parte de su colección.

2.8. El proceso de revisión

Un proceso de revisión es importante para asegurar que los datos en el repositorio estén correctamente descritos, categorizados, permitidos y que sean confiables. Las revisiones pueden sacar a la luz posibles discrepancias en los datos lingüísticos.

Los elementos pueden estar incompletos. Por ejemplo, pueden ser solo fragmentos de texto o su origen puede ser ambiguo. Aun así, estos elementos pueden ser contribuciones importantes. El proceso de revisión proporciona forma útil de medir su relevancia.

Los datos aportados al repositorio pueden provenir de muy diversas fuentes. A veces las personas, ya sea con buena o mala intención, envían información que es especulativa o no genuina.

Algunos idiomas no han sido estudiados ni documentados previamente. Su vocabulario, fonología, reglas gramaticales, entre otros, no se conocen formalmente. Los datos que se recopilen del idioma se utilizarán para extraer su estructura y terminología. Cuantos más datos lingüísticos se tengan como ejemplo de una gran diversidad de situaciones, mayor será la capacidad de digitalizar el idioma con precisión. Sin embargo, si alguno de los datos recopilados no es realmente representativo del idioma, el proceso de digitalización adecuado puede verse afectado. Los errores u omisiones pueden ocasionar una larga sucesión de conclusiones incorrectas. La probabilidad de cometer estos errores se reduce mediante un proceso de revisión.

Por lo tanto, es importante contar con un proceso en el que las contribuciones al repositorio se acepten fácilmente y que después sean revisadas por hablantes nativos, miembros de la comunidad, profesionales de la lengua y otros expertos. Los revisores pueden evaluar, comentar e incluso cuestionar los materiales en cuanto a su autenticidad, precisión e interpretación. También pueden proporcionar contexto adicional, informar sobre el uso adecuado de estos y corregir errores. El proceso de revisión puede terminar en la realización de correcciones o en una recomendación para realizar una búsqueda de ejemplos adicionales o específicos de información. También puede suceder que se le pida al colaborador que proporcione información sobre la procedencia o la licencia de los datos que facilitó. La revisión también puede confirmar si existe alguna infracción legal, como imágenes o discursos prohibidos. El proceso de revisión puede ser reiterativo.

Hasta acabar una revisión, únicamente las personas que tienen el rol de revisor pueden visualizar cada elemento. Este proceso garantiza que los otros usuarios del repositorio solo puedan ver la información que ya ha sido examinada. Si los revisores tienen preguntas sobre algún elemento, se ponen en contacto con el colaborador hasta resolver todas las dudas. Luego, el elemento puede recibir el estado de *publicado* y pasa a ser visible para todos.

2.9. Creación del proceso de recopilación de datos lingüísticos

El primer paso es reunir un equipo con las habilidades y herramientas necesarias para cada una de las tareas de recopilación de datos. Deberá crear pautas, definir responsabilidades y planificar su flujo de trabajo para respaldar su recopilación de datos. Tenga en consideración las siguientes preguntas:

- ➢ Recopilación de materiales
  - ¿Ha planificado un flujo de trabajo o proceso para invitar, aceptar y digitalizar materiales?
  - ¿Quiénes son las posibles fuentes de datos lingüísticos?
  - ¿Qué tipos de materiales de datos lingüísticos tiene a su disposición?
  - ¿Tiene alguna manera de conservar y archivar material físico de datos lingüísticos?
  - ¿Cuenta con las herramientas y habilidades para convertir los materiales a formato digital?
  - ¿Conoce los derechos de uso y publicación de los datos lingüísticos de su región?
- ➢ Alojamiento de datos lingüísticos digitales en un repositorio
  - ¿Ha instalado y configurado un repositorio de datos?
  - ¿Cuenta con las habilidades informáticas para administrar el repositorio?
- ➢ Verificación de datos lingüísticos
  - ¿Ha creado un equipo de expertos en revisión de datos?
  - ¿Cuenta con las pautas para realizar anotaciones y verificar los datos?
  - ¿Cuenta con las pautas de los requisitos necesarios para aprobar la publicación de los elementos de estos datos?
  - ¿Cuenta con las pautas para pedir a los colaboradores más información o cuestionar respetuosamente la relevancia, autenticidad, derechos de publicación u otros asuntos referentes a los materiales aportados?
  - ¿Cuenta con las pautas para resolver disputas sobre problemas con los materiales aportados?
  - ¿Ha planificado un flujo de trabajo o proceso para subir, revisar, anotar y corregir los datos digitales en el repositorio?

➢ Manejo de su equipo

- ¿Los miembros de su equipo entienden cuáles son sus tareas y responsabilidades?

2.10. Resumen de la sección

La recopilación de datos lingüísticos es un proceso continuo y reiterativo. Los elementos de datos se agregan y se revisan a medida que están disponibles. Un resumen del proceso de recopilación de datos lingüísticos es el siguiente:

- Invitar a fuentes confiables de datos.
- Aceptar el aporte de datos.
- Convertir los materiales a formato digital.
- Subir los materiales al repositorio.
- Verificar los derechos de uso y publicación.
- Anotar.
- Revisar (corregir, anotar y autenticar).
- Publicar.

3. CÓMO SE UTILIZAN LOS DATOS LINGÜÍSTICOS EN LOS SISTEMAS INFORMÁTICOS

Los datos lingüísticos se utilizan de muchas formas para respaldar la digitalización de un idioma. Está fuera del alcance de este documento detallarlas todas. Sin embargo, estas son algunas de las formas más básicas en que los lingüistas y los expertos en tecnología utilizan los datos para respaldar la digitalización de un idioma.

