RéseauURFIST

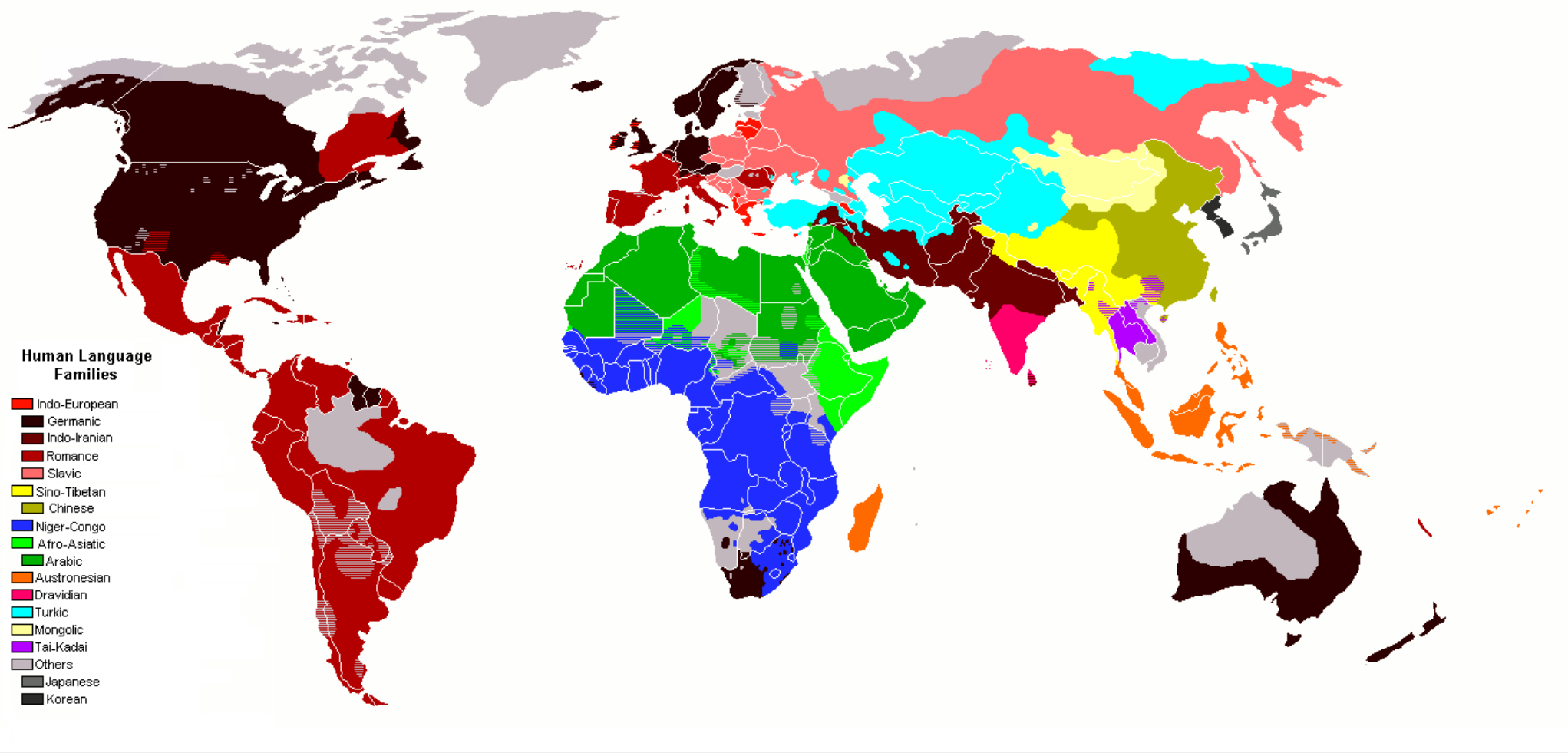
語音詞典

Recording voices and local languages with Lingualibre

Hugo Lopez
hugo.lopez@univ-toulouse.fr

Outline

- Context
- LinguaLibre & objectives
- Demo of the tool (10mins)
- Current progresses, limits & biases
- Q&A


Human Language Families
Indo-European
Germanic
Indo-Iranian
Romance
Slavic
Sino-Tibetan
Chinese
Niger-Congo
Afro-Asiatic
Arabic
Austronesian
Dravidian
Turkic
Mongolic
Tai-Kadai
Others
Japanese
Korean

Hugo Lopez

- Elearning and language professional
- Open education resources
- Teach new technologies to PhD
- Gascon, not French !

