Collecte de Corpus sur le Terrain : Methodes et Applications

Laurent Besacier, Elodie Gauthier

Linguistic Fieldwork, Language Documentation and Description

- In linguistic fieldwork, notions of documentation and description have been used often interchangeably
- **Documentation** gathers primary documentation (recording of audio and visual data, their transcription and translation)
- **Description** obtains secondary data of increasingly interpretative nature (description of underlying structures of the language such as lexicons and grammars)
- Language documentation is not only for field linguistics but also for the language community itself (future generations may want to learn about their ancestors' language)

Linguistic fieldwork needs means and tools

- Bird (2009): need renewal in computational linguistics technics, to accelerate the documentation and description of the world's endangered linguistic heritage
- Liberman (2011): towards very large scale phonetics
- Adda (2012): speech technology as instrument for phonetics and linguistics
- Thieberger (2013, 2016): need for tools appropriate to a discipline; need for renewal in data access, analysis and publication; reflection about methods for language documentation
- Holton *et al.* (2017): tools and standards for processing linguistic data archiving are still lacking

Linguistic fieldwork needs means and tools

Conferences and workshops

- VLSP 2011 - Very Large Scale Phonetics workshop[1]
- Digital Tools Summit in the Humanities[2]
- Spoken Language Technologies for Under-resourced Languages (SLTU)[3]
- Language Documentation Tools Summit[4]
- Computational Methods for Endangered Language Documentation and Description (CMLD)[5]

[1] http://www.phon.ox.ac.uk/jcoleman/MiningVLSP.pdf
[2] http://www.iath.virginia.edu/dtsummit
[3] http://www.mica.edu.vn/sltu/
[4] https://sites.google.com/site/ldtoolssummit/home
[5] http://lattice.cnrs.fr/cmld/

Linguistic fieldwork needs means and tools

Workgroups

- Special Interest Group: Under-resourced Languages (SIGUL)[6]

Summer schools

- CIDLeS 2014: Coding for Language Communities[7]
- BigDataSpeech 2018[8], Summer school for using computational tools in linguistic, phonetic research

- SENELANGUES 2015[9], Spring school for West African language description

[6] www.elra.info/en/sig/sigul
[7] www.cidles.eu/summer-school-coding-for-language-communities-2014
[8] http://bigdataspeech.alwaysdata.net
[9] senelangues2015.ucad.sn

This talk

My (very biased) experience from the BULB (Breaking the Unwritten Language Barrier) project

- Focus on unwritten languages (some of them endangered)
- Data collection methodologies (**this talk**)
- Data processing methodologies

Close collaboration between computer scientists and linguists

- July 3rd 2015: Workshop on Natural Language Processing Technology for Linguists in Paris at LPP
- January 25th and 26th 2016: Linguistic training workshop for language technology experts in Paris
- ICPhS 2019 Special Session on Computational Approaches for Documenting and Analyzing Oral Languages

Some Language Documentation Challenges

- Endangered languages corpora are often small, scattered and mostly not freely available which - in addition to prevent interesting cross lingual or cross dialect studies - limits their wide use by other linguists or computer scientists,
- Classical computational linguistics (CL) and speech processing mostly rely on supervised algorithms and will not scale to the endangered languages targeted which are poorly described and un-annotated,
- Most endangered languages are unwritten so the main raw material is the speech signal, often recorded in poor conditions (remote rural areas, elderly speakers, etc.)

Our Vision

field linguist

Our Vision

field linguist

cyber field linguist

Computational Language Documentation and Description

- Recast language documentation and description as highly inter-disciplinary research

Computational Language Documentation and Description

- Recast language documentation and description as highly inter-disciplinary research
- Where field linguistics leverage computational models and machine learning

Computational Language Documentation and Description

- Recast language documentation and description as highly inter-disciplinary research
- Where field linguistics leverage computational models and machine learning
- Focus on endangered and/or unwritten languages (**speech!**)

Computational Language Documentation and Description

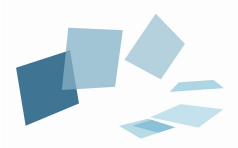
- Recast language documentation and description as highly inter-disciplinary research
- Where field linguistics leverage computational models and machine learning
- Focus on endangered and/or unwritten languages (**speech!**)
- Relies on
  - Large and (as much as possible) naturalistic speech corpora ...
  - ... automatically processed ....
  - ... to provide a (radically) new machine-assistance to the field linguist / dialectologist

The BULB Project

French-German 3 years project

The BULB Project

French-German 3 years project



The BULB Project

French-German 3 years project

Zentrum für Allgemeine Sprachwissenschaft

Universität Stuttgart

The BULB Project

French-German 3 years project

The BULB Project

French-German 3 years project

Zentrum für Allgemeine Sprachwissenschaft

Universität Stuttgart

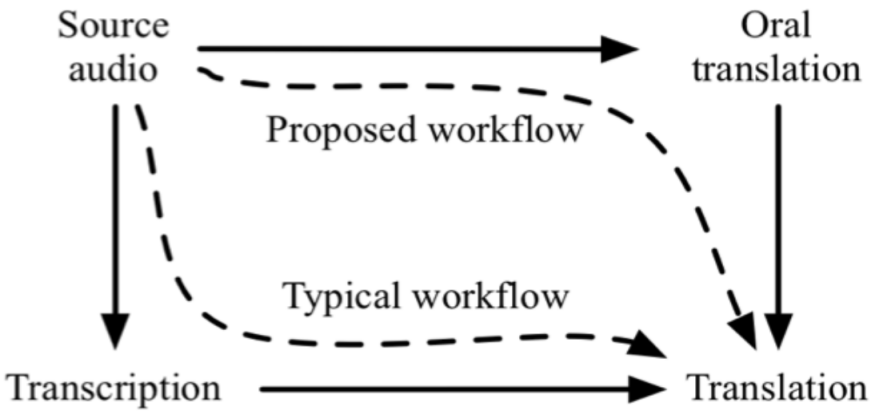
The BULB Project

French-German 3 years project

Recording Methodology

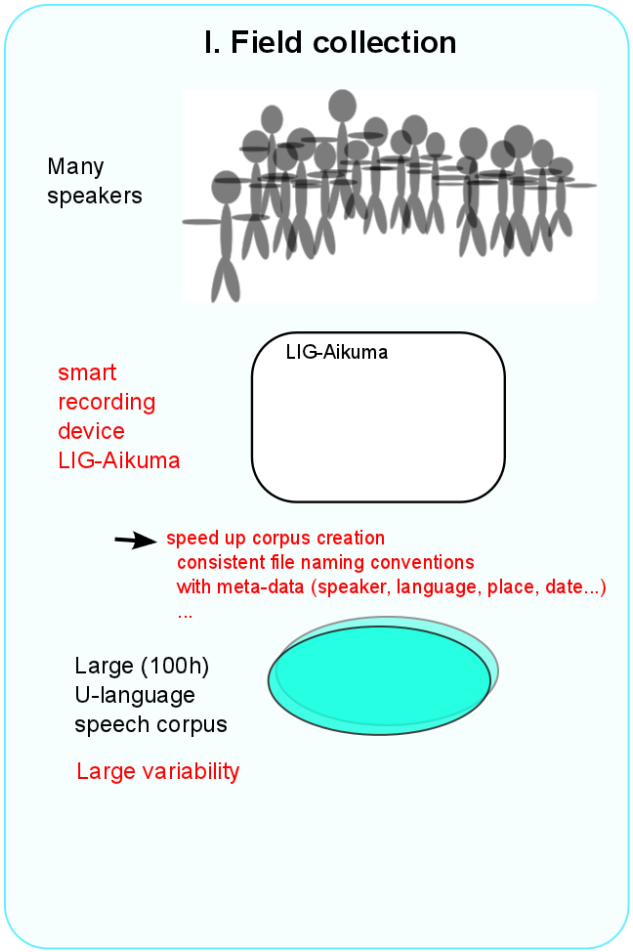
New methodology, inspired by pioneering work by Steven Bird and Mark Liberman.

