Low-resource Languages Toolkit

Versión 0.1

Alp Öktem, Col·lectivaT

20 de octubre de 2022

Índice

1. Introducción **3**
- 1.1. ¿Para quién es este documento? . . . 3
- 1.2. ¿Cómo puedo colaborar? . . . 4
- 1.3. Autores . . . 4

2. Los idiomas y la era digital **5**
- 2.1. Extinción digital . . . 5
- 2.2. Presencia online de un idioma . . . 6
- 2.3. Proyectos de documentación lingüística . . . 7
- 2.4. Tecnología «Tradicional» . . . 10

3. Tecnologías del lenguaje basadas en datos **17**
- 3.1. Traducción automática . . . 18
- 3.2. Reconocimiento automático del habla . . . 19
- 3.3. De texto a voz . . . 20

4. Datos lingüísticos **23**
- 4.1. Corpus de texto . . . 23
- 4.2. Datos paralelos (bitext) . . . 24
- 4.3. Corpus del habla . . . 26

5. Estudios de caso **29**
- 5.1. Judeoespañol: Conectando las dos orrillas del Mediterráneo . . . 29
- 5.2. Comunidades de base PLN . . . 32
- 5.3. Campañas de Common Voice . . . 33
- 5.4. Otras iniciativas . . . 33

6. Prácticas más idóneas **37**

7. Licencia **39**

Figura 1: Este proyecto está financiado por la Unión Europea.

En este documento cubrimos el papel de la tecnología, con un enfoque en la tecnología basada en la inteligencia artificial, en la preservación de lenguajes en peligro de extinción.

Este es un documento vivo que fue creado como parte del proyecto «Judeoespañol: Conectando las dos orrillas del Mediterráneo» llevado a cabo por Col·lectivaT y el Centro Sefardí de Estambul (SKAD) en el marco del «Programa de Subvenciones para el Patrimonio Cultural Común: Preservación y Diálogo entre Turquía y la UE-II (CCH-II)» implementado por el Ministerio de Cultura y Turismo de la República de Turquía con el apoyo financiero de la Unión Europea.

Nota: Este documento también está disponible en inglés y turco.

Bu doküman Türkçe de erişilebilir.

Este documento también está disponible en español.

Figura 2: Este proyecto está financiado por la Unión Europea.

CAPÍTULO 1

Introducción

Actualmente existe una creciente brecha digital entre los idiomas que cuentan con suficientes recursos y los idiomas con menos recursos, lo que agrava aún más el peligro de extinción digital para ellos. Para la mayoría de los idiomas, el proceso de generar herramientas y recursos útiles es mucho más fácil debido a su gran presencia en la web. Sin embargo, muchos idiomas minoritarios no tienen suficientes recursos materiales ni humanos para impulsar la creación de esas herramientas. La falta de apoyo estatal, de la visibilidad pública, y de la opresión social e institucional son las causas directas de que estos idiomas no tengan prioridad en los actuales espacios digitales.

Los esfuerzos sobre la preservación de idiomas se centran principalmente en la documentación lingüística, la enseñanza y la construcción de comunidad. Un área que se pasa por alto es la creación de herramientas basadas en inteligencia artificial. Herramientas como la traducción automática, la síntesis de voz y el reconocimiento de voz son ahora importantes contrapartes en la creación de interfaces humano-máquina. Además, estas herramientas pueden ayudar a modelar el conocimiento de los idiomas en peligro de extinción, y preservarlos para las generaciones futuras.

1.1 ¿Para quién es este documento?

Este documento es para ti si eres:

- *Activista lingüístico/a* que está interesado/a en ampliar herramientas y recursos en su idioma
- *Lingüista* que está interesado/a en recopilar datos para la investigación y la construcción de tecnología lingüística
- *Investigador/a de procesamiento de lenguaje natural (PLN)* que está interesado/a en aumentar los datos para su idioma de interés
- *Aliados/as de activistas de idiomas* que quieren apoyar la revitalización de los idiomas con recursos insuficientes

1.2 ¿Cómo puedo colaborar?

Este es un documento vivo con una licencia abierta (CC-BY). Su archivo fuente se comparte públicamente en https://github.com/CollectivaT-dev/language-toolkit, donde puedes extraer una versión para trabajar por tu cuenta, y luego enviar tu contribución. Puede variar desde corregir errores tipográficos hasta añadir una traducción, detallar una sección o explicar tu estudio de caso. Si tienes dudas no dudes en escribirnos a *info-arroba-collectivat.cat*.

1.3 Autores

- Alp Öktem - (Web, Twitter)

Figura 1: Este documento fue creado con el apoyo financiero de la Unión Europea. El contenido de este sitio web es responsabilidad exclusiva de Col·lectivaT y SKAD, y no refleja necesariamente las opiniones de la Unión Europea.

Figura 2: Este proyecto está financiado por la Unión Europea.

CAPÍTULO 2

Los idiomas y la era digital

La era digital ha traído muchas nuevas oportunidades y ventajas. Sin duda, ha conectado la humanidad de maneras inimaginables dentro de poco tiempo. No obtstante, ninguna innovación viene con sus desafíos y amenazas. World wide web es un recurso accesible (casi) a todas las personas, pero está dominado por solo unos cuantos idiomas.

El inglés, por ejemplo, es hablado por el 15 % del mundo, pero actualmente el 54 % de todo el contenido de la web es en inglés. Por otro lado, los idiomas como el ruso, el chino y el español representan entre el 5 - 6 % del contenido de la web, a pesar de su dominio geopolítico.

¿Dónde deja esto a los muchos idiomas del mundo, amenazados y en vías de extinción? Tristemente, los deja en los márgenes de esta imagen.

Los idiomas prosperan en comunidades y pasan a las generaciones con el uso diario. Cuanto más conectadas están nuestras vidas a través de los medios digitales, menos expuestos estamos a nuestras lenguas maternas que no están representadas en línea. Esto tarde o temprano conduce a una disminución de su uso por las generaciones más jóvenes.

Nota: ¿Cómo? Para ilustrar, un hablante kurdo en Turquía accede al sitio web de servicios de salud de su gobierno, y ve que todo está en turco. O va a ver la última plataforma de redes sociales en línea y ve que por defecto está disponible en inglés. Este tipo de pequeños encuentros llevan a pensar que para encontrar su camino y abordar sus necesidades, hablar su lengua materna no es suficiente. Tienes que conocer el idioma mayoritario más cercano, y muchas veces tienes que saber aún más idiomas.

2.1 Extinción digital

Según la UNESCO, se estima que unas 3.500 lenguas estén extintas a finales de este siglo y no podemos negar el papel de la tecnología en esto. Kornai afirma que la gran mayoría (más del 95 %) de los idiomas ya han perdido la capacidad de ascender digitalmente. El ascenso digital requiere el uso de un idioma en una amplia variedad de contextos digitales, desde el mantenimiento de una página de Wikipedia, hasta la disponibilidad de clases de idiomas, y la creación de datos de tecnología lingüística.

Por supuesto, no deberíamos caer en la trampa de hacer de la tecnología el culpable por la pérdida de un idioma. Es simplemente una representación de las dinámicas de poder ya existentes en la sociedad. Los Estados que privan o incluso

oprimen ciertos idiomas comienzan su transformación digital excluyendo todos estos idiomas en su infraestructura digital. Ahora, las grandes empresas tecnológicas norteamericanas y eurocéntricas siguen un enfoque de "primero-inglés" por defecto.

La tecnología también se puede utilizar para formar comunidades en torno a la preservación del idioma, el intercambio de conocimientos y la documentación lingüística.

2.2 Presencia online de un idioma

Tradicionalmente, las responsabilidades de una persona activista lingüística eran hablar el idioma activamente, transmitirlo a las generaciones más jóvenes, formar comunidades de aprendizaje y habla del idioma, negociar con instituciones públicas para la inclusión de su idioma, colaborar con lingüistas para la documentación de su idioma, etc.

Hoy en día, el reto no es sólo hacer que el idioma viva en el mundo físico, sino también en línea. Se relaciona de dos maneras para la supervivencia de un idioma:

1. Los intercambios y la visibilidad en línea despiertan interés, y ayudan a involucrar a las existentes o nuevas estudiantes de idiomas.
2. Lo que se almacena en línea a su vez es un registro digital para el idioma que ayuda a la documentación y el desarrollo tecnológico.

A continuación, describimos algunas formas en que Internet se está haciendo multilingüe y plural, al tiempo que ayuda a reanimar idiomas en peligro de extinción.

2.2.1 El libre acceso al conocimiento

Una de las iniciativas más populares que traen los idiomas en línea es Wikipedia. Wikipedia es una enciclopedia en línea abierta, redactada y mantenida por una comunidad de voluntarios/as a través de un sistema abierto de colaboración y revisión.

El objetivo general de Wikipedia es democratizar el acceso al conocimiento. Naturalmente, la cultura construida en torno a este ethos va de la mano con la multilingüidad. Aunque comenzó solo con el inglés, se expandió rápidamente a muchos idiomas del mundo. Probablemente podemos decir que es la plataforma lingüísticamente más diversa en Internet, con 326 idiomas (al día 03.05.22) y sigue aumentando.

Nota: La primera edición de un artículo de Wikipedia no inglés fue en catalán el 16 de marzo de 2001. Actualmente, este idioma destaca por su gran cantidad de artículos de calidad, que ilustra la importante presencia en línea de la lengua catalana a pesar de ser una lengua minoritaria. La Wikipedia en catalán es actualmente la 20ª Wikipedia más grande.

Hacer que un nuevo idioma esté disponible en Wikipedia no es una tarea fácil (como se explica aquí), pero definitivamente es una gran manera de hacer que el conocimiento sea accesible en línea y construir una comunidad virtual a su alrededor.

