Lingua Libre

What is it and how to take part in it?

A workshop by [[User:WikiLucas00|Lucas Prégaldiny]]

WIKIMANIA

Lingua Libre

alias "LiLi"

1. The project   2. Demonstration   3. Impact   4. Your turn!

1. The project in a nutshell

Goals of Lingua Libre

- Document oral linguistic diversity
- Feed the Wikimedia projects with audio content
- Bring new linguistic communities to Wikimedia projects

Origins: a multidisciplinary team

- The chapter Wikimédia France
- Contributors from the French-speaking Wiktionary (*Wiktionnaire*)
- Linguists
- Speakers of some French regional languages
- The creator of Shtooka, a PC software for recording pronunciations

- A Project Grant provided to 0x010C
- Funds from the French Ministry of Culture

How does it work?

- Lingua Libre is a wiki, runs on MediaWiki (1.35) and BlueLL is its custom skin
- A Wikibase is installed, allowing SPARQL queries
- The core of LiLi is a MW extension based on Voice Activity Detection (RecordWizard)
- The files are sent to Commons via OAuth on behalf of the user (Upload2Commons)
- Datasets can also be downloaded in .zip files (Work in progress)

Advantages

Without Lingua Libre

RECORD	PROCESS	UPLOAD	USE
"I want to make some pronunciation records" → RAW audio files	→ Clean audio files	→ Files are on Commons	→ DONE!
Use a dedicated software to manually start, stop, name each record	Use a dedicated software to cut the blanks, equalize, transcode, export	upload each files write descriptions add categories	Edit individually each Wikipedia/Wiktionnary pages

With

LINGUA LIBRE	BOT
"I want to make some pronunciation records" → Files are on Commons	→ DONE!
One easy-to-use tool, no installation needed, no sound-editing skills to have	Edit automatically Wikipedia/Wiktionnary pages

Lingua Libre Bot

What is its purpose?

- Integrates Commons-hosted Lingua Libre files to local wikis
- Coded in Python, see [GitHub](#)

How to have it on your wiki?

- Fill in a request for your wiki on the [LL:Bot](#) page
- Start a community discussion on your wiki and a vote for its Bot status

8
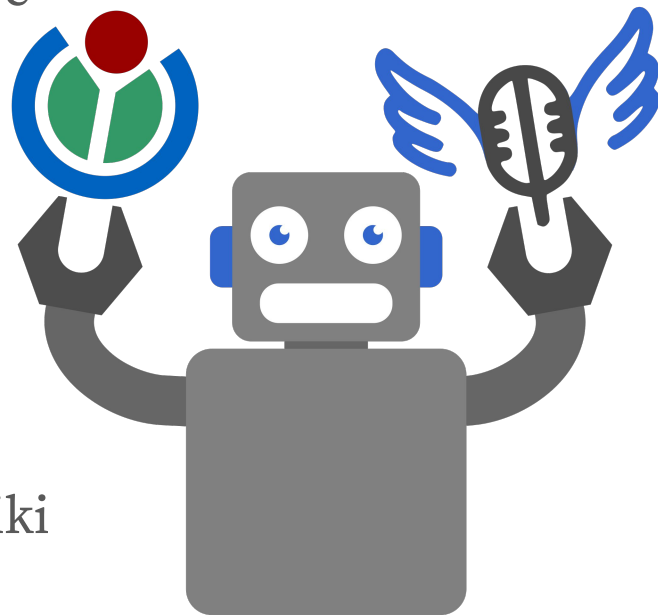
2. Demonstration

Logging in

A Wikimedia account is required, as Lingua Libre uses an OAuth token

It means that it will edit Commons on your behalf

<u>Access</u> page

Test your microphone

Record a voice

1 **Tutorial**
2 Speaker
3 Details
4 Studio
5 Publish
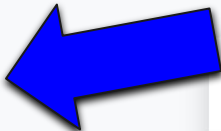
Check your microphone

Here's how to do it:

1. When you're ready, press the button above to start a record trial.
2. Pronounce a word aloud, for example "the sun".
3. Listen to your record to check that it's correct.
4. Adapt your microphone's configuration if needed.
5. If everything is correct, you can move on to the next step!

! To learn how to configure your microphone, visit this help page.

When you are ready, press the button below to start the test record.

