

---

# Low-resource Languages Toolkit

*Release 0.1*

Alp Öktem, Col·lectivaT

Oct 20, 2022



## CONTENTS

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Who is this document for? . . . . .	3
1.2	Can I contribute? . . . . .	3
1.3	Authors . . . . .	4
<b>2</b>	<b>Languages and the digital age</b>	<b>5</b>
2.1	Digital extinction . . . . .	5
2.2	Online presence of a language . . . . .	6
2.3	Language documentation projects . . . . .	7
2.4	“Traditional” technology . . . . .	7
<b>3</b>	<b>Data-driven language technologies</b>	<b>15</b>
3.1	Machine translation . . . . .	16
3.2	Automatic speech recognition . . . . .	17
3.3	Text-to-speech . . . . .	18
<b>4</b>	<b>Language data</b>	<b>21</b>
4.1	Text corpus . . . . .	21
4.2	Parallel data (bitext) . . . . .	22
4.3	Speech corpus . . . . .	24
<b>5</b>	<b>Case studies</b>	<b>27</b>
5.1	Judeo-Spanish: Connecting the two ends of the Mediterranean . . . . .	27
5.2	Grassroots NLP communities . . . . .	30
5.3	Common Voice campaigns . . . . .	31
5.4	Other initiatives . . . . .	32
<b>6</b>	<b>Good practices</b>	<b>33</b>
<b>7</b>	<b>License</b>	<b>35</b>





Fig. 1: This project is funded by the European Union.

In this document we cover the role of technology, with a focus on artificial intelligence-based technology, on the preservation of endangered languages.

This is a living document that was created as part of the project “Judeo-Spanish: Connecting the two ends of the Mediterranean” carried out by Col·lectivaT and Sephardic Center of Istanbul (SKAD) within the framework of the “Grant Scheme for Common Cultural Heritage: Preservation and Dialogue between Turkey and the EU–II (CCH-II)” implemented by the Ministry of Culture and Tourism of the Republic of Turkey with the financial support of the European Union.

---

**Note:** This document is also available in Turkish and Spanish.

Bu doküman Türkçe de erişilebilir.

Este documento también está disponible en español.

---



Fig. 2: This project is funded by the European Union.



## INTRODUCTION

There is currently a growing digital divide between languages with sufficient resources and languages with fewer resources, further exacerbating the danger of digital extinction for them. For the majority languages, the process of generating useful tools and resources is much easier due to their large web-presence. However, many minority languages do not have sufficient material and human resources to power the creation of such tools. Lack of state support, public visibility, societal and institutional oppression are direct causes of these languages being deprioritized in the digital spaces of today.

Efforts on preservation of languages focus mainly on language documentation, teaching, and physical community building. One area that is overlooked is creation of tools based on artificial intelligence. Tools like machine translation, speech synthesis, and speech recognition are now important counterparts in creating human-machine interfaces. Also, these tools can help model the knowledge of dying languages and preserve them for future generations.

### 1.1 Who is this document for?

This document is for you if you are:

- *Language activist* who's interested in extending tools and resources in their language
- *Linguist* who is interested in collecting data for research and building language technology
- *Natural language processing (NLP) researcher* who is interested in augmenting data for their language of interest
- *Language activist allies* who want to support the revitalization of under-resourced languages

### 1.2 Can I contribute?

This is a living document with an open license (CC-BY). Its source file is shared publicly in <https://github.com/CollectivaT-dev/language-toolkit> where you can pull a version to work on your own and then submit your contribution. It can range from correcting typos to adding a translation, detailing a section and explaining your case study. If you have doubts please feel free to write to us at [info@collectivat.cat](mailto:info@collectivat.cat).

## 1.3 Authors

- Alp Öktem - ([Web](#), [Twitter](#))
- 



Fig. 1: This document was created with the financial support of the European Union. The content of this website is the sole responsibility of Col-lectivaT and SKAD and does not necessarily reflect the views of the European Union.



Fig. 2: This project is funded by the European Union.



## LANGUAGES AND THE DIGITAL AGE

Digital age has brought many new opportunities and advantages. It has undoubtedly connected humankind in unimaginable ways in a short amount of time. Although, no innovation comes with its challenges and threats. World wide web is a resource accessible (almost) to all but it is dominated by a handful of languages.

English for example is spoken by 15% of the world but it currently holds 54% of all the content on the web. On the other hand, languages like Russian, Chinese, and Spanish each represent around 5 to 6% of the content on the web, in spite of their geopolitical dominance.

Where does this leave the world's many endangered and dying languages? Sadly, on the very margins of this picture.

Languages thrive in communities and pass onto generations with day-to-day usage. The more our lives are connected through digital mediums, we are less and less exposed to our mother tongues which are not represented online. This eventually leads to a decline of their usage by younger generations.

---

**Note:** How? To illustrate, a Kurdish speaker in Turkey, accesses their government health services website and sees that everything is in Turkish, or goes to check the latest online social media platform and sees that it's by default available in English. These kinds of small encounters lead to thinking that in order to find their way and address their needs, speaking their mother tongue is not enough. You have to know the nearest majority language and many times even more.

---

### 2.1 Digital extinction

According to UNESCO, around 3,500 languages are expected to be extinct by the end of this century and we cannot deny the role of technology in this. Kornaï states that the vast majority (over 95%) of languages have already lost the capacity to ascend digitally. Digital ascent requires use in a broad variety of digital contexts ranging from maintaining a Wikipedia page, to making language classes available and creating language technology data.

We shouldn't of course fall into the trap of making technology the culprit of language loss. It is merely a representation of already existing power dynamics in the society. States which deprioritize or even oppress certain languages start their digital transformation by excluding all these languages in their digital infrastructure. Now, North American and Eurocentric big technology companies follow a English-first approach by default.

Technology can also be utilized to form communities around language preservation, knowledge sharing, language documentation.

## 2.2 Online presence of a language

Traditionally, the responsibilities of a language activist has been actively speaking the language, passing it to younger generations, forming language learning and speaking communities, negotiating with public institutions for the inclusion of their language, collaborating with linguists for the documentation of their language and so on.

Nowadays, the challenge is not just making the language alive in the physical world but also online. It is related in two ways for the survival of a language:

1. Exchanges and visibility online strikes interest and helps engage existing and new language learners.
2. What's stored online in turn is a digital record for language which helps documentation and technology development.

Below, we will describe some ways the internet is becoming multilingual and plural while helping revive endangered languages.

### 2.2.1 Access to knowledge

One of the most popular initiatives on bringing languages online is Wikipedia. Wikipedia is open an online encyclopedia written and maintained by a community of volunteers through an open collaboration and reviewing system.

Wikipedia's broad aim is to democratize access to knowledge. Naturally, the culture built around this ethos goes hand in hand with multilinguality. Although it started only with English, it quickly expanded to the world's many languages. We can probably say that it's the most language diverse platform on the internet with 326 languages (as of 03.05.22) and counting up.

**Note:** The first edit on a non-English Wikipedia was made in Catalan on March 16, 2001. Today, it is most notable for its large number of quality articles, which illustrates Catalan language's important online presence despite being a minority language. Catalan wikipedia is currently the 20th largest Wikipedia.

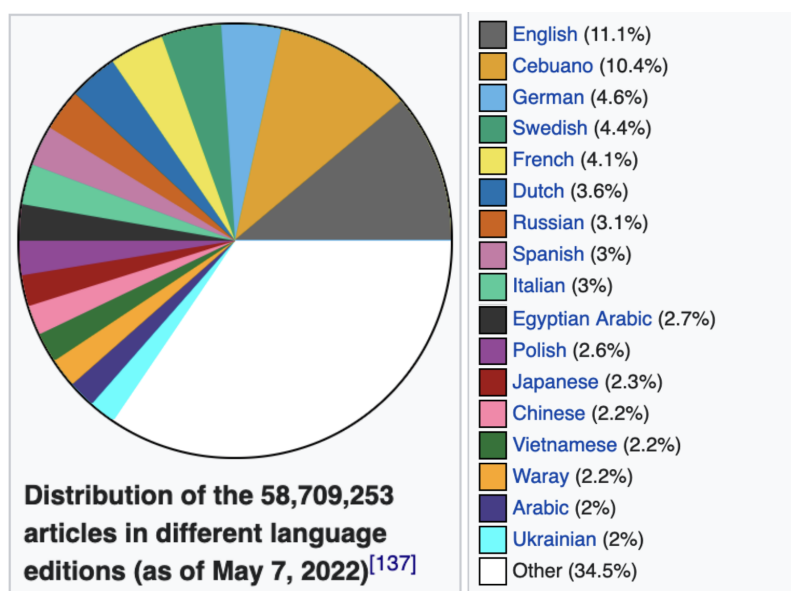


Fig. 1: Relative sizes of different wikipedias ([source](#))

Making a new language available in Wikipedia is no easy task (as explained [here](#)), but it is definitely a great way to make knowledge accessible online and build a virtual community around it.