Attendence

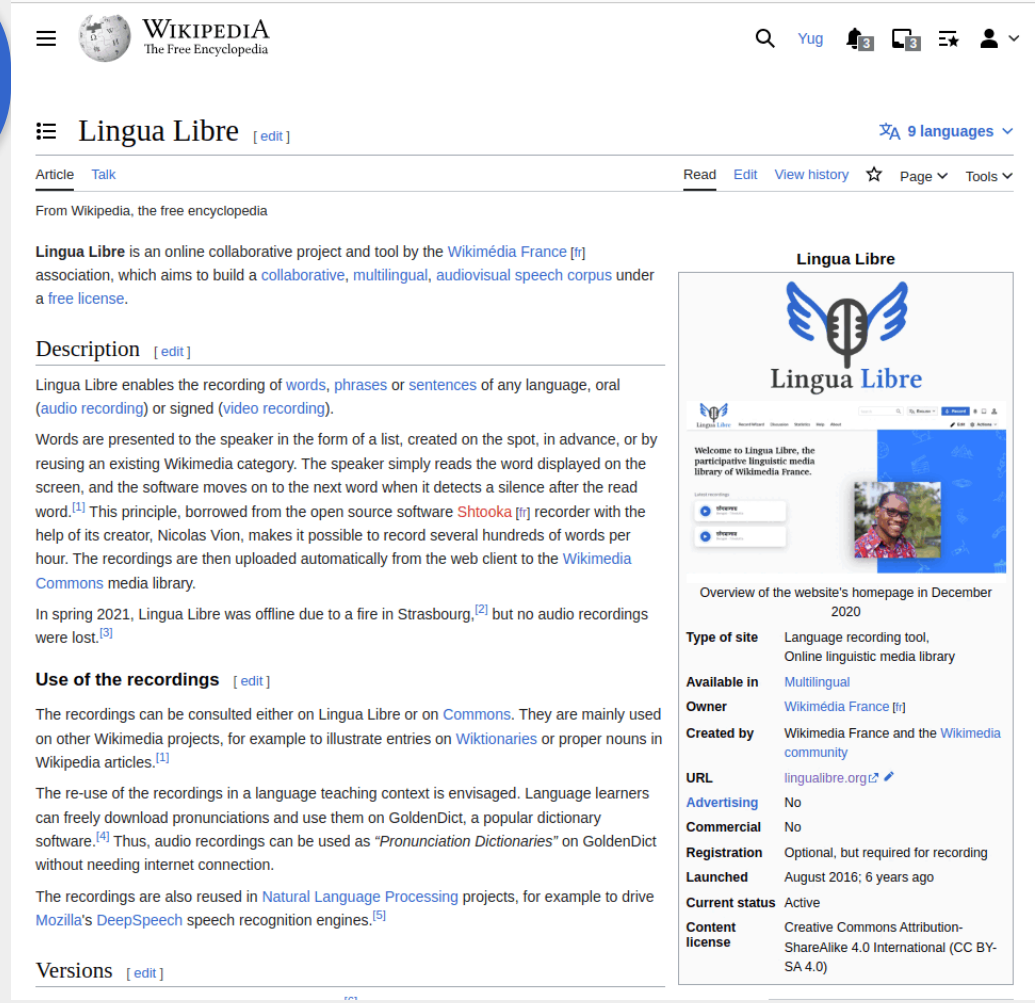
What affinity do you have with :

- Language diversity and minority languages ?
- Web applications ?
- Open source & its community management ?

LINGUA LIBRE

What is Lingua Libre

- Wikimedia's open source recording tool
- ...to document languages diversity
- 193 languages, 900,000 records
- For language e-learning services



WIKIPEDIA
The Free Encyclopedia

Yug

Lingua Libre [edit]

9 languages

Article Talk | Read Edit View history Page Tools

From Wikipedia, the free encyclopedia

Lingua Libre is an online collaborative project and tool by the Wikimédia France [fr] association, which aims to build a collaborative, multilingual, audiovisual speech corpus under a free license.

Description [edit]

Lingua Libre enables the recording of words, phrases or sentences of any language, oral (audio recording) or signed (video recording).