Start the test

Cancel

Next >

Fill a speaker profile

Record a voice

- ✓ **Tutorial**
- ② **Speaker**
- ③ Details
- ④ Studio
- ⑤ Publish
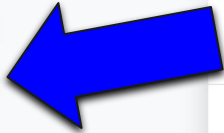
Speaker profile

⚠ This information will be publicly displayed and associated with your records.

Speaker profile to use	Test ⌄
Name to display	Test
Gender	**Male** Female Other
Spoken languages	French ✕ ＊
Place of residence	Paris

License

I, Test, irrevocably grant anyone the right to use my audio records created with this tool under the following license:

Creative Commons Attribution ShareAlike 4.0 ⌄

You agree that they may be automatically used on several Wikimedia projects, including Wikipedia, Wiktionary and Wikidata.

Cancel ‹ **Previous** **Next** ›

13

If a language is not available
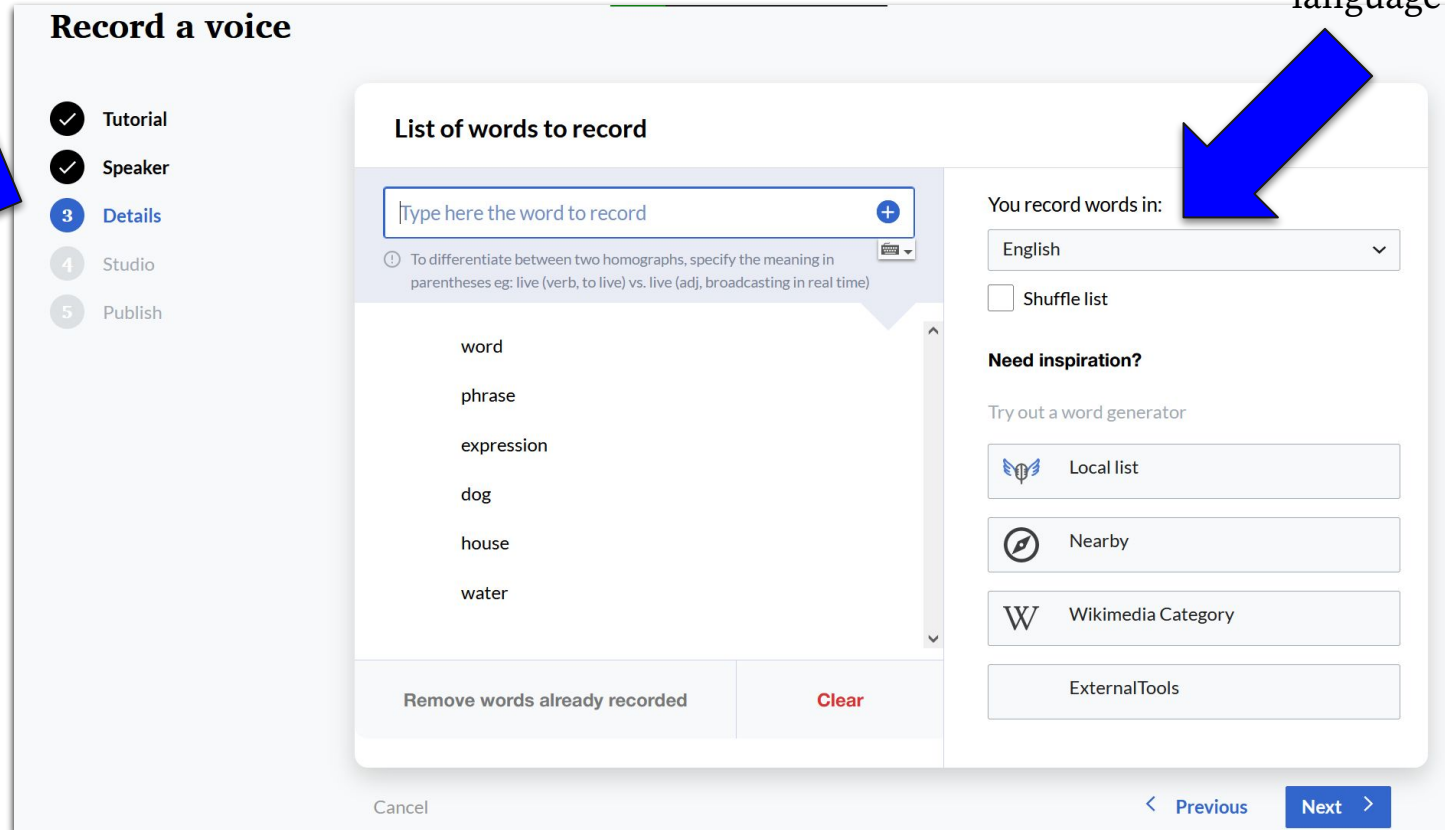
→ Ask an administrator via their talk page or the chat room, they will add it rapidly