## 2.2.2 Immersion into language

One great example of a language revitalization with the help of technology is Yiddish. After the holocaust, the number of Yiddish speakers decreased drastically from around 10 million speakers. Whoever survived were forced to assimilate the language of their new lands to avoid persecution. In the last century, the use of Yiddish had almost disappeared except for small and dispersed Hasidic communities.

With the rise of the internet and popularity of online forums, Yiddish speakers used these platforms to converse in their language. Over time, the virtual world became the primary meeting point for Yiddish speakers in forums like The Idishe Velt (The Jewish World) and Kave Shtiebel (The Coffee House).

## 2.3 Language documentation projects

The rapid loss of languages in the last century has powered many initiatives for language documentation and revitalization. One of these initiatives is The Endangered Languages Project (<https://www.endangeredlanguages.com/>) which is a web-based platform that acts as a collaborative hub of language enthusiasts, linguists and industry partners to help strengthen endangered languages. Users of the website act as contributors by uploading language samples in text, audio, link or video format using a unique geotagging system that allows for easy searchability.

Similarly, [Wikitongues](#), which started in 2014, collects recordings and resources of the world's languages. Currently it holds videos in over 700 languages, lexicons in 200 languages and links to hundreds of external resources.


## 2.4 “Traditional” technology

Preserving a language isn't just a matter of recording words or phrases and digitising them to be held in an online vault. Language is inherently about people, culture and identity. In order to keep a language alive, it needs to be spoken by many, immersed in everyday culture and actively passed onto future generations. These days, the internet, social media, software, platforms occupy a large space in our everyday lives. In this section, we will list some of the core tools needed for a technology to thrive digitally.

### 2.4.1 Unicode-supported font

A digital font is the way computers know how to display the characters in your language. Unicode, formally the Unicode Standard, is an information technology standard for the consistent encoding, representation, and handling of text expressed in most of the world's writing systems. The standard, which is maintained by the Unicode Consortium, defines 144,697 characters covering 159 modern and historic scripts, as well as symbols, emoji, and non-visual control and formatting codes.

You can check if your language is supported by a font by going to [Google Noto](#) and searching among fonts that represent more than 500 writing systems. If it is not there, you can create your own with the help from a font designer and install it manually on your computer.



**ELP**  
 Endangered  
 Languages  
 Project  
Supporting and celebrating global linguistic diversity

[Map](#) | [Languages](#) | [Resources](#) | [Submit](#) | [Blog](#) | [Download](#) | [About](#)

# Ladino

[aka *Judeo-Spanish*, *Sephardic*, *Hakitia*]

Classification: Indo-European • at risk


[Description](#) | [Resources](#) | [Activity](#) | [Revitalization](#) | [Bibliography](#) | [Suggest a Change](#) | [Subscribe](#)


## Language metadata

The name Ladino refers most commonly to the written/literary form of the language. Most speakers refer to the spoken language as Judeo-Spanish.

ALSO KNOWN AS	Judeo-Spanish, Sephardic, Hakitia, Haketia, Judeo Spanish, Sefardi, Dzhudezmo, Judezmo, Spanyol, Haquetiya
CLASSIFICATION	Indo-European, Italic, Romance, Western Romance
CODE AUTHORITY	ISO 639-3
LANGUAGE CODE	lad
VARIANTS & DIALECTS	<ul style="list-style-type: none"> <li>• Ladino</li> <li>• Judezmo</li> <li>• Haquetiya</li> </ul>
DOWNLOAD	As <a href="#">csv</a>
MORE RESOURCES	<a href="#">OLAC search</a>

[Map](#) | [Hybrid](#)



**LOCATION INFORMATION**  
 COORDINATES 40.0,33.0  
 COMPARE SOURCES (2) 

Information from: "The World Atlas of Language Structures". Bernard Comrie and David Gil and Martin Haspelmath and Matthew S. Dryer · Oxford University Press

Fig. 2: Ladino in Endangered Languages Project

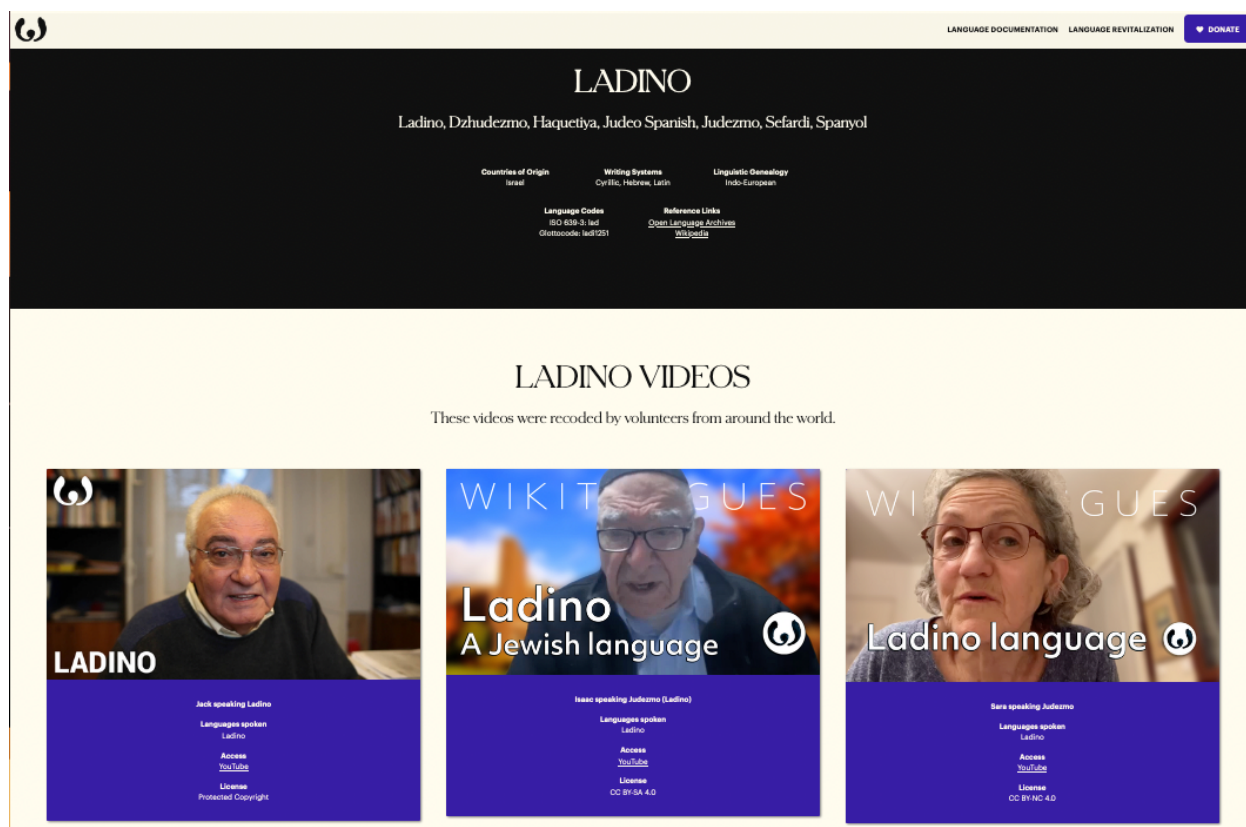


Fig. 3: Ladino videos in Wikitongues



## 2.4.2 Keyboard

Until the day we will be speaking naturally with computers, the most common interface for interacting with them will be the keyboard. It is a technology easily taken for granted for the world's many languages but unfortunately it is not available in all of the world's writing systems. If a keyboard is not present or not developed enough for a language, its speakers tend to prefer other alphabets or even languages to communicate. For example, speakers of Ethiopian languages like Amharic, Tigrinya and Oromo switch to using English as Ge'ez script is not pre-installed in their smartphones. Young arabic speakers in many countries have invented their own chat-alphabet [Arabizi](#) consisting of latin characters and numerals to account for the lack of Arabic script support in early mobile and web technology.

