

Lessons learned after development and use of a data collection app for language documentation (Lig-Aikuma)

Laurent Besacier¹, Elodie Gauthier², Sylvie Voisin³

¹LIG, Grenoble, France ²LORIA, Nancy, France ³DDL, Lyon, France

Abstract

Lig-Aikuma is a free Android app running on various mobile phones and tablets. It proposes a range of different speech collection modes (recording, respeaking, translation and elicitation) and offers the possibility to share recordings between users. More than 250 hours of speech in 6 different languages from sub-Saharan Africa (including 3 oral languages in the process of being documented) have already been collected with Lig-Aikuma. This paper presents the lessons learned after 3 years of development and use of Lig-Aikuma. While significant data collections were conducted, this has not been done without difficulties. Some mixed results lead us to stress the importance of design choices, data sharing architecture and user manual. We also discuss other potential uses of the app, discovered during its deployment: data collection for language revitalisation, data collection for speech technology development (ASR) and enrichment of existing corpora through the addition of spoken comments.

1. Introduction

Mobile apps can be now easily produced and authors such as (Drude et al., 2013) believe that an upcoming technological revolution is on the way and that we could face "a great transformation of the field triggered by an exponential increase in the use of smartphones and tablets .../... even in less developed regions of the world. .../... The development of simple and intuitive app interfaces for smartphones and tablets is having a democratizing effect, allowing for the engagement of user groups who were unable to participate in earlier phases of the digital revolution."

Using apps on mobile devices, it is now possible to collect audio and video recordings from large number of speakers with lower supervision of a researcher. Apps lower the pressure of defining the best sampling selection process, which speakers and what data exactly to collect. Moreover, additional meta informations can be collected automatically from mobile devices (geographic coordinates, movement patterns, images, time codes). For instance, images (photos taken by potentially hundred of users) can be used to enrich lexical databases or, conversely, these images can be used to elicit speech.

With such a technology, we may envision oral language documentation collections growing very large with many speakers and material to study a bunch of linguistic phenomena, from acoustic-phonetics to discourse analysis, including phonology, morphology and lexicon, grammar, prosody and tonal information. Large scale data collection also allows to collect statistically significant data, for instance on dialectal and socio-linguistic variation. However, large data collections require well organized repositories to access the content, with efficient file naming and metadata conventions that should also facilitate automatic processing.

Contribution. This paper presents the lessons learned after 3 years of development and use of a data collection app (LIG-AIKUMA). While significant data collections were conducted, this has not been done without difficulties. Some mixed results lead us to stress the importance of design choices, data sharing architecture and user manual. We also discuss other potential uses of the app, discovered

during its deployment: data collection for language revitalisation, data collection for speech technology development (ASR) and enrichment of existing corpora through the addition of spoken comments.

2. The app and its evolution

LIG-AIKUMA is an improved version of the Android application *Aikuma* initially developed by Steven Bird and colleagues (Bird et al., 2014). Features were added to the app in order to facilitate the collection of parallel speech data in line with the requirements of several field linguists interviewed. The resulting app, called LIG-AIKUMA, runs on various mobile phones and tablets and proposes a range of different speech collection modes (recording, respeaking, translation and elicitation).

The application LIG-AIKUMA has been successfully tested on different devices (including Samsung Galaxy SIII, Google Nexus 6, HTC Desire 820 smartphones and a Galaxy Tab 4 tablet). It can be downloaded from a dedicated website.¹

Table 1 presents the main features of the app. *Recording* lets simply record speech. *Respeaking*, initially introduced by Woodbury (Woodbury, 2003), involves listening to an original recording and repeating what was heard carefully and slowly. This results in a secondary recording that is much easier to transcribe later on (transcription by a linguist or by a machine). In this recording mode, parallel audio data mapping is captured (between source recording and respoken recording). *Translating* is a translation of an original recording. In *Elicitation* mode, the user can load a text, an image or a video from the device and then record read speech or comment on images/videos.

In addition to those recording modes, the app has the following features: smart generation and handling of speaker metadata (age, languages spoken, geolocalisation) ; automatic backup of interrupted sessions ; data sharing between users ; automatic generation of a consent form (from speaker's metadata) ; export to Elan software.² The inter-

¹<http://lig-aikuma.imag.fr>

²<https://tla.mpi.nl/tools/tla-tools/elan/>

Table 1: Main features of LIG-AIKUMA

FEATURES	AIKUMA	LIG-AIKUMA
Recording and documentation	✓	✓
Respeaking and oral translation	✓	✓
<i>Extras</i> : Sync. and Sharing, Geolocalisation, Textless interface	✓	✓
Elicitation (text-image-video) mode	✗	✓
User profiles, Consent form, Metadata	✗	✓
Automatic backup of interrupted sessions	✗	✓
Multilingual interface and User feedback	✗	✓
Documentation (samples, tutorial, ...)	✗	✓
Export to Elan	✗	✓

face and the documentation are available in 3 different languages (English, French and German).