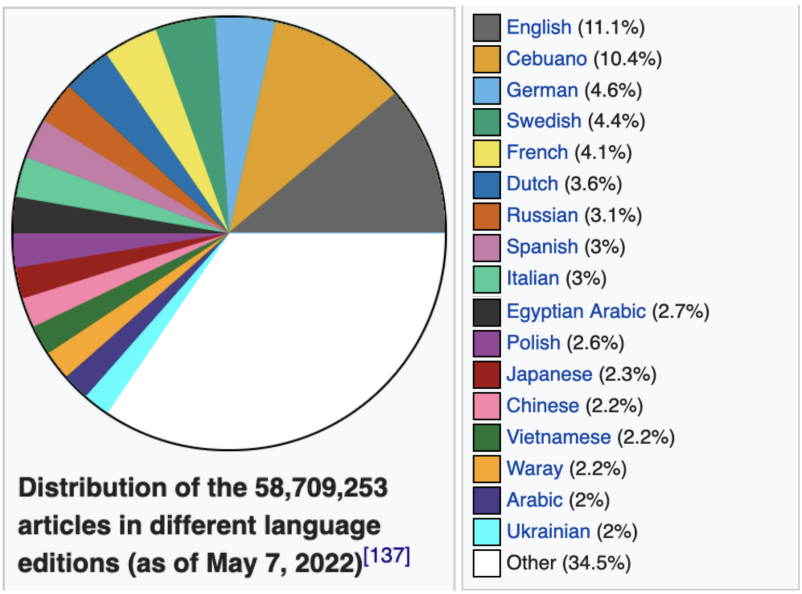

Figura 1: Tamaños relativos de diferentes wikipedias (fuente)

2.2.2 Inmersión en el idioma

Un gran ejemplo de revitalización del idioma con la ayuda de la tecnología es el yiddish. Después del holocausto, el número de hablantes de yiddish disminuyó drásticamente de alrededor de 10 millones de hablantes. Quienes sobrevivieron se vieron obligados a adoptar el idioma de sus nuevas tierras para evitar la persecución. En el siglo pasado, el uso del yiddish casi había desaparecido, excepto para las pequeñas y dispersas comunidades jasídicas.

Con el auge de Internet y la popularidad de los foros en línea, los hablantes de yiddish utilizaron estas plataformas para conversar en su idioma. Con el tiempo, el mundo virtual se convirtió en el principal punto de encuentro para los hablantes de yiddish en foros como The Idishe Velt (El mundo judío) y Kave Shtiebel (La cafetería).

2.3 Proyectos de documentación lingüística

La rápida pérdida de idiomas en el siglo pasado ha impulsado muchas iniciativas para la documentación lingüística y la revitalización. Una de estas iniciativas es The Endangered Languages Project, una plataforma basada en la web que actúa como un centro de colaboración de entusiastas de los idiomas, lingüistas, y socios de la industria para ayudar a fortalecer los idiomas en peligro de extinción. Los y las usuarias del sitio web actúan como colaboradores al subir muestras de idioma en formatos de texto, audio, enlace o video utilizando un sistema de geoetiquetado único que permite una búsqueda fácil.

Del mismo modo, Wikitongues, que comenzó en 2014, recopila grabaciones y recursos de los idiomas del mundo. Actualmente tiene vídeos en más de 700 idiomas, lexicones en 200 idiomas y enlaces a cientos de recursos externos.

Figura 2: Ladino en el proyecto Endangered Languages Project

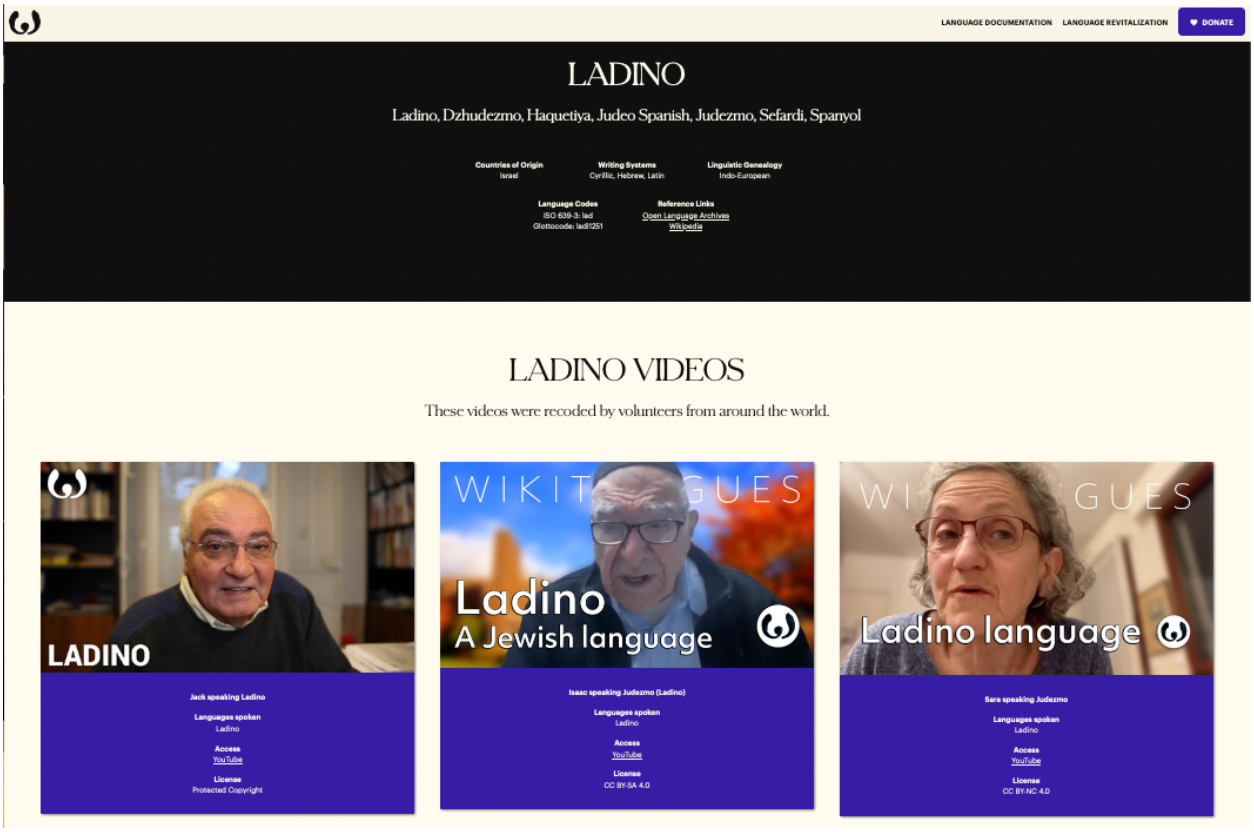

Figura 3: Ladino vídeos en Wikitongues

2.4 Tecnología «Tradicional»

Preservar un idioma no es solo una cuestión de registrar palabras o frases y digitalizarlas para que se guarden en un sótano en línea. El idioma es inherentemente acerca de las personas, la cultura y la identidad. Para mantener viva una lengua, ésta tiene que ser hablada por mucha gente, inmersa en la cultura cotidiana y transmitida activamente a las generaciones futuras. En estos días, internet, las redes sociales, el software, y las plataformas ocupan un gran espacio en nuestra vida cotidiana. En esta sección, enumeraremos algunas de las herramientas básicas necesarias para que una tecnología prospere digitalmente.

2.4.1 Fuente compatible con Unicode

Una fuente digital es la forma en que los ordenadores saben cómo mostrar los caracteres en tu idioma. Unicode, formalmente el Unicode Standard, es una tecnología informática estándar para la codificación, representación y manejo consistentes del texto expresado en la mayoría de los sistemas de escritura del mundo. El estándar, mantenido por el Consorcio Unicode, define 144.697 caracteres que cubren 159 scripts modernos e históricos, así como símbolos, emojis y códigos de control y formato no visuales.

Para comprobar si una fuente admite tu idioma, accede a Google Noto y busca entre fuentes que representan más de 500 sistemas de escritura. Si no está allí, puedes crear su propia fuente con la ayuda de un/a diseñador/a de fuentes e instalarla manualmente en tu ordenador.

2.4.2 Teclado

Hasta el día en que hablemos de forma natural con los ordenadores, la interfaz más común para interactuar con ellos será el teclado. Es una tecnología que fácilmente se da por descontada para muchos idiomas del mundo, pero desafortunadamente no está disponible en todos los sistemas de escritura del mundo. Si un teclado no existe o no está suficientemente desarrollado para un idioma, sus hablantes tienden a preferir otros alfabetos o incluso idiomas para comunicarse. Por ejemplo, los y las hablantes de idiomas etíopes como el amhárico, el tigrinya y el oromo cambian a usar el inglés porque el script ge'ez no está preinstalado en sus móviles inteligentes. Los jóvenes hablantes de árabe en muchos países han inventado su propio alfabeto de chat Arabizi que consiste en números y caracteres latinos para responder por la falta de soporte de escriptura árabe en la tecnología móvil y web temprana.

Si un teclado no está disponible, por ejemplo en tu teléfono u ordenador, aquí hay algunos recursos para buscar o ayudar a crear tu propio teclado:

- Teclado móvil de Google, Gboard
- Keyman soporta 2000 idiomas
- Microsoft Keyboard Layout Creator
- Ukelele es un editor de diseño de teclado Unicode para MacOS

2.4.3 Diccionario en línea

Un diccionario, o léxico, es una forma sólida de documentar un lenguaje, ya que actúa como una referencia de las palabras y sus significados. Un diccionario en línea nunca se agota, ya que es accesible desde cualquier dispositivo con conexión a Internet. Además, los diccionarios de código abierto pueden vivir y crecer en colaboración como un esfuerzo comunitario que involucra tanto a hablantes de la lengua, lingüistas y tecnólogos.

Living Dictionaries es una plataforma de creación de diccionarios en línea creada por Living Tongues Institute for Endangered Languages. Proporciona herramientas tecnológicas en línea completas y gratuitas que ayudan a las comunidades lingüísticas en los esfuerzos de conservación y revitalización. También permite el registro de palabras y frases.