3.1. Muestras de escritura

Estas foras se utilizan inicialmente para establecer los símbolos de la escritura (letras, dígitos, acentos, tonos, puntuación y otros caracteres) que se usan en una lengua. A medida que se determina este conjunto de caracteres, las muestras de escritura se pueden utilizar para descubrir cómo se traza a mano cada carácter y para crear fuentes con esos caracteres. También se requiere el conjunto de caracteres para definir la disposición del teclado o el método de introducción utilizado en un sistema digital.

Recuerde que los caracteres a menudo se trazan de diversas maneras. Tenga en consideración que se pueden estilizar con y sin trazos terminales o en cursiva y que tienen muchas variaciones. En algunos idiomas, cambian de forma en función de su posición en una palabra o según el carácter que tienen al lado.

Por otro lado, también existen caracteres que rara vez se utilizan. Que se utilizan únicamente en ciertas ceremonias o en versiones antiguas del idioma. Es por esto que, cuantas más muestras se recopilen, más confiable, completa y útil será la digitalización de su idioma.

3.2. Oraciones y párrafos

Los datos que contienen oraciones y párrafos completos pueden revelar convenciones fonológicas, ortográficas, tipográficas, gramaticales u otras convenciones lingüísticas de un idioma que son necesarias para el procesamiento de textos. Los ejemplos incluyen:

- Reglas gramaticales.
- Uso de guiones, separación de palabras, uso de mayúsculas, énfasis y puntuación.
- Justificación y dirección de la escritura.
- Pronunciación.
- Formas de tratamiento (por ejemplo, títulos honoríficos y orden de los nombres).

Los datos también pueden revelar convenciones de escritura únicas y formatos usados para representar fechas, horas, eras, números, porcentajes, etc.

3.3. Listas terminológicas

Las listas terminológicas pueden obtenerse a partir de muestras de escritura, grabaciones de audio y otros datos. Estas listas de palabras y frases son utilizadas por correctores ortográficos, correctores automáticos, texto predictivo, reconocimiento óptico de caracteres (OCR) y otras funciones digitales.

3.4. Diccionarios

Los diccionarios monolingües y bilingües proporcionan definiciones, partes de la oración, pronunciación, etimología, traducción y más información. Esta información puede ser útil para el procesamiento de textos, separación de palabras, revisión ortográfica y gramatical, traducción automática y otros aspectos de la digitalización.

Diseñar y dar formato a un diccionario puede ser complicado. Los lexicógrafos prestan especial atención, por ejemplo, a los enfoques para *lemas* en idiomas polisintéticos, donde las palabras u oraciones complejas se construyen a partir de muchas partes y a la alfabetización de las palabras. Algunos de los recursos disponibles para elaborar diccionarios de lenguas indígenas incluyen:

- Nick Thieberger, *"La lexicografía de las lenguas indígenas en Australia y el Pacífico"*
- Antonia Cristinoi y François Nemo, *"Challenges in endangered language lexicography"* *(Desafíos en la lexicografía de las lenguas en peligro de extinción)*

- Paul V. Kroskrity, *"Designing a Dictionary for an Endangered Language Community" (Diseñar un diccionario para una comunidad lingüística en peligro de extinción)*
- Frawley, Hill y Munro, *"Making Dictionaries: Preserving Indigenous Languages of the Americas" (Elaboración de diccionarios: preservación de las lenguas indígenas de las Américas)*
- Sarah Ogilvie, *"Linguistics, Lexicography, and the Revitalization of Endangered Language"* (*Lingüística, lexicografía y la revitalización de los idiomas en peligro de extinción*)

3.5. Grabaciones

Se pueden utilizar grabaciones de audio y video para descubrir las normas de pronunciación. Esta información permite utilizar capacidades de texto a voz y de voz a texto. La conversión de texto a voz es útil para personas con discapacidades visuales o de lectura y de bajo nivel de alfabetización. El reconocimiento de voz habilita las capacidades de comando de voz y también es útil para las personas con discapacidades que no pueden escribir en un teclado o pantalla táctil.

3.6. Datos bilingües

Los datos bilingües sirven para muchos propósitos. Por ejemplo, los diccionarios de traducción, los videos subtitulados y los documentos traducidos permiten la creación de diccionarios de búsqueda de palabras en línea, herramientas de traducción de voz, traducción automática y otras herramientas. Además, la comparación de terminología entre idiomas revela diferencias matizadas.

3.7. Volumen

Por lo general, cuantos más datos se recopilen, mejor será la calidad de la digitalización. La gramática y las definiciones de las palabras, entre otras, se vuelven más precisas y matizadas. Se pueden documentar modismos y términos de uso poco frecuente.

Existe terminología que se utiliza únicamente cuando está asociada con ciertos temas o áreas particulares, por ejemplo, salud, agricultura, leyes, etc. Cuanto mayor sea el volumen de datos recopilados, mayor será la probabilidad de cubrir más áreas.

Además, algunas aplicaciones de idiomas, como, por ejemplo, la traducción automática, funcionan correctamente solo cuando hay un volumen significativo de datos lingüísticos disponibles para entrenar el sistema de traducción.

3.8. Variedad

Todos los tipos de datos lingüísticos son útiles para establecer un soporte sólido que permita digitalizar su idioma. Nunca excluya fuentes de datos lingüísticos que parezcan arcaicas, informales, formales, oficiales, para gente joven o sin educación, hiperbólicas (publicidad) o inverosímiles (leyendas o cuentos históricos). Cualquiera de los siguientes puede ser de utilidad:

- Cuentos infantiles y libros ilustrados.
- Materiales educativos.
- Correspondencia, cartas personales, notas y mensajes.
- Documentos legales (certificados de nacimiento, matrimonio, defunción, etc.).
- Diccionarios (monolingües y bilingües).
- Libros.
- Periódicos.
- Señalizaciones.
- Carteles.
- Escritura arcaica.
- Historias y tradiciones orales.
- Dibujos y otras obras de arte.
- Conversaciones cotidianas (orales, escritas, formales e informales).