[List of admins](#)

Several ways to choose the words you will record

A. Add the words one by one

double-check the language you will record

Record a voice

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

List of words to record

Type here the word to record ⊕

⊘ To differentiate between two homographs, specify the meaning in parentheses eg: live (verb, to live) vs. live (adj, broadcasting in real time)

word

phrase

expression

dog

house

water

Remove words already recorded | Clear

You record words in:

English ⌄

☐ Shuffle list

Need inspiration?

Try out a word generator

🕊 Local list

🧭 Nearby

W Wikimedia Category

ExternalTools

Cancel

‹ Previous | Next ›

B. Use an existing local list

List of words to record

2

Local list

Done

Title ⓘ

Eng/Lemmas-without-audio-sorted-by-number-of-wiktionaries

Exclude words you have already recorded ⬤

⌨ ▾

ish

huffle list

ecord words in:

inspiration?

t a word gen

1
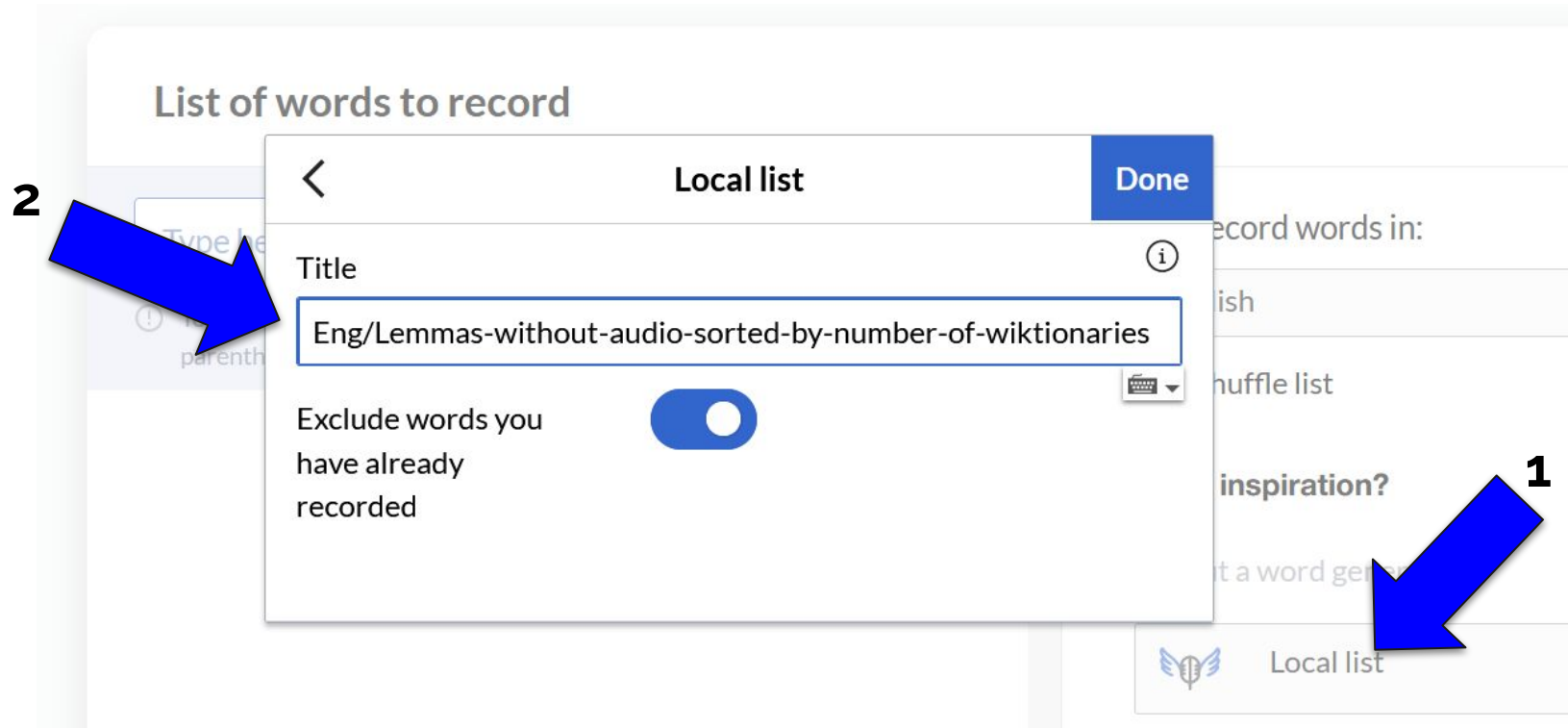
🦅 Local list

Local lists can be manually created, or by bot,
like [Olaf's bot lists for +70 languages](#)

C. Use a category from Wikipedia or Wikt.
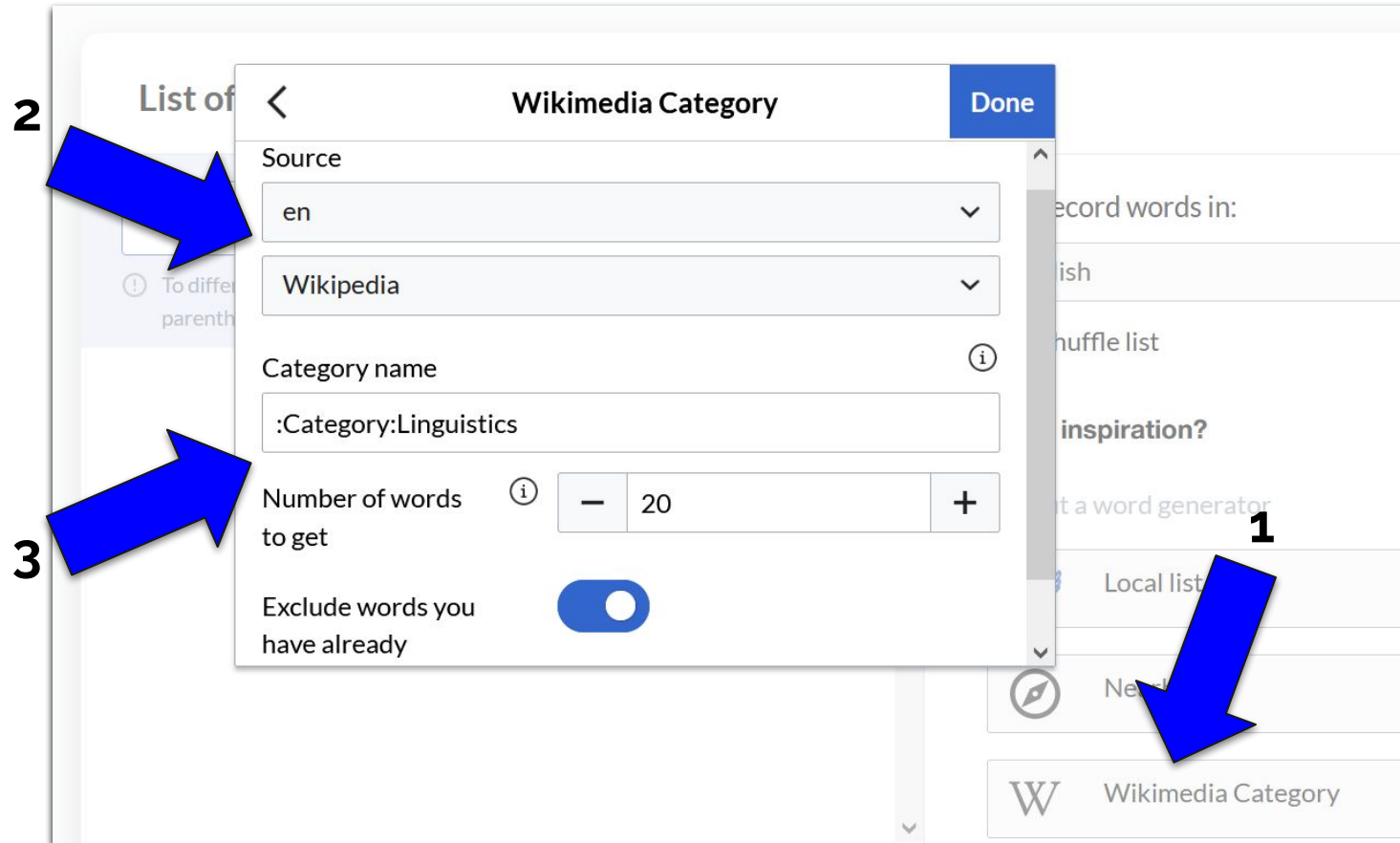
Recording studio

Record a voice

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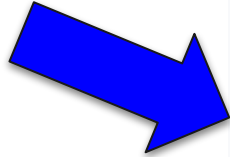
Studio