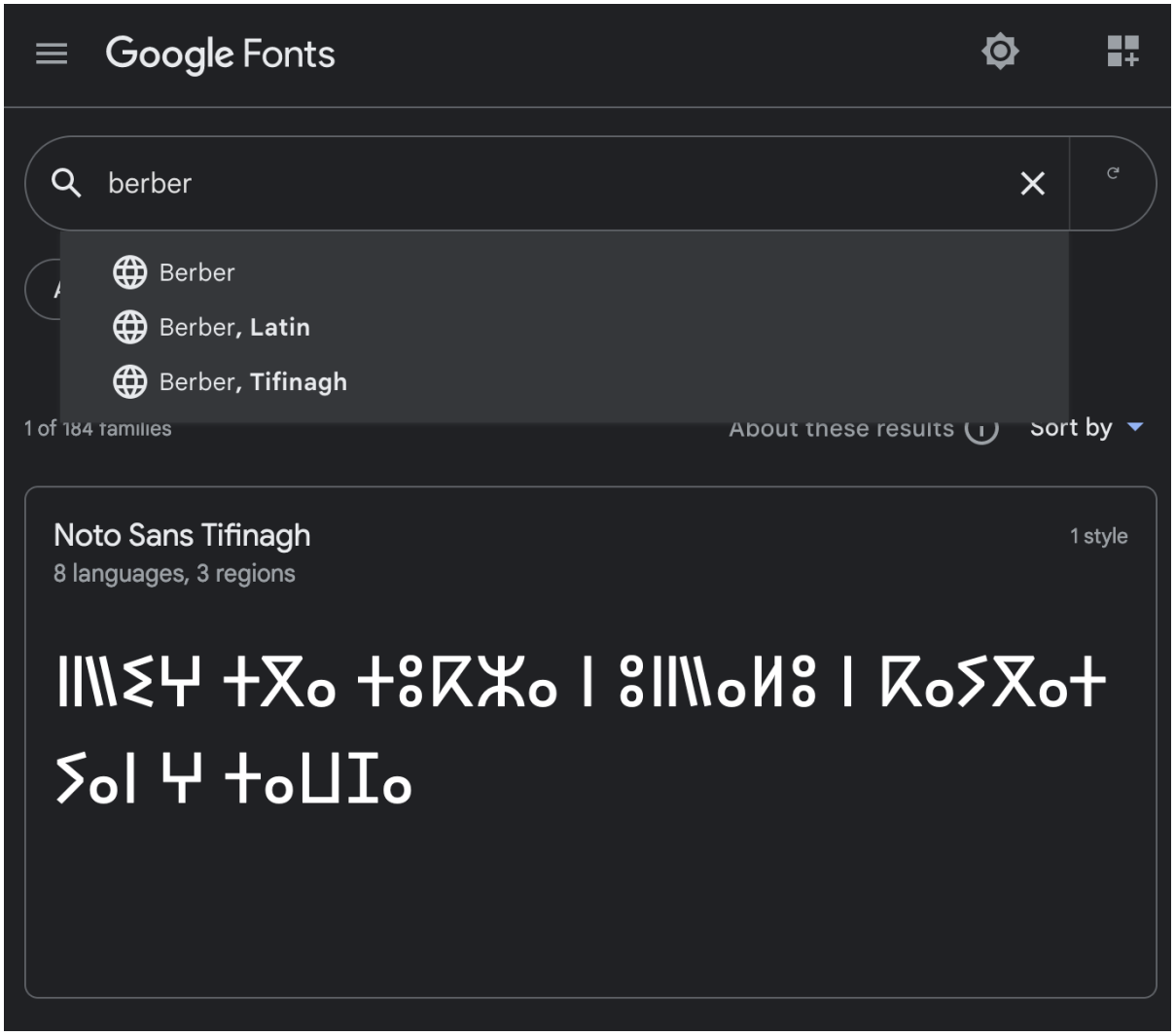

Figura 4: Alfabeto tifinagh de bereber marroquí en fuentes de Google

A partir de mayo de 2022, soporta 237 idiomas. Para iniciar su idioma en los diccionarios vivos, puede utilizar sus Lista de elicitación y ver los tutoriales en su canal de YouTube.

SIL Dictionary App Builder "le ayuda a crear aplicaciones de diccionario personalizadas para teléfonos inteligentes y tabletas Android e iOS. Especifique el archivo de datos de léxico que desea usar, el nombre de la aplicación, las fuentes, los colores, el audio, las ilustraciones y los iconos. Dictionary App Builder empaquetará todo y creará la aplicación personalizada para ti. A continuación, puede instalarlo en su teléfono, enviarlo a otros por Bluetooth, compartirlo en tarjetas de memoria microSD y publicarlo en tiendas de aplicaciones en Internet».

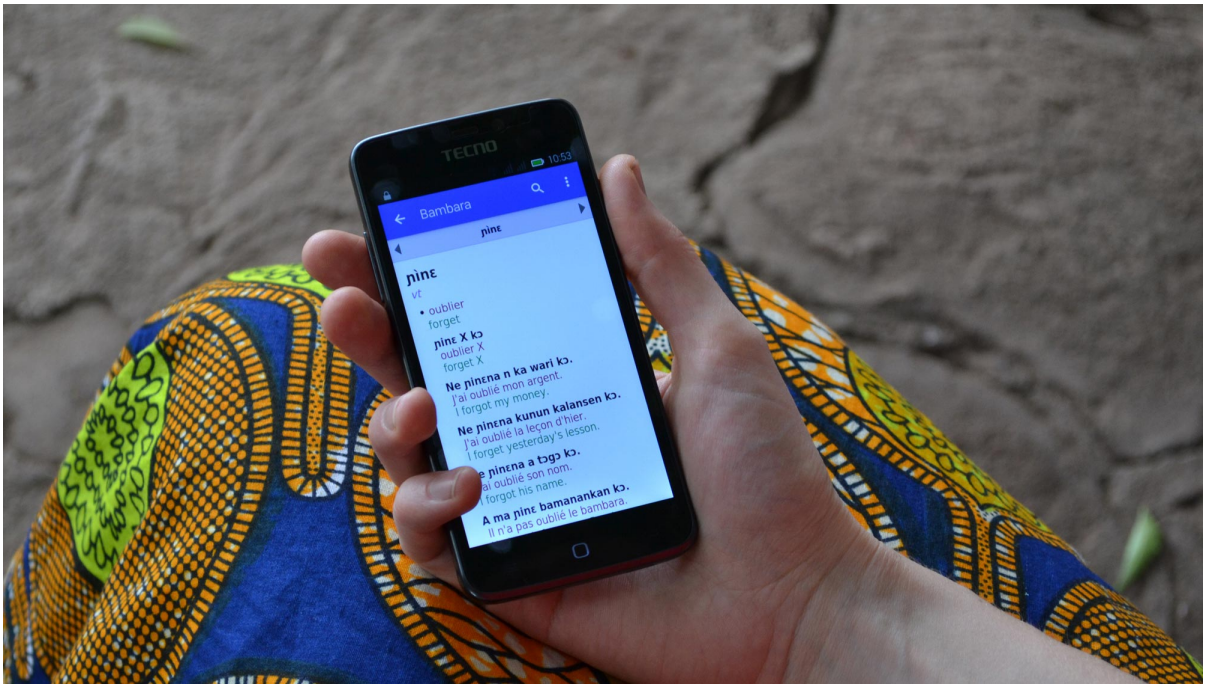

Figura 5: Una mujer usando el diccionario bambara en su teléfono móvil (crédito de la imagen SIL International)

2.4.4 Aplicaciones de aprendizaje de idiomas

La disponibilidad de plataformas educativas en línea ha revolucionado la forma en que muchas personas abordan el aprendizaje de idiomas en la actualidad. A pesar de que no sustituyen a un profesor, complementan las clases tradicionales o son la única opción en los contextos de algunos idiomas. También ofrecen muchas ventajas, como permitir que las personas aprendan en cualquier dispositivo (móvil o de escritorio), a su propio ritmo y horario. Estas aplicaciones sirven clases de idiomas y ejercicios en sprints cortos, divertidos y digeribles, permiten a los estudiantes hacer un seguimiento de su progreso e incluso chatear o contratar tutores de idiomas en espacios comunitarios en línea.

Muchos de los grupos amenazados, minoritarios y de escasos recursos del mundo aún no tienen una presencia significativa en línea ni documentación lingüística adecuada para crear cursos en línea. Sin embargo, hay gracias al impulso de las comunidades lingüísticas y una mayor sensibilidad para el aprendizaje de idiomas indígenas en todo el mundo, un mayor interés por parte de las empresas que desarrollan estas aplicaciones para invertir en idiomas minoritarios y en peligro de extinción. Idiomas como maorí, gaélico escocés, hawaiano, quechua, navajo y lakota se están abriendo paso en plataformas educativas conocidas como Duolingo, Babbel y uTalk.

Podemos categorizar estas plataformas de cuatro maneras:

- **Basado en módulos:** El uso de estas aplicaciones se siente más o menos como tomar una clase en una escuela o una universidad, donde los usuarios siguen un plan de estudios modular planificado por los educadores. Permite a los alumnos hacer un seguimiento de su progreso, recibir notificaciones y ganar puntos. Algunos ejemplos

notables son: Duolingo , Babbel, uTalk , Master Any Language. Desafortunadamente, no es posible que las propias comunidades lingüísticas decidan e implementen un nuevo idioma en estas plataformas. Sin embargo, es posible «presionar» y participar en la creación de nuevos módulos a través de las comunidades que algunas de estas plataformas proporcionan. Además, hay que señalar que la mayoría de estas plataformas son plataformas con fines de lucro y el trabajo de las comunidades lingüísticas sigue sin ser compensado.

