

語音詞典

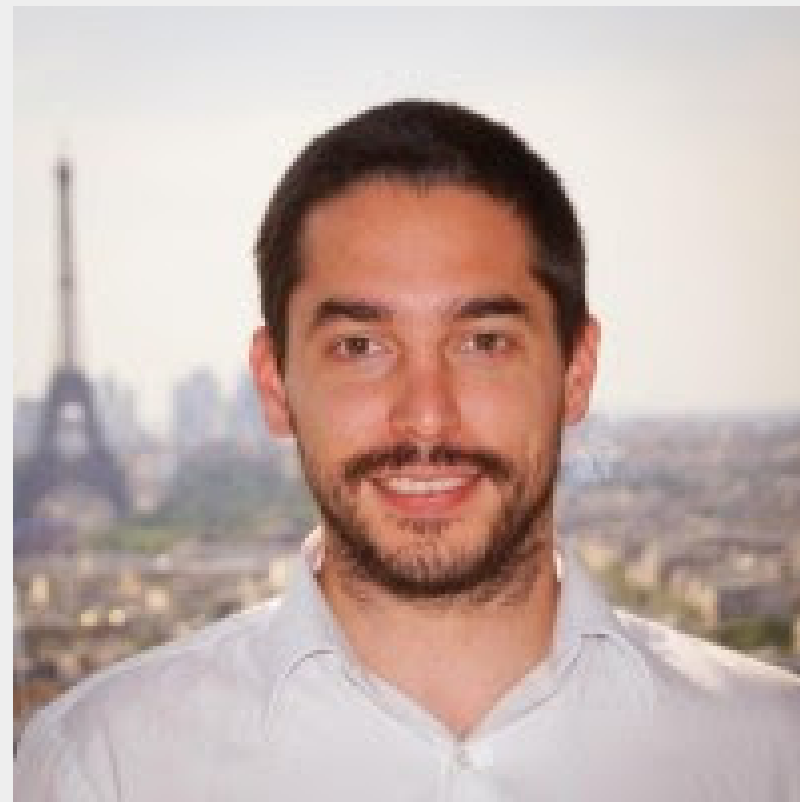
Recording voices and local languages with Lingualibre

Hugo Lopez

hugo.lopez@univ-toulouse.fr

Hugo Lopez

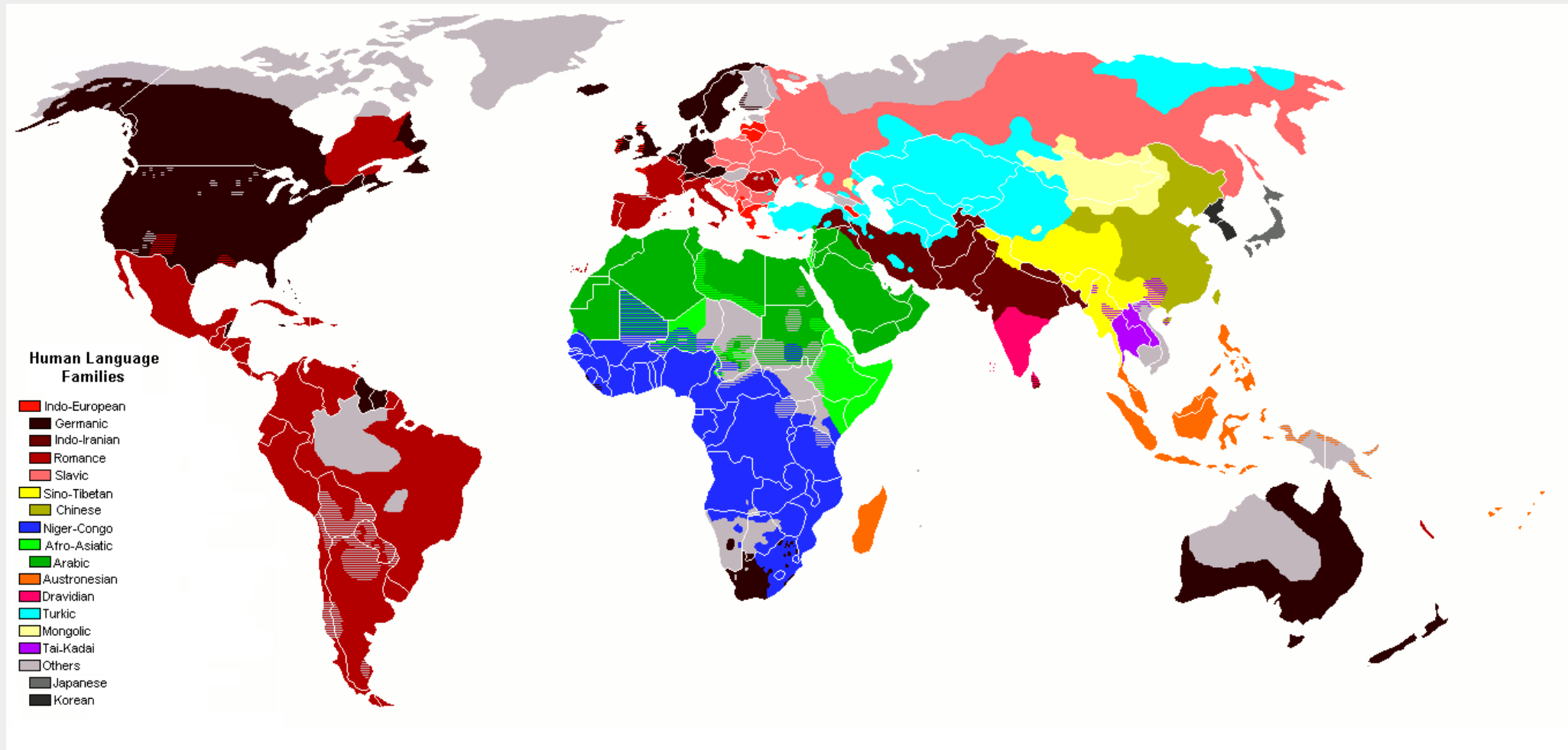
- Elearning and language professional
- Open education resources
- Wikimédian in University
- Occitan, not French.