If a keyboard is not available, let's say in your phone or computer, here are some resources to search or help create your own keyboard:

- Google's mobile keyboard, [Gboard](#)
- [Keyman](#) supports 2000 languages
- [Microsoft Keyboard Layout Creator](#)
- [Ukelele](#) is a Unicode Keyboard Layout Editor for the MacOS

## 2.4.3 Online dictionary

A dictionary, or lexicon, is a solid way of documenting a language since it acts as a reference of words and their meanings. An online dictionary never goes out-of-print as it is accessible from any device with an internet connection. Also, open-source dictionaries can live and grow collaboratively as a community effort involving both speakers of the language, linguists and technologists.

[Living Dictionaries](#) is an online dictionary-builder platform created by Living Tongues Institute for Endangered Languages. It provides comprehensive, free online tech tools that assist language communities in conservation and revitalization efforts. It also allows recording of words and phrases. As of May 2022, it support 237 languages. To initiate your language in Living Dictionaries, you can get use of their [Elicitation lists](#) and watch tutorials on their [YouTube channel](#).

[SIL Dictionary App Builder](#) “helps you to build customized dictionary apps for Android and iOS smartphones and tablets. You specify the lexicon data file to use, the app name, fonts, colors, the ‘about box’ information, the audio, illustrations and the icons. Dictionary App Builder will package everything together and build the customized app for you. You can then install it on your phone, send it to others by Bluetooth, share it on microSD memory cards and publish it to app stores on the Internet.”

## 2.4.4 Language learning applications

The availability of online educational platforms has revolutionized the way many people approach language learning today. Even though they don't replace a teacher, they either complement traditional classes or are the only choice in some languages' contexts. They also give many advantages like letting people learn on any device (mobile or desktop), at their own pace and schedule. These apps serve language classes and exercises in short, fun and digestible sprints, let the students track their progress, and even chat with or hire language tutors in online community spaces.

Many of the world's endangered, minority and under-resourced do not yet have a significant presence online or adequate language documentation to create online courses. There are, however, thanks to push from language communities and increased sensibility for learning indigenous languages worldwide, increased interest by companies who develop these apps for investing into endangered and minority languages. Languages like Maori, Scots Gaelic, Hawaiian, Quechua, Navajo and Lakota are making their way into well-known educational platforms like Duolingo, Babbel and uTalk.

We can categorize these platforms in four ways:



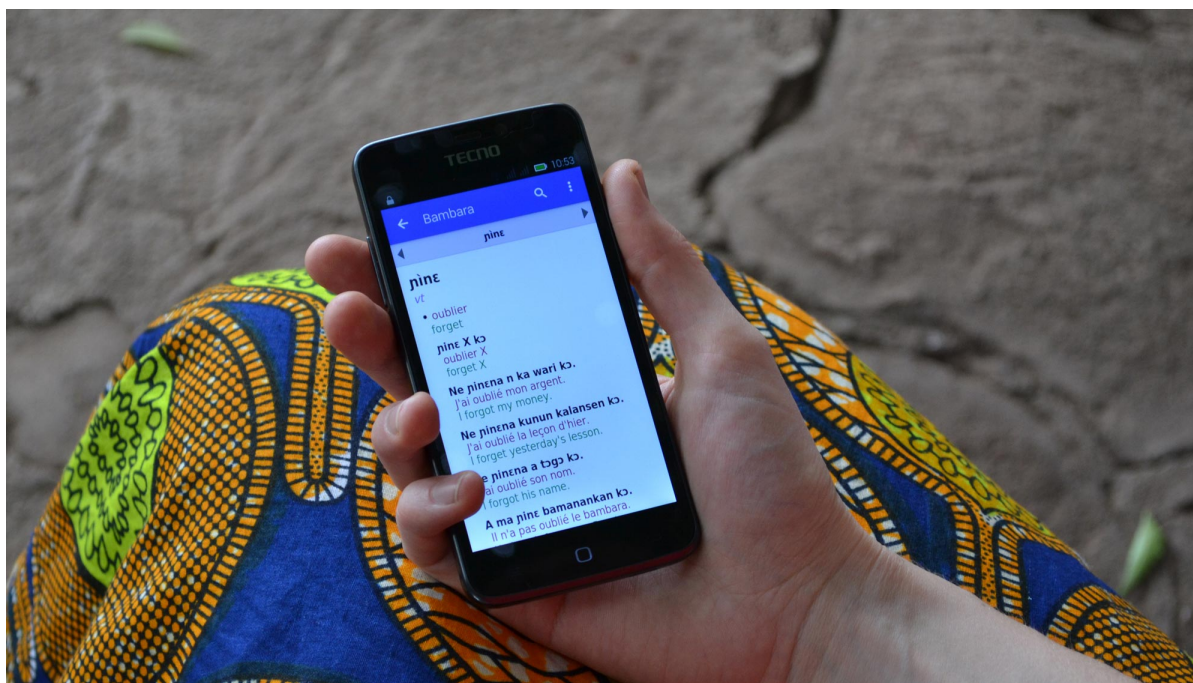


Fig. 5: A woman using Bambara dictionary on her mobile phone (image credit SIL International)

- Module-based:** Using these apps feels more or less like taking a class in a school or a college, where users follow a modular curriculum planned by educators. It allows learners to track their progress, receive notifications, and earn points. Some notable examples are: [Duolingo](#) , [Babbel](#) , [uTalk](#) , [Master Any Language](#). Unfortunately, it is not possible for language communities themselves to decide and implement a new language in these platforms. However, it is possible to “lobby” and participate in the creation of new modules through the communities some of these platforms provide. Also, it has to be noted that most of these platforms are in-profit platforms and the work by language communities remain uncompensated.
- Game-based:** These applications feed the learner with “question and answer” pairs and can evaluate how well the person memorizes the pairing over time. This method is popular online for visual as well as auditory learners of languages, and also for many other areas of study such as math and science. [Memrise](#) may be the most compelling for indigenous and other under-resourced minority languages. It is an online educational platform that uses memory techniques such as flashcards, SRS (spaced repetition software) and visual aids to optimise language learning. Memrise has content available for over 170 languages, a much greater number than most other language-learning platforms. Notably, the site has an impressive, community-created selection of content for indigenous languages and dialects from around the globe such as Yup’ik (see first screenshot below), Cherokee (see second screenshot below), Algonquian, Alutiiq, Choctaw, Greenlandic, Inuktitut, Lakota, Nahuatl, Yucatec Maya, K’iche’, Quechua, Guarani, Ainu, Jeju and many other medium-sized tongues spoken in Europe, Africa, the Middle East, Asia and the Pacific. The Memrise platform has a DIY, democratic and grassroots feeling to it. The platform is innovative because: 1) users can follow existing courses and flashcard sets and seamlessly upload their own mnemonic aids to help recall words and phrases as they progress through a course; 2) the site provides a engaging way for users to connect with language content through repetition, little quizzes, short videos, funny images and recordings made by fluent speakers, and 3) the platform allows community users to easily create their own language courses that others can use as well. Similar platforms include: [AnkiApp](#), [Language Drops](#), and [MosaLingua](#).
- Chat-based:** These apps allow learners to connect with speakers of the language they are interested in through a live interactive chat. This provides a stress-free and social environment for learners. Some examples like [HiNative](#) and [HelloTalk](#) have recently exploded in popularity especially in Asian countries.