Words are presented to the speaker in the form of a list, created on the spot, in advance, or by reusing an existing Wikimedia category. The speaker simply reads the word displayed on the screen, and the software moves on to the next word when it detects a silence after the read word.[1] This principle, borrowed from the open source software Shtooka [fr] recorder with the help of its creator, Nicolas Vion, makes it possible to record several hundreds of words per hour. The recordings are then uploaded automatically from the web client to the Wikimedia Commons media library.

In spring 2021, Lingua Libre was offline due to a fire in Strasbourg,[2] but no audio recordings were lost.[3]

Use of the recordings [edit]

The recordings can be consulted either on Lingua Libre or on Commons. They are mainly used on other Wikimedia projects, for example to illustrate entries on Wiktionaries or proper nouns in Wikipedia articles.[1]

The re-use of the recordings in a language teaching context is envisaged. Language learners can freely download pronunciations and use them on GoldenDict, a popular dictionary software.[4] Thus, audio recordings can be used as *"Pronunciation Dictionaries"* on GoldenDict without needing internet connection.

The recordings are also reused in Natural Language Processing projects, for example to drive Mozilla's DeepSpeech speech recognition engines.[5]

Versions [edit]

Lingua Libre

Lingua Libre

Welcome to Lingua Libre, the participative linguistic media library of Wikimedia France.

Overview of the website's homepage in December 2020

Type of site	Language recording tool, Online linguistic media library
Available in	Multilingual
Owner	Wikimédia France [fr]
Created by	Wikimedia France and the Wikimedia community
URL	lingualibre.org
Advertising	No
Commercial	No
Registration	Optional, but required for recording
Launched	August 2016; 6 years ago
Current status	Active
Content license	Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

Language diversity

- Languages are core and center of our identity, human experience, memories, kindnes.
- 7,000+ languages
- 3,500 (half) will die out this century

Language diversity

- Languages are often segregated, politized
- Case of France vs Taiwan

Language diversity

> The diversity of world languages, their words, expressions, voices, are poorly documented and accessible. Lingualibre.org allows us to record languages vocabulary and audio dictionaries at large scale, in an easy and quick fashion (800 audio/hour).

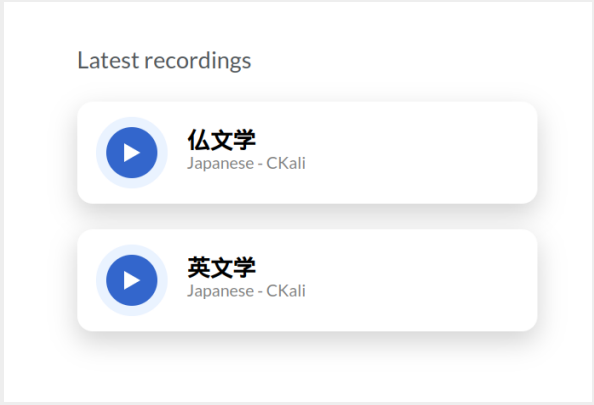
LINGUA LIBRE (10MINS)