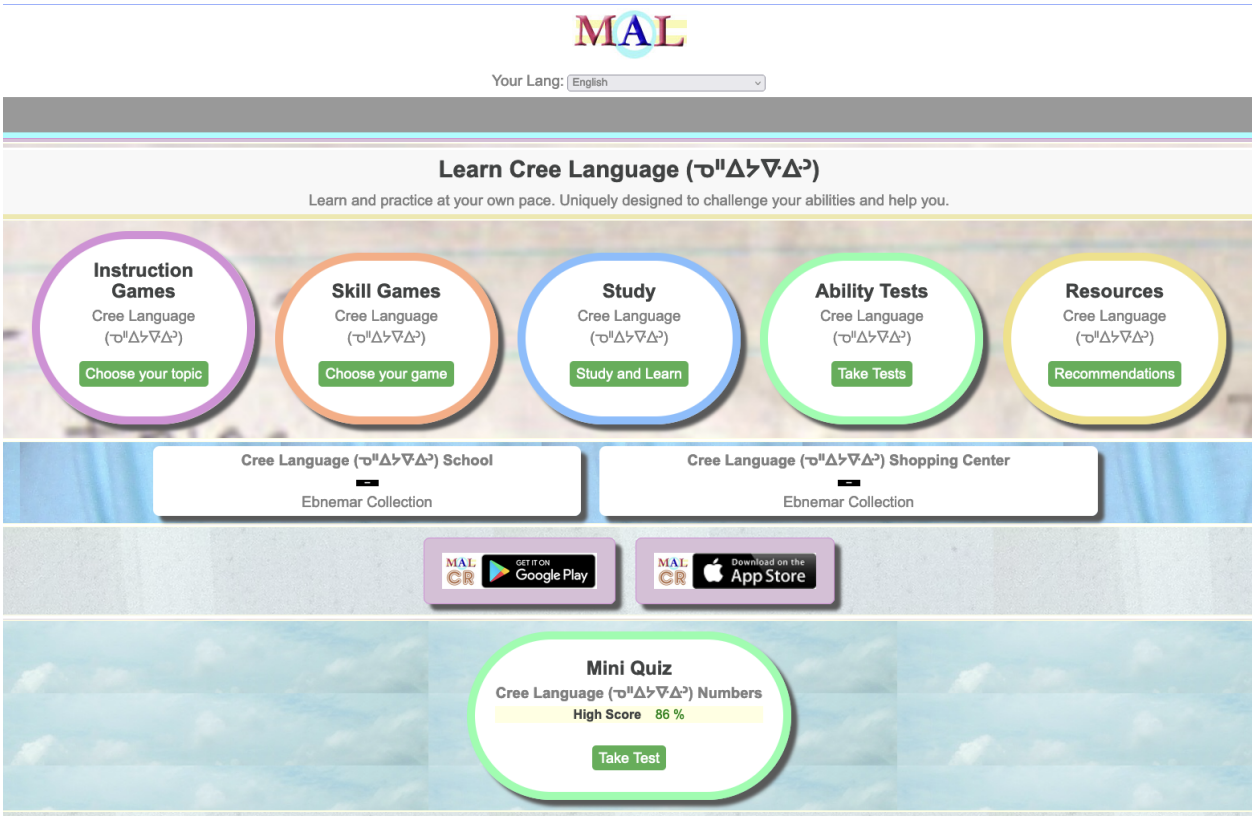

Figura 6: Contenido para aprender el idioma cree en la plataforma Master Any Language

- **Basado en juegos:** Estas aplicaciones alimentan al alumno con pares de «preguntas y respuestas» y pueden evaluar qué tan bien la persona memoriza el emparejamiento a lo largo del tiempo. Este método es popular en línea para los estudiantes visuales y auditivos de idiomas, y también para muchas otras áreas de estudio como las matemáticas y las ciencias. Memrise puede ser el más atractivo para las lenguas indígenas y otras lenguas minoritarias con recursos insuficientes. Se trata de una plataforma educativa online que utiliza técnicas de memoria como flashcards, SRS (software de repetición espaciada) y ayudas visuales para optimizar el aprendizaje de idiomas. Memrise tiene contenido disponible para más de 170 idiomas, un número mucho mayor que la mayoría de las otras plataformas de aprendizaje de idiomas. En particular, el sitio tiene una impresionante selección de contenido creado por la comunidad para lenguas y dialectos indígenas de todo el mundo, como Yup'ik (ver la primera captura de pantalla a continuación), Cherokee (ver la segunda captura de pantalla a continuación), Algonquian, Alutiiq, Choctaw, Groenlandia, Inuktitut, Lakota, Nahuatl, Yucatec Maya, K' iche', Quechua, Guaraní, Ainu, Jeju y muchas otras lenguas de tamaño mediano habladas en Europa, África, Oriente Medio, Asia y el Pacífico. La plataforma Memrise tiene un sentimiento de bricolaje, democrático y popular. La plataforma es innovadora porque: 1) los usuarios pueden seguir los cursos y conjuntos de tarjetas existentes y cargar sin problemas sus propias ayudas mnemotécnicas para ayudar a recordar palabras y frases a medida que avanzan a través de un curso; 2) el sitio proporciona una forma atractiva para que los usuarios se conecten con el contenido del idioma a través de la repetición, pequeños cuestionarios, videos cortos, imágenes divertidas y grabaciones hechas por hablantes fluidos, y 3) la plataforma permite a los usuarios de la comunidad crear fácilmente sus propios cursos de idiomas que otros también pueden usar. Las plataformas similares incluyen: AnkiApp, Language Drops <https://languagedrops.com/languages >`_y `MosaLingua.

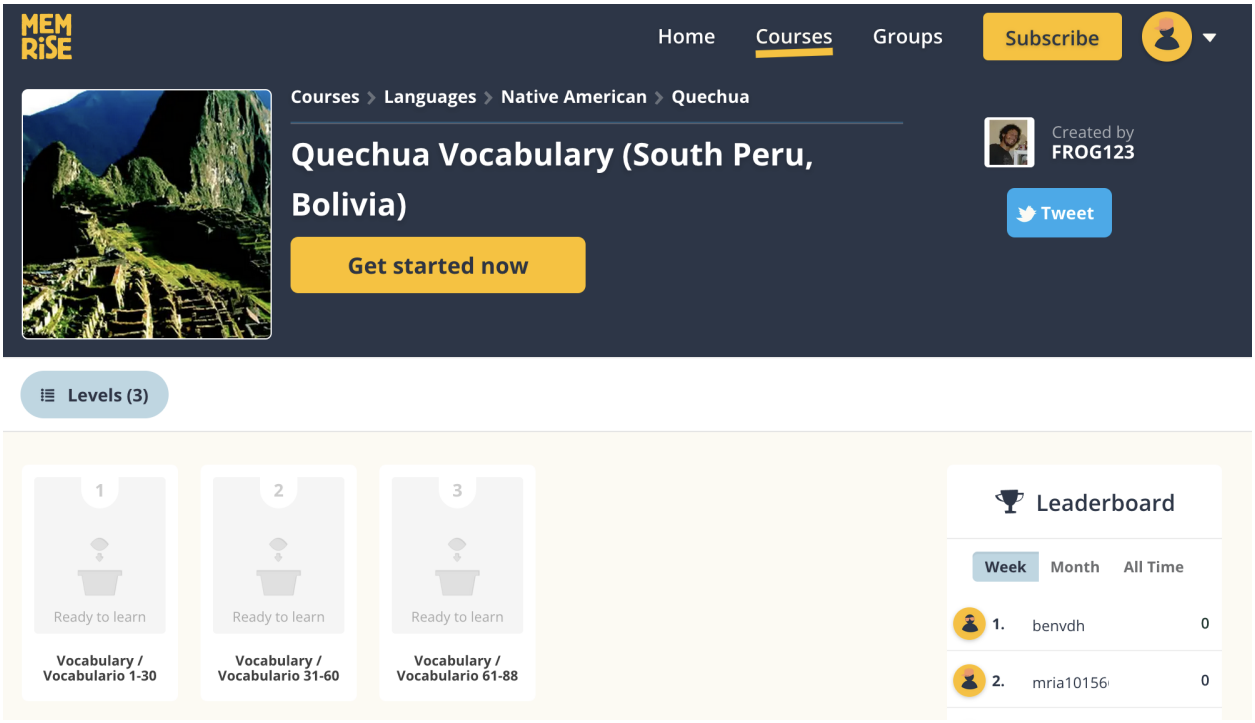

Figura 7: Ejercicios de aprendizaje de quechua creados por la comunidad en Memrise

- **Basado en chat:** Estas aplicaciones permiten a los alumnos conectarse con hablantes del idioma que les interesa a través de un chat interactivo en vivo. Esto proporciona un entorno social y libre de estrés para los alumnos. Algunos ejemplos como HiNative y HelloTalk han aumentado recientemente en popularidad, especialmente en países asiáticos.
- **Plataformas de tutoría estudiantil en línea:** Para aquellos estudiantes que prefieren la relación clásica profesor-alumno pero no tienen acceso a los maestros en su entorno, plataformas como iTalki y Verbling ayudan a configurar clases en línea. Esto también contribuye directamente a la comunidad lingüística, ya que genera ingresos directos para los profesores.

2.4.5 Fuentes

- Wikipedia catalana en Wikipedia
- Cómo la tecnología puede ahorrar idiomas que mueren rápidamente
- Unicode en Wikipedia
- Language sustainability toolkit
- Learn Language Online by Living Tongues
- Aprendizaje asistido por ordenador en Wikipedia

Figura 8: Este documento fue creado con el apoyo financiero de la Unión Europea. El contenido de este sitio web es responsabilidad exclusiva de Col·lectivaT y SKAD, y no refleja necesariamente las opiniones de la Unión Europea.

Figura 9: Este proyecto está financiado por la Unión Europea.

CAPÍTULO 3

Tecnologías del lenguaje basadas en datos

La revolución digital está aquí con nosotros y la Inteligencia Artificial (IA) es un facilitador tecnológico clave. Ofrece una serie de nuevas oportunidades para derribar las barreras existentes al desarrollo humano y la inclusión social. Un área impulsada por la IA es la tecnología del lenguaje que permite interactuar con nuestros teléfonos a través de asistentes digitales, traducir sitios web y documentos con unos pocos clics, aumentar la accesibilidad de videos con subtítulos automáticos, etc.

El motor principal detrás de estos es el avance del campo **Procesamiento del Lenguaje Natural (PLN)**. Pero, ¿qué implica la PLN? Aquí hay una lista de tecnologías básicas que caen en el área de este campo:

Basado en texto:

- Traducción automática
- Recuperación de información
- Extracción de la información
- Análisis de sentimiento
- Responder a la pregunta
- Resumen de texto
- Reconocimiento de nombres de entidades

Basado en el habla:

- Reconocimiento automático del habla
- Síntesis de texto a voz

El aspecto revolucionario de estas tecnologías es que están *basadas en datos*, lo que significa que la inteligencia que se crea con estas herramientas se recopila a partir de grandes volúmenes de información, o simplemente *datos*. Por ejemplo, en el caso de la traducción automática, el motor «modela» la traducción de un idioma al otro mirando una colección de documentos y oraciones traducidos por humanos. Del mismo modo, un *análisis de los sentimientos* aprende a etiquetar si un tweet dice bueno o malo sobre una empresa a partir de miles de tweets etiquetados por los humanos como portadores de un sentimiento bueno o malo.