⚙ ⌨

- ▶ **language**
- lexicon
- outline of linguistics
- Noam Chomsky
- linguistic description
- theoretical linguistics
- pleonasm
- metalanguage
- Object language
- gloss
- rhema
- conditional sentence
- topic–comment
- possible world

Click on $1 below, then read the word aloud

language

Skip ›

⬛ ▮▮▮▮▮▮▮▮▮▮▮▮ ◉   0 / 20

Cancel

‹ Previous   Next ›

Check-listen before you submit

Verify the uploading, retry if needed

Record a voice

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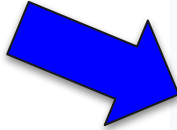
Publication

- ⚠ language
- lexicon
- ▶ **language classification**
- ✓ language documentation
- ✓ phrase
- ✓ sentence

Check your records before starting their publication.

language classification

0:02 / 0:02

3 / 4
⚠ 1

Cancel   🔥 **Check your uploads on Wikimedia Commons**   ↻ **retry failed upload**   › **Record more words**

Check your uploads on Commons

In case a mispronounced recording has been send, you only need to record it again on Lingua Libre

Categorized by language and user on Commons

https://commons.wikimedia.org/wiki/Category:Lingua_Libre_pronunciation-ron

3. Impact

A. The community

Map of Lingua Libre's speakers (>550)
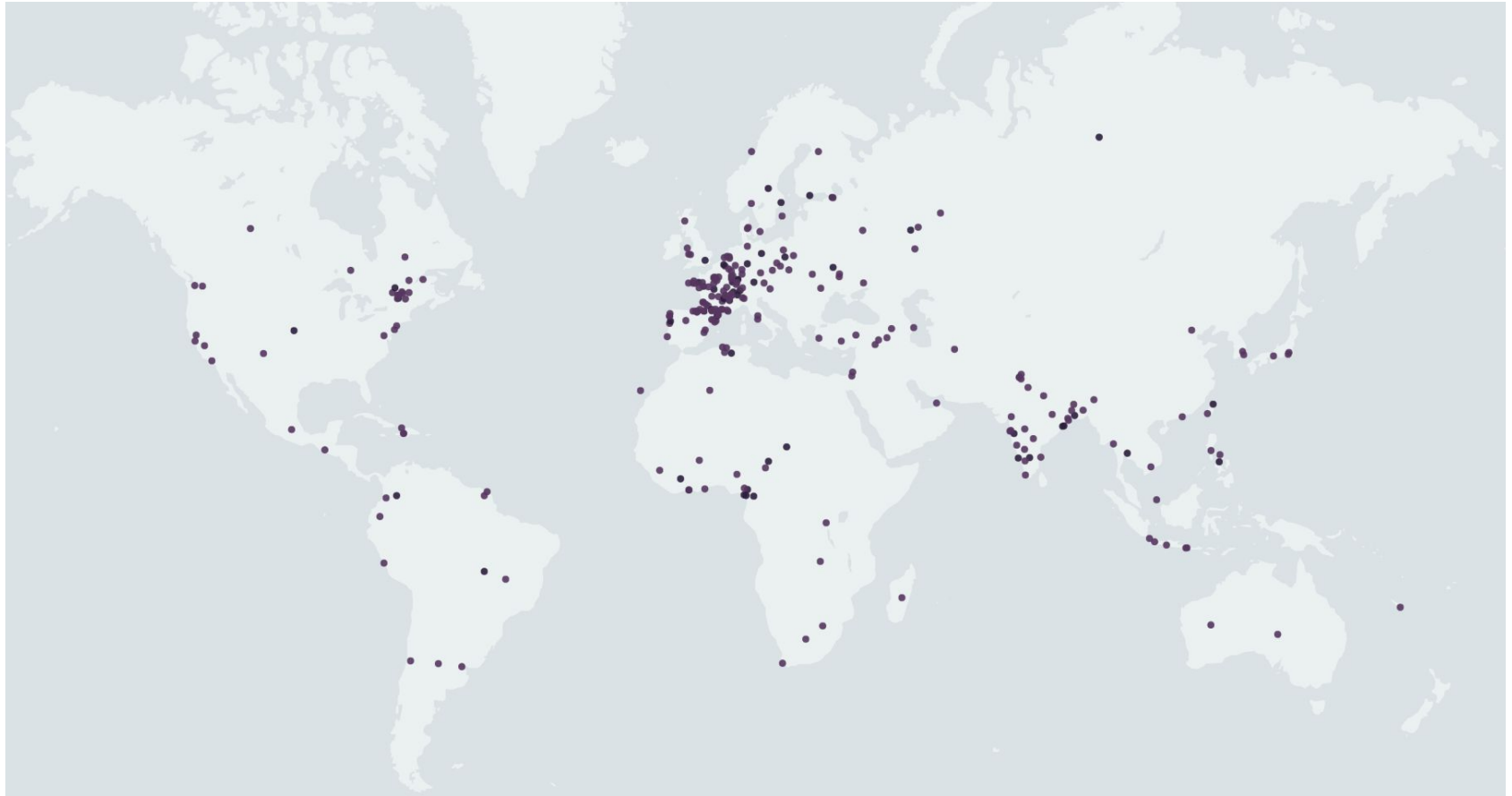
An international community