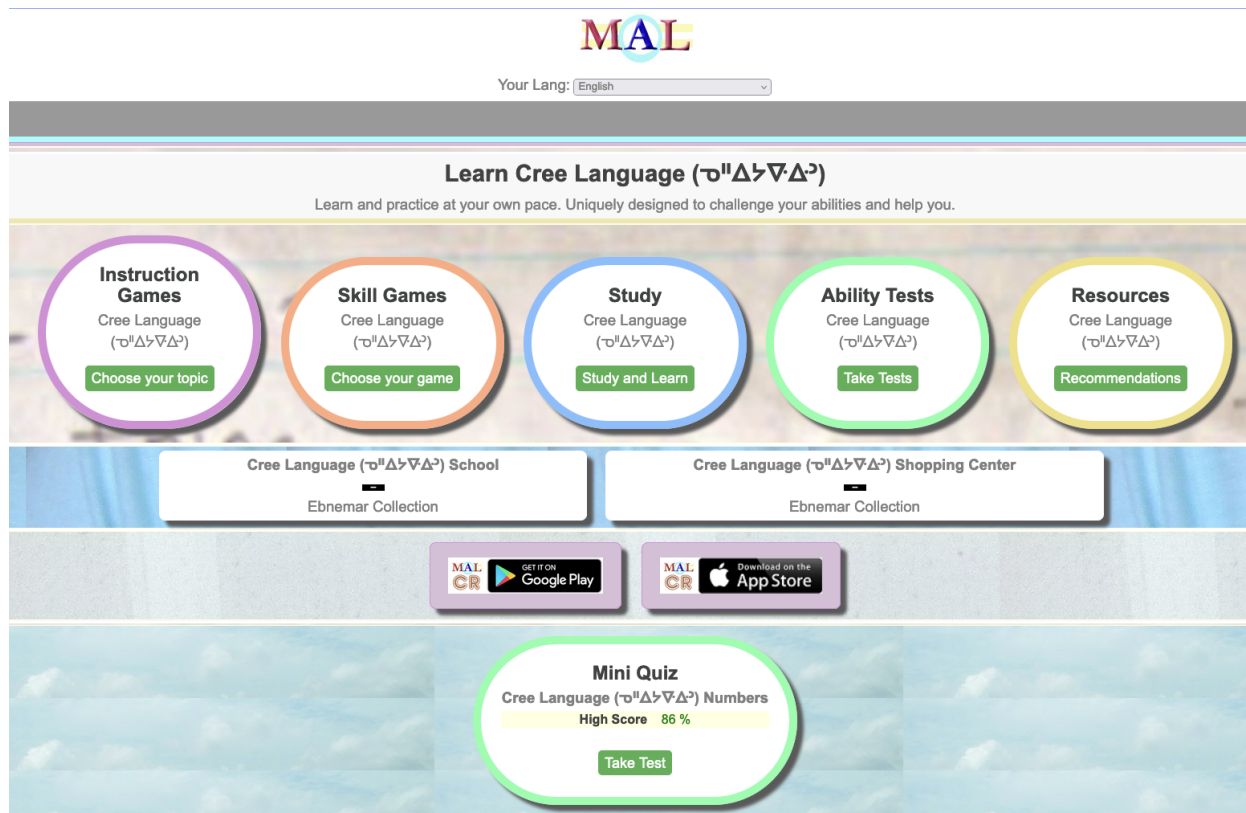


Fig. 6: Content for learning Cree language in Master Any Language platform

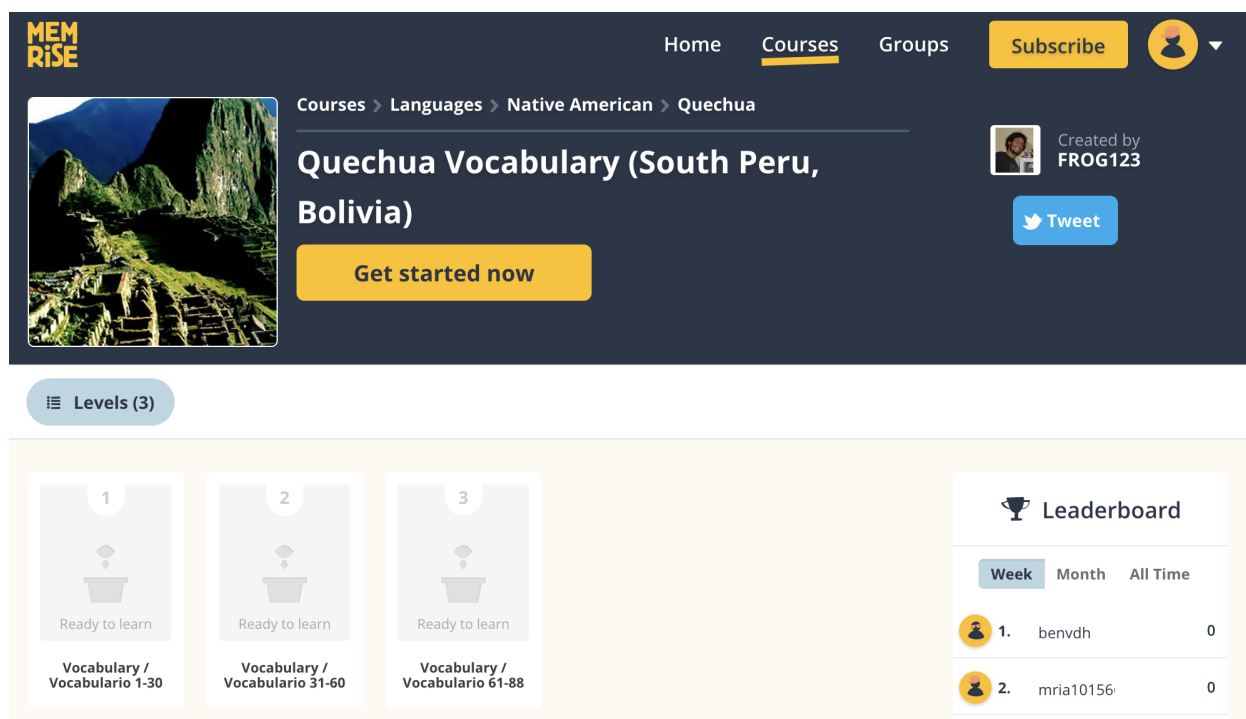


Fig. 7: Community created Quechua learning exercises in Memrise

- **Online student-tutor platforms:** For those learners who prefer classic teacher-student relationship but do not have access to teachers in their vicinity, platforms like [iTalki](#) and [Verbling](#) help setup online classes. This also contributes directly to the language community as it generates direct income for the teachers.

### 2.4.5 Sources

- [Catalan wikipedia on Wikipedia](#)
  - [How Technology Can Save Rapidly Dying Languages](#)
  - [Unicode on Wikipedia](#)
  - [Language sustainability toolkit](#)
  - [Learn Language Online by Living Tongues](#)
  - [Computer assisted language learning in Wikipedia](#)
- 



Fig. 8: This document was created with the financial support of the European Union. The content of this website is the sole responsibility of Col-lectivaT and SKAD and does not necessarily reflect the views of the European Union.



Fig. 9: This project is funded by the European Union.

## DATA-DRIVEN LANGUAGE TECHNOLOGIES

The digital revolution is here with us and Artificial Intelligence (AI) is a key technological enabler. It offers a range of new opportunities to break down existing barriers to human development and social inclusion. One area that is powered by AI is language technology which makes it possible to interact with our phones through digital assistants, translate websites and documents with a few clicks, increase accessibility of videos with automatic captioning etc.

The main motor behind these is the advancement of the field **Natural Language Processing (NLP)**. But what does NLP entail? Here's a list of core technologies that fall in the area of this field:

Text-based:

- Machine translation
- Information retrieval
- Information extraction
- Sentiment analysis
- Question answering
- Text summarization
- Named-entity recognition

Speech-based:

- Automatic speech recognition
- Text-to-speech synthesis

The revolutionary aspect of these technologies is that they are *data-driven*, which means that the intelligence that is created with these tools are collected from large volumes of information—or simply *data*. For example, in the case of machine translation, the engine “models” translation from a language to the other by looking at a collection of human-translated documents and sentences. Similarly a *sentiment analyzer* learns how to label if a tweet says good or bad about a company from thousands of tweets labelled by humans as carrying a good or bad sentiment.

This dependency on data is what makes these technologies accessible to some languages and not to the others. The available resources for a language directly influences the possibility of developing an application for a language. As the greatest resource of textual data is the internet, and it is dominated by a few languages, these technologies tend to focus on only a handful of *dominant languages* e.g. English, Spanish, Chinese, Arabic etc.

The diagram below by [Microsoft Research Labs India](#) illustrates the hierarchy created by this “power-law” among languages.