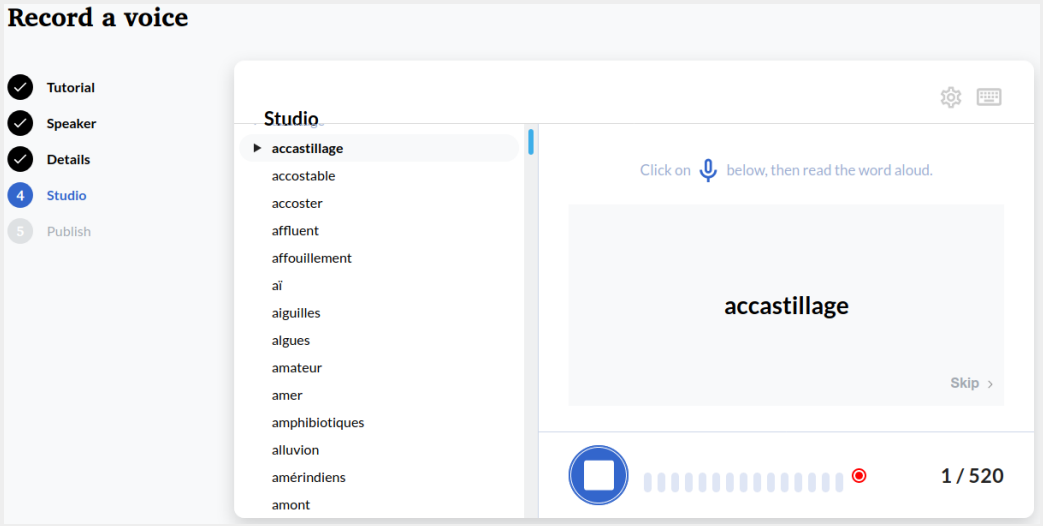
RéseauURFIST

Oral languages' learning chain

Learners ← **Speakers**

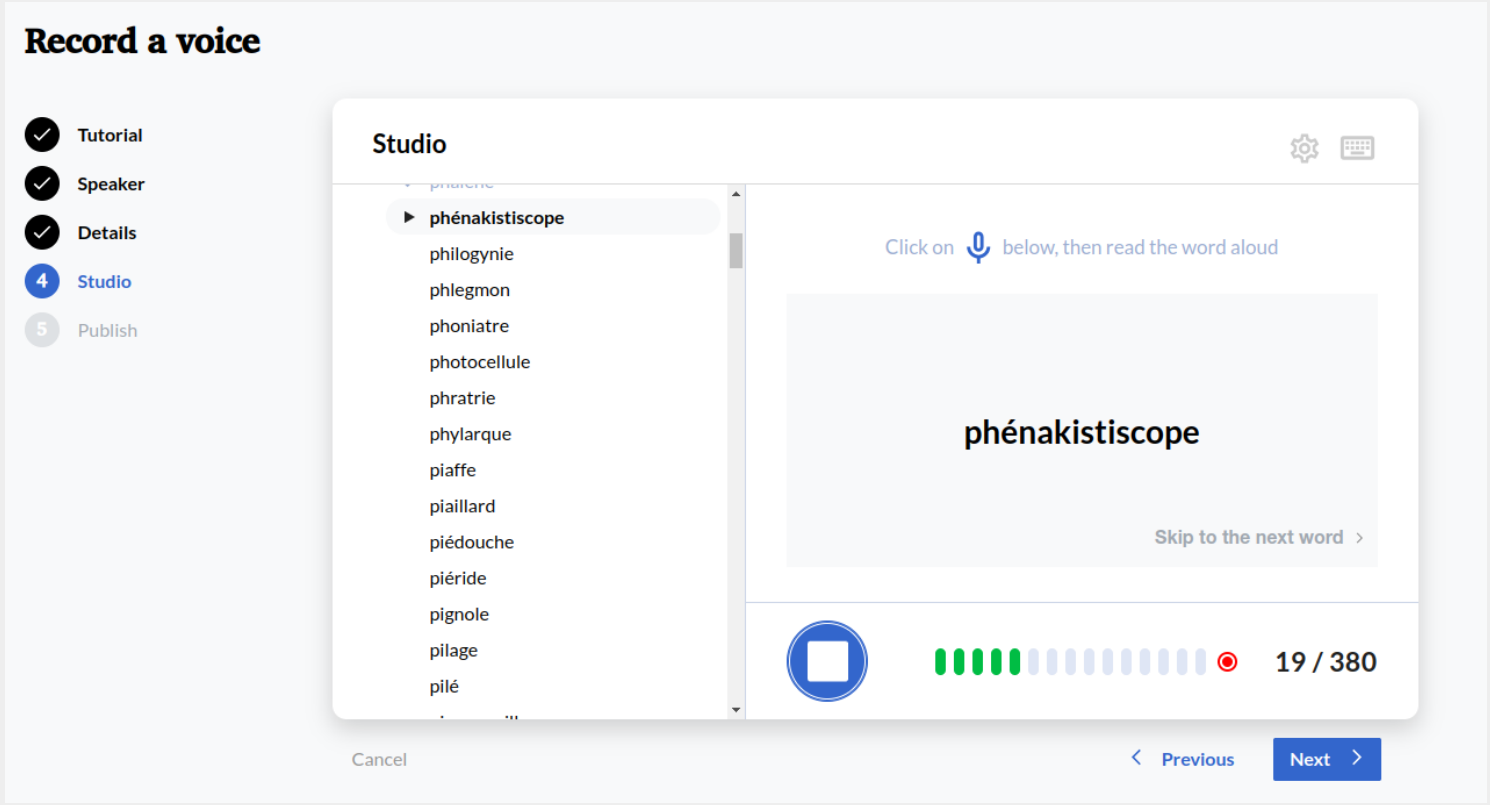
e-services | Data | Lingualibre Studio





Learning | Audio & text | Sharing

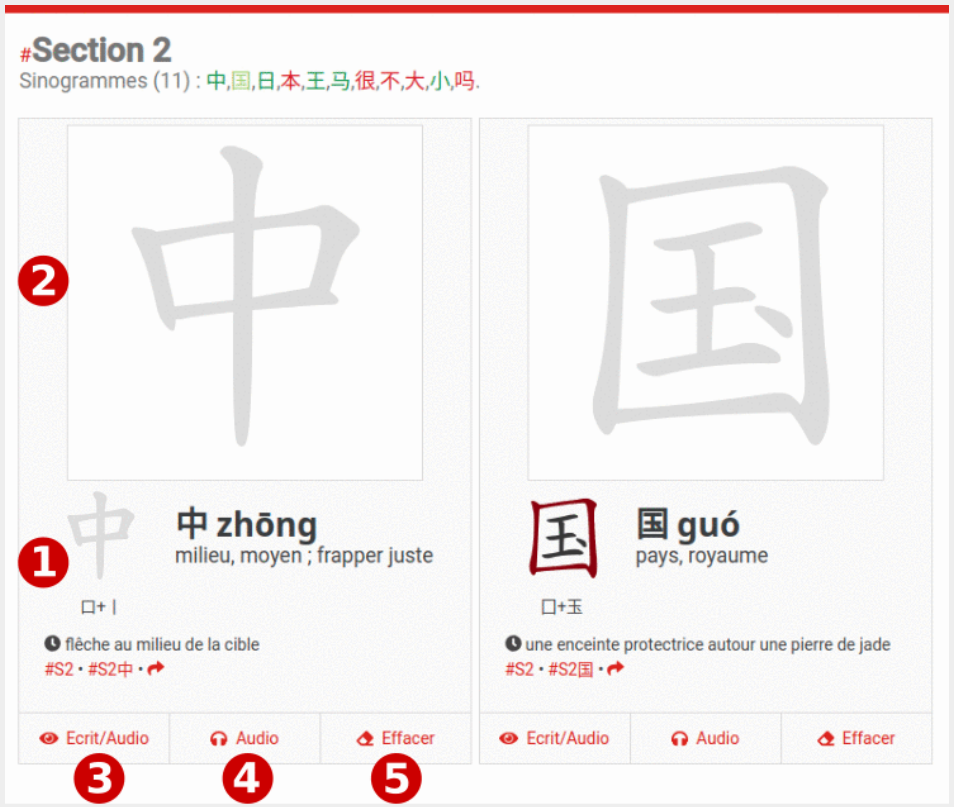
Lingua Libre: audio recording studio



Page: Lingualibre.org Recording Studio

RéseauURFIST

Lingua Libre: Apps



萌典 [edit source]

Add languages

Page Discussion 汉 漢 不转换 Tools

From Wikipedia

Open View it!

萌典是一部由台灣自由軟體程式設計師唐鳳開發的數位化漢語詞典，是台灣開源社群g0v零時政府的專案之一。作爲一部數位化漢語詞典，萌典除了收錄了十六萬筆的中華民國國語词条之外，還收錄了兩萬筆台灣閩南語、一萬四千筆台灣客家語词条，以及提供了漢語與英語、法語以及德語的對照。網站作者唐鳳將其以創用CC0協議釋放至公有領域[2]。除在線版外，萌典還提供有適用於Windows、macOS和Linux的桌面版， 以及使用於Android和iOS的手機版。

萌典

g0v.tw

网站类型 数位化漢語詞典

List: Lingua Libre Apps. Future: Moedict.tw ?

STATED OBJECTIVES

> Document languages diversity and voices.