Recording session in Atikamekw (Canada)
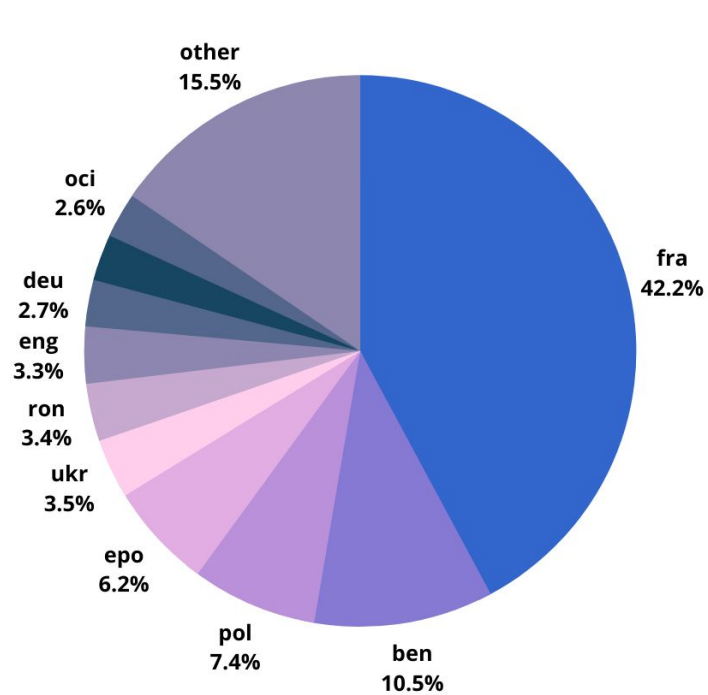Photo by Benoit Rochon, CC BY-SA 3.0

Recording some languages of Cameroon
Photo by Geugeor
[original pdf], CC BY-SA 4.0
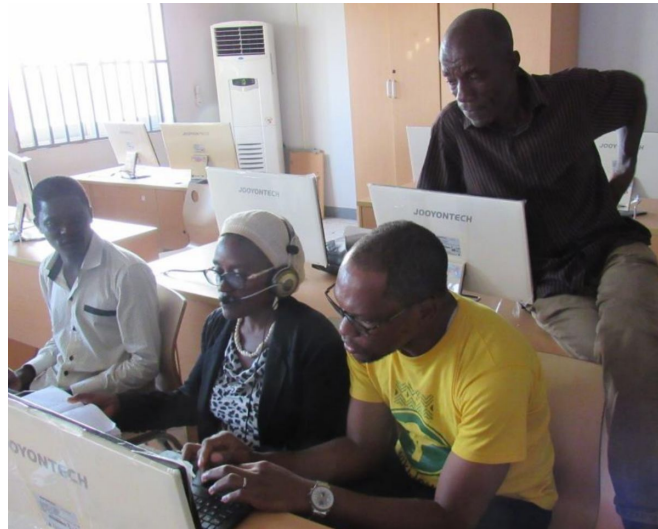
Team of Wikimedia France and LiLi contributors
Photo by 0x010C, CC BY-SA 4.0