Esta dependencia de los datos es lo que hace que estas tecnologías sean accesibles a algunos idiomas y no a los demás. Los recursos disponibles para un idioma influyen directamente en la posibilidad de desarrollar una aplicación para un idioma. Como el mayor recurso de datos textuales es Internet, y está dominado por unos cuantos idiomas, estas tecnologías tienden a centrarse solo en un puñado de *idiomas dominantes*, por ejemplo, inglés, español, chino, árabe, etc.

El siguiente diagrama de Microsoft Research Labs India ilustra la jerarquía creada por esta «ley del poder» entre los idiomas.

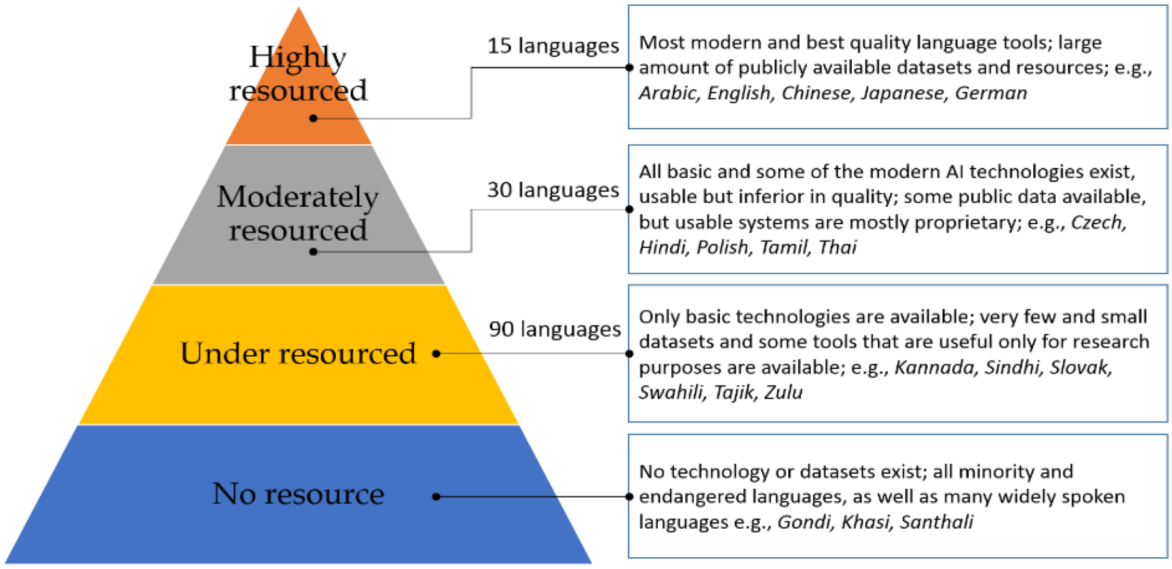

Figura 1: Clasificación de idiomas según la disponibilidad de tecnología, herramientas y recursos lingüísticos

3.1 Traducción automática

La traducción automática (MT) se define como la conversión automática de una secuencia de símbolos en un idioma a una secuencia de símbolos en otro idioma. Ha evolucionado a lo largo de los años, pasando de enfoques basados en normas a enfoques estadísticos, que modelaban las probabilidades de mapeo entre subfrases entre traducciones. Estas probabilidades se aprenden de manera estadística a partir de textos paralelos donde las traducciones alineadas con oraciones están disponibles en los idiomas involucrados (referidos como idiomas de origen y destino). El siguiente diagrama ilustra el modelado de la traducción de la palabra «seguro» en inglés a español utilizando traducciones hechas en el Parlamento de la ONU.

Los servicios de traducción automática como Google Translate y DeepL se han abierto camino en herramientas confiables para traductores y también gente regular en los últimos años. Esto se debe en gran parte al avance de las técnicas de *aprendizaje profundo* que revolucionaron el campo de la inteligencia artificial. Esta nueva forma de modelar introducida en 2014 cometió un 50 % menos de errores de orden de palabras, un 17 % menos de errores léxicos y un 19 % menos de errores gramaticales en comparación con los modelos anteriores.

Los usos de la traducción automática son los siguientes:

1. **Asimilación**, emulando un determinado documento en otro idioma. Este caso de uso permite, por ejemplo, leer un sitio de noticias o un documento técnico en un idioma que no entendemos. Sabemos que no es una traducción 100 % precisa, pero da la esencia para explorar más.
2. **Comunicación**, que permite la comunicación entre individuos y organizaciones, por ejemplo, en chat, turismo, comercio electrónico, reduciendo la necesidad de una lingua franca.

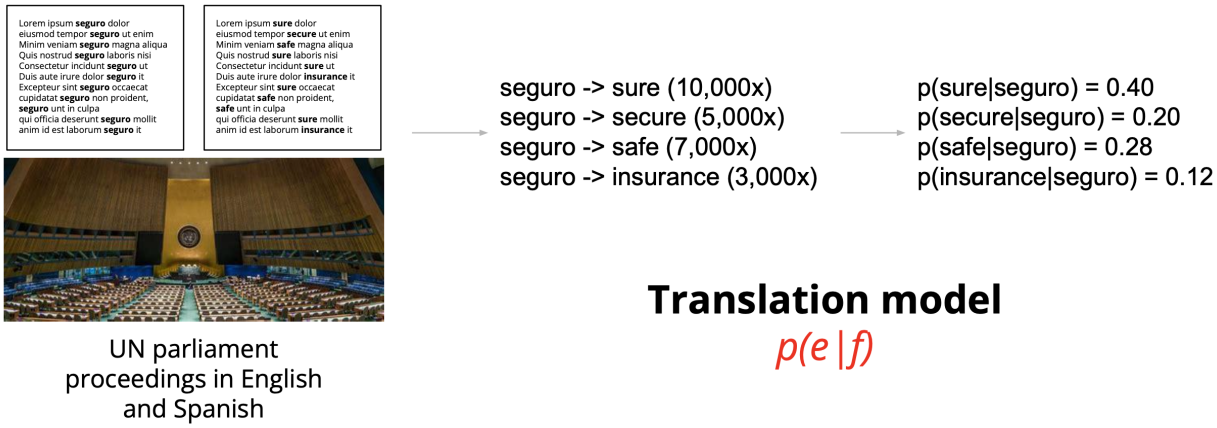

Figura 2: Extracción de estadísticas de traducciones de datos paralelos

3. **Monitoreo**, que permite el rastreo de información en documentos multilingües a gran escala, por ejemplo, descubriendo tendencias internacionales en Twitter.
4. **Asistencia**, mejorando los flujos de trabajo de traducción, por ejemplo, traducción asistida por ordenador, posedición.

La TA también se ha convertido en una herramienta esencial en el aprendizaje de idiomas. Un trabajo reciente de Duke University estudia su uso entre los estudiantes de idiomas de nivel universitario junto con otras herramientas clásicas como diccionarios y tesauros. Informan que el 76 % de los estudiantes matriculados en una clase de español utilizan herramientas de MT basadas en la web como Google Translate durante sus estudios.

Finalmente, Bird y Chiang también han propuesto la traducción automática como una herramienta de documentación y preservación para idiomas en peligro de extinción en su documento Traducción automática para la preservación del idioma. Citando directamente de su artículo: «… cuando los textos originales se traducen a un idioma mundial importante, garantizamos que la documentación del idioma será interpretable incluso después de que el idioma haya caído fuera de uso. En segundo lugar, cuando un orador sobreviviente puede identificar errores en la salida de un sistema de TA, tenemos evidencia oportuna de aquellas áreas de gramática y léxico que necesitan una mejor cobertura mientras todavía hay tiempo para recopilar más. Estas tareas de producción y corrección de traducciones pueden ser realizadas por hablantes del idioma sin depender de la intervención de lingüistas externos. Además, eludimos la necesidad de recursos lingüísticos como bancos de árboles y redes de palabras, que son caros de crear y que dependen de la existencia de análisis morfológicos, sintácticos y semánticos de la lengua «.

Esta forma innovadora de documentación del lenguaje reduce el esfuerzo en la recopilación de oraciones traducidas, ya que el desarrollo de MT se basa en este tipo de datos. *(Más información sobre los datos paralelos en la siguiente sección)*

3.2 Reconocimiento automático del habla

El reconocimiento automático de voz (ASR) es la conversión del habla en su forma acústica en una forma simbólica como palabras o letras. Es el modelado probabilístico de la pregunta «¿Cuál es la secuencia de palabras más probable entre todas las secuencias de palabras posibles dada una entrada acústica?». El siguiente diagrama ilustra este proceso. La señal de voz capturada por un micrófono se codifica primero en una secuencia de vectores de características acústicas. A continuación, los vectores de características acústicas se decodifican en las palabras que representan la información lingüística que se encuentra en la señal de voz.

El desarrollo de un sistema automático de reconocimiento de voz para un idioma depende del siguiente tipo de datos:

Figura 3: Un diagrama básico de reconocimiento automático de voz

1. Colección de grabaciones cortas de audio de voz de muchos locutores y sus transcripciones de texto
2. Un gran corpus de texto
3. Diccionario de pronunciación fonética (opcional en tecnologías más modernas)

La ASR ha progresado significativamente en la última década nuevamente gracias al advenimiento del aprendizaje profundo. En septiembre de 2017, Microsoft anunció sus resultados para un sistema de reconocimiento de voz en inglés que podría lograr un rendimiento mejor que el humano en la transcripción del habla. Su sistema se basó en un conjunto de datos de 200 MILLONES de palabras transcritas del discurso conversacional. Estos desarrollos ya han tenido un gran impacto ya que los asistentes virtuales se han convertido en una aplicación cotidiana, búsqueda por voz y transcripción automática de audio.

3.3 De texto a voz

Texto a voz (o síntesis de voz) implica la producción de un discurso similar al humano dada una entrada de texto con un ordenador. Antes del advenimiento del aprendizaje profundo, había dos enfoques principales para la síntesis de texto a voz (TTS): TTS concatenativo y TTS paramétrico. Concatenative TTS, también llamada selección de unidades, combina cortos clips de audio pregrabados llamados unidades para sintetizar el texto deseado. Concatenative TTS puede proporcionar un buen rendimiento en términos de calidad del habla, pero el procedimiento de corte y puntada significa una falta de naturalidad. El TTS paramétrico se basa en métodos estadísticos mediante la generación de voz con una combinación de parámetros como F0 y energía, modelando la producción de voz humana.