- Collect a large speech corpus of the source language (U-language):
- Re-speaking in the source language by a reference speaker.
- Spoken translation into a target language and processing with high-quality spoken language technologies (language in contact and lingua franca).



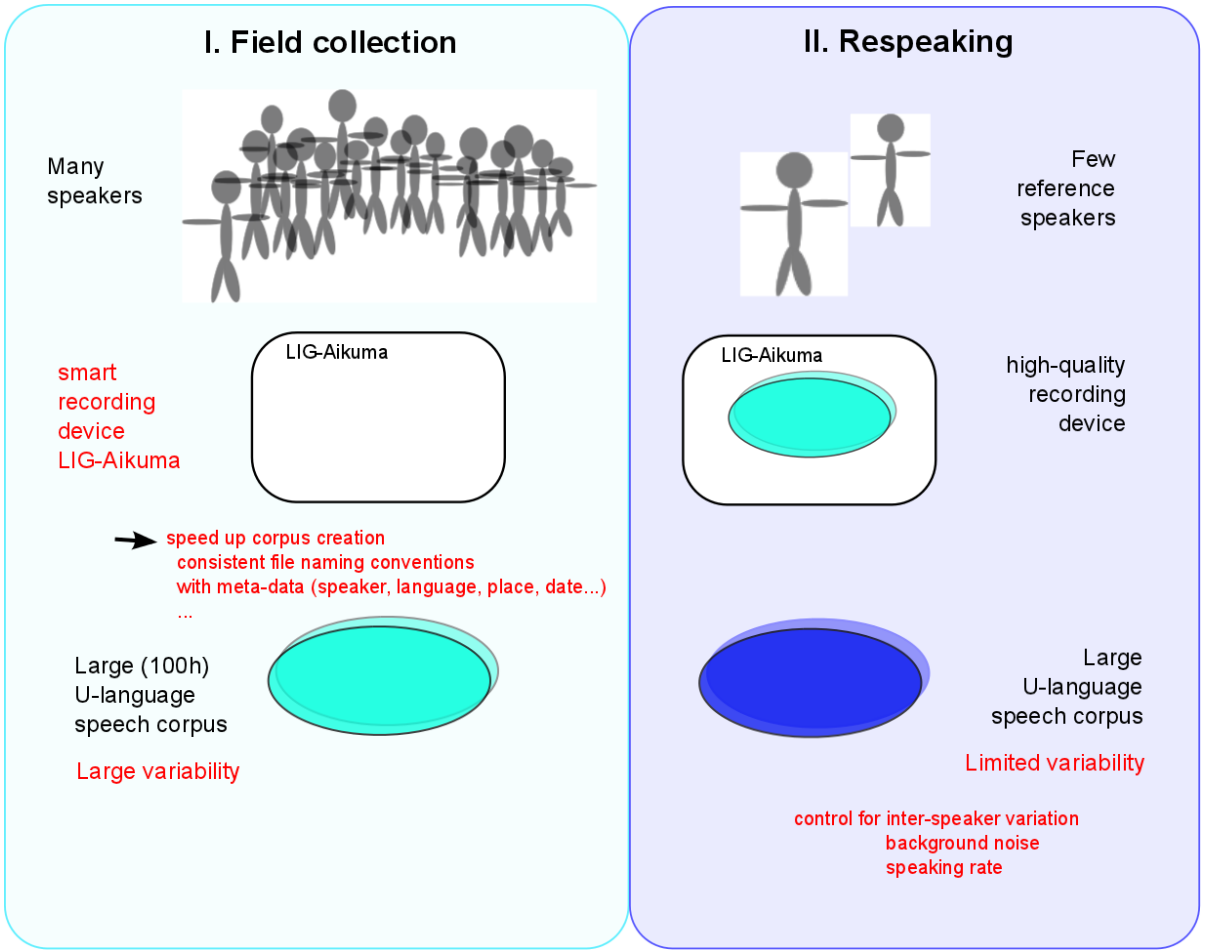
Recording Methodology

(Fig. are courtesy of Gilles and Martine Adda)



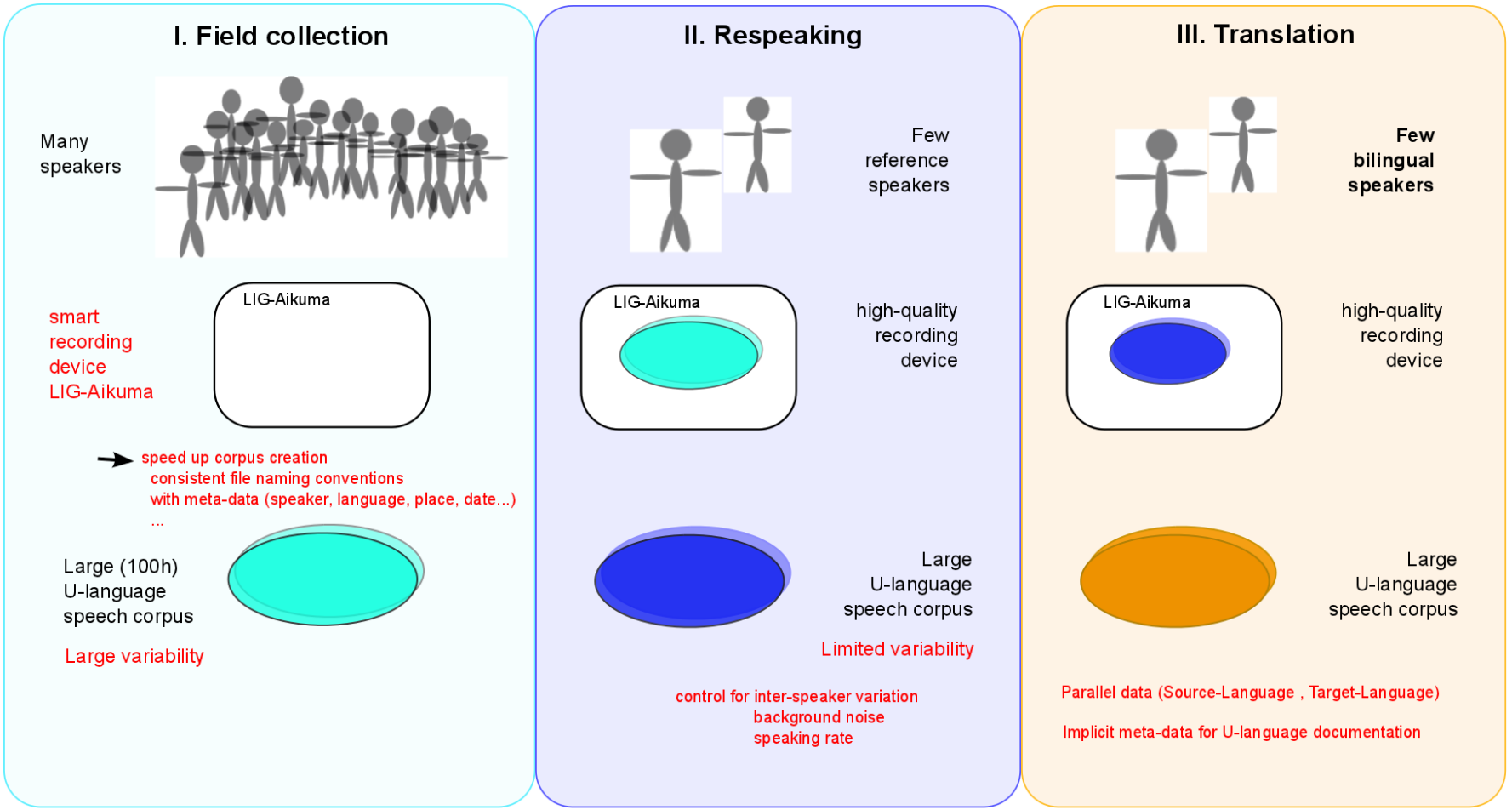
Recording Methodology

(Fig. are courtesy of Gilles and Martine Adda)