3. Data collection of three oral Bantu languages

So far, LIG-AIKUMA was used to collect data in three unwritten African Bantu languages in close collaboration with three major European language documentation groups (LPP, LLACAN in France; ZAS in Germany).

Basaa, which is spoken by approximately 300,000 speakers (SIL, 2005) from the “Centre” and “Littoral” regions of Cameroon, is the best studied of our three languages. The earliest lexical and grammatical description of Basaa goes back to the beginning of the twentieth century (Rosenhuber, 1908) and the first Basaa-French dictionary was developed over half a century ago (Lemb and de Gastines, 1973). Several dissertations have focused on various aspects of Basaa (Bot ba Njock, 1970; Makasso, 2008) and the language also benefits from recent and ongoing linguistic studies (Dimmendaal, 1988; Hyman, 2003; Hamlaoui and Makasso, 2015).

Myene, a cluster of six mutually intelligible varieties (Adyumba, Enenga, Galwa, Mpongwe, Nkomi and Orungu), is spoken at the coastal areas and around the town of Lambarene in Gabon. The current number of Myene speakers is estimated at 46,000 (Lewis et al., 2013). The language is presently considered as having a “vigorous” status, but the fact that no children were found that could participate in a study on the acquisition of Myene suggests that the language is already endangered. A basic grammatical description of the Orungu variety (Ambourou, 2007) is available, as well as a few articles on aspects of the phonology, morphology and syntax of Myene ((Van de Velde and Ambourou, 2011) and references therein).

Our third and last language, **Embosi** (or alternatively Mbochi), originates from the “Cuvette” region of the Republic of Congo and is also spoken in Brazzaville and in

the diaspora. The number of Embosi speakers is estimated at 150,000 (Congo National Inst. of Statistics, 2009). A dictionary (Beapami et al., 2000) is available and, just like Basaa and Myene, the language benefits from recent linguistic studies (Amboulou, 1998; Embanga Aborobongui, 2013).

From a linguistic perspective, the three languages display a number of features characteristic of the Bantu family: (i) a complex morphology (both nominal and verbal), (ii) challenging lexical and postlexical phonologies (with processes such as vowel elision and coalescence, which bring additional complexities in the recovery of individual words), and (iii) tones that serve establishing both lexical and grammatical contrasts.

As shown in Table 2, 239h of speech data in 3 languages were collected with LIG-AIKUMA. The corpus is composed of 65h of recorded speech, 83h of respoken speech and 69h of French translations of the respoken utterances. Speech was also elicited from images or texts (22h).

Table 2: Overview of data collections for 3 oral Bantu languages made with LIG-AIKUMA

Language	Record.	Respeak.	Translat.	Elicitat.
Basaaá	23h	24h	34h	8h
Mboshi	33h	30h	30h	14h
Myene	9h	29h	5h	x
Total	65h	83h	69h	22h

More details on these data collections can be found in (Rialland et al., 2018; Hamlaoui et al., 2018) for Mbochi and Basaa respectively. A subset of 5k utterances in Mbochi was also provided to the community for computational language documentation experiments (Godard et al., 2018) and is distributed by ELRA.³

4. Lessons learned

4.1. Recording outdoors and indoors

A frequently asked question from linguists about the mobile app concerns the quality of the recordings obtained. Our experience, after recording several hundreds hours of speech, shows that modern smartphones and tablets are equipped with good microphones that provide good quality recordings for further human or automatic analysis. It is important, however, to control the recording environment and following recommendations can be made on this aspect. First, the smartphone (or tablet) must be placed on a table rather than being handled at the time of the recording, in order to avoid unwanted noises due to phone manipulation. Secondly, recording should be done preferably indoors to limit environmental noises (sounds of children or animals, footsteps, discussions, wind, etc). Even indoors, to prevent reverberation and echo, it is preferable not to stand too close to a wall and, if the room is poorly furnished, the walls should be covered with a heavy fabric to muffle the sound. Should the recording occur outdoors, the

³Available for free at: <http://catalogue.elra.info/en-us/repository/browse/ELRA-S0396/>

use of a lapel microphone is strongly recommended in order to be as close as possible to the speech source. Finally, another issue faced with the mobile app use outdoors is the reflection of the sun on the tablet/smartphone screen, which sometimes makes it difficult to view videos or images for the elicitation mode (especially with elderly speakers having vision problems).

