Parallel Corpora in Mboshi (Bantu C25, Congo-Brazzaville)

Annie Rialland, Martine Adda-Decker, Guy-Noël Kouarata, Gilles Adda, Laurent Besacier, Lori Lamel, Elodie Gauthier, Pierre Godard, Jamison Cooper-Leavitt

Parallel Corpora in Mboshi (Bantu C25, Congo-Brazzaville)

Annie Rialland[1], Martine Adda-Decker [1], Guy-Noël Kouarata[1], Gilles Adda[2], Laurent Besacier[3], Lori Lamel[2], Elodie Gauthier[3], Pierre Godard[2], Jamison Cooper-Leavitt[2]
[1] LPP, CNRS-Paris 3/Sorbonne Nouvelle, France, [2] LIMSI, CNRS, Université Paris-Saclay, France, [3] Laboratoire d'Informatique de Grenoble (LIG)/GETALP group, France

{annie.rialland, martine.adda-decker}@univ-paris3.fr, guy_kouarata@yahoo.com,
{gilles.adda, lamel, pierre.godard, cooperleavitt}@limsi.fr
{laurent.besacier, elodie.gauthier}@univ-grenoble-alpes.fr

Abstract
This article presents multimodal and parallel data collections in Mboshi, as part of the French-German BULB project. It aims at supporting documentation and providing digital resources for less resourced languages with the help of speech and language-based technology. The data collection specifications thus have to meet both field linguists' and computer scientists' requirements, which are large corpora for the latter and linguistically dense data for the former. Beyond speech, the collection comprises pictures and videos documenting social practices, agriculture, wildlife and plants. Visual supports aimed at encouraging people to comment on objects which are meaningful in their daily lives. Speech recordings are composed of the original speech in Mboshi, a respoken version and a translated version to French. These three speech streams remain time-aligned thanks to LIG-AIKUMA, which adds new features to a previous AIKUMA application. The speech corpus includes read material (5k sentences, Bible), verb conjugations and a large part of spontaneous speech (conversations, picture descriptions) resulting in over 50 hours of Mboshi speech, of which 20 hours are already respoken and orally translated to French. These parallel oral data are intended for linguistic documentation (tonology, phonology...) and automatic processing (corpus annotation, alignment between Mboshi speech and French translations).

Keywords: fieldwork, ASR, Bantu, Mboshi, French, speech, parallel corpus, pictures

1. Introduction

According to UNESCO, 43% of the 6000 estimated languages of the World are endangered. Even languages spoken by more than a million people can be threatened due to lack of transmission from one generation to the next one. Archiving languages (as well as the knowledge associated to them) is currently an emergency and an overwhelming task. The challenge is even greater when the languages are under-resourced, often lacking writing system, written texts and translated corpora. The BULB (Breaking the Unwritten Language Barrier) project aims at providing tools to language documentation and description for unwritten languages (or languages with scarce textual material) with the help of language-based technologies (in section 2). The data collection specifications thus have to meet both field linguists' and computer scientists' requirements, which are large corpora for the latter and linguistically dense data for the former. By linguistically dense, we mean calibrated material to speed up the grammar development, typically conjugations and sentence lists as proposed by Bouquiaux and Thomas (1976) among others.

The first step of the methodology of the BULB project involves the collection of parallel corpora (speech, respeaking and oral translation). They could be recorded with an application developed within the BULB project: LIG-AIKUMA, a user friendly device for linguists and speakers of the language community (in section 3). In this article, we will focus on the Mboshi language (in section 4), on the parallel corpora collected for this language (in section 5) and the first linguistic analysis done on this corpus with the help of automatic processing (section 6).

2. The BULB project

BULB (Breaking the Unwritten Language Barrier) is a French-German project supported by the French ANR and the German DFG and began in 2015. It relies on a cooperation between linguists and speech researchers in both countries. The institutions involved are KIT (Karlsruhe), University of Stuttgart, ZAS (Berlin) in Germany, LIMSI (Orsay), Laboratoire de Phonétique et Phonologie (Paris), LLACAN (Villejuif), LIG (Grenoble) in France.