Recording Methodology

(Fig. are courtesy of Gilles and Martine Adda)



Aikuma: The origin

Initial Android app **dedicated to speakers** of endangered languages
Developed by Steven Bird and Florian Hanke (Hanke and Bird, 2013)

> *Goal*: Collecting speech at a large scale with self-recording

features

- Recording
- Respeaking
  - Concept introduced by Woodbury (2003)
  - Firstly experienced by Bird (2010)
- Translation

Lig-Aikuma

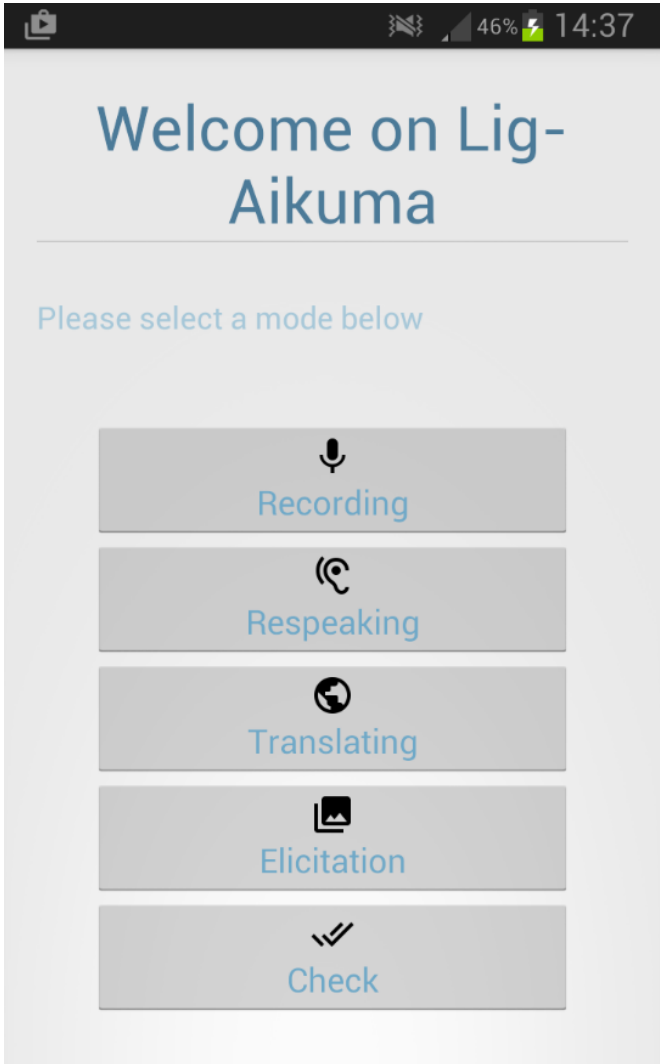
- From Aikuma to Lig-Aikuma
  - New branch specified with BULB field linguists
  - From self-recording to more controlled recording
- Lig-Aikuma V3
  - `https://lig-aikuma.imag.fr` and *playstore*
  - **Contact:** laurent.besacier@imag.fr and elodie.gauthier@imag.fr
  - Available *open source* on LIG Forge
  - Online video tutorial and manuals

LIG-AIKUMA

Features	**Aikuma**	LIG-AIKUMA
Recording and documentation	✔	✔
Respeaking and oral translation	✔	✔
Extras : Sync. and Sharing, Geolocalisation, Textless interface	✔	✔
Elicitation (text-image-video) mode	✘	✔
Check mode	✘	✔
User profiles, Consent form, Metadata	✘	✔
Automatic backup of interrupted sessions	✘	✔
Multilingual interface and User feedback	✘	✔
Documentation (samples, tutorial, . . .)	✘	✔
Export to Elan	✘	✔

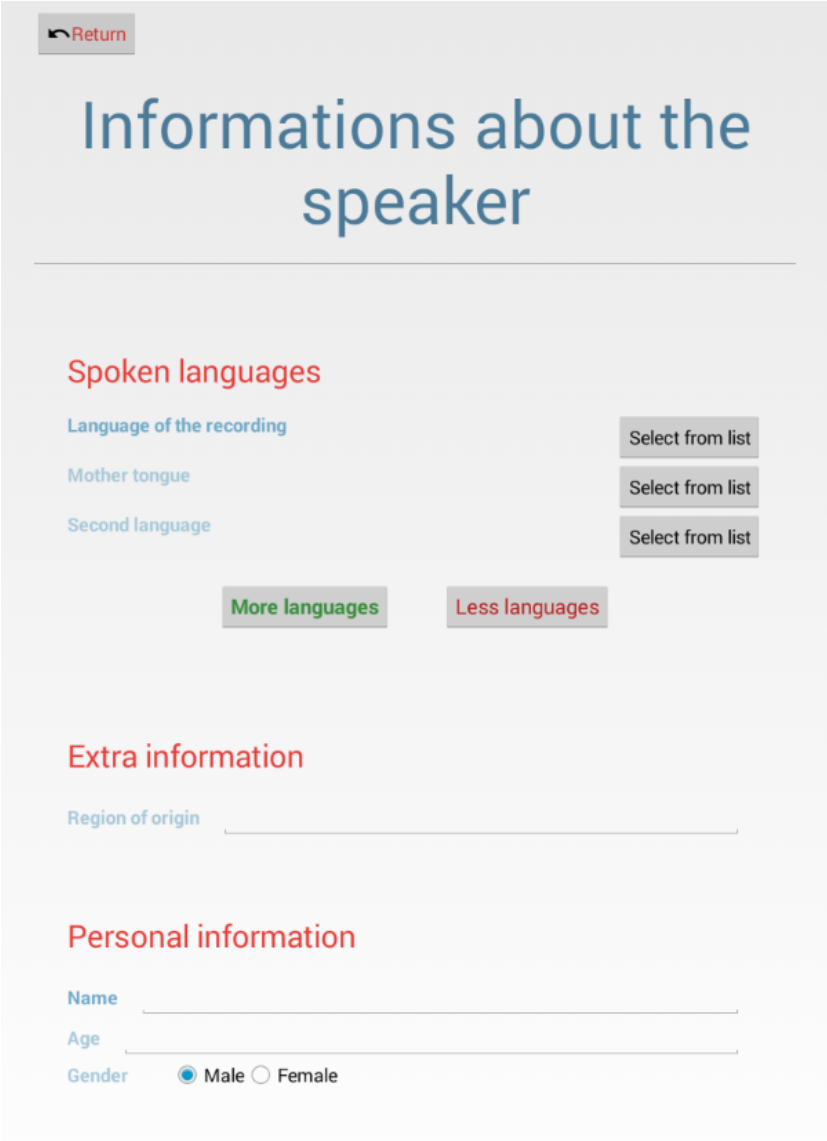
LIG-AIKUMA

- **Recording** speech in a very simple way
- **Respeaking** an existing recording or an external audio file
- **Translation** is the same as respeaking, except the language changes
- **Elicitation** records speech based on external resource (text, images)
- **Check** mode (in progress)