3.9. Creación de terminología nueva

La creación de nueva terminología, o neologismos, es necesaria para expresar ideas, invenciones y actividades novedosas. Esto es especialmente cierto cuando se prepara un idioma para la digitalización. Por ejemplo, la terminología que se utiliza en la interfaz de usuario de software y hardware debe definirse o adaptarse en el idioma nativo. Se incluyen términos como menú, botón, menú desplegable, archivo, editar, ayuda, salir, aceptar, cancelar, hacer clic, descargar, etc.

Los equivalentes del idioma nativo no siempre se eligen o se inventan de manera fácil u obvia. Las traducciones literales pueden no ser la mejor opción. Por ejemplo, el término "página principal" en algunos idiomas se traduce como "página de inicio". Además, un término puede tener múltiples usos y requerir diferentes traducciones según el contexto. Por ejemplo, a veces "cancelar" puede tener el significado de "abortar" y otras veces puede significar "deshacer". Por ejemplo, en inglés, *orange* es tanto un color como el nombre de una fruta, pero en otros idiomas hay términos separados para cada uno.

Es posible que una comunidad lingüística necesite crear un proceso para generar y acordar su terminología para la digitalización informática. Consulte las ***Normas de terminología De cero a digital*** para obtener más información sobre la creación de nuevos términos.

3.10. Resumen de la sección

Cuanto mayor sea el volumen y la variedad de datos lingüísticos de su colección, más precisa y completa será la digitalización de su idioma. La siguiente tabla ilustra algunas formas básicas sobre cómo son utilizados los datos lingüísticos en los sistemas informáticos.

Tabla 2. Uso de datos lingüísticos en sistemas informáticos

Tipo de datos lingüísticos	Valor lingüístico	Usos de programas de software
Muestras de escritura	Revelar símbolos y pautas de escritura	Determinar el conjunto de caracteres, fuente, distribución del teclado, método de introducción, etc.
Datos de la oración, párrafos	Revelar pautas fonológicas, ortográficas, tipográficas, gramaticales y otras pautas lingüísticas	Procesamiento de palabras y voz
Listas terminológicas	Revelar vocabulario	Correctores ortográficos, corrección automática, texto predictivo y reconocimiento óptico de caracteres
Diccionarios	Proporcionar definiciones y partes de la oración	Correctores ortográficos y gramaticales, procesamiento de textos, conversión de texto a voz y traducción automática
Grabaciones de audio y video	Revelar las pautas de pronunciación y los patrones del lenguaje coloquial	Conversión de texto a voz y reconocimiento de voz
Datos bilingües	Proporcionar traducciones, correlaciones entre idiomas y sistemas de escritura	Diccionarios bilingües, videos subtitulados, documentos y software traducidos, herramientas de traducción de voz y traducción automática
Neologismos	Establece términos que son necesarios para los sistemas digitales	Menús de software, comandos, diálogos, etc.

4. UTILIDADES ESPECIALIZADAS DEL LENGUAJE

4.1. Repositorio de datos de configuración regional común de Unicode

El Repositorio de datos de configuración regional común de Unicode (CLDR) proporciona componentes clave para la internacionalización y localización de software en apoyo de los idiomas del mundo. Como su nombre lo indica, el CLDR representa una gran colección de datos locales. Es utilizado por una amplia gama de empresas para adaptar su software y apoyar las convenciones de distintos idiomas.

Por ejemplo, el CLDR contiene traducciones aceptadas de muchos nombres propios que son necesarios para las aplicaciones de software, tales como los meses, los días de la semana, los países y sus subdivisiones, los nombres de idiomas, las unidades de medida y las divisas. También proporciona expresiones de codificación que el software puede usar para dar formato a los datos según las convenciones locales de fecha, hora, número, medida y moneda, entre otros.

Si su idioma aún no está representado en el CLDR, considere enviar información para que se incluya. Esto permitirá que los desarrolladores de aplicaciones de software puedan incluir su idioma en sus sistemas más fácilmente. Una revisión a los contenidos del CLDR también puede ayudarlo a determinar la información y la terminología que necesitan los sistemas digitales y que deben crearse para su idioma.

4.2. Traducción automática

La traducción automática es una aplicación lingüística cada vez más importante porque puede traducir grandes volúmenes de texto de manera rápida y eficiente.

Sin embargo, para establecer sistemas de traducción automática se necesita una gran cantidad de datos bilingües. Idealmente, los datos deben estar alineados en corpus paralelos donde cada segmento (u oración) del idioma de origen se empareja con un segmento del idioma de destino correspondiente. Para obtener más información sobre la importancia de la traducción automática y sus requisitos específicos, consulte el *Apéndice C: Requisitos para la traducción automática*. Esta guía proporciona una descripción

general de la traducción automática y los requisitos de datos para crear un sistema de dicha traducción.

4.3. Herramientas de procesamiento del lenguaje natural

Esta sección trata sobre herramientas que ayudan a analizar su idioma y a generar datos lingüísticos. Un ejemplo es una herramienta que escanea un documento y extrae una lista de palabras en su idioma. Inicialmente, debe identificar herramientas que le ayuden con tareas inmediatas. Estas herramientas pueden necesitar alguna personalización para que puedan trabajar con su idioma.