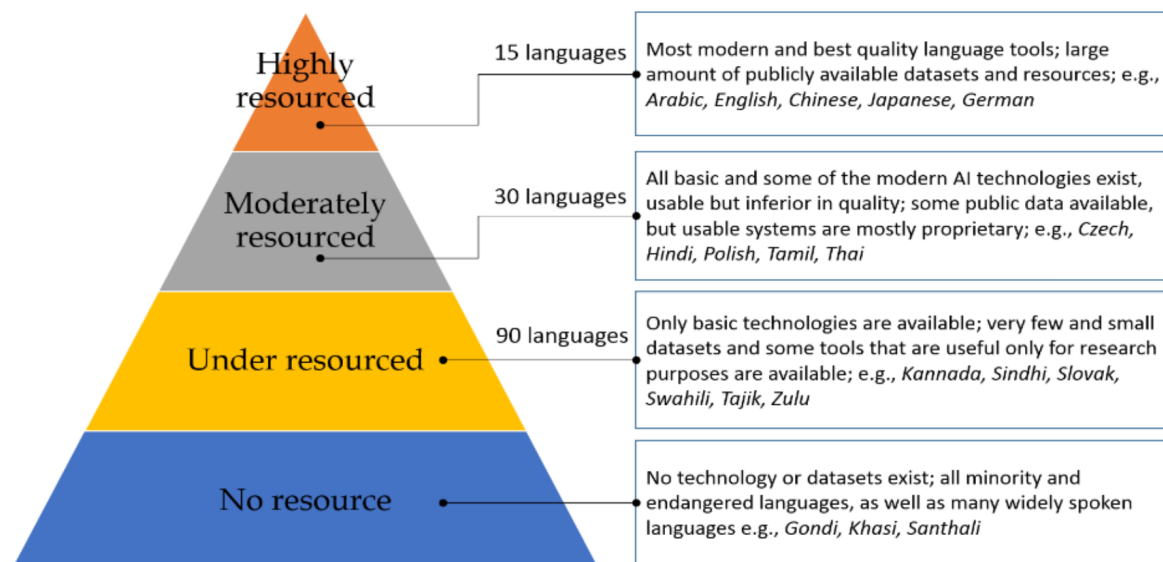


Fig. 1: Classification of languages according to the availability of language technology, tools and resources

### 3.1 Machine translation

Machine Translation (MT) is defined as the automatic conversion of a sequence of symbols in one language to a sequence of symbols in another language. It has evolved through years from rule-based to statistical approaches, which modeled the probabilities of mappings between sub-phrases between translations. These probabilities are learned in a statistical fashion from parallel texts where sentence-aligned translations are available in the languages involved (referred as source and target languages). The diagram below illustrates the modelling of translating the word “sure” in English to Spanish using translations made in the UN Parliament.

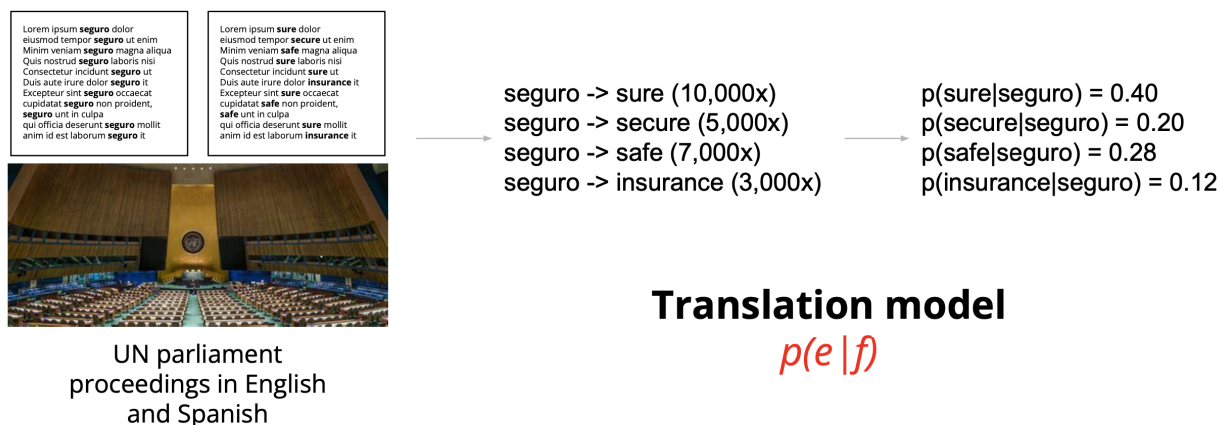


Fig. 2: Extracting statistics of translations from parallel data

Machine translation services like Google Translate and DeepL have made their way into reliable tools for translators and also regular folk in the recent years. This is largely due to the advancement of *deep learning* techniques that revolutionized the artificial intelligence field. This new way of modelling introduced in 2014 made 50% fewer word order mistakes, 17% fewer lexical mistakes, 19% fewer grammar mistakes compared to earlier models.

The uses of machine translation are as follows:

1. **Assimilation**, emulating a certain document in another language. This use-case enables e.g. reading a news site or technical paper in a language that we don't understand. We know that it's not a 100% accurate translation, but it gives the gist to explore further more.
2. **Communication**, enabling the communication between individuals and organizations e.g. in chat, tourism, e-commerce, lowering the need for a lingua franca.
3. **Monitoring**, enabling tracing of information in large-scale multilingual documents e.g. discovering international trends in Twitter.
4. **Assistance**, improving translation workflows e.g. computer-assisted translation, post-editing.

MT has also become an essential tool in language learning. [A recent work by Duke University](#) studies their usage among university level language learners beside other classic tools like dictionaries and thesauri. They report that 76% of the students enrolled in a Spanish class use web-based MT tools like Google Translate during their studies.

Finally, MT has also been proposed as a documentation and preservation tool for endangered languages by Bird and Chiang in their paper [Machine translation for language preservation](#). Directly quoting from their paper: "... when source texts are translated into a major world language, we guarantee that the language documentation will be interpretable even after the language has fallen out of use. Second, when a surviving speaker can identify errors in the output of an MT system, we have timely evidence of those areas of grammar and lexicon that need better coverage while there is still time to collect more. These tasks of producing and correcting translations can be performed by speakers of the language without depending on the intervention of outside linguists. Furthermore, we sidestep the need for linguistic resources like treebanks and wordnets, which are expensive to create and which depend on the existence of morphological, syntactic, and semantic analyses of the language."

This innovative way of language documentation reduces the effort into translated sentence collection as MT development relies on this type of data. (*More on parallel data in the next section*)

## 3.2 Automatic speech recognition

Automatic speech recognition (ASR) is the conversion of speech in its acoustic form into a symbolic form such as words or letters. It is the probabilistic modelling of the question "What is the most probable word sequence among all possible word sequences given an acoustic input?". Diagram below illustrates this process. Speech signal captured by a microphone is first encoded into a sequence of acoustic feature vectors. Following, the acoustic feature vectors are decoded into the words that represent the linguistic information that lies in the speech signal.

Developing an automatic speech recognition system for a language is dependant on the following type of data:

1. Collection of short speech audio recordings from many speakers and their text transcriptions
2. A large text corpus
3. Phonetic pronunciation dictionary (This is optional in more modern technologies)

ASR has progressed significantly in the last decade again thanks to the advent of deep-learning. In September 2017, Microsoft announced [their results](#) for an English speech recognition system that could achieve better-than-human performance in speech transcription. Their system was based on a dataset of 200M transcribed words from conversational speech. These developments have had great impact already as virtual assistants have become a cotidian application, voice search and automatic audio transcription.

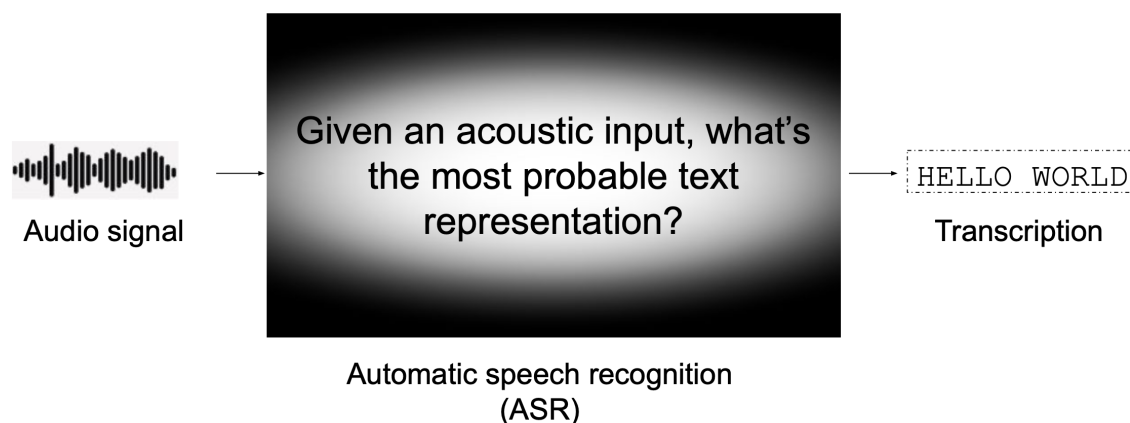


Fig. 3: A simple diagram of automatic speech recognition

### 3.3 Text-to-speech

Text-to-speech (or speech synthesis) involves production of a human-like speech given a text input with a computer. Before the advent of deep learning, there were two main approaches to text-to-speech (TTS) synthesis: concatenative TTS, and parametric TTS. Concatenative TTS, also called unit selection, combines short pre-recorded audio clips called units to synthesise the desired text. Concatenative TTS can provide a good performance in terms of speech quality but the cut and stitch procedure means a lack of naturalness. Parametric TTS relies on statistical methods by generating speech with a combination of parameters like F0 and energy, modelling the human speech production.