- Languages
- Accents
- Voices
- Genders

PROGRESSES, LIMITS & BIASES

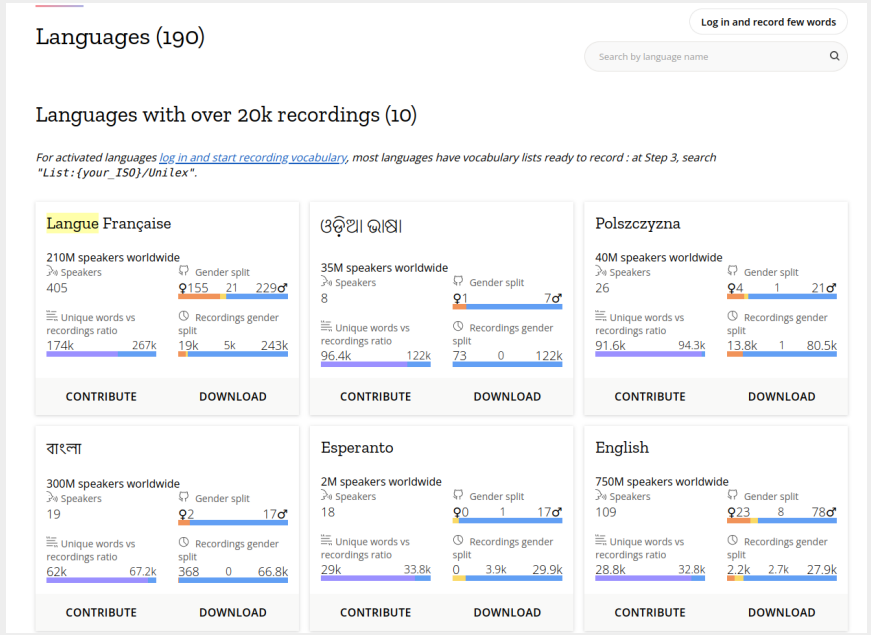
After 5 year and 900,000+ recordings, we would like to share past progresses, current analysis and future actions.