Muchas aplicaciones de software se basan en componentes de Procesamiento de Lenguaje Natural (PLN) para trabajar con datos lingüísticos. Cuando haya alcanzado las etapas clave de la recopilación lingüística, es posible que desee invertir en la formación de modelos lingüísticos y la actualización de las bibliotecas de software utilizadas por la industria. Proporcionar modelos que han sido previamente preparados para su idioma puede hacer que muchas aplicaciones lo adopten más rápidamente.

Aunque los idiomas son diversos, muchos de ellos se pueden agrupar por patrones similares de sintaxis y gramática, entre otras características. Las herramientas que reconocen los patrones utilizados por su idioma pueden ayudar a procesar los datos lingüísticos. Por ejemplo, hay herramientas que pueden analizar texto para ciertos tipos de idioma y crear listas de palabras. Existen herramientas más sofisticadas que usan modelos del idioma para procesar los datos lingüísticos. Los datos lingüísticos también se pueden usar para enseñar y crear modelos del idioma cada vez más refinados. Existen algunas herramientas que están específicamente diseñadas para trabajar con idiomas poco documentados.

También existen varios sitios web de código abierto que alojan herramientas de modelado de lenguaje y PLN. Considere seleccionar una herramienta que admita idiomas gramaticalmente similares al suyo y, luego, personalícela y entrénela con sus datos lingüísticos. A medida que lo va logrando, puede agregar el modelo entrenado de su idioma a estos sitios web. Los desarrolladores de software podrán admitir su idioma más

fácilmente y, posteriormente, usted podrá invitar a las principales empresas de software a agregar su idioma a sus bibliotecas.

Los ejemplos de estas herramientas incluyen:

- Vectores de palabras como word2vec (un modelo que identifica asociaciones de palabras a partir de un corpus de texto grande).
- Modelos entrenados como BERT y GPT (captura de funciones gramaticales, significados y asociaciones de palabras).
- Modelos para el reconocimiento de entidades nombradas (NER) (nombres, geografía, fechas, etc.).
- Modelos para seguimiento de las partes de la oración (sustantivo, verbo, etc.).
- Modelos para análisis sintáctico o de dependencia (sujeto/objeto/..., sintagmas verbales y nominales, etc.).
- "Transformers" de Huggingface (extracción de información, traducción y otras funciones de PLN).
- Biblioteca SpaCy (extracción de información y otras funciones de PLN).
- Natural Language ToolKit (NLTK) (bibliotecas de procesamiento de texto).

Muchas aplicaciones se basan en bibliotecas de identificación de idiomas con el fin de configurarse para el idioma actual. No soportan idiomas que estas bibliotecas no reconocen. El aportar información a estas bibliotecas para que reconozcan su idioma puede acelerar la aceptación y el soporte del mismo.

Los ejemplos incluyen:

- CLD3 de Google (Detector de lenguaje compacto v3).
- fastText de Facebook.

4.4. Resumen de la sección

Estas utilidades pueden ayudar con el análisis de su idioma, respaldar su digitalización, promover la adopción por parte de los desarrolladores de software y acelerar la traducción de materiales desde y hacia él.

- Repositorio de datos de configuración regional común de Unicode.
- Traducción automática.
- Herramientas de procesamiento del lenguaje natural.

5. GLOSARIO

Término	Descripción
Carácter	Una letra, logograma, signo, marca o símbolo que se utiliza en la escritura.
Corpus	Una colección grande o completa de escritos.
Datos locales	Información utilizada para personalizar las interfaces de usuario para un idioma y una cultura determinadas de una región.
Diacrónico	Relacionado con la forma en que algo, especialmente el lenguaje, evoluciona con el tiempo.
Dialectal	Perteneciente o relativo a un dialecto de un idioma.
Digitalizar	Convertir a un formato digital que pueda ser procesado por una computadora.
Escanear	Copiar y almacenar información en forma digital.
Fonológico	Relativo a los sonidos en un idioma o idiomas en particular o al estudio de los mismos.
Formato de archivo	Un método para almacenar información en un archivo informático. El método varía según el tipo de datos que se almacenan y generalmente se puede identificar por la extensión del archivo. (ej., Html para una página web).
Fuente	Representación gráfica del texto. Una colección de símbolos de escritura con un diseño gráfico similar.
Glosario	Lista alfabética de palabras y sus definiciones relacionadas con un tema específico.
Indígena	Originario de una región específica.
Lexicografía	La práctica de compilar diccionarios.

Medios de comunicación	1. Medios de comunicación masiva (radiodifusión, publicación e Internet) considerados colectivamente. 2. Dispositivos de almacenamiento de datos.
NER	Reconocimiento de entidad nombrada.
NLTK	Herramientas del idioma natural.
Ortográfico	Conjunto de pautas para la escritura de un idioma. Incluye pautas de ortografía, unión con guiones, separación de palabras, uso de mayúsculas, énfasis y puntuación.
PLN	Procesamiento del lenguaje natural.
Polisintético	Que denota o es relativo a un idioma caracterizado por palabras complejas que constan de varios morfemas en los que una sola palabra puede funcionar como una oración completa.
Repositorio	Una ubicación central en la que se almacenan y se administran datos.
Segmento	Una unidad significativa discreta de texto o lenguaje hablado. La segmentación de texto es el proceso de dividir el texto escrito en unidades significativas, como palabras, oraciones o temas. La segmentación del habla es el proceso de identificar los límites entre palabras, sílabas o fonemas en los lenguajes naturales hablados.
Subir	Transferir (datos) de un equipo informático a otro, por lo general más grande o más distante del usuario o que funciona como un servidor.
TA	Traducción automática.
Texto a voz	Tecnología de asistencia que lee el texto digital en voz alta al usuario.
Tipográfico	Arte y técnica en la organización y selección de tipos para hacer que el lenguaje escrito sea legible y atractivo cuando se visualiza.
Trazos terminales	Una ligera proyección que remata el trazo de una letra en determinados tipos de letra.