Its central goal is to support language documentation and elaboration of resources for unwritten languages or less-resourced languages with the help of language-based technology, particularly automatic speech recognition and machine translation (Adda and al., 2016, Stüker and al., 2016). The methodology of the project follows Bird and al. (2014). Its steps are the following:

Collection of multimodal and parallel corpora

- collection of a large oral corpus (source language)
- respeaking of this corpus by a reference native speaker (to eliminate noises, hesitations, speaker variations)
- parallel and oral translation into a language that has access to speech technology, particularly ASR (French, the target language)
- collection of pictures and videos documenting local life and culture

Automatic treatments

- automatic time-alignment

- automatic phone annotation of source and respeaking
- automatic segmentation into words/morphemes
- automatic alignment between words/morphemes of the source and target languages.

Within the BULB project, three example languages are being addressed : Myene (Gabon), Basaa (Cameroun) and Mboshi (Congo-Brazzaville), the data collection of which are directed by LLACAN, ZAS and LPP respectively. Most of these corpora were created with LIG-AIKUMA, an application developed within the BULB project and presented section in 3.

3. LIG-AIKUMA

The initial smartphone application AIKUMA was developed by Bird and al. [2014] for the purpose of saving time in the process of language documentation and providing rather quickly, large amount of resources which could be processed and analysed later on. It enables time-aligned recording of speech with respeaking or oral translation. LIG-AIKUMA improved AIKUMA in various ways (Gauthier and al. 2016). It includes a full pipeline between recording, respeaking and translation, allowing to obtained time-aligned data between the three modes. It also implemented additional features : elicitation of speech from texts, images or videos. More detailed meta-information for speakers could be introduced and the naming of files was clarified, being based on the date and time of the recordings. Feedbacks to the user were also implemented and the application was adapted for tablets. It is downloadable from the following address : https://lig-aikuma.imag.fr and from Google Play.

4. The Mboshi language

4.1 Mboshi as an under-resourced language

Mboshi is a Bantu language (C25) of Congo Brazzaville. It is spoken in the «Cuvette region» and in Brazzaville, as well as in the diaspora. The estimated number of speakers in the «Cuvette region» is 140000 (Embanga Aborobongui, 2013). The number of speakers in Brazzaville and in the diaspora is unknown. A writing system was developed by missionaries but there is no standardised form of the orthography. There are very few texts in Mbochi, mainly translations of the Bible.

Mboshi has linguistic resources, that is, grammatical studies (Fontaney, 1988, 1989; Amboulou, 1998; Embanga Aborobongui, 2013; Kouarata, 2014) a Mboshi-French dictionary (Beapami and al., 2000) and a Mboshi-English dictionary (Ndongo Ibara, 2012). Meanwhile, it has no digital resources. Thus, Mboshi has no on-line dictionary yet.

4.2 Some linguistic characteristics of Mboshi

Mboshi is a typical Bantu language. It is classified C25, in zone C which belongs to the Western part of the Bantu domain. It has a seven vowel system (i, e, ɛ, a, ɔ, o, u) with an opposition between long and short vowels. Its consonantal system includes the following phonemes: p, t, k, b, d, β, l, r, m, n, ɲ, mb, nd, ndz, ng, mbv, f, s, ɣ , pf, bv, ts, dz, w, j. We can notice the absence of /g/, the presence of a voiced bilabial fricative / β/ and a set of prenasalized consonants (mb, nd, ndz, ng, mbv) which are common in Bantu languages (Embanga Aborobongui 2013, Kouarata 2014). The Mboshi prosodic system involves two tones and an intonational organisation without downdrift (Rialland and Embanga Aborobongui 2016). The syllable structures are simple (V, CV, CVV, Cj/wV, Cj/wVV).