RéseauURFIST

In numbers

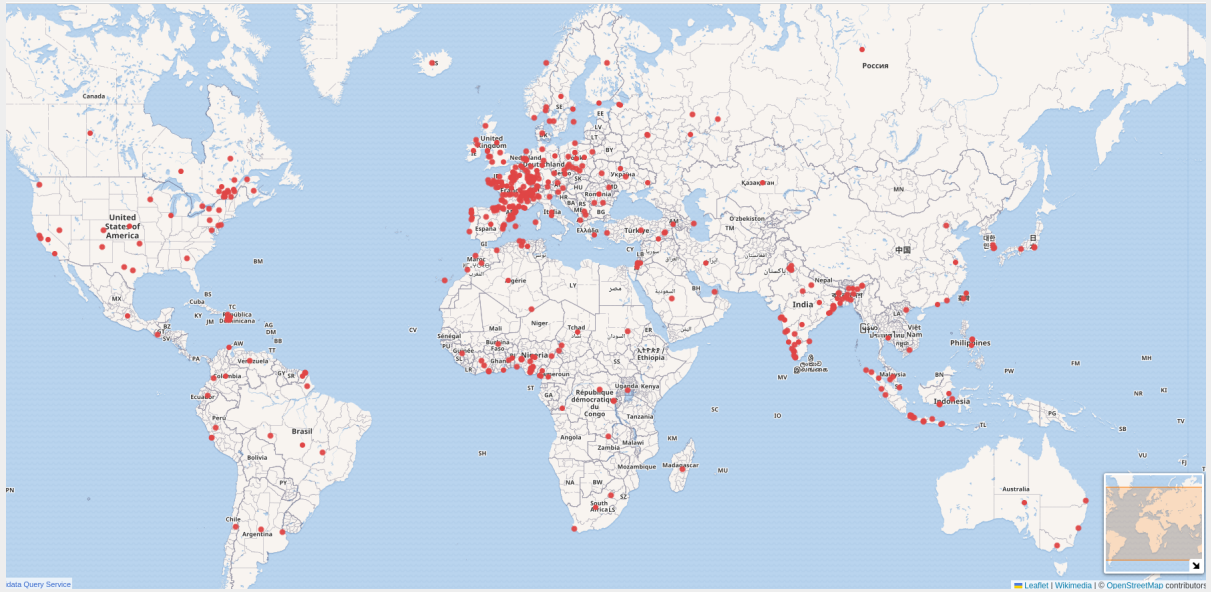
Production

Languages gallery



Contributors

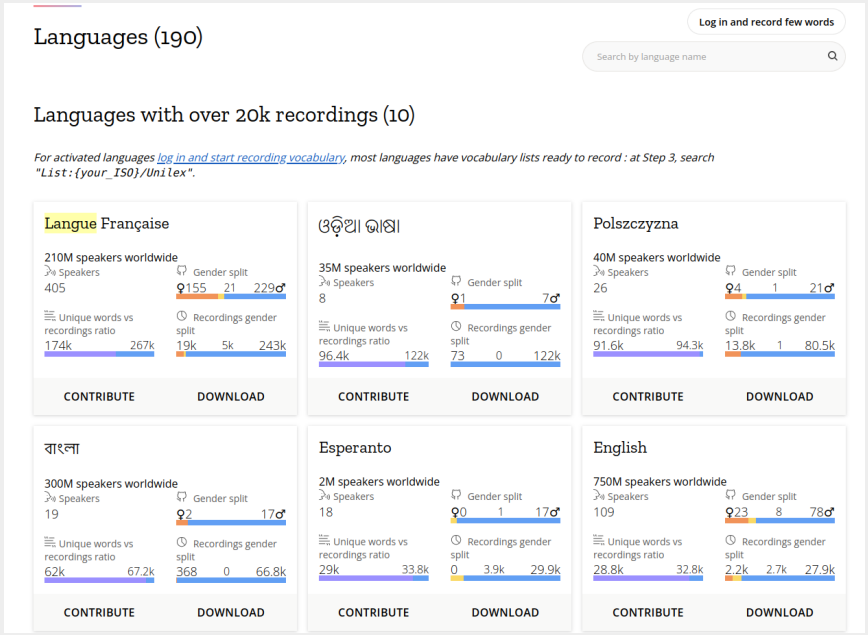
Lingua Libre map



In numbers

Production

Languages gallery



Reuses

2022 review

Lingua Libre Bot, March 2023 (g)

Local wiki	First edit	Edit count	%	Groups	Region of most beneficiaries
Wiktionaries					
fr.wiktionary.org	12 June 2018	308,193	54.7%	bot	Europe/France
ku.wiktionary.org	30 November 2021	42,820	7.6%	bot	Asia
oc.wiktionary.org	16 December 2018	20,606	3.7%	bot	Europe/France
shy.wiktionary.org	8 September 2021	1,930	0.34%	bot	Africa
or.wiktionary.org	10 January 2023	249	0.04%	—	Asia/India
All other projects	—	0	0%	—	World
Technical projects					
www.wikidata.org	10 June 2018	62,045		bot	Unclear
meta.wikimedia.org	10 June 2018	6		—	

Olafbot, March 2023 (g)

Local wiki	First edit	Edit count	%	Groups	Region of most beneficiaries
Wiktionaries					
pl.wiktionary.org	4 March 2020	189,157	33.6%	bot	Europe/Poland
Technical projects					
lingualibre.org	26 February 2021	5,208		bot	Unclear

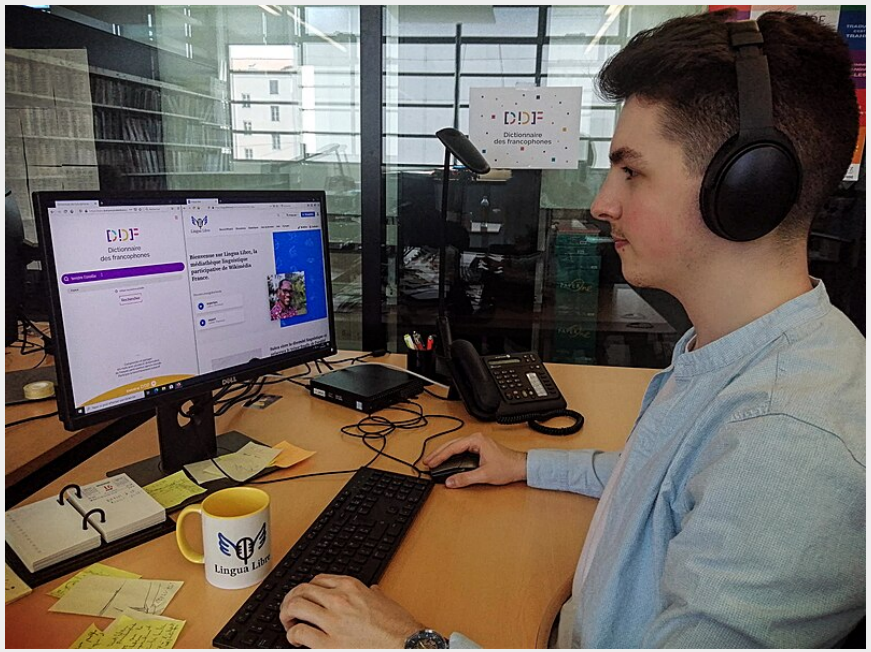
Qualitative

- Per language: large vs minorities
- Per gender
- Per age
- Per per area, income, etc.