Currently, most modern TTS systems rely on deep-learning methods. The deep neural networks are trained using a large amount of recorded speech and the associated text transcriptions. In contrast to ASR training data, they are usually collected from a single speaker. The resulting TTS system is able to replicate the voice of this particular speaker.

TTS is important in making computers accessible to blind or partially sighted people as it enables them to “read” from the screen. TTS technology can be linked to any written input in a variety of languages, e.g. automatic pronunciation of words from an online dictionary, reading aloud of a text, interface for a voice assistant etc.

In the case of endangered and minority languages, TTS can aid language learning and language documentation. Students who don’t have access to speakers can study how a sentence is pronounced without assistance from a tutor. It is a permanent record of the language as it will persist even after the moment there are no speakers left for the language.



Fig. 4: This document was created with the financial support of the European Union. The content of this website is the sole responsibility of Col-lectivaT and SKAD and does not necessarily reflect the views of the European Union.



Fig. 5: This project is funded by the European Union.





## LANGUAGE DATA

Artificial intelligence tools opens up a new area for language resource creation for endangered and minority languages. Compared to the “classic” language resources created to preserve languages like lexica, grammar documentation, language maps etc., these require less linguistic expertise but are usually only useful in large volumes.

In this section we will explain the types of data that power the creation of artificial intelligence-based language technology explained in the previous chapter. Also, we will cover some ways to collect them and get the most out of them even if they are not of large volumes usually required by these applications.

### 4.1 Text corpus

In linguistics, a corpus (plural corpora) or text corpus is a language resource consisting of a large and structured set of texts in a language in digital format. They are useful in corpus linguistics for doing statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory. For language technology, they are an essential part in creating statistical language models that are used in applications such as optical character recognition, handwriting recognition, machine translation, spelling correction and assisted writing.

Text corpora by themselves are of the type *unlabeled data*. That is, they are a mere collection of data (in this case text) without any annotations or labelling. Language models store the probabilities of sequences words in order to get an “understanding” for the language. Also, text corpora can be annotated with following information in order to create *labeled data* for different NLP tasks:

- **Part-of-speech** (Noun, verb, Adjective etc.)
- **Named entities** (Person, location, personally identifiable information, organization, time etc.)
- **Lemmas** (roots of the words e.g. break for broken)
- **Dependency and phrase structure** (syntactic tree)

#### 4.1.1 Sourcing text corpora

The most common way of sourcing text corpora is through *crawling* the world wide web. This technique parses the whole web for collecting text in a certain language or many languages at once. Wikipedia publishes its content in different languages which can be used to create text corpora. [Common Crawl](#) initiative collects web site data and freely provides petabytes of data. OSCAR distributes this data classified into 166 languages.

Another common resource used by resourceful languages are books. [BookCorpus](#) consists of 11,038 books from the web containing 74 Million sentences and 984 Million words and is known to have powered many influential language models by big tech companies.

---

**Note:** Language models that are created from data in the wild represent what they see, and nothing else. Language in the web and books contain as well biases, toxic language which eventually gets replicated in these models. For an analysis of potential risks of building language models out of big language corpora, refer to [this paper](#) by Bender et al.

---

## 4.2 Parallel data (bitext)

The type of data that is needed to build a machine translation system is parallel data, which consists of a collection of sentences in a language together with their translations. Historically, parallel data were sourced from translations in multilingual public spaces like United Nations, European Parliament. Now, the greatest resource of parallel text is the multilingual web.

In order to train machine translation models it is not enough just to have translated documents. The texts need to be segmented to sentences and aligned. Parallel text alignment is the identification of the corresponding sentences in both sides of the parallel text. The resulting documents either have to correspond line by line or contain the original sentences and their translations in the same line. [Hunalign](#) helps in creating sentence alignments from translated documents. Translation memories (TMX files) also make great parallel data as they are already sentence segmented.

### 4.2.1 Sourcing parallel data

[OPUS](#) is a collection of almost all publicly available parallel data. It is the go-to point for many researchers to publish their parallel data or source data for development of MT models.

Some common sources for parallel data are: - Multilingual web sites (e.g. international news outlets), - Movie subtitles (see [OpenSubtitles](#)), - Holy texts, - Parliament proceedings, - Software localization data.

### 4.2.2 Crowdsourcing parallel data with Tatoeba.org

Tatoeba is a free collection of example sentences with translations geared towards foreign language learners. It is written and maintained by a community of volunteers through a model of open collaboration. It is hosted by Tatoeba Association, a French non-profit organization funded through donations. It currently holds 10,397,308 sentences in 412 supported languages.


Users can search for words in any language to retrieve sentences that use them. Each sentence in the Tatoeba database is displayed next to its likely translations in other languages; direct and indirect translations are differentiated. Sentences are tagged for content such as subject matter, dialect, or vulgarity; they also each have individual comment threads to facilitate feedback and corrections from other users and cultural notes. Sentences can be browsed by language, tag, and other criteria.

Registered users can add new sentences or translate or proofread existing ones, even if their target language is not their native tongue. However, users are encouraged to add original sentences or translations in their native or strongest language.




The entire Tatoeba database is published under a Creative Commons Attribution 2.0 license. It is also very easy to download parts of corpora in monolingual or parallel format from its [downloads page](#).

Random sentence

Sentence #1533839 — belongs to GrizaLeono




Ĉu vi scias, kiu verkis tiun romanon?








Translations


>






Weißt du, wer diesen Roman verfasst hat?


>






Μήπως γνωρίζετε ποιος έγραψε αυτό το μυθιστόρημα;

>




Sais-tu qui a écrit ce roman ?








Translations of translations


>






Tezriđ anwa i yuran ungal-a ?

>



Weißt du, wer diesen Roman geschrieben hat?

SHOW 7 MORE TRANSLATIONS

Fig. 1: A sentence and its translations from Tatoeba

## 4.3 Speech corpus

A speech corpus is a collection of speech audio files usually accompanied with their text transcriptions. In speech technology, speech corpora are used to create acoustic models for tasks like automatic speech recognition, text-to-speech synthesis and also speaker identification.

Speech corpora can contain read (e.g. audiobooks, news, read numbers and words) or spontaneous speech (dialogues). Corpora adequate for training ASR models contain samples from as many speakers as possible and in various acoustic settings (e.g. noisy, from far). In contrast, training data for TTS contains usually recordings from one speaker in an acoustically optimal setting.

OpenSLR lists many publicly available speech corpora.

### 4.3.1 Common Voice

Common Voice is a crowdsourcing project started by Mozilla to create a free database for making speech recognition accessible to everyone. The project is supported by volunteers who record sample sentences with a microphone and review recordings of other users. The voiced samples are released in regular intervals under the public domain license CC0 ([public domain](#)). This license ensures that developers can use the database for voice-to-text applications without restrictions or costs.

---

**Note:** As of May 2022, Common Voice supports 63 languages with 68 new on the way. Check here the current list of languages: <https://commonvoice.mozilla.org/en/languages>.

---

### 4.3.2 Adding a language to Common Voice

Common Voice works as a community platform where each language has their own community. The procedure for adding a new language into Common Voice is as follows:

1. **Find a community manager for the language** ([Information on roles](#))
2. **Localization request to Mozilla** This is done using [this template](#) on their github page. This will start the localization process of Common Voice to the desired language by placing it on Pontoon.
3. **Localization on Pontoon** ([user manual](#)) Every string on Common Voice platform needs to be translated to the language respecting the style guide. In total there are 663 strings. Translations can be made by any speaker who registers to the platform, but they need to be reviewed by the community manager.
4. **Sentence collection** A minimum of 5000 public domain sentences needs to be collected and entered to [Common Voice sentence collector](#).
5. **Reviewing sentences** Each collected sentence needs to be reviewed manually by at least two users on sentence collector.
6. **Wait for next CV release** Once localization is complete and there are 5000 reviewed sentences, next CV release should contain your language. Releases are done twice a month with schedules listed on [their github repository](#).



Fig. 2: Recording a Kurmanji Kurdish sentence on Common Voice

### 4.3.3 Found data

It is also possible to source voice data from broadcast radio shows, movies and other recorded material like interviews. This type of data is called “found data” as it is not originally intended to serve for building voice technology but it is *repurposed* to do so. Found data requires to be processed in order to obtain short audio segments and their transcriptions.