Actualmente, la mayoría de los sistemas TTS modernos se basan en métodos de aprendizaje profundo. Las redes neuronales profundas se entrenan utilizando una gran cantidad de voz grabada y las transcripciones de texto asociadas. A diferencia de los datos de capacitación de ASR, generalmente se recopilan de un solo orador. El sistema TTS resultante es capaz de replicar la voz de este orador en particular.

TTS es importante para hacer que los ordenadores sean accesibles a las personas ciegas o con visión parcial, ya que les permite «leer» desde la pantalla. La tecnología TTS se puede vincular a cualquier entrada escrita en una variedad de idiomas, por ejemplo, pronunciación automática de palabras de un diccionario en línea, lectura en voz alta de un texto, interfaz para un asistente de voz, etc.

En el caso de lenguas en peligro y minoritarias, TTS puede ayudar al aprendizaje de idiomas y la documentación

lingüística. Los estudiantes que no tienen acceso a oradores pueden estudiar cómo se pronuncia una oración sin la ayuda de un tutor. Es un registro permanente de la lengua, ya que persistirá incluso después del momento en que no queden hablantes para la lengua.


Figura 4: Este documento fue creado con el apoyo financiero de la Unión Europea. El contenido de este sitio web es responsabilidad exclusiva de Col·lectivaT y SKAD, y no refleja necesariamente las opiniones de la Unión Europea.

Figura 5: Este proyecto está financiado por la Unión Europea.

CAPÍTULO 4

Datos lingüísticos

Las herramientas de inteligencia artificial abren una nueva área para la creación de recursos lingüísticos para las lenguas minoritarias y en peligro de extinción. En comparación con los recursos lingüísticos «clásicos» creados para preservar idiomas como la léxica, la documentación gramatical, los mapas lingüísticos, etc., estos requieren menos conocimientos lingüísticos, pero por lo general solo son útiles en grandes volúmenes.

En esta sección explicaremos los tipos de datos que impulsan la creación de tecnología del lenguaje basada en inteligencia artificial explicada en el capítulo anterior. Además, cubriremos algunas formas de recopilarlos y obtener el máximo provecho de ellos, incluso si no son de grandes volúmenes que suelen requerir estas aplicaciones.

4.1 Corpus de texto

En lingüística, un corpus (corpus plural) o corpus de texto es un recurso lingüístico que consiste en un conjunto grande y estructurado de textos en un idioma en formato digital. Son útiles en lingüística de corpus para hacer análisis estadísticos y pruebas de hipótesis, verificar ocurrencias o validar reglas lingüísticas dentro de un territorio lingüístico específico. Para la tecnología del lenguaje, son una parte esencial en la creación de modelos de lenguaje estadísticos que se utilizan en aplicaciones como el reconocimiento óptico de caracteres, el reconocimiento de escritura a mano, la traducción automática, la corrección ortográfica y la escritura asistida.

Los corpus de texto por sí mismos son del tipo * datos sin etiquetar *. Es decir, son una mera recopilación de datos (en este caso texto) sin anotaciones ni etiquetado. Los modelos de lenguaje almacenan las probabilidades de secuencias de palabras para obtener una «comprensión» del lenguaje. Además, los corpus de texto se pueden anotar con la siguiente información para crear * datos etiquetados * para diferentes tareas de PLN:

- **Parte del discurso** (sustantivo, verbo, adjetivo, etc.)
- **Entidades nombradas** (Persona, ubicación, información de identificación personal, organización, hora, etc.)
- **Lemas** (raíces de las palabras, por ejemplo, venir por venga)
- **Estructura de dependencia y frase** (árbol sintáctico)

4.1.1 Obtención de corpus de texto

La forma más común de obtener corpus de texto es a través de *rastreo* de la World Wide Web. Esta técnica analiza toda la web para recopilar texto en un idioma determinado o en muchos idiomas a la vez. Wikipedia publica su contenido en diferentes idiomas que se pueden utilizar para crear corpus de texto. Common Crawl initiative recopila datos del sitio web y proporciona libremente petabytes de datos. OSCAR distribuye estos datos clasificados en 166 idiomas.

Otro recurso común utilizado por los idiomas ingeniosos son los libros. BookCorpus consta de 11.038 libros de la web que contienen 74 millones de frases y 984 millones de palabras y se sabe que ha impulsado muchos modelos de lenguaje influyentes por las grandes empresas tecnológicas.

Nota: Los modelos de lenguaje que se crean a partir de datos en la naturaleza representan lo que ven, y nada más. El lenguaje en la web y los libros también contienen sesgos, lenguaje tóxico que eventualmente se replica en estos modelos. Para un análisis de los riesgos potenciales de construir modelos de lenguaje a partir de grandes corpus de lenguaje, consulte este artículo de Bender et al.

4.2 Datos paralelos (bitext)

El tipo de datos que se necesitan para construir un sistema de traducción automática son datos paralelos, que consisten en una colección de oraciones en un idioma junto con sus traducciones. Históricamente, los datos paralelos se obtuvieron de traducciones en espacios públicos multilingües como las Naciones Unidas y el Parlamento Europeo. Ahora, el mayor recurso de texto paralelo es la web multilingüe.

Con el fin de entrenar a los modelos de traducción automática no es suficiente sólo tener documentos traducidos. Los textos deben segmentarse en oraciones y alinearse. La alineación de texto paralelo es la identificación de las oraciones correspondientes a ambos lados del texto paralelo. Los documentos resultantes deben corresponder línea por línea o contener las oraciones originales y sus traducciones en la misma línea. Hunalign ayuda a crear alineaciones de oraciones a partir de documentos traducidos. Las memorias de traducción (archivos TMX) también hacen grandes datos paralelos, ya que ya están segmentados por oraciones.

4.2.1 Obtención de datos paralelos

OPUS es una recopilación de casi todos los datos paralelos disponibles públicamente. Es el punto de referencia para que muchos investigadores publiquen sus datos paralelos o datos de origen para el desarrollo de modelos de TA.

Algunas fuentes comunes de datos paralelos son: - sitios web multilingües (por ejemplo, medios de comunicación internacionales), - subtítulos de películas (véase OpenSubtitles), - textos sagrados, - procedimientos parlamentarios, - datos de localización de software.

4.2.2 Crowdsourcing de datos paralelos con Tatoeba.org

Tatoeba es una colección gratuita de oraciones de ejemplo con traducciones dirigidas a estudiantes de idiomas extranjeros. Está escrito y mantenido por una comunidad de voluntarios a través de un modelo de colaboración abierta. Está organizado por la Asociación Tatoeba, una organización francesa sin fines de lucro financiada mediante donaciones. Actualmente tiene 10.397.308 sentencias en 412 idiomas admitidos.

Los usuarios pueden buscar palabras en cualquier idioma para recuperar oraciones que las usen. Cada oración en la base de datos de Tatoeba se muestra junto a sus traducciones probables en otros idiomas; las traducciones directas e indirectas se diferencian. Las oraciones se etiquetan por contenido como tema, dialecto o vulgaridad; también tienen hilos de comentarios individuales para facilitar la retroalimentación y las correcciones de otros usuarios y notas culturales. Las oraciones se pueden examinar por idioma, etiqueta y otros criterios.

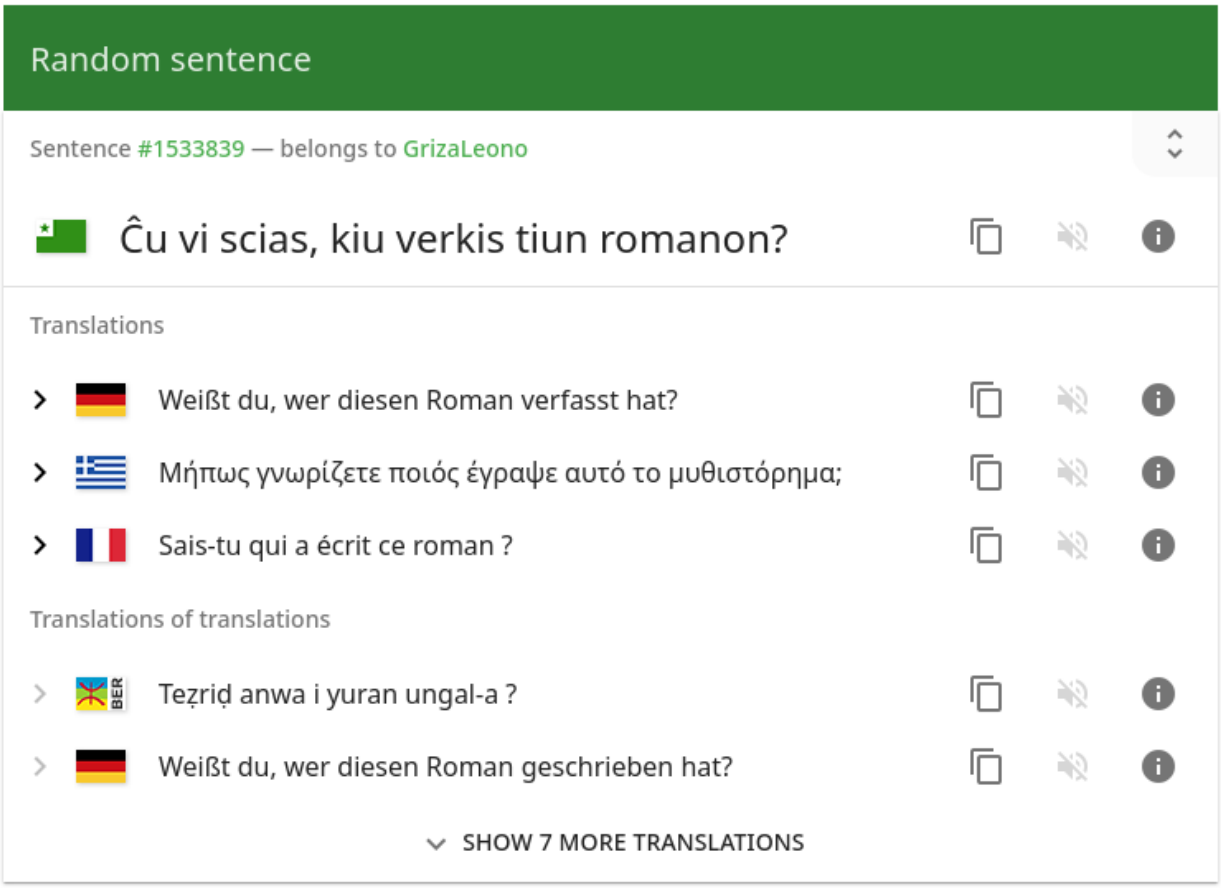

Figura 1: Una frase y sus traducciones de Tatoeba

Los usuarios registrados pueden añadir nuevas frases o traducir o corregir las existentes, incluso si su idioma de destino no es su lengua materna. Sin embargo, se anima a los usuarios a añadir oraciones originales o traducciones en su idioma nativo o más fuerte.