Unicode	Un estándar de codificación internacional que admite texto digital en diferentes idiomas y presentaciones. A cada letra, dígito, símbolo u otro carácter se le asigna un valor numérico único y funciona de manera consistente en diferentes plataformas y programas.
URL	Uniform Resource Locator. (Localizador Uniforme de Recursos). La dirección de una página u otra información en la red.
UTF-8	Codificación de caracteres de ancho variable basada en Unicode utilizada para la comunicación electrónica de texto.
Voz a texto	Un programa de reconocimiento de voz que convierte el lenguaje hablado en texto escrito.
XLIFF	Un formato de archivo basado en XML para intercambiar datos localizables.

Apéndice A: Ejemplos de aplicaciones digitales

Estas actividades comunes pueden estar disponibles en su lengua nativa en sistemas digitales. También es posible crear nuevas aplicaciones que sean específicas para las necesidades de su comunidad.

Comunicación

- Enviar y recibir mensajes de texto.
- Enviar y recibir correos electrónicos.
- Enviar y recibir medios (imágenes, audio y video).
- Traducir texto automáticamente; traducción automática.
- Convertir automáticamente los mensajes de voz para visualizarlos como texto y viceversa.

Publicación, documentación y procesamiento de texto

- Publicar y acceder a información en sitios web.
- Crear y compartir documentos, libros, medios de información, señalización, carteles y materiales educativos.
- Crear diccionarios impresos y en línea.
- Escanear documentos para convertirlos a texto digital.
- Crear una fuente para su presentación.
- Comprar y vender cosas en línea.
- Localizar sitios web y aplicaciones en su idioma.
- Crear aplicaciones en su idioma nativo.
- Corrector ortográfico.
- Revisión gramatical y corrección automática.

Interfaces de usuario y asistencia para personas con discapacidad

- Reconocimiento de voz (útil para personas con discapacidades físicas).
- Utilización de comandos de voz para controlar los dispositivos.

- Conversión de texto a voz y lectores de pantalla (útil para personas con discapacidades visuales o personas con bajo nivel de alfabetización).
- Voz a texto y subtitulación en tiempo real (útil para personas con discapacidad auditiva).

Apéndice B: Requisitos de datos para aplicaciones tecnológicas

La tabla de requisitos de datos para aplicaciones tecnológicas es una guía visual para ayudarlo a determinar un buen punto de partida en la recopilación de datos lingüísticos con respecto a sus objetivos con el fin de habilitar aplicaciones tecnológicas en su idioma.

La tabla ayuda a las comunidades desde dos ángulos: una comunidad puede buscar las aplicaciones informáticas deseadas y descubrir los tipos de datos lingüísticos que son necesarios para respaldar su creación; o en función de los tipos de datos lingüísticos que están disponibles o que se pueden adquirir, una comunidad puede determinar las aplicaciones que se pueden lograr en el corto plazo.

La tabla muestra gran parte de la misma información descrita en este documento y ayuda a los usuarios a tener expectativas realistas para lograr sus objetivos de digitalización.

Tabla 3. Requisitos de datos para aplicaciones tecnológicas

Tipos de datos	Símbolos de escritura	Terminología	Traducción	Habla
Contenido de ejemplo	Caracteres, letras, dígitos, signos de puntuación, fuentes, etc.	Listas de términos, diccionarios monolingües	Diccionarios bilingües, redes de palabras, corpus	Fonología, reglas de pronunciación, grabaciones de audio y video
Aplicaciones digitales				
Comunicación				
Enviar/recibir mensajes de texto	x			
Enviar/recibir correos electrónicos	x			
Traducir texto automáticamente, traducción automática	x	x	x	
Tipos de datos	Símbolos de escritura	Terminología	Traducción	Habla

Convertir automáticamente los mensajes de voz para visualizarlos como texto y viceversa	x	x		x
Publicación, documentación y procesamiento de palabras				
Publicar y acceder a información en sitios web	x			
Crear y compartir documentos, libros, medios de información, señalización, carteles y materiales educativos	x			
Crear diccionarios impresos y en línea	x	x	x	
Escanear documentos para convertirlos a texto digital (OCR)	x			
Crear una fuente para su presentación	x			
Comprar y vender cosas en línea	x			
Sitios web y aplicaciones localizados en su idioma	x	x	x	
Crear aplicaciones en su idioma nativo	x			
Corrector ortográfico	x	x		
Revisión gramatical y corrección automática	x	x		

Interfaces de usuario y asistencia para personas con discapacidad				
Tipos de datos	Símbolos de escritura	Terminología	Traducción	Habla
Reconocimiento de voz (útil para personas con discapacidades físicas, también para hablar con Siri o Alexa)		x		x
Conversión de texto a voz y lectores de pantalla (útil para personas con discapacidades visuales o personas con bajo nivel de alfabetización)	x	X		x
Voz a texto y subtitulación en tiempo real (útil para personas con discapacidad auditiva)	x	X		x
Leyenda	**x**	La marca de verificación indica el tipo de datos que más probablemente admita la creación de este tipo de aplicación.		
Nota		Esta tabla no es una lista completa de los tipos de datos o aplicaciones.		