4.2. Importance of metadata

Filling in metadata is sometimes considered a tedious task by users. This sometimes resulted in incomplete or too quickly filled forms. A consequence is that missing data is found on return from the field trip. For this reason, we developed lately a new feature called *speaker profiles* which saves all speakers' metadata automatically (when metadata of a new speaker is filled in, it is saved - in a so called *profile* - and can be retrieved with an *import* button). Moreover, the possibility of using speaker metadata to automatically generate consent sheets has been proposed and implemented: a *pdf* file is automatically generated according to a template prepared by the user and then filled in with the speaker's information (age, name, etc.). During the use of the app, we also realized the need to add other metadata / informations *at the end of* a recording session (about recording conditions, speaker behavior, etc.) but this feature does not exist yet in our code and is left for future improvements. Finally, the need to annotate non-speech segments, expressed by the users, lead us to introduce this possibility for the *respeaking* and *translation* modes. The time codes of the non speech segments are then stored in a file placed in the same folder as the recording.

4.3. Limited autonomy of mobile devices

The application runs on a mobile terminal with non infinite autonomy (usage time and memory). It is thus essential to plan recording sessions that do not exceed this autonomy and to plan "rest" periods used to recharge the phone's battery and export the recorded data to a laptop or to an external disk. This is all the more important as the mobile phone (or tablet) may be used for other functions such as capturing images and videos. Consequently, when developing the app, we paid particular attention to optimizing the code to preserve the autonomy of the mobile device. In the event of a phone shutdown or crash in the middle of a recording session, we also provide a complete backup of the session history in order to avoid losing data and to be able to return to the exact point of the current session after recharging the phone.

4.4. Global architecture for data collection

An engineer was responsible for gathering data from all the different linguists in 3 bantu languages, backing it up on a single server and checking its integrity. This process must be facilitated if we want to scale up to 100 languages. For instance, the data collected on each deployed mobile device must be regularly deposited (synchronized) on a back-end server that ensures data backup and integrity. Such a global architecture for data collection still needs to be designed. Ideally, we would like an architecture that keeps track of all recordings distributed through N autonomous

mobile devices, that addresses internet connectivity issues, that verifies data integrity and facilitates automatic quality control. It should also provide less labour-intensive data uploading and compiling.

4.5. Need for documentation and tutorials

In the first few months of the project, misunderstandings about how to use the app led us to write documentation in three languages (French, English, German). We have also added a video tutorial (in French with English subtitles) to these written documents, as well as a quick introduction to the application in the form of a 90mn practical exercise.⁴

5. Future extensions and opportunities

During the use of LIG-AIKUMA, we discovered several extensions and opportunities, not identified at the beginning of the initial project. These are briefly described in this section.

5.1. Data collection for language revitalisation

For Mbochi language, we also recorded, with the app, 1500 pictures illustrating plants, artifacts, animals and everyday activities to be included later on in an Encyclopedia or to be archived as culturally sensitive or to be included in an image book for language teaching. These pictures (see figure 1) were commented by 2, 3 or 4 speakers. Each comment lasted between 20 seconds to 3 minutes. Each image is therefore associated with a recording corresponding to a discussion, between several speakers, about that image.

Figure 1: Example of local pictures used for speech elicitation. Listening to the corresponding recordings, one clearly distinguishes several repetitions of a word corresponding to the main object of the image (left: *ambamba* ; right: *don-godongo*)



5.2. Data collection for speech technology development (ASR)

Originally intended for language documentation and data collection in the field, our app has also been useful for collecting speech for technological development purposes targeting under resourced languages. For instance, 10h of read speech in Fongbe (spoken especially in Benin, Togo, and Nigeria) were collected using the *elicitation* (from text) mode of the app. The first ever ASR system for this language was trained using this initial corpus (Laleye et al., 2016). Similarly, 7h40 of speech in Amharic (Ethiopia) was recorded after translating the Basic Traveler Expression Corpus (BTEC) under a normal working environment

⁴see <https://lig-aikuma.imag.fr/tutorial/> for more details

(Melese et al., 2017). An ASR system was trained to recognize basic travel expressions in Amharic. These two data collections, accelerated by the use of the app, allowed two African doctoral students to quickly record a dataset for their researches in automatic speech recognition (ASR). An ASR system for Wolof (Senegal) was also developed in (Gauthier et al., 2017) and used for analyzing vowel length contrast in different dialectal variants of Wolof.

5.3. Enrichment of existing corpora through the addition of spoken comments

In discussions with several linguists, we realized that many corpora have already been collected to document the world's languages. These data, which exist in different formats (digital or not), are undoubtedly valuable resources that must be safeguarded and enriched. There is a risk that these corpora will disappear with the linguist who recorded them. We think that the *respeaking* mode of the app could allow the linguist to enrich recordings with an additional tier of spoken comments. If speakers of the language are available, it may also be possible to make them repeat part of the dataset under more favorable acoustic conditions (using the *respeaking* mode).