Outline

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

LANGUAGES DIVERSITY



Languages diversity : dimensions

Document languages diversity and voices.

- Languages
- Accents
- Voices
- Genders

Languages diversity : mission

“

The diversity of our languages, their words, expressions, voices, are poorly documented and accessible. We want to record, share and make visible those expressions at large scale, in an easy and quick fashion (800 audio/hour).

”

LINGUALIBRE

What is Lingua Libre

The case of Alsatian !

- Wikimedia's open source recording tool
- ...to document Alsatian
- 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



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)

Oral languages' learning chain

Learners



Speakers

e-services

Data

Lingualibre Studio

排灣語

簡介

檢索 詞項列表 逐詞註解

精準查詢條件: 排灣語 中文為「穿」之查詢結果 (共1筆)

kemava

解釋1 穿衣服

- inika sun a kemava tua 'uba ayau, maya sa vuceleljan!
天氣這麼冷, 你還不穿大衣哦!

解釋2 穿(衣服)

- nu kemava / mitung a ku kaka pusaladjan ni kina.
弟弟穿衣服時媽媽會幫忙。

詳細 →

Latest recordings

仏文学
Japanese - CKali

英文学
Japanese - CKali

Record a voice

- Tutorial
- Speaker
- Details
- 4 Studio
- 5 Publish

Record a voice

Studio

- ▶ accastillage
- accostable
- accoster
- affluent
- affouillement
- aï
- aiguilles
- algues
- amateur
- amer
- amphibiotiques
- alluvion
- amérindiens
- amont

Click on below, then read the word aloud.

accastillage

Skip >

1 / 520

Learning

Audio & text

Sharing

DEMONSTRATION (10MINS)

Lingua Libre: audio recording studio


Record a voice

- ✓ Tutorial
- ✓ Speaker
- ✓ Details
- 4 Studio**
- 5 Publish

Studio



phénakistiscope

- philogynie
- phlegmon
- phoniatre
- photocellule
- phratrie
- phylarque
- piaffe
- piaillard
- piédouche
- piéride
- pignole
- pilage
- pilé

Click on  below, then read the word aloud

phénakistiscope

Skip to the next word >

  19 / 380

Cancel < Previous Next >

Page: [Lingualibre.org](https://lingualibre.org) Recording Studio

Lingua Libre: Apps

#Section 2
Sinogrammes (11) : 中,国,日,本,王,马,很,不,大,小,吗.