Statistics: Languages and Genders

Languages
Total of 530k recordings, 127 languages
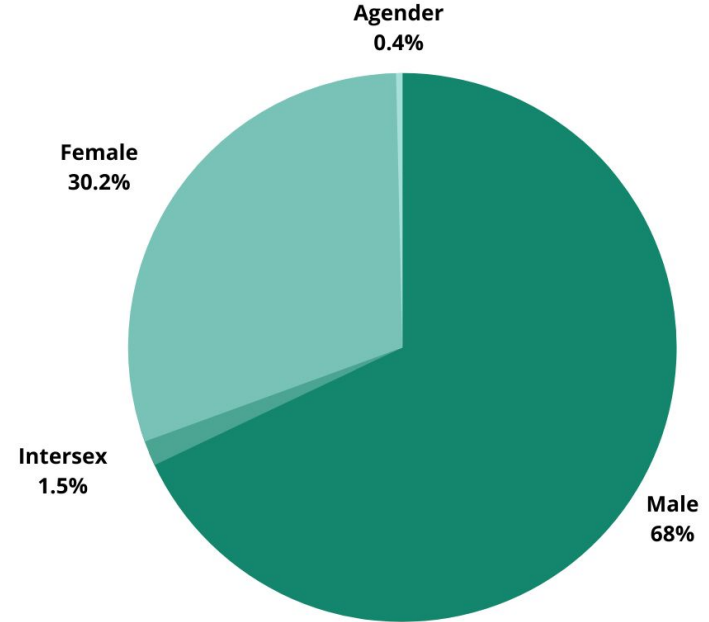
Genders
Total of 550 speakers