Tipos de datos	Normas del idioma y la disposición		Datos lingüísticos masivos	
Contenido de ejemplo	Ortografía, reglas gramaticales, separación de palabras, uso de mayúsculas, puntuación, sentido de la escritura, justificación	Normas de escritura: fechas, horas, eras, números, porcentajes, etc.	Oraciones, párrafos, corpus textuales monolingües	
Aplicaciones digitales				
Comunicación				
Enviar/recibir mensajes de texto				

Enviar/recibir correos electrónicos				
Traducir texto automáticamente, traducción automática	x	x	x	
Tipos de datos	Normas del idioma y la disposición		Datos lingüísticos masivos	
Convertir automáticamente los mensajes de voz para visualizarlos como texto y viceversa	x	x	x	
Publicación, documentación y procesamiento de palabras				
Publicar y acceder a información en sitios web	x	x		
Crear y compartir documentos, libros, medios de información, señalización, carteles y materiales educativos	x	x		
Crear diccionarios impresos y en línea	x	x		
Escanear documentos para convertirlos a texto digital (OCR)			X	
Crear una fuente para su presentación	x		X	
Comprar y vender cosas en línea	x	x		
Sitios web y aplicaciones localizados en su idioma	x	x		
Crear aplicaciones en su idioma nativo	x	x		
Corrector ortográfico				
Revisión gramatical y corrección automática	x	x		

Interfaces de usuario y asistencia para personas con discapacidad				
Tipos de datos	Normas del idioma y la disposición		Datos lingüísticos masivos	
Reconocimiento de voz (útil para personas con discapacidades físicas, también para hablar con Siri o Alexa)				
Conversión de texto a voz y lectores de pantalla (útiles para personas con discapacidades visuales o personas con bajo nivel de alfabetización)	x	x	X	
Voz a texto y subtitulación en tiempo real (útil para personas con discapacidad auditiva)	x	x	x	
Leyenda	La marca de verificación indica el tipo de datos que más probablemente admita la creación de este tipo de aplicación.			
Nota	Esta tabla no es una lista completa de los tipos de datos o aplicaciones.			

Apéndice C: Requisitos para los datos de traducción automática

¿Qué es la traducción automática?

La traducción automática (TA) es el uso de software para traducir texto o voz de un idioma a otro.

Para producir traducciones de alta calidad, esta es más que una sustitución mecánica de palabra por palabra. La traducción automática utiliza algoritmos avanzados y requiere grandes cantidades de datos lingüísticos como ejemplo para configurar un sistema productivo.

Los sistemas de traducción automática no alcanzan los niveles de calidad de las traducciones humanas. Con el fin de mejorar la calidad, estos sistemas de traducción automática se personalizan según el dominio o la profesión para limitar el alcance del contenido.

La traducción automática es útil como herramienta de apoyo para los traductores humanos y en algunos casos puede generar resultados que pueden ser usados ta l cual.

Creación de sistemas de traducción automática

Para automatizar la traducción entre un par de idiomas, se debe crear un sistema de traducción automática específico para ese par. Esto implica elegir una tecnología de traducción automática y utilizar datos lingüísticos para ambos idiomas con el fin de configurar el sistema.

La creación de sistemas de traducción automática para nuevas combinaciones lingüísticas solo es posible después de haber creado una base digital segura para cada idioma. También debe haber una proliferación natural y creciente de recursos lingüísticos en un idioma recién digitalizado.

Actualmente existen muchos idiomas con millones de hablantes que no cuentan con sistemas de traducción automática utilizables, ya sea porque no se dispone de los recursos de datos adecuados o porque no se ha realizado el esfuerzo necesario. Si bien la tecnología central de desarrollo de la traducción automática continúa mejorando y es cada vez más fácil construir nuevos sistemas con menos datos, debe entenderse desde el principio que se requieren volúmenes significativos de datos para producir *buenos* sistemas de traducción automática.

Los datos utilizados para construir sistemas de traducción automática se denominan datos de entrenamiento.

Si bien es posible desarrollar un sistema básico en un tiempo relativamente corto si algunos de estos datos se encuentran disponibles, con el tiempo se pueden agregar a un sistema de traducción automática existente para promover mejoras y rendimiento continuos.

Adicionalmente, los sistemas de traducción automática están en constante evolución y mejoría a medida que se incorporan nuevos datos, nuevas técnicas y comentarios correctivos continuos. Es necesario planificar evaluaciones y actualizaciones periódicas de estos sistemas.

¿Para qué es útil la traducción automática?

La traducción automática es útil para poner a disposición rápidamente grandes volúmenes de información y recursos de conocimiento a un costo relativamente bajo. Sin embargo, también debemos entender que, en la actualidad, la traducción automática no llega a ser una traducción humana competente.

Sin embargo, la traducción automática es rápida, por lo general ofrece una buena aproximación y puede implementarse para que millones de personas la usen libremente en la red después de que se haya creado un sistema de traducción automática. La existencia de esta permite que millones de personas accedan a información que de otra manera no sería posible hacerlo. Si bien abundan las historias de contratiempos y errores, para muchos es cada vez más evidente que es crucial aprender a utilizar y ampliar las

capacidades de esta tecnología con éxito. Si bien es poco probable que la traducción automática sustituya a los seres humanos en cualquier tarea en la que la calidad sea primordial, hay un número creciente de casos que muestran que la TA es adecuada para:

- Contenido muy repetitivo.
- Contenido que simplemente no se podría traducir de otra manera.
- Contenido que no está al alcance de la traducción humana.
- Contenido de alto valor que cambia cada hora y todos los días.
- Contenido de conocimiento que facilita y mejora la difusión global del conocimiento crítico.
- Contenido creado para mejorar y acelerar la comunicación con clientes globales que prefieren un modelo de autoservicio.
- Contenido que no necesita ser perfecto, sino apenas aproximadamente comprensible.