Toda la base de datos Tatoeba se publica bajo una licencia Creative Commons Atribución 2.0. También es muy fácil descargar partes de corpus en formato monolingüe o paralelo desde su página de descargas.

4.3 Corpus del habla

Un corpus de voz es una colección de archivos de audio de voz que generalmente se acompañan con sus transcripciones de texto. En la tecnología del habla, los corpus del habla se utilizan para crear modelos acústicos para tareas como el reconocimiento automático de voz, la síntesis de texto a voz y también la identificación de altavoces.

Los corpus de habla pueden contener lectura (por ejemplo, audiolibros, noticias, números y palabras leídas) o habla espontánea (diálogos). El adequeato de corpus para los modelos ASR de entrenamiento contiene muestras de tantos altavoces como sea posible y en diversos entornos acústicos (por ejemplo, ruidoso, de lejos). Por el contrario, los datos de entrenamiento para TTS generalmente contienen grabaciones de un altavoz en un entorno acústicamente óptimo.

OpenSLR lista muchos corpus de discursos disponibles públicamente.

4.3.1 Common Voice

Common Voice es un proyecto de crowdsourcing iniciado por Mozilla para crear una base de datos gratuita para hacer que el reconocimiento de voz sea accesible para todos. El proyecto cuenta con el apoyo de voluntarios que registran oraciones de muestra con un micrófono y revisan las grabaciones de otros usuarios. Las muestras con voz se publican a intervalos regulares bajo la licencia de dominio público CC0 (dominio público). Esta licencia garantiza que los desarrolladores puedan utilizar la base de datos para aplicaciones de voz a texto sin restricciones ni costes.

Nota: A partir de mayo de 2022, Common Voice admite 63 idiomas con 68 nuevos en camino. Consulta aquí la lista actual de idiomas: https://commonvoice.mozilla.org/en/languages.

4.3.2 Añadir un idioma a Common Voice

Common Voice funciona como una plataforma comunitaria donde cada idioma tiene su propia comunidad. El procedimiento para añadir un nuevo lenguaje en Common Voice es el siguiente:

1. **Encuentre un administrador de la comunidad para el idioma** (Información sobre roles)
2. **Solicitud de localización a Mozilla** Esto se hace utilizando esta plantilla en su página de github. Esto iniciará el proceso de localización de Common Voice al idioma deseado colocándolo en Pontoon.
3. **Localización en Pontoon** (manual de usuario) Cada cadena en la plataforma Common Voice necesita ser traducida al idioma respetando la guía de estilo. En total hay 663 cuerdas. Las traducciones pueden ser realizadas por cualquier orador que se registre en la plataforma, pero deben ser revisadas por el administrador de la comunidad.
4. **Recopilación de oraciones** Se debe recopilar un mínimo de 5000 oraciones de dominio público e ingresar a Common Voice sentence Collector.
5. **Revisar frases** Cada frase recopilada debe ser revisada manualmente por al menos dos usuarios en el recopilador de oraciones.
6. **Espere a la próxima versión del CV** Una vez que se complete la localización y haya 5000 oraciones revisadas, la próxima versión del CV debe contener su idioma. Los lanzamientos se realizan dos veces al mes con los horarios listados en su repositorio de github.

Figura 2: Grabación de una sentencia kurda Kurmanji en Common Voice

4.3.3 Datos encontrados

También es posible obtener datos de voz de programas de radiodifusión, películas y otros materiales grabados, como entrevistas. Este tipo de datos se denomina «datos encontrados», ya que originalmente no está destinado a servir para construir tecnología de voz, sino que *se reutiliza* para hacerlo. Los datos encontrados requieren ser procesados para obtener segmentos de audio cortos y sus transcripciones.

4.3.4 Fuentes

- Corpus de texto en Wikipedia
- Tatoeba en Wikipedia
- Speech corpus in Wikipedia
- Common Voice en Wikipedia

Figura 3: Este documento fue creado con el apoyo financiero de la Unión Europea. El contenido de este sitio web es responsabilidad exclusiva de Col·lectivaT y SKAD, y no refleja necesariamente las opiniones de la Unión Europea.

Figura 4: Este proyecto está financiado por la Unión Europea.

CAPÍTULO 5

Estudios de caso

En esta sección, enumeramos algunas iniciativas y trabajos relacionados con la tecnología del lenguaje que consideramos ejemplares o inspiradores para las lenguas en peligro y minoritarias. Comenzaremos describiendo el proyecto que dio fruto a este documento «Judeo-Español: Conectando los dos extremos del Mediterráneo» y continuaremos con proyectos que involucran las lenguas anatólica, ibérica y africana.

5.1 Judeoespañol: Conectando las dos orrillas del Mediterráneo

Col·lectivaT y el Centro Sefardí de Estambul se han unido para llevar a cabo un conjunto diverso de actividades que van desde la creación de contenidos en redes sociales hasta el desarrollo de tecnología lingüística avanzada que ayuda al judeoespañol (ladino) hasta la era digital. También pretendía dar a conocer esta lengua como patrimonio común entre Turquía y España.

5.1.1 Creación de contenidos audiovisuales para redes sociales

El proyecto creó videos cortos de aprendizaje de idiomas que ayudarían a ganar visibilidad en las plataformas de redes sociales y atraerían a las generaciones jóvenes a aprender ladino. En estos videos cortos, se presenta una frase judeoespañola con su traducción en turco, inglés y español con un audio que ayuda a aprender su pronuciación.

5.1.2 Ladino Data Hub y conjuntos de datos abiertos de lenguaje

El proyecto lanzó Ladino Data Hub que actuará como un archivo web centralizado dedicado a alojar datos en idioma ladino y otros recursos que ayudan a documentar la cultura sefardí. Su objetivo es permitir a investigadores y periodistas de todo el mundo acceder y compartir conjuntos de datos que ayudarían a impulsar la investigación y el desarrollo de Ladino.

El proyecto creó y reenvasó conjuntos de datos ya existentes y compartidos en este portal. Estos son:

- Un corpus de texto que consta de frases arrastradas desde el periódico Şalom.
- Un corpus de texto paralelo con audio que contiene frases de Una Fraza al diya y su audio

Figura 1: Promoción de Ladino en la página de Instagram de SKAD con vídeos de Fraza del diya (Una frase al día)

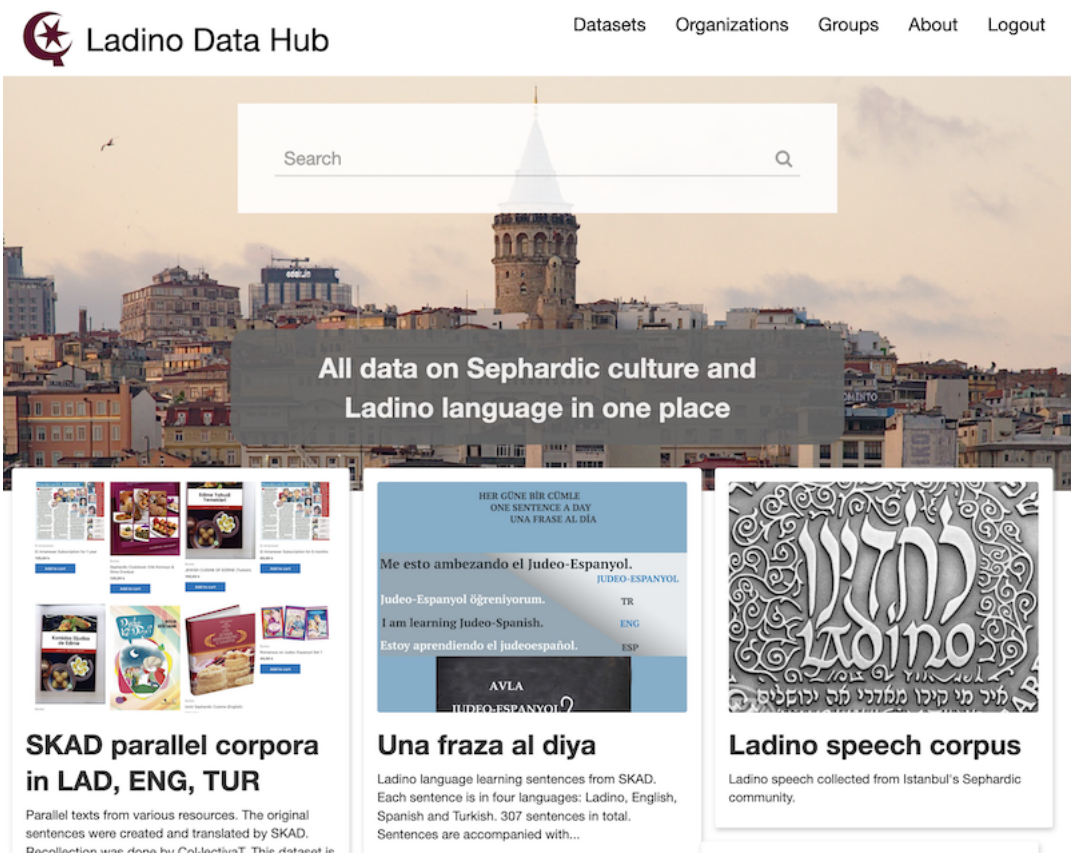

Figura 2: Ladino Data Hub alberga datos relacionados con la cultura ladina y sefardí

- Ladino speech corpus creado por SKAD
- Conjunto de datos de entrenamiento de síntesis del habla limpia consiste en material de habla leída por Karen Şarhon
- Un léxico ladino español basado en el diccionario de Güler, Portal i Tinoco
- Textos paralelos en ladino, inglés y turco a partir de las traducciones realizadad en SKAD
- Corpus paralelo producido sintéticamente para los modelos de MT de referencia de entrenamiento

5.1.3 Aplicación web para traducción automática y síntesis de voz

El resultado final y más importante del proyecto es una aplicación web que es capaz de traducir entre ladino y tres idiomas relacionados turco, español e inglés. El objetivo es ayudar a los estudiantes de idiomas, investigadores y lingüistas que quieran estudiar judeo-español. El back-end de traducción automática se construyó con la ayuda de un sistema de traducción automática basado en reglas que puede convertir del español al ladino, utilizando la sintaxis similar entre idiomas pero cambiando la ortografía y el vocabulario con un conjunto de reglas derivadas de diccionarios y libros de gramática. La aplicación web también puede sintetizar oraciones ladinas con una aplicación TTS que se creó con el conjunto de datos de capacitación TTS.