RéseauURFIST

Languages typology & specifics

Large	Medium	Minorities
Resourced	Low resources	Written ?



RéseauURFIST

Languages typology & specifics

Large	Medium	Minorities
Resourced	Low resources	Written ?

Languages typology coverage

	World languages		Lili languages				
Demographic[1]	**Number**	**Ratio**	**Number**	**Coverage**	**Supported language's profile**	**Examples**	**Community's presence**
Major (>30M)	30	0.5%	20	66%	Mostly major Western or Indian languages.	FRA, SPA, BEN	**Solid:** Several productive speakers. Sustained or periodic.
Large (1~30M)	350	5%	100	33%	Mostly Western languages, other notable languages	NLD, AFR, CAT	**Emerging:** One productive speaker, few not-retained speakers. Fragile.
Marginalized (<1M)	6500	94%	40	<1%	Mostly larger minorities in Western countries.	ATJ, BRE, EUS	**Contact point:** No productive speaker, one not-retained speaker. Below fragile.

Key needs

- Mobile **e-dictionaries** for local communities
- For **revitalisation** ! Not documentation.
- Outreach to 6,500 local communities ?

Lingua Libre		[Collapse]
Lingualibre	Repositories	Record Wizard (Recording Studio) · SPARQL2DATA
	Documentations	{Helps}
	Technical helps	{Technicals} · Help:SPARQL series
	Reports	Winter 2021-2022 Public Relations Campaign · Lingua Libre/2022 wishlist#Approach · 2022 Review · Lingua_Libre/Supports/Melody#10. langues régionales et minoritaires_dans les autres pays · Lingua Libre/Supports
	Referents	Github, Phabricator, Userscripts (Yug, Poslovitch, Pamputt) · Bots (Pamputt, Poslovitch, Lepticed7, Yug) · Onboarding (Yug, WikiLucas, Pamputt) · Events, Outreach (Yug, Adélaide) · Reports (Yug, Poslovitch) · Funding (Adélaide)
Lingua Libre/SignIt	Repositories	Record Wizard (Recording Studio with video track) · Lingualibre/SignIt (Firefox Web Extension)
	Documentations	Minimal video recording tutorial for elegant signed videos
	Reports	2022.09.16 : 2022/Phase_1 light diagnosis · 2022.09.16 : 2022/Phase_2 volunteer coding report and detailed strategy · 2023.04.25 : 2023/Phase_1 freelance coding report and learning patterns · 2023.05-12 : 2023/Phase_2 outreach campaign challenges
	Referents	Édouard Lopez (concept design) · User:0x010C (creator, developer) · Yug (expansion, coordination)

Q&A

Keep in touch

Role	Contacts
Lead, dev	User:Yug
Wikimedia France	User:Adélaïde Calais WMFr
Home	Lingualibre.org
Code	github.com/lingua-libre

THANK !

Credits

The following Wikimedia Commons images have been used:

File	Licence	Author
Human_Language_Families.png	CC-BY-SA	JFDP13
WikiLucas00_à_l'Institut_international_pour_la_Francophonie.jpg	CC-BY-SA	WikiLucas00
Daramlagon_and_Maerui-sama_session_on_Bikol_Wiktionary_and_Lingua_Libre_03.jpg	CC-BY-SA	Daramlagon
LinguaLibre_2022_Paris_Surui_training-03.jpg	CC-BY-SA	Yug
Forom_des_langues,_Toulouse,_2023-01.jpg	CC-BY-SA	Yug
Séance_Lingua_Libre_à_Cotonou_en_Mars_2021_-_Photo_17.jpg	CC-BY-SA	Fawaz.tairou
Lingua_Libre_Atikamekw_at_Wikimania_2017_Montreal.jpg	CC-BY-SA	Benoit Rochon