¿Qué tipo de datos son necesarios para desarrollar un sistema de traducción automática?

Todas las tecnologías modernas de desarrollo de traducción automática están basadas en datos, es decir, los equipos informáticos analizan grandes cantidades de datos de traducción acumulados para *aprender* a traducir de un idioma a otro. Estos datos lingüísticos que se utilizan para desarrollar sistemas de traducción automática se denominan datos de entrenamiento. La tecnología de traducción automática actual que se implementa más ampliamente es la traducción automática neuronal (NMT) y está reemplazando lentamente las muchas instalaciones de un enfoque anterior llamado traducción automática estadística (Stat MT o SMT). Ambos son enfoques para desarrollar sistemas de traducción automática utilizando los siguientes tipos de datos:

- Textos bilingües.
- Glosarios de traducción.
- Datos monolingües en el idioma de destino.
- Datos monolingües en el idioma de origen o en idiomas estrechamente relacionados**.**

Datos bilingües del idioma de origen y de destino

Las grandes colecciones de textos traducidos frase por frase se llaman corpus paralelos. Se puede crear un motor de traducción inicial con un mínimo de 100 000 segmentos (frases) traducidos bilingües. Un segmento de traducción puede ser una oración completa o un grupo de palabras que traducen términos y frases cruciales. Idealmente, debe haber al menos 1 000 000 de segmentos. Algunos sistemas de traducción automática están construidos con miles de millones de segmentos. En general, podemos decir que los volúmenes más grandes de segmentos bilingües de alta calidad generarán un resultado de traducción automática de mayor calidad.

Muchas comunidades lingüísticas no cuentan con grandes volúmenes de este tipo de datos. La fase de adquisición de datos suele requerir un esfuerzo concertado y a largo plazo, además de la colaboración entre las agencias gubernamentales, los establecimientos educativos y la comunidad en general. Mientras tanto, existe una tecnología que permite utilizar una cantidad menor de segmentos, con comentarios humanos que brindan mejoras incrementales a medida que se traducen los datos.

Los corpus que se utilizan como conjuntos de datos de entrenamiento para los algoritmos de traducción automática generalmente se extraen a partir de grandes cuerpos de fuentes similares, como bases de datos de artículos de noticias escritos en los idiomas de origen y de destino que describen eventos similares.

Sin embargo, los fragmentos extraídos pueden contener ruido, con elementos adicionales insertados en cada corpus. Las técnicas de extracción pueden diferenciar entre elementos bilingües representados en ambos corpus y elementos monolingües representados en un solo corpus para extraer fragmentos paralelos más limpios a partir de los elementos bilingües. Los corpus comparables se utilizan para obtener directamente el conocimiento para fines de traducción. Sin embargo, los datos paralelos de alta calidad son difíciles de obtener, especialmente en el caso de idiomas con pocos recursos.

Los datos de entrenamiento que se utilizan para crear un sistema de traducción automática suelen ser memorias de traducción (un archivo de traducciones anteriores) u otros recursos de traducción heredados que se han recopilado durante cierto tiempo. Esta

información definirá lo que el sistema de traducción automática aprenderá a traducir mejor. A menudo, existen límites en el volumen de datos disponibles. En tales casos, se deben hacer esfuerzos especiales para enseñar al sistema a aprender el material en el que es más probable que se concentre en traducir. Recuerde que aquello sobre lo que entrena será lo que su sistema traducirá mejor. Por lo tanto, un sistema de traducción automática que se utilizará para traducir contenido médico se entrenará mejor con memorias de traducción y glosarios médicos.

Cuando se proporcionan datos bilingües para entrenamiento, los datos deben estar **alineados**: la fuente y el destino deben ser traducciones directas entre sí. No es adecuado utilizar textos traducidos que sean resúmenes o comentarios del texto original. Los datos deben examinarse cuidadosamente antes de utilizarse para asegurarse de que serán útiles para los fines de entrenamiento.

Los glosarios y diccionarios de terminología clave con sus traducciones permiten una mayor precisión de traducción.

También merece la pena desarrollar una estrategia de metadatos integral a largo plazo para los datos lingüísticos que se recopilan. En las fases iniciales de la adquisición de datos, el enfoque suele estar en encontrar datos siempre que sea posible para abordar la necesidad de la masa de datos crítica necesaria para comenzar. Sin embargo, a medida que los motores de traducción automática maduran, puede haber beneficios de rendimiento significativos si se utiliza el tipo de datos adecuado para construir los motores. Por lo tanto, un sistema de traducción automática se puede optimizar para contenido relacionado con el dominio médico o para contenido relacionado con la tecnología informática, en lugar de tener un único sistema que lo haga todo. Esta especialización a menudo dará como resultado sistemas de traducción automática de mejor rendimiento.

Formatos de datos bilingües

Los datos bilingües tienen tres características importantes que los hacen ser útiles como datos de entrenamiento para traducción automática. Deben estar en un formato de archivo que los sistemas de traducción automática puedan importar. Los datos de texto

deben estar en codificación en caracteres Unicode UTF-8. Y los datos paralelos deben estar alineados con cada segmento del idioma de origen y emparejados con un segmento del idioma destino correspondiente.