While Mboshi has unremarkable vocalic and consonantal systems, it has quite complex phonological rules. A process is particularly frequent: the deletion of a vowel before another vowel, which occurs at 40% of word junctions (Rialland and al. 2015) as exemplified in (1) and (2).

(1) o-kondzi + áseri → okondzáseri
cl1.chief + 3sg.say.REC.
«The chief said.»

(2) o-yúlu + álámbi → o-yúlálámbi
cl1.femme + 3sg.cook.REC.
«The woman cooked.»

This process, which is common in Bantu languages, tends to obscure word segmentation and introduces an additional challenge for automatic processing, particularly automatic word segmentation and dictionary creation.

Mboshi has a noun class prefix system, which is another typical feature of Bantu languages. However, it has an unusual rule of deletion targeting the consonant of prefixes: a prefix consonant drops if the root begins with a consonant (Rialland and al. 2015). This rule triggers the formation of many words beginning by a vowel as shown by the following nouns involving the class prefix -ba.

(3) ba+kondzi → akondzi «chiefs»

(4) ba+kúsu → akúsu «tortoises»

(5) ba+ ási → bási «wives»

(6) ba+ ána → bána «children»

This type of rule is shared by a small group of languages in this area. The structure of the verb in Mboshi is also characteristic of a Bantu verb. Its structure is the following : Subject Marker – Tense/Aspect/Mood Marker/ – root-derivative extensions – Final Vowel. A verb can be very short or quite long, depending of the markers involved as shown by the examples (7) and (8):

(7) a-bva-í [bvé] «he falls»
SM3sg – fall - FV

(8) í-mi-ding-im-a [ímidingima] «he was loved»
SMcl5- PERF-love-PASS-FV

5. Composition and collection of the parallel corpora

The parallel corpora collection involves three phases : 1) collection of audio corpora, 2) Respeaking, 3) Translation. All of them were recorded with LIG-AIKUMA. A small part was also transcribed manually.

5.1 Phase 1: Composition and collection of audio corpora

Oral corpora include: a) read sentences, b) debates, c) conjugations, d) reading of Bible, e) comments on pictures

a) Read sentences.

A total of 5178 read sentences were collected. Among these, 3706 sentences were extracted from the Mboshi-French dictionary (Beapami and al., 2000) and 1472 were Mboshi translations of examples from Bouquiaux and Thomas's corpus designed for fieldwork (Bouquiaux and Thomas, 1976). These two sets of sentences were recorded using the text-based elicitation function of LIG-AIKUMA. An Mboshi written sentence associated with its translation in French was displayed on the tablet screen. The tablet was given to speakers who could manipulate it by themselves, after a short period of instruction. The recording of these sentences was divided between 3 speakers and its was made in Congo-Brazzaville. The total duration of this recording is 4h51.

This 5k read sentences corpus was used for acoustic-phonetic studies as described in section 6 (Cooper-Leavitt J. and al. 2017 a, b). It provided also a testing ground for machine learning studies e.g. unsupervised word discovery from speech and recently, it was used during the Jelinek Summer Workshop on Speech and Language Technology (JSALT) 2017 in CMU, Pittsburgh (Godard and al. 2016, Godard and al. 2018).

b) Debates

Debates were preferred to monologues or life stories as they provide an interaction which favors spontaneous speech and natural exchanges. They were moderated by one of the co-authors of this article: Guy-Noël Kouarata, who is also a native speaker and a linguist. There are 67 debates, each of them involving between 2 and 5 speakers. Altogether there were 19 participants in these debates. To avoid any overlapping, speakers were instructed not to speak at the same time. The debates deal with a variety of topics concerning traditions or reflecting current concerns, such as the death of an old person, the structure of a song, regional mushrooms, alimentation before and after the building of a good road to access the region, sorcellery, deseases, immigration… Some of these debates