Metadata

- **Spoken languages**
  - language of the recording
  - mother tongue
  - extra languages
  - quick input using language codes (en, fra)
- **Extra information**
  - geographical region of origin of the speaker
- **Personal information**
  - name
  - age
  - gender



Metadata

Figure: (1)

1. Click "Select from list"
2. A list appears, with a **filter search box** at the top
3. Type in the language

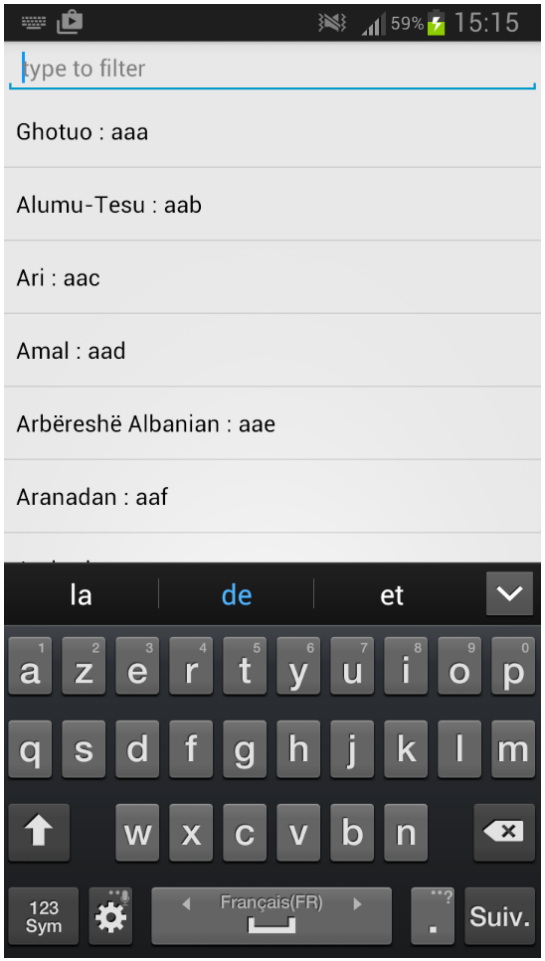


Figure: (2)

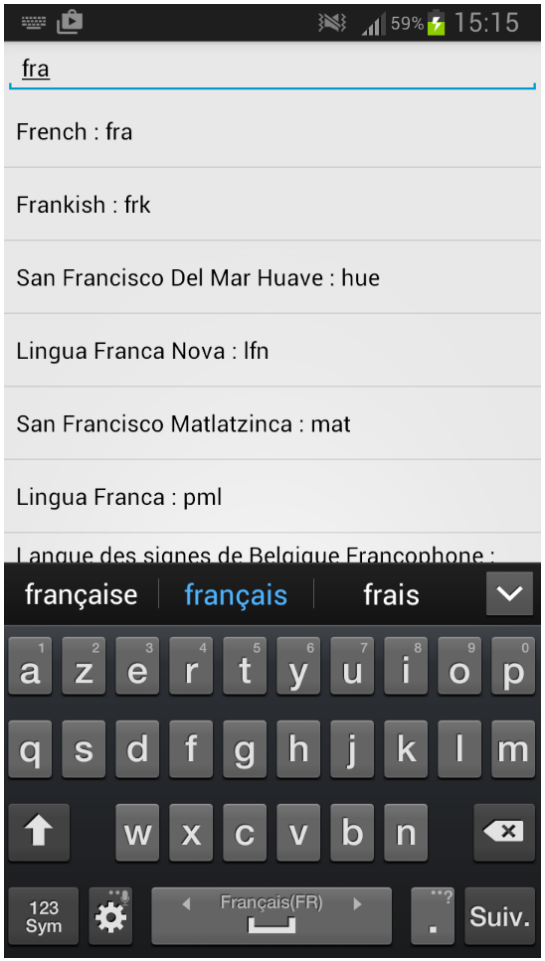


Figure: (3)

Recording mode

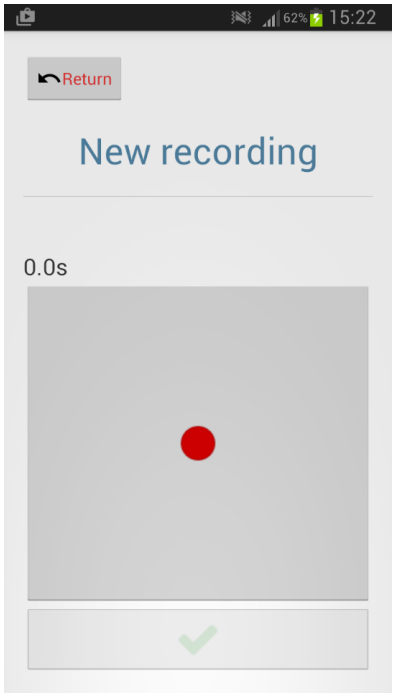


Figure: (1)

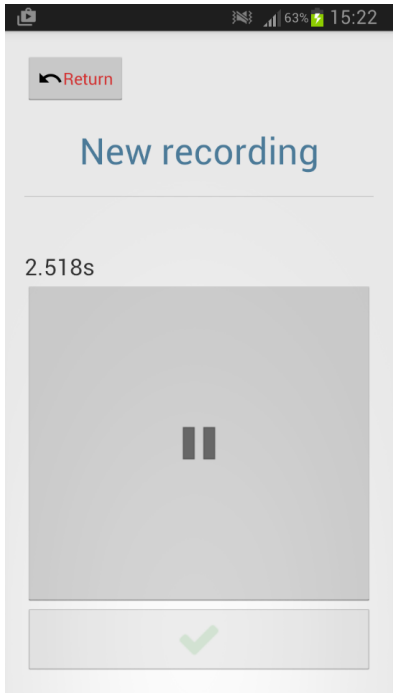


Figure: (2)

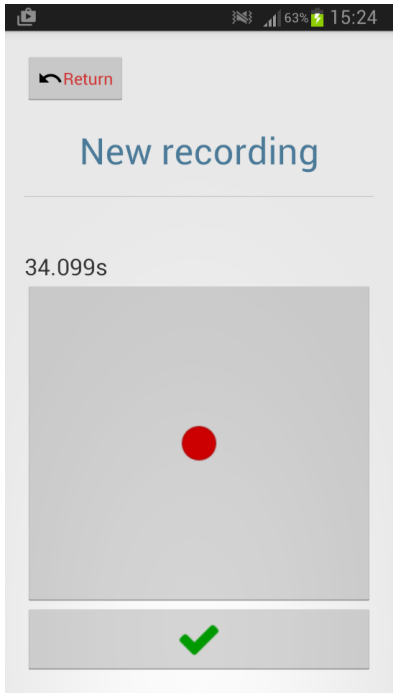


Figure: (3)