Statistics
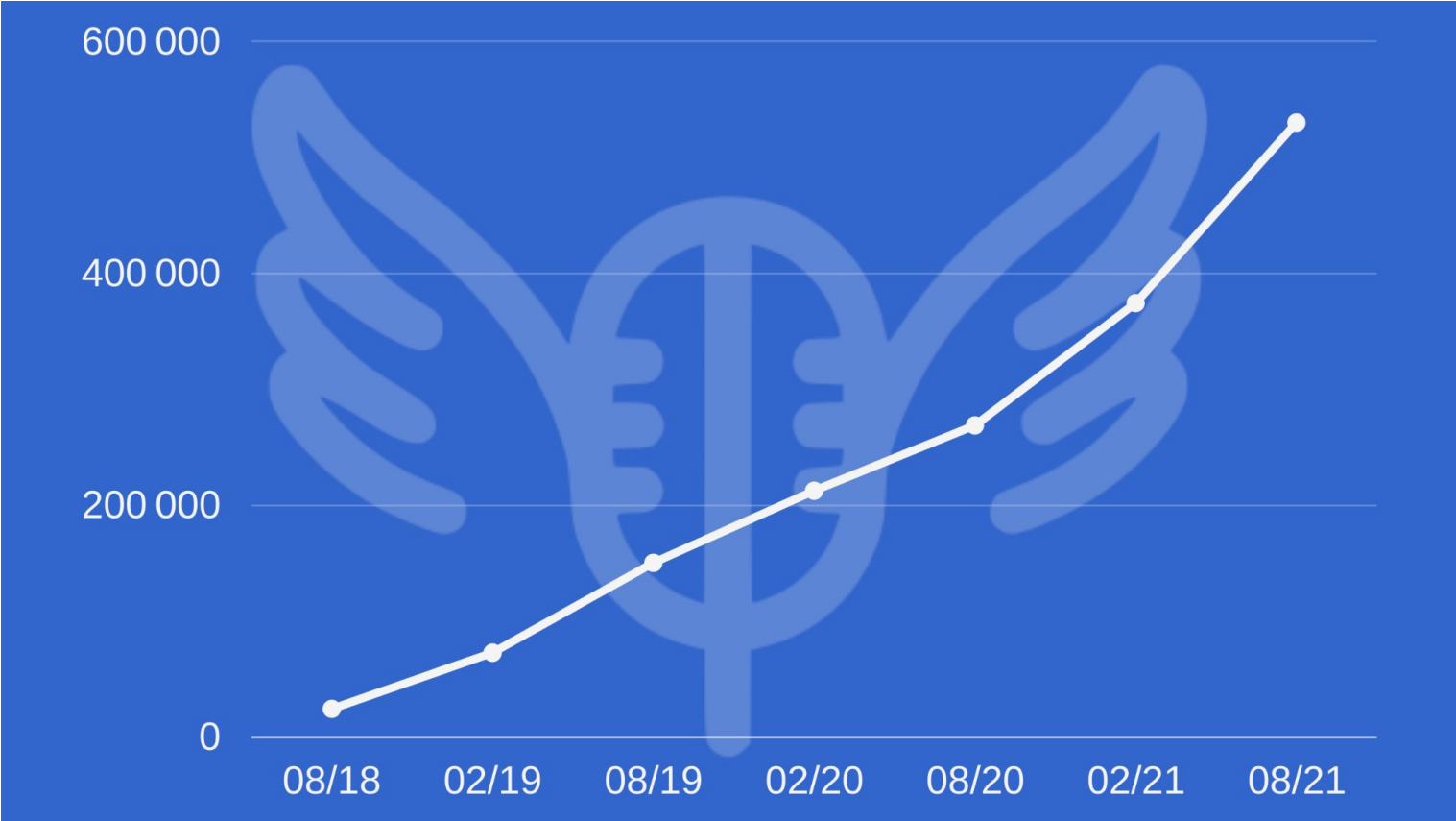
B. Use of the recordings

Examples of use on the Wikimedia projects

Illustrate the pronunciation of words and names on different projects:
Wikipedias, Wiktionaries, Wikidata, Wikivoyage, Wikisource…

Numbers from 1 to 20 in Kaqchikel on Wikisource

Entry forêt on the French speaking Wiktionary
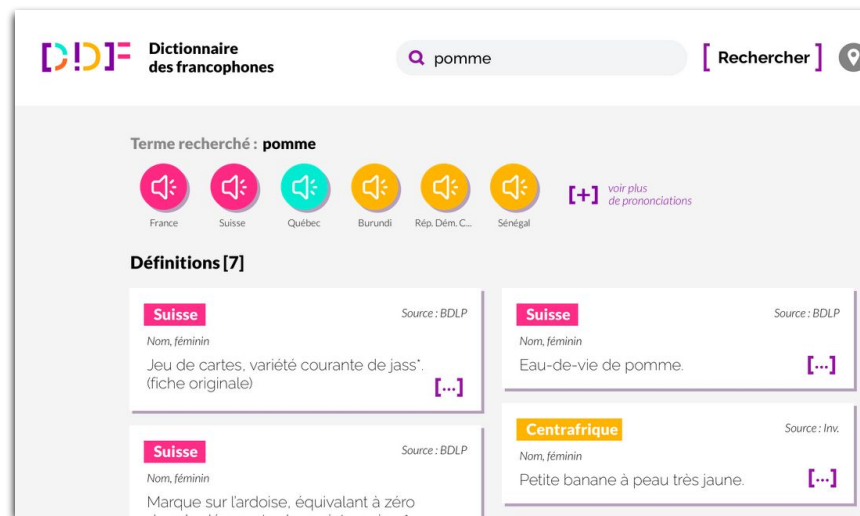
A lexeme in Bengali on Wikidata

Examples of use outside of Wikimedia projects

- Some recordings are used as training data for Mozilla's speech to text engine, DeepSpeech

- Le Dico des Ados, a libre dictionary for young people, records and uses some recordings.

- The Dictionnaire des francophones will soon be including Lingua Libre recordings to highlight the accents of the French-speaking world.

- Perspectives for language-learning applications

Thank you!
Do you have questions?

5-10 minutes Q&A before the interactive session

4. We need you!

Interactive session

A. Get involved in Lingua Libre

Several ways of getting involved

- The whole content can be translated, learn more at Help:Translate
- Create lists in your own language
- Any question, suggestion, news? The Chat Room is yours
- For technical discussions, see the Tech Board
- Help us improving the code via GitHub and Phabricator
- Request Lingua Libre Bot on a specific wiki here
- Spread the word on your own wikis
- **Record your voice!**

B. Record your voice

Record your voice: advice

- Lingua Libre works better on Firefox for the moment
- Please mind the noises around you (and turn off the sound of this session!) to avoid interferences
- Follow the steps of the Record Wizard or read the [demonstration slides](#)
- Please listen to your recording before uploading them to Commons
- Watch your files on Commons, and on the main page of Lingua Libre!

THANK YOU!

WIKIMANIA

Credits