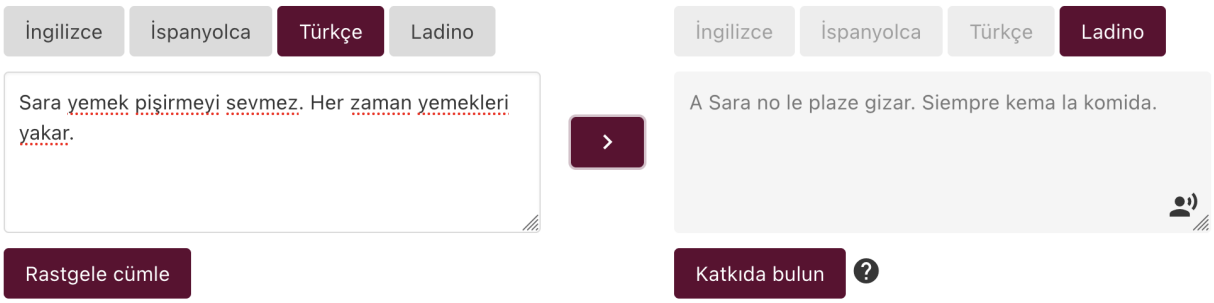

Figura 3: Aplicación web de traducción ladina con síntesis de voz

La aplicación web también permite la contribución de datos paralelos. Los usuarios pueden cargar una oración aleatoria y enviar su traducción corregida para extender los datos paralelos para Ladino.

Nota: Para obtener un informe técnico detallado de este proyecto, consulte el documento «Preparando una Lengua en Peligro para la Era Digital: El Caso del Judeo-Español» presentado en el Taller sobre Recursos y Tecnologías para las Lenguas Indígenas, en Peligro y de Menores Recursos en Eurasia (EURALI): *Enlace que se colocará pronto*

5.2 Comunidades de base PLN

Contra la investigación de la PLN centrada en un puñado de idiomas, las comunidades de investigación de base se organizan para llevar los idiomas del mundo a la vanguardia de la tecnología. Dos ejemplos de estas iniciativas son Masakhane y Turkic Interlingua.

Como se define en su página web, «Masakhane es una organización de base cuya misión es fortalecer y estimular la investigación de PLN en idiomas africanos, para los africanos, por los africanos.» Es una iniciativa abierta a todos para «construir juntos» como sugiere el significado de su nombre. Investigadores africanos y no africanos de todo el

mundo llevan a cabo muchas actividades simultáneas con el objetivo de representar los 2000 idiomas de África en la investigación sobre tecnología lingüística. Algunos trabajos destacados de Masakhane son:

- Traductor Masakhane apoyando 6 idiomas africanos: Yoruba, Shona, Lingala, Swahili, Tshiluba
- Lanfrica catalogación de recursos lingüísticos africanos para contrarrestar las dificultades encontradas en el descubrimiento de obras lingüísticas africanas
- BibleTTS desbloquear el desarrollo de modelos de texto a voz de alta calidad para diez idiomas hablados en el África subsahariana: ewe, hausa, kikuyu, lingala, luganda, luo, chichewa, akuapem twi, asante twi, yoruba.
- Traducción automática para preservar el idioma y la cultura Oshiwambo
- MasakhaNER Conozca nuestros nombres creando conjuntos de datos de reconocimiento de entidades con nombre (Ner) hechos a mano para varios idiomas africanos

Masakhane también organiza talleres anuales para publicar investigaciones relacionadas con la PLN africana y participa en fondos de recopilación de datos como Lacuna.

Turkic Interlingua (TIL) es una «comunidad de investigadores, ingenieros, entusiastas del lenguaje y líderes comunitarios cuya misión es desarrollar tecnologías lingüísticas (desde correctores ortográficos hasta modelos de traducción), recopilar diversos conjuntos de datos y explorar fenómenos lingüísticos a través de la lente de la investigación académica para las lenguas túrquicas» como Altai, Azerbaijani, Bashkir, Shor, Crimean Tatar, Chuvash, Gagauz, Karakalpak, Khakas, Kazakhstan, Karachay-Balkar, Kumyk, Kirghiz, Sakha (Yakut), Salar, Turkmen, Turkish, Tatar, Tuvinian, Uighur, Urum y Uzbek.

5.3 Campañas de Common Voice

Varias comunidades lingüísticas se han embarcado en la movilización de la participación en Common Voice. Estas iniciativas son preparadas por grupos que van desde individuos individuales hasta gobiernos locales. Algunos ejemplos son:

- Common Voice Türkçe
- Kurdish crowdsource
- Comunidad Igwebuike para Igbo
- Proyecto AINA para catalán

También nos gustaría hacer una mención especial a la comunidad catalana por su contribución a Common Voice. Siendo una lengua minoritaria apátrida en España, es la cuarta lengua más grande (a partir de mayo de 2022) en Common Voice gracias a las increíbles contribuciones de los activistas y también a una campaña de movilización por parte de la iniciativa de IA del gobierno local catalán.

5.4 Otras iniciativas

Otras iniciativas:

- Col·lectivaT ha creado open speech data and text corpora for Catalan utilizando emisiones de televisión pública y procedimientos parlimantarios.
- Catotron es el primer motor de síntesis de voz de código abierto basado en una red neuronal para catalán construido con el apoyo del Departamento de Cultura de Cataluña.
- Cebuano y Waray-Waray, idiomas de Filipinas, tiene una de las páginas de Wikipedia más grandes gracias al uso competitivo de la traducción automática (source)

Figura 4: Cartelera con «Es hora de que Internet hable catalán» que se exhibe en Times Square de Nueva York (foto de Aina Martí)

- El pueblo maorí ha rechazado las iniciativas privadas y de código abierto para recopilar datos de voz en su idioma a fin de «mantener el derecho a la libre determinación» (fuente)
- Un Manifiesto para la Tecnología del Lenguaje Abierto
- ELLORA habilitando idiomas de bajos recursos por Microsoft Research India

---

Figura 5: Este documento fue creado con el apoyo financiero de la Unión Europea. El contenido de este sitio web es responsabilidad exclusiva de Col·lectivaT y SKAD, y no refleja necesariamente las opiniones de la Unión Europea.

Figura 6: Este proyecto está financiado por la Unión Europea.

CAPÍTULO 6

Prácticas más idóneas

Nota: Esta sección se finalizará con los comentarios recogidos en los talleres organizados durante el proyecto.

Si bien es evidente que un gran número de idiomas en el mundo requieren una inversión intensiva en la creación de recursos para la habilitación tecnológica, parece muy poco probable que esa inversión pueda realizarse fácil y rápidamente en un corto período de tiempo. Dados estos recursos limitados, las comunidades lingüísticas deben estar facultadas para determinar el futuro de sus idiomas. En este documento, hemos presentado cómo la representación digital es parte de este proceso.

Para decidir por dónde empezar, sugerimos adoptar la metodología del diseño 4-D pensando en *Descubrir, diseñar, desarrollar e implementar* como lo introdujo Bali et al. en su iniciativa ELLORA. Este enfoque centrado en el usuario es el siguiente:

1. Descubre lo que más necesita la comunidad dewl idioma,
2. Diseñar para los usuarios y su lengua prestando atención a la diversidad de la lengua y evitando un enfoque que se aleje de una lengua mayoritaria,
3. Desarrollar e implementar con frecuencia de manera interactiva mejorando constantemente y detectando fallas desde el principio.

Incluso cuando la comunidad no ha desarrollado una perspectiva del desarrollo de la tecnología del lenguaje, es una buena práctica tener en cuenta el valor de los datos al realizar actividades de preservación del lenguaje. Algunos ejemplos de estos son:

- Organizar eventos y crear contenido para aumentar la conciencia de los datos en la comunidad lingüística,
- Introducir idiomas a las plataformas de crowdsourcing,
- Organizar datathons para la recopilación de datos de idioma,
- Traducir cuentos populares y cuentos infantiles que residen en el dominio público,
- Almacenar versiones de texto plano o documentos de material publicado con el fin de crear corpus de texto,
- Guarde y comparta abiertamente las memorias de traducción para ayudar a otros traductores y para crear datos paralelos,

- Almacenar grabaciones de material de radiodifusión (por ejemplo, programas de radio) y transcribirlas si es posible para que puedan convertirse en datos de voz,
- Guarde el contenido publicado en las redes sociales en un lugar permanente para que no se pierdan en las líneas de tiempo.

¿Tiene alguna sugerencia o pregunta? Escríbenos a info-arroba-collectivat.cat

---


Figura 1: Este documento fue creado con el apoyo financiero de la Unión Europea. El contenido de este sitio web es responsabilidad exclusiva de Col·lectivaT y SKAD, y no refleja necesariamente las opiniones de la Unión Europea.

CAPÍTULO 7

Licencia

Este documento está licenciado con Atribución 4.0 Internacional (CC BY 4.0)

Figura 1: Este documento fue creado con el apoyo financiero de la Unión Europea. El contenido de este sitio web es responsabilidad exclusiva de Col·lectivaT y SKAD, y no refleja necesariamente las opiniones de la Unión Europea.