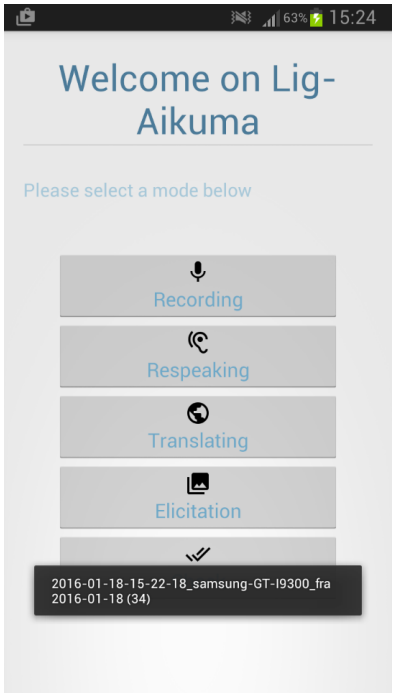
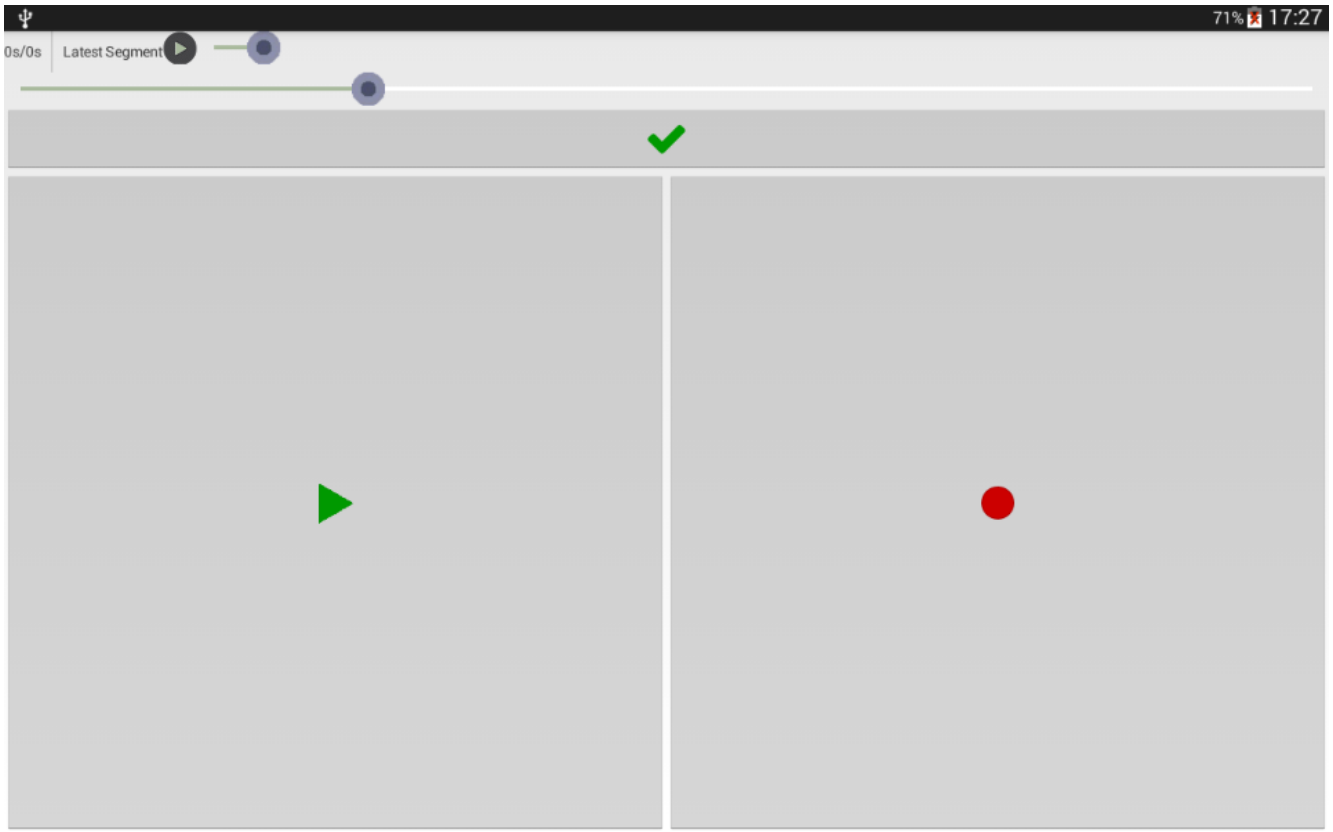


Figure: (4)

1. Start a recording
2. Currently recording
3. Recording is done, now save it !
4. Recording saved (see popup), back to home view

Respeaking mode

- Alternatively play and record audio segments
- Listen to the latest recorded segment
- Validate once it is done

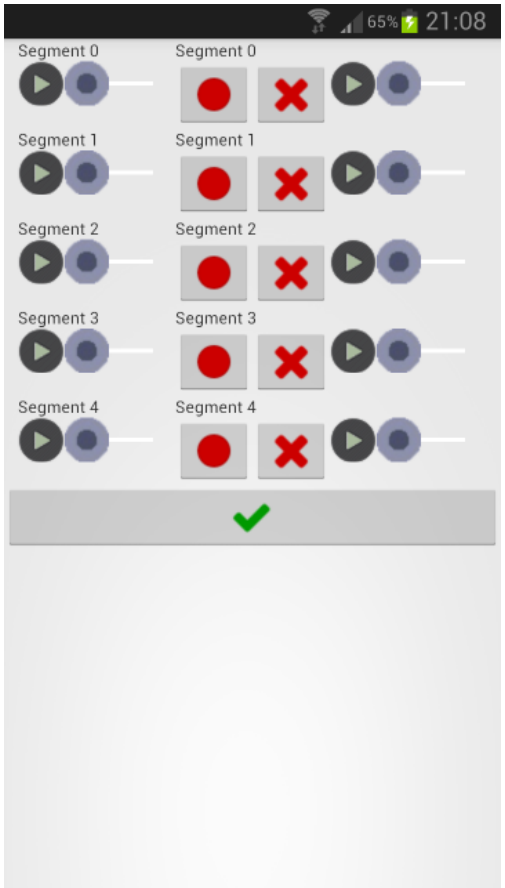


Respeaking mode

(Optionally) Play and Record any Segment

- Every pair of original and respoken segments are aligned on rows
- Both segments can be played
- Respoken segments can be recorded again

Then Validate! Popup view informs about the success or failure of the respeaking



Translation mode

Translation works the same way as the respeaking mode

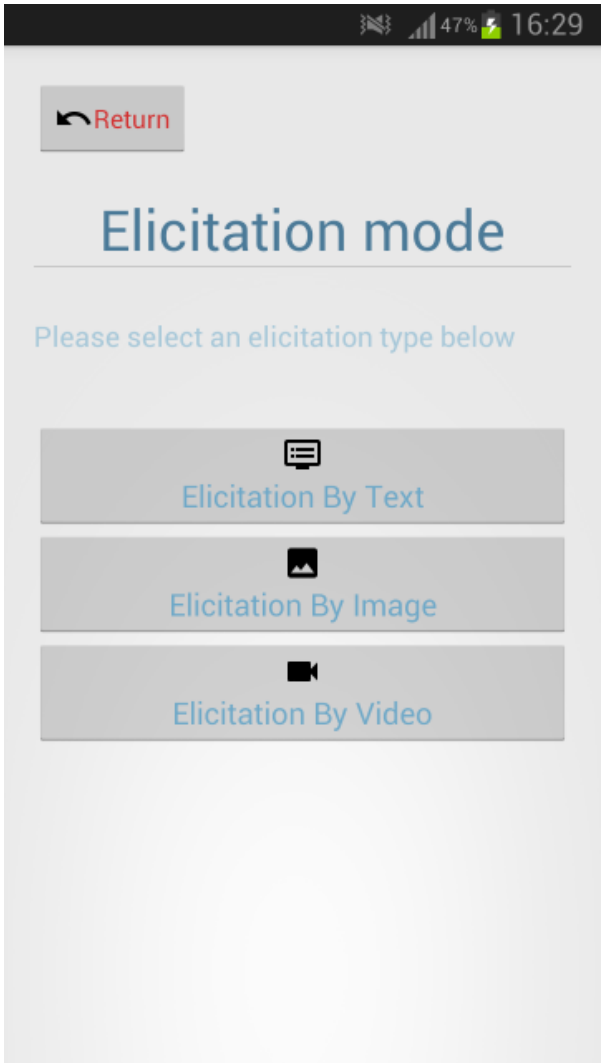
The only difference relies in the produced name file