Los datos se pueden entregar en los siguientes formatos de archivo y se enumeran en un orden de preferencia tentativo:

- Memoria de traducción (TMX, TBX, XLIFF, CSV): formato preferido.
- Texto sin formato (TXT).
- Contenido web (HTML).
- Estructurado (XML).
- Microsoft Office (DOC, DOCX, PPT, PPTX, XLS, XLSX).
- Formatos editoriales o DTP (TTX, PDF, FrameMaker).
- Reconocimiento Óptico de Caracteres (OCR), (TIFF, PNG, JPEG, etc.).

Los datos ya pueden estar alineados y emparejados entre el idioma de origen y el de destino o pueden entregarse en su forma original como, por ejemplo, documentos de Microsoft Word o HTML. Si los datos no están alineados, será necesario utilizar herramientas que permitan el alineamiento de datos con altos niveles de precisión.

Datos lingüísticos monolingües

Si bien las traducciones de archivos y el texto bilingüe son quizás los datos más importantes para crear un motor de traducción automática, también es imprescindible contar con datos monolingües de buena calidad en el idioma de destino. Estos datos se usan para aprender la estructura gramatical correcta durante la traducción e influyen estadísticamente en el resultado para leer en el estilo de escritura deseado. Esto es especialmente cierto cuando se construyen sistemas estadísticos de traducción automática.

Los datos monolingües son mucho más fáciles de adquirir que los datos bilingües. Puede ser útil recopilar direcciones URL de sitios web que tengan un dominio o estilo gramatical similar. Se pueden extraer sus datos lingüísticos y aprovechar el conocimiento lingüístico allí contenido.

Por lo general, es más difícil obtener datos monolingües para las lenguas indígenas. Existen muchas fuentes de datos monolingües para los idiomas más comunes. Sin embargo, un caso práctico en el que los datos lingüísticos monolingües pueden ser importantes es en la creación de un sistema de traducción automática para medicina. El rastreo de sitios web médicos y relacionados con este contenido específico puede identificar información crítica del idioma para construir glosarios y memorias de traducción.

Los datos de texto deben estar en codificación en caracteres Unicode UTF-8. Los datos se pueden entregar en los siguientes formatos y se enumeran en orden de preferencia tentativo:

- Texto sin formato (TXT).
- Memoria de traducción (TMX, TBX, XLIFF, CSV).
- URL de contenidos web.
- Contenido web (HTML).
- Estructurado (XML).
- Microsoft Office (DOC, DOCX, PPT, PPTX, XLS, XLSX).
- Formatos editoriales o DTP (TTX, PDF, FrameMaker)**.**

Apéndice D: Beneficios de la digitalización de una lengua

Esperamos que la serie de pautas *De cero a digital* ofrezca a todas las comunidades lingüísticas interesadas un camino claro hacia la digitalización si desean disfrutar de capacidades informáticas completas en sus idiomas nativos.

Los beneficios de representar un idioma en formato digital dependerán de los objetivos de su comunidad de hablantes. Estos pueden ser celebrar la belleza del idioma, mantener sistemas de conocimiento, difundir sus valores, crear aplicaciones y productos, divulgar relatos e historias, facilitar la administración ambiental y el liderazgo intelectual y la expansión del comercio, la educación, el empleo, el entretenimiento, la salud y la seguridad. La digitalización permite que una comunidad se beneficie de un conjunto en constante expansión de herramientas informáticas para el mantenimiento, la revitalización y la educación de la lengua. Una fuerte presencia digital en línea y, por lo tanto, una mayor visibilidad, pueden ayudar a influir en las políticas gubernamentales en apoyo de las comunidades indígenas e influenciar a las empresas hacia la inclusión. Dada la presencia generalizada de los teléfonos inteligentes entre los jóvenes, la digitalización puede ser una vía natural para involucrarlos con su lengua nativa. Una mayor exposición a través de plataformas digitales puede generar más oportunidades y demanda de hablantes nativos.

Cuando las comunidades lingüísticas ocupan su lugar en el escenario global a través de plataformas digitales, el mundo en general se beneficia. Sus experiencias, conocimientos y cosmovisiones únicas serán una contribución significativa para el resto del mundo y las sinergias resultantes pueden traer nuevas soluciones a los problemas globales. La digitalización facilita la conservación y publicación de esta información, haciendo que la amplitud y la naturaleza del lenguaje humano estén disponibles para el mundo y que se preserven para el beneficio de la humanidad en maneras que aún no podemos predecir.

En resumen, algunos de los beneficios obtenidos por la digitalización de las lenguas son:

- Permitir que los hablantes monolingües accedan, creen e intercambien fácilmente contenido en su lengua nativa con personas o grupos grandes, incluso a largas distancias.
- Aumentar el acceso a la información médica y de atención de la salud.
- Apoyar a las comunicaciones de emergencia y de desastre para salvar vidas.
- Ampliar el comercio local y el electrónico.
- Crear nuevas vías para divulgar arte, liderazgo intelectual y filosofías.
- Desarrollar materiales educativos en lengua nativa.
- Mejorar las relaciones y la comunicación con los vecinos.
- Mejorar la resolución de disputas.
- Facilitar el apoyo y el acceso a los procedimientos legales y gubernamentales en la lengua nativa.
- Aumentar el acceso a la información en Internet para la educación, el comercio y la participación, ya sea en la lengua nativa o en otras lenguas a medida que se desarrollan las herramientas de traducción.
- Conferir el reconocimiento del idioma, la cultura y la sabiduría nativas a los demás.
- Ampliar la función y la visibilidad de las comunidades indígenas a nivel mundial.
- Permitir que los grupos marginados o minorizados mantengan o revitalicen su idioma, aunque se vean absorbidos por los grupos dominantes.
- Preservar los sistemas nativos de conocimiento, cultura, historia, arte, medicina, sabiduría, valores y cosmovisión.