6. Conclusion

This article summarized three years of development and testing of a mobile application for collecting speech in the field: LIG-AIKUMA. A significant amount of speech could be collected to document three Bantu languages. This shows the potential of the app. However, we also presented, in this article, its limitations and the problems encountered during data collection. We also discussed other uses of the app, discovered during its deployment: data collection for language revitalisation or for speech technology development (ASR), enrichment of existing corpora through the addition of spoken comments, etc. LIG-AIKUMA can be downloaded on <https://lig-aikuma.imag.fr> and its source code is also available on <https://gricad-gitlab.univ-grenoble-alpes.fr/besaciel/lig-aikuma>.

7. Bibliographical References

- Amboulou, C. (1998). Le Mbochi: langue bantoue du Congo Brazzaville (Zone C, groupe C20). Ph.D. thesis, INALCO, Paris.
- Ambourou, O. (2007). Eléments de description de l'orungu, langue bantou du Gabon (B11b). Ph.D. thesis, Université Libre de Bruxelles.
- Beapami, R. P., Chatfield, R., Kouarata, G., and Waldschmidt, A. (2000). Dictionnaire Mbochi - Français. SIL-Congo, Brazzaville.
- Bird, S., Hanke, F. R., Adams, O., and Lee, H. (2014). Aikuma: A mobile app for collaborative language documentation. ACL 2014, page 1.
- Bot ba Njock, H.-M. (1970). Nexus et nominaux en bàsàa. Ph.D. thesis, Université Paris 3 Sorbonne Nouvelle.
- Dimmendaal, G. (1988). Aspects du basaa. Peeters/SELAF. [translated by Luc Bouquiaux].
- Drude, S., Birch, B., Broeder, D., Withers, P., and Wittenburg, P. (2013). Crowdsourcing and apps in the field of linguistics: Potentials and challenges of the coming technology. Technical report, The Language Archive, Max Planck Institute for Psycholinguistics.
- Embanga Aborobongui, G. M. (2013). Processus segmentaux et tonals en Mbondzi – (variété de la langue embosi C25). Ph.D. thesis, Université Paris 3 Sorbonne Nouvelle.
- Gauthier, E., Besacier, L., and Voisin, S. (2017). Machine assisted analysis of vowel length contrasts in wolof. In Proceedings of Interspeech, Stockholm, Sweden, August 2017.
- Godard, P., Adda, G., Adda-Decker, M., Benjumea, J., Besacier, L., Cooper-Leavitt, J., Kouarata, G., Lamel, L., Maynard, H., Müller, M., Rialland, A., Stüker, S., Yvon, F., and Boito, M. Z. (2018). A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments. In Proc. LREC, Miyazaki, Japan.
- Hamlaoui, F. and Makasso, E.-M. (2015). Focus marking and the unavailability of inversion structures in the Bantu language Bàsàá. Lingua, 154:35–64.
- Hamlaoui, F., Makasso, E., Müller, M., Engelmann, J., Adda, G., Waibel, A., and Stüker, S. (2018). Bulbasaa: A bilingual basaa-french speech corpus for the evaluation of language documentation tools. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.
- Hyman, L. (2003). Basaa (A43). In Derek Nurse et al., editors, The Bantu languages, pages 257–282. Routledge.
- Laleye, F. A. A., Besacier, L., Ezin, E. C., and Motamed, C. (2016). First automatic fonbe continuous speech recognition system: Development of acoustic models and language models. In Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016, Gdańsk, Poland, September 11-14, 2016., pages 477–482.
- Lemb, P. and de Gastines, F. (1973). Dictionnaire Basaa-Français. Collège Libermann, Douala.
- Paul M Lewis, et al., editors. (2013). Ethnologue: Languages of the World. SIL International, Dallas, Texas, seventeenth edition.
- Makasso, E.-M. (2008). Intonation et mélismes dans le discours oral spontané en bàsàa. Ph.D. thesis, Université de Provence (Aix-Marseille 1).
- Melese, M., Besacier, L., and Meshesha, M. (2017). Amharic-english speech translation in tourism domain. In Proceedings of the Workshop on Speech-Centric Natural Language Processing, SCNLP@EMNLP 2017, Copenhagen, Denmark, September 7, 2017, pages 59–66.
- Rialland, A., Adda-Decker, M., Kouarata, G.-N., Adda, G., Besacier, L., Lamel, L., Gauthier, E., Godard, P., and Cooper-Leavitt, J. (2018). Parallel corpora in mboshi (bantu c25, congo-brazzaville). In LREC.
- Rosenhuber, S. (1908). Die Basa-Sprache. MSOS, 11:219–306.
- van de Velde, M. and Ambourou, O. (2011). The

grammar of Orungu proper names. Journal of African Languages and Linguistics, 23:113–141.

Woodbury, A. C., (2003). Defining documentary linguistics, volume 1, pages 35–51. Language Documentation and Description, SOAS.