- 160118-152218_**fra**_58a_**rspk**
- 160118-152218_**eng**_58a_**trsl**

Elicitation mode

Dedicated to elicitation of resources

- ✓ Text: speak written text
  - words or sentences
- ✓ Image, Video: implemented recently



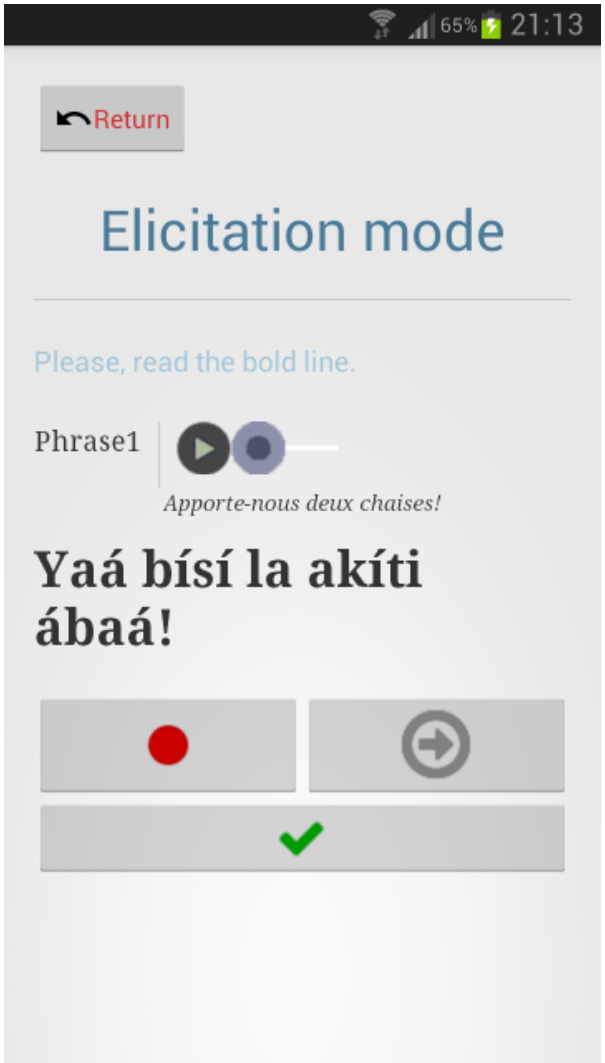
Elicitation from text

Key Features

- Record your speech
- Listen to the record
- Go to next item or validate and/or quit

Text items

- Items: words or sentences
- Loaded from the imported text file
- One line at a time
  - ending by "##"



Elicitation from image

Key Features

- Visualize an image
- Record your speech
- Go to next item or validate and/or quit

Image items

- Loaded from a directory containing **only** images
  - JPG or PNG formats

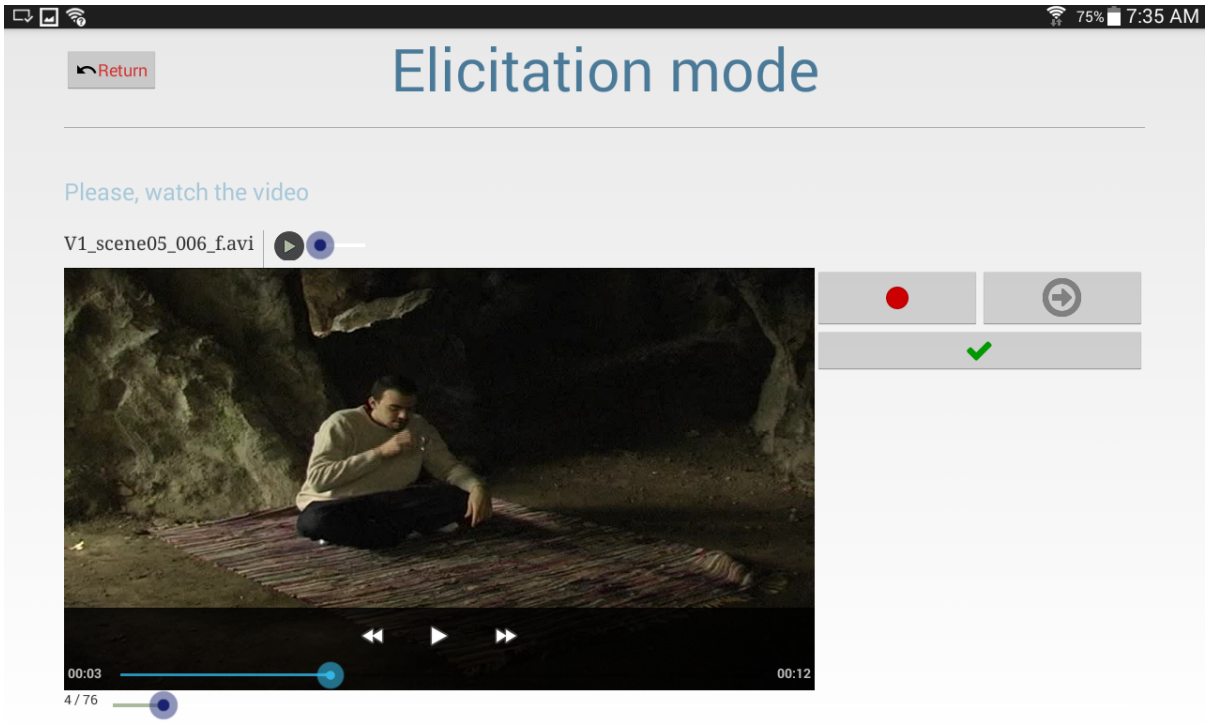
Elicitation from video

Key Features

- Watch a video
- Record your speech
- Go to next item or validate and/or quit

Video items

- Loaded from a directory containing **only** video
  - AVI or MP4 formats



Download and Install

Download URL `https://lig-aikuma.imag.fr/download/` **from the Android device**

- In a web browser, type in the url address
- A popup window may appear to require you the install confirmation
- You must have authorized the install from other sources than Play Store. To do so:
  → Settings → Security tab → Check the "Unknown sources" button

From a computer

- Download the application on your computer
- Email the application file (MainActivity.apk) to an account synced on the Android device
- On your Android device, open the joined file, an installation popup window should appear

Documentation of Bantu Languages

- Three northwestern Bantu:

Documentation of Bantu Languages

- Three northwestern Bantu:
  1. Basaa (A43, Cameroon) 300,000 speakers,
  2. Myene (B10, Gabon) 46,000 speakers,
  3. Embosi (C25,Congo-Brazzaville) 150,000 speakers,

(Fig. is courtesy of Gilles and Martine Adda)

Documentation of Bantu Languages

- Three northwestern Bantu:
  1. Basaa (A43, Cameroon) 300,000 speakers,
  2. Myene (B10, Gabon) 46,000 speakers,
  3. Embosi (C25,Congo-Brazzaville) 150,000 speakers,
- Well-described, competent native-speaker linguists, basic electronic resources.