2



1 中 zhōng
milieu, moyen ; frapper juste
口+丨
flèche au milieu de la cible
#S2 · #S2中 · ↗

Ecrit/Audio Audio Effacer

3 **4** **5**

国 guó
pays, royaume
口+玉
une enceinte protectrice autour une pierre de jade
#S2 · #S2国 · ↗

Ecrit/Audio Audio Effacer

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。作為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
萌典
g0v.tw



網站类型 數位化漢語詞典

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

PROGRESSES, LIMITS & BIASES

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

Languages typology & specifics

Large

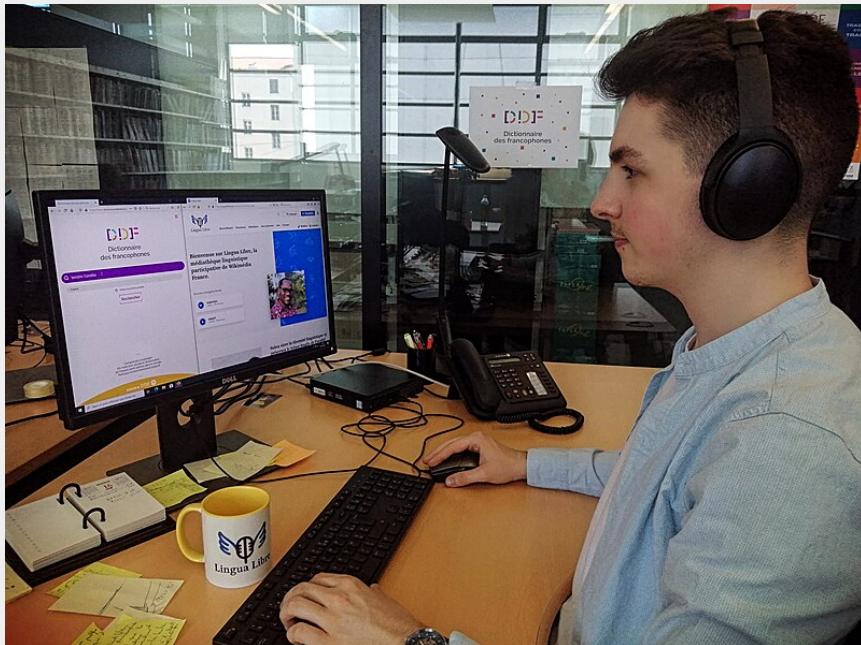
Medium

Minorities

Resourced

Low resources

Written ?



Languages typology & specifics

Large

Medium

Minorities

Resourced

Low resources

Written ?



In numbers

Production

Languages gallery

Contributors

Lingua Libre map

Log in and record few words

Search by language name

Languages (190)

Languages with over 20k recordings (10)

For activated languages [log in and start recording vocabulary](#), most languages have vocabulary lists ready to record : at Step 3, search "List::your_ISO/Unix".

<p>Langue Française</p> <p>210M speakers worldwide</p> <p>Speakers: 405</p> <p>Gender split: 9155 ♂ 21 229 ♀</p> <p>Unique words vs recordings ratio: 174k 267k</p> <p>Recordings gender split: 19k 5k 243k</p> <p>CONTRIBUTE DOWNLOAD</p>	<p>ଓଡ଼ିଆ ଭାଷା</p> <p>35M speakers worldwide</p> <p>Speakers: 8</p> <p>Gender split: 91 ♂ 7 ♀</p> <p>Unique words vs recordings ratio: 96.4k 122k</p> <p>Recordings gender split: 73 0 122k</p> <p>CONTRIBUTE DOWNLOAD</p>	<p>Polszczyzna</p> <p>40M speakers worldwide</p> <p>Speakers: 26</p> <p>Gender split: 94 ♂ 1 21 ♀</p> <p>Unique words vs recordings ratio: 91.6k 94.3k</p> <p>Recordings gender split: 13.8k 1 80.5k</p> <p>CONTRIBUTE DOWNLOAD</p>
<p>বাংলা</p> <p>300M speakers worldwide</p> <p>Speakers: 19</p> <p>Gender split: 92 ♂ 17 ♀</p> <p>Unique words vs recordings ratio: 62k 67.2k</p> <p>Recordings gender split: 368 0 66.8k</p> <p>CONTRIBUTE DOWNLOAD</p>	<p>Esperanto</p> <p>2M speakers worldwide</p> <p>Speakers: 18</p> <p>Gender split: 90 ♂ 1 17 ♀</p> <p>Unique words vs recordings ratio: 29k 33.8k</p> <p>Recordings gender split: 0 3.9k 29.9k</p> <p>CONTRIBUTE DOWNLOAD</p>	<p>English</p> <p>750M speakers worldwide</p> <p>Speakers: 109</p> <p>Gender split: 923 ♂ 8 78 ♀</p> <p>Unique words vs recordings ratio: 28.8k 32.8k</p> <p>Recordings gender split: 2.2k 2.7k 27.9k</p> <p>CONTRIBUTE DOWNLOAD</p>