### 4.3.4 Sources

- Text corpus in Wikipedia
  - Tatoeba in Wikipedia
  - Speech corpus in Wikipedia
  - Common Voice in Wikipedia
- 



Fig. 3: This document was created with the financial support of the European Union. The content of this website is the sole responsibility of Col-lectivaT and SKAD and does not necessarily reflect the views of the European Union.



Fig. 4: This project is funded by the European Union.

## CASE STUDIES

In this section, we list some language technology-related initiatives and works that we deem exemplary or inspiring for endangered and minority languages. We will start with describing the project which gave fruit to this document “Judeo-Spanish: Connecting the two ends of the Mediterranean” and continue with projects that involve Anatolian, Iberian and African languages.

### 5.1 Judeo-Spanish: Connecting the two ends of the Mediterranean

Col·lectivaT and Sephardic Centre of Istanbul have come together for this project to carry out a diverse set of activities ranging from social media content creation to development of advanced language technology that assist Judeo-Spanish (Ladino) to the digital age. It also aimed to form awareness on this language as a common heritage between Turkey and Spain.

#### 5.1.1 Creation of audio-visual content for social media

The project created short language learning videos that would help gain visibility in social-media platforms and attract young generations to learn Ladino. In these short videos, a Judeo-Spanish phrase is presented with its translation in Turkish, English and Spanish with an audio helping to learn its pronunciation.

#### 5.1.2 Ladino Data Hub and open language datasets

The project launched [Ladino Data Hub](#) which will act as a centralized web archive dedicated to host Ladino language data and other resources that help document Sephardic culture. It aims to enable researchers, journalists from all over the world to access and share datasets that would help boost research and development for Ladino.

The project created and repackaged already existing datasets and shared in this portal. These are:

- A text corpus consisting of sentences crawled from [Şalom newspaper](#).
- A parallel text corpus with audio that contains Una Fraza al diya sentences and their audio
- Ladino speech corpus created by SKAD
- Clean speech synthesis training dataset consisting of read speech material by Karen Şarhon
- A Spanish Ladino lexicon based on dictionary by Güler, Portal i Tinoco
- Parallel texts in Ladino English and Turkish from translations made within SKAD
- Synthetically produced parallel corpus for training baseline MT models

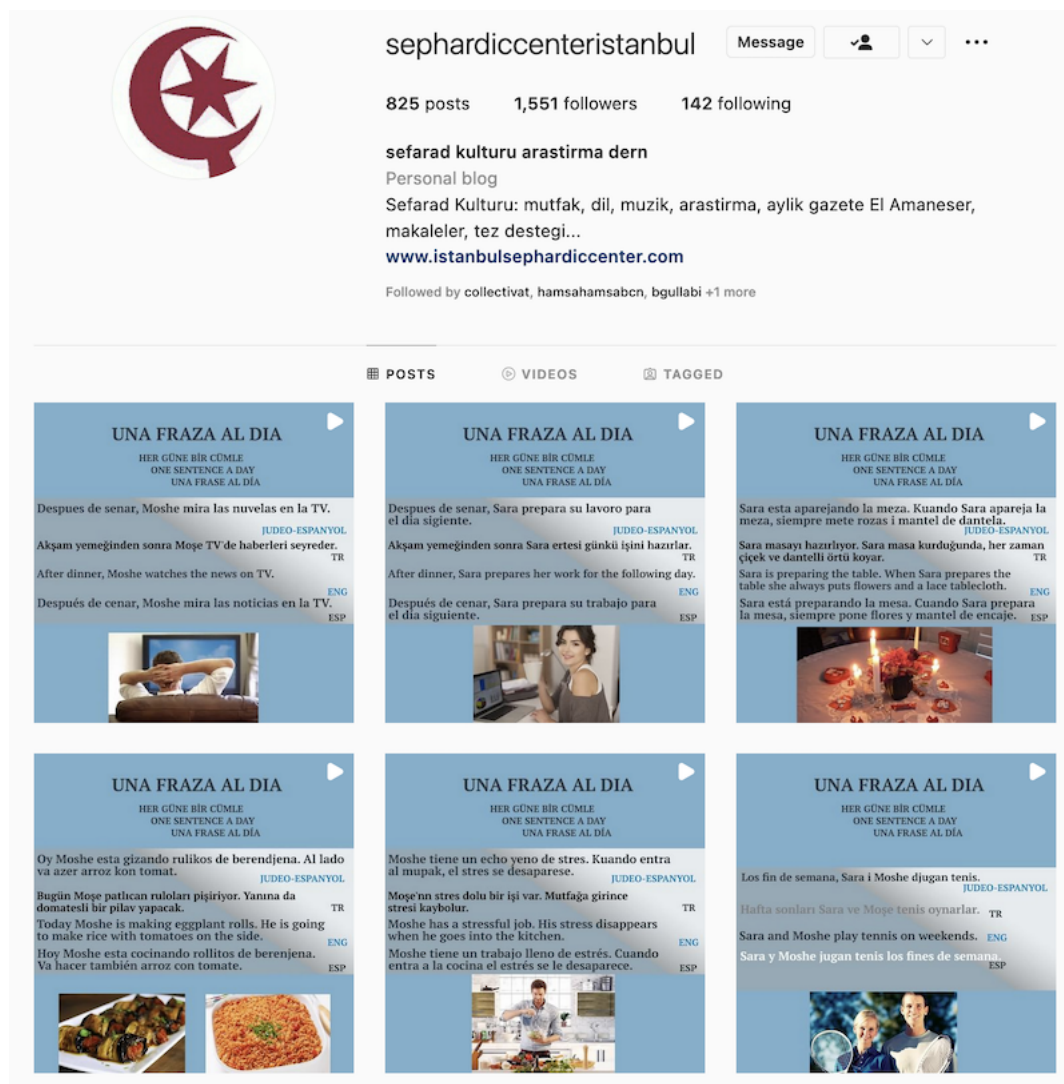


Fig. 1: Promoting Ladino on SKAD's instagram page with Frazza del diya (One sentence a day) videos



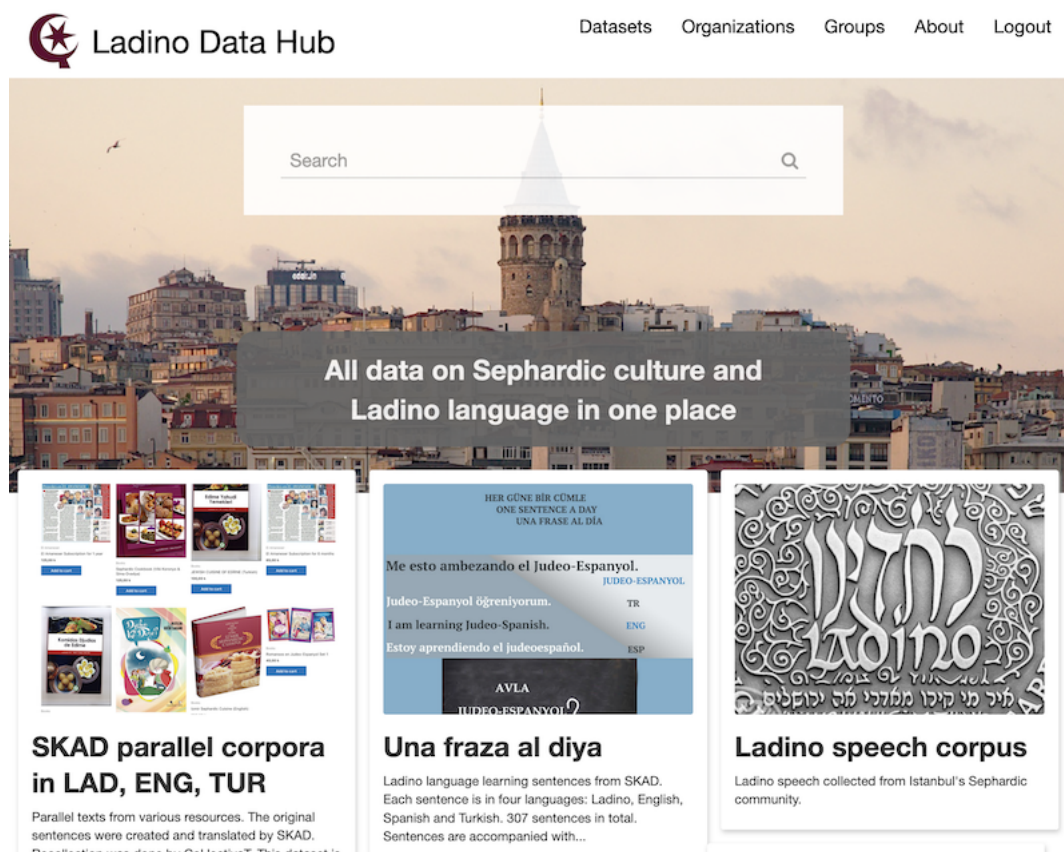


Fig. 2: Ladino Data Hub hosts data related to Ladino and Sephardic culture

### 5.1.3 Web application for machine translation and speech synthesis

The final and most important output of the project is a web application that is able to translate between Ladino and three related languages Turkish, Spanish and English. The objective is to help language learners, researchers and linguists who want to study Judeo-Spanish. The machine translation back-end was built with the help of a [rule based machine translation system](#) that can convert from Spanish to Ladino, getting use of the similar syntax between languages but changing orthography and vocabulary with a set of rules derived from dictionaries and grammar books. The web app is also able to synthesize Ladino sentences with a TTS application which was built with the [TTS training dataset](#).