Documentation of Bantu Languages

- Three northwestern Bantu:
  1. Basaa (A43, Cameroon) 300,000 speakers,
  2. Myene (B10, Gabon) 46,000 speakers,
  3. Embosi (C25,Congo-Brazzaville) 150,000 speakers,
- Well-described, competent native-speaker linguists, basic electronic resources.
- A complex morphology (both nominal and verbal).
- Challenging lexical an postlexical phonologies.
- Contrastive tones both at lexical and grammatical levels.

Field recordings

Basaa

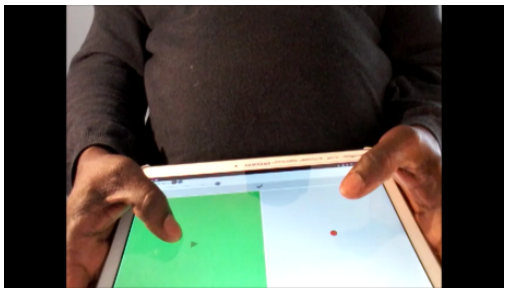
Embosi

Sound samples

Basaa

Embosi

re-speaking



translating

Data collected with LIG-AIKUMA

Table: *Source (clean) and translated speech collected so far*

Language	Source (clean)	Translated
Mboshi	55h	20h
Myene	45h	44h
Basaa	55h	25h

- Released a *Mboshi5k* corpus for computational language documentation exp[10] - used at JSALT Workshop 2017 at CMU
- P. Godard et.al. *A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments* LREC18
- A. Rialland et.al. *Parallel Corpora in Mboshi (Bantu C25, Congo-Brazzaville)* LREC18
- F. Hamalaoui et.al. *BULBasaa: A Dataset for Comp. Ling.* LREC18

[10] `http://github.com/besacier/mboshi-french-parallel-corpus`

Data Collection in Mboshi with Lig-Aikuma

Type of corpus	#speakers	quantity	dur. (h)	respoken	translated	manually transcribed
Debates	19	67	25h18	x (20h22)	x (20h22) (oral)	x (1h10)
Comments on pictures	8	1500 pictures	~15h			
Conjugations	1	50verbs*15TAM* 18subjects	5h56			x pre-existing conjugation tables
Read sentences **See Corpus 5K**	3	5178 sentences	4h51		x (written)	x pre-existing written sentences
Bible reading	6		4h06			

Figure: More details in a LREC 2018 paper (session P59 yesterday) - A. Rialland et al. *Parallel Corpora in Mboshi (Bantu C25, Congo-Brazzaville)* LREC 2018

Technologies Needed

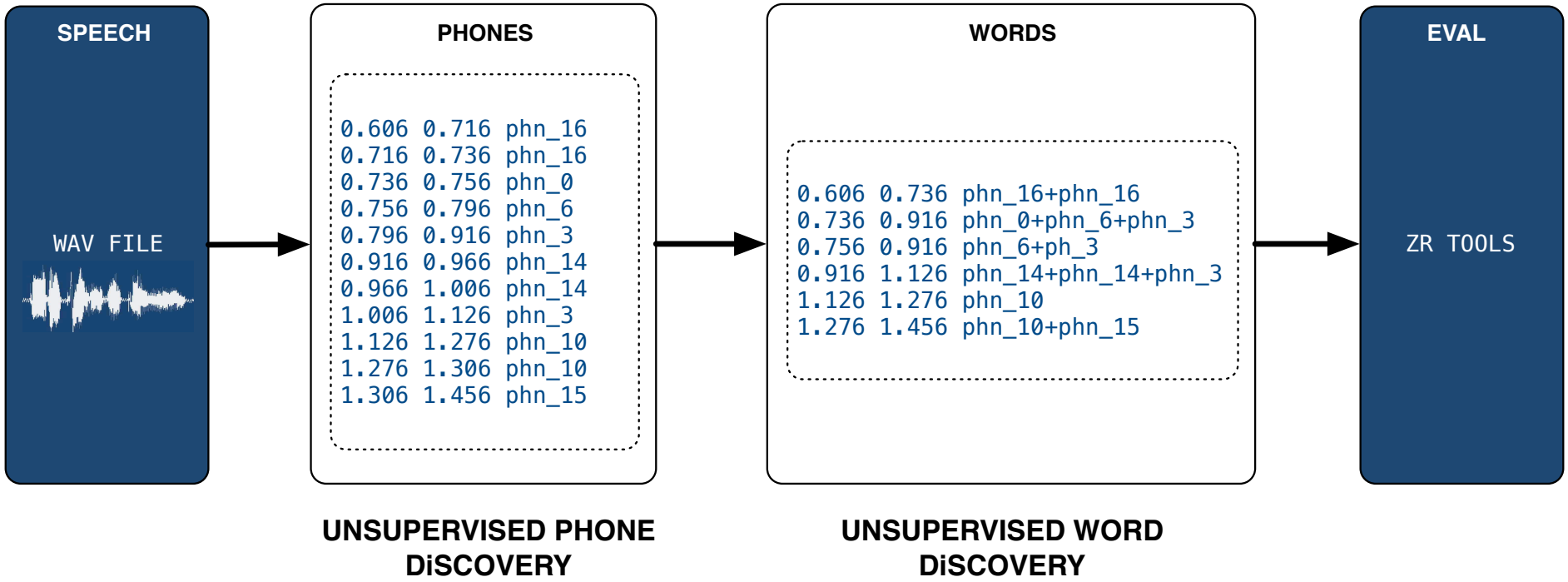
- French LVCSR (adapted to domain) - **LIMSI**
- Unsupervised Word Segmentation from Speech - **KIT,LIG,LIMSI**
  - Unsupervised Phone Discovery (UPD)
  - Unsupervised Word Discovery (UWD)
- Alignment of French word level sentences and phoneme sequences in Basaa/Myene/Embosi **LIG+LIMSI**

Unsupervised Word Discovery

- Output time-stamps delimiting stretches of speech, associated with class labels, corresponding to real words
- Task 2 of the Zero Resource Speech Challenge 2017 [11]
- Evaluation: measure quality of discovered boundaries with respect to the gold standard (forced alignments) **?**
- Interest for language documentation: provides a useful segmentation of speech into lexical units
- Interest for speech technology: unsupervised word discovery from raw speech

[11] http://zerospeech.com/2017

Unsupervised Word Discovery Pipeline



ZR TOOLS: evaluation tools for word discovery from speech **?**

Unsupervised Phone Discovery from Speech (KIT)

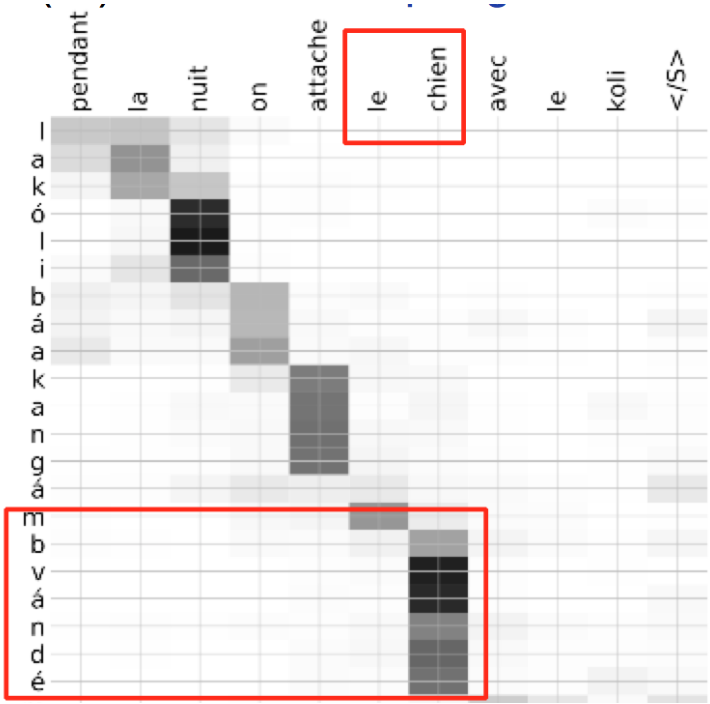
- Multilingual phoneme recognition suboptimal, since not all phonemes might be covered
- Three-step approach
  - Find phoneme boundaries using a BLSTM **?**
  - Extract articulatory features (AF) within the segments **?**
  - Cluster segments into phoneme-like units **?**

Unsupervised Word Unit Discovery (LIG-LIMSI)

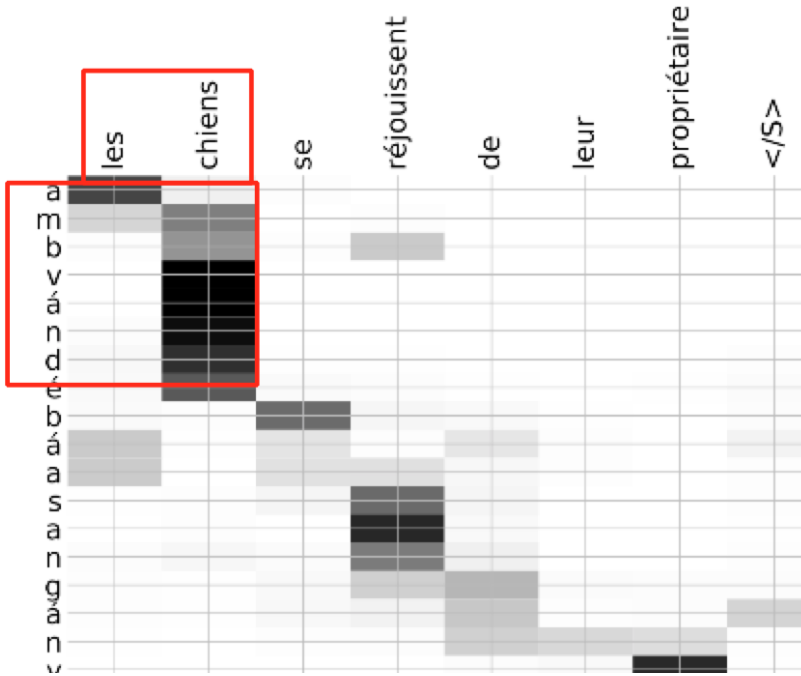
- Monolingual
  - Experiments with various non-parametric Bayesian word segmentation models
  - Study of the impact of tones on word segmentation
  - Preliminary experiments with Adaptor Grammars
  - Character or phone sequences vs. Lattice inputs
  - Pipeline scoring using (language independent) phoneme transcription
  - Ondel et.al. *Bayesian Models for Unit Discovery on a Very Low Resource Language*. In IEEE ICASSP 2018
- Cross-Lingual (using neural End-to-End approaches)
  - Zanon Boito et.al. *Unwritten languages demand attention too! word discovery with encoder-decoder models*. In IEEE ASRU 2017.
  - Godard et.al *Unsupervised Word Segmentation from Speech with Attention*. Submitted to INTERSPEECH 2018.

Seq2Seq Models for Word Segmentation

the dog
mbvande



the dogs
ambvande



Conclusion

- LIG-AIKUMA available for language documentation projects: `https://lig-aikuma.imag.fr/`

Conclusion

- LIG-AIKUMA available for language documentation projects: `https://lig-aikuma.imag.fr/`
- Several recording campaigns in Africa

Conclusion

- LIG-AIKUMA available for language documentation projects: `https://lig-aikuma.imag.fr/`
- Several recording campaigns in Africa
- Ongoing development of new features

Conclusion

- LIG-AIKUMA available for language documentation projects: `https://lig-aikuma.imag.fr/`
- Several recording campaigns in Africa
- Ongoing development of new features
- Side use for ASR technology development

Conclusion

- LIG-AIKUMA available for language documentation projects: `https://lig-aikuma.imag.fr/`
- Several recording campaigns in Africa
- Ongoing development of new features
- Side use for ASR technology development
  - 10h of read speech quickly collected in Fongbe (Benin)

Conclusion

- LIG-AIKUMA available for language documentation projects: `https://lig-aikuma.imag.fr/`
- Several recording campaigns in Africa
- Ongoing development of new features
- Side use for ASR technology development
  - 10h of read speech quickly collected in Fongbe (Benin)
  - 7h30 of translated speech (8k sentences from BTEC corpus) quickly collected in Amharic (Ethiopia)

Conclusion

- LIG-AIKUMA available for language documentation projects: `https://lig-aikuma.imag.fr/`
- Several recording campaigns in Africa
- Ongoing development of new features
- Side use for ASR technology development
  - 10h of read speech quickly collected in Fongbe (Benin)
  - 7h30 of translated speech (8k sentences from BTEC corpus) quickly collected in Amharic (Ethiopia)
  - small scale data collection in Sereer (spoken in Senegal)

Conclusion

- LIG-AIKUMA available for language documentation projects: `https://lig-aikuma.imag.fr/`
- Several recording campaigns in Africa
- Ongoing development of new features
- Side use for ASR technology development
  - 10h of read speech quickly collected in Fongbe (Benin)
  - 7h30 of translated speech (8k sentences from BTEC corpus) quickly collected in Amharic (Ethiopia)
  - small scale data collection in Sereer (spoken in Senegal)
- Access to the source code on demand (*Laurent.Besacier@imag.fr* and *Elodie.Gauthier@imag.fr*)

Conclusion

- Ergonomic, specialized data collection tool for documentary linguists in the field
  - Lig-Aikuma
- Collected valuable data in three Bantu languages
  - Partly processed yet
  - Partly published yet (see LREC 2018 publications)
  - More to become available in the future
- Progress on language technologies for documenting unwritten languages
  - Automatic phone transcription and phone set discovery
  - Automatic word discovery, mono- and crosslingual
  - Current process of making (first) technologies available to the public, e.g., via VM or Web services

Tools for field linguists

- Close collaboration with computer scientists to analyze their needs
  - First steps taken in the project
  - Organization of training activities
    - July 3rd 2015: Workshop on Natural Language Processing Technology for Linguists in Paris at LPP
    - January 25th and 26th 2016: Linguistic training workshop for language technology experts in Paris
    - ICPhS 2019 Special Session on Computational Approaches for Documenting and Analyzing Oral Languages
- Built prototypes of tools
  - Some technologies now on the brink of usability
  - LigAikuma as standalone project already available at production grade
  - But ...

Areas of improvement and future

- Only 3 languages documented in 3 years, is that so impressive ?
- Weaknesses
  - Project would not scale to 100 languages
  - Global architecture for data collection still to be done (decentralized architecture that keeps track of all recordings, verify data integrity and handles asynchronous deletions or insertions of new recordings, robust backup, etc.)
  - Transcription bottleneck: no off-the-shelf (and robust) language independent phone recognizer yet[12]
- LigAikuma could be used to enrich existing corpora and help for larger distribution to the community (for reproducible linguistic studies)

[12] other collaboration with Florian Metze at CMU to build an universal phonetic recognizer

Lab Overview

- LIG-AIKUMA in 90mn:
  `https://lig-aikuma.imag.fr/lig-aikuma-in-90mn/`

Questions?

Thank you