In numbers

Production

Languages gallery

Reuses

2022 review

Languages (190)

Log in and record few words

Search by language name

Languages with over 20k recordings (10)

For activated languages [log in and start recording vocabulary](#), most languages have vocabulary lists ready to record : at Step 3, search "List: {your_ISO}/Unilex".

<p>Langue Française</p> <p>210M speakers worldwide</p> <p>Speakers: 405</p> <p>Gender split: 9155 (21) 229 (2)</p> <p>Unique words vs recordings ratio: 174k / 267k</p> <p>Recordings gender split: 19k (5k) 243k (243k)</p> <p>CONTRIBUTE DOWNLOAD</p>	<p>ଓଡ଼ିଆ ଭାଷା</p> <p>35M speakers worldwide</p> <p>Speakers: 8</p> <p>Gender split: 91 (7) 7 (0)</p> <p>Unique words vs recordings ratio: 96.4k / 122k</p> <p>Recordings gender split: 73 (0) 122k (122k)</p> <p>CONTRIBUTE DOWNLOAD</p>	<p>Polszczyzna</p> <p>40M speakers worldwide</p> <p>Speakers: 26</p> <p>Gender split: 94 (1) 21 (0)</p> <p>Unique words vs recordings ratio: 91.6k / 94.3k</p> <p>Recordings gender split: 13.8k (1) 80.5k (80.5k)</p> <p>CONTRIBUTE DOWNLOAD</p>
<p>বাংলা</p> <p>300M speakers worldwide</p> <p>Speakers: 19</p> <p>Gender split: 92 (17) 17 (0)</p> <p>Unique words vs recordings ratio: 62k / 67.2k</p> <p>Recordings gender split: 368 (0) 66.8k (66.8k)</p> <p>CONTRIBUTE DOWNLOAD</p>	<p>Esperanto</p> <p>2M speakers worldwide</p> <p>Speakers: 18</p> <p>Gender split: 90 (1) 17 (0)</p> <p>Unique words vs recordings ratio: 29k / 33.8k</p> <p>Recordings gender split: 0 (3.9k) 29.9k (29.9k)</p> <p>CONTRIBUTE DOWNLOAD</p>	<p>English</p> <p>750M speakers worldwide</p> <p>Speakers: 109</p> <p>Gender split: 923 (8) 78 (0)</p> <p>Unique words vs recordings ratio: 28.8k / 32.8k</p> <p>Recordings gender split: 2.2k (2.7k) 27.9k (27.9k)</p> <p>CONTRIBUTE DOWNLOAD</p>

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.



Languages typology coverage

Demographic ^[1]	World languages		Lili languages		Supported language's profile	Examples	Community's presence
	Number	Ratio	Number	Coverage			
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 ?

V · T · E		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, Adélaïde) · Reports (Yug, Poslovitch) · Funding (Adélaïde)	
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)	

Wikimedia Movement

Best global network to support languages diversity



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
File:Human_Language_Families.png	CC-BY-SA	JFDP13
File:WikiLucas00_à_l'Institut_international_pour_la_Francophonie.jpg	CC-BY-SA	WikiLucas00
File:Daramlagon_and_Maerui-sama_session_on_Bikol_Wiktionary_and_Lingua_Libre_03.jpg	CC-BY-SA	Daramlagon
File:LinguaLibre_2022_Paris_Surui_training-03.jpg	CC-BY-SA	Yug
File:Forom_des_langues,_Toulouse,_2023-01.jpg	CC-BY-SA	Yug
File:éance_Lingua_Libre_à_Cotonou_en_Mars_2021_-_Photo_17.jpg	CC-BY-SA	Fawaz.tairou
File:Lingua_Libre_Atikamekw_at_Wikimania_2017_Montreal.jpg	CC-BY-SA	Benoit Rochon