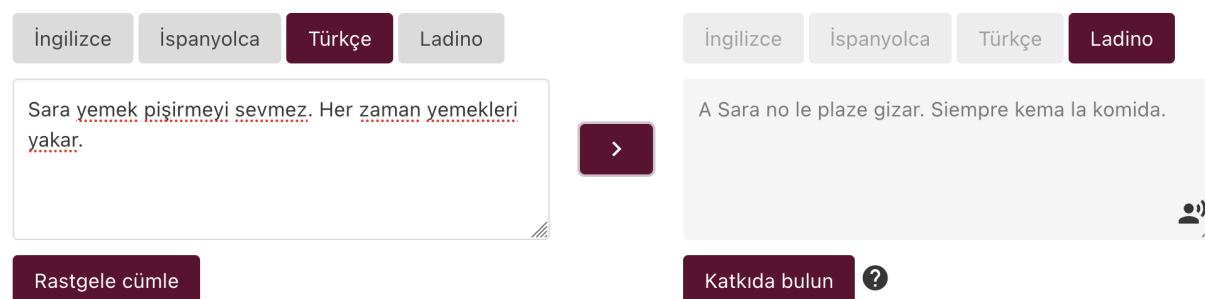


Fig. 3: Ladino translation web app with speech synthesis

The web application also allows contribution of parallel data. Users can load a random sentence and submit their corrected translation in order to extend the parallel data for Ladino.

**Note:** For a detailed technical report of this project, please refer the paper “Preparing an Endangered Language for the Digital Age: The Case of Judeo-Spanish” presented in the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia ([EURALI](#)): *Link to be placed soon*

## 5.2 Grassroots NLP communities

Against the NLP research focusing on a handful of languages, grassroots research communities organize together to bring the world’s languages into the forefront of technology. Two examples of these initiatives are [Masakhane](#) and [Turkic Interlingua](#).

As defined in their web-page, “Masakhane is a grassroots organisation whose mission is to strengthen and spur NLP research in African languages, for Africans, by Africans.” It’s an open-for-all initiative for “building together” as the meaning of their name suggests. Many concurrent activities take place by African and non-African researchers all over the world with the objective of representing the 2000 languages of Africa in language technology research. Some highlighting work from Masakhane are:

- [Masakhane translator](#) supporting 6 African languages: Yoruba, Shona, Lingala, Swahili, Tshiluba
- [Lanfrica](#) cataloguing African language resources to counter difficulties encountered in discovering African language works
- [BibleTTS](#) unlocking the development of high-quality Text-to-Speech models for ten languages spoken in Sub-Saharan Africa: Ewe, Hausa, Kikuyu, Lingala, Luganda, Luo, Chichewa, Akuapem Twi, Asante Twi, Yoruba.
- [Machine translation for preserving Oshiwambo language and culture](#)

- [MasakhaNER Know our names](#) creating hand-crafted named entity recognition (NER) datasets for various African languages

Masakhane also organizes annual workshops for publishing research related to African NLP and participates in data collection funds like [Lacuna](#).

**Turkic Interlingua (TIL)** is a “community of researchers, engineers, language enthusiasts and community leaders whose mission is to develop language technologies (from spell checkers to translation models), collect diverse datasets, and explore linguistic phenomena through the lens of academic research for Turkic languages” like Altai, Azerbaijani, Bashkir, Shor, Crimean Tatar, Chuvash, Gagauz, Karakalpak, Khakas, Kazakh, Karachay-Balkar, Kumyk, Kirghiz, Sakha (Yakut), Salar, Turkmen, Turkish, Tatar, Tuvian, Uighur, Urum, and Uzbek.

## 5.3 Common Voice campaigns

Various language communities have embarked on mobilizing participation into Common Voice. These initiatives are prepared by groups ranging from single individuals to local governments. Some examples are:

- [Common Voice Türkçe](#)
- [Kurdish crowdsource](#)
- [Igwebuike community](#) for Igbo
- [Projecte AINA](#) for Catalan

We also would like to give a special mention to the Catalan community for their contribution to Common Voice. Being a stateless minority language in Spain, it’s the 4th largest language (as of May 2022) in Common Voice thanks to the amazing contributions by activists and also a mobilization campaign by the [AI initiative of Catalan local government](#).



Fig. 4: Billboard with “It’s time the Internet speaks Catalan” being displayed on New York’s Times Square (photo by Aina Martí)

## 5.4 Other initiatives

Some other initiatives worth mentioning:

- Col-lectivaT has created [open speech data and text corpora for Catalan](#) using public television broadcasts and parliamentary proceedings.
- [Catotron](#) is the first open-source, neural network-based speech synthesis engine for Catalan built with support from Department of Culture of Catalonia.
- Cebuano and Waray-Waray, languages of Phillipines, has one of the biggest Wikipedia pages thanks to competitive usage of machine translation ([source](#))
- Maori people have rejected both private and open-source initiatives for collecting voice data in their language in order to “maintain the right to self-determination” ([source](#))
- [A Manifesto for Open Language Technology](#)
- [ELLORA Enabling Low Resource Languages by Microsoft Research India](#)



Fig. 5: This document was created with the financial support of the European Union. The content of this website is the sole responsibility of Col-lectivaT and SKAD and does not necessarily reflect the views of the European Union.



Fig. 6: This project is funded by the European Union.

## GOOD PRACTICES

---

**Note:** This section will be finalized with the feedback collected in the workshops organized during the project.

---

While it is clear that a large number of languages in the world require intensive investment in resource creation for technology enablement, it seems highly unlikely that such an investment can be delivered readily and easily in a short span of time. Given these limited resources, language communities should be empowered to determine the future of their languages. In this document, we have presented how digital representation is part of this process.

For deciding where to start, we suggest adopting the methodology of 4-D design thinking of *Discover, Design, Develop and Deploy* as introduced by Bali et al. in their [ELLORA initiative](#). This user-centric approach is as follows:

1. Discover what is most needed by the language community,
2. Design for the users and their language giving attention to diversity of the language and avoiding an approach parting from a majority language,
3. Develop and deploy frequently in an iterative manner constantly improving and detecting failures from the beginning.

Even when the community hasn't developed a perspective of language technology development, it is good practice to keep the value of data in mind when doing language preservation activities. Some examples of these are:

- Organize events and create content to raise data-awareness in the language community,
- Introduce languages to crowdsourcing platforms,
- Organize datathons for language data collection,
- Translate folk tales and children's stories which reside in public domain,
- Store plain text or document versions of published material in order to create text corpora,
- Save and openly share translation memories for helping other translators and for creating parallel data,
- Store recordings of broadcast material (e.g. radioshows) and transcribe if possible so that they can be converted into speech data,
- Save content published in social media to a permanent place so that they don't get lost in timelines.

*Do you have suggestions or questions? Write us at [info-at-collectivat.cat](mailto:info-at-collectivat.cat)*

---



Fig. 1: This document was created with the financial support of the European Union. The content of this website is the sole responsibility of Col-lectivaT and SKAD and does not necessarily reflect the views of the European Union.

## LICENSE

This document is licensed with [Attribution 4.0 International \(CC BY 4.0\)](#)

---



Fig. 1: This document was created with the financial support of the European Union. The content of this website is the sole responsibility of Col-lectivaT and SKAD and does not necessarily reflect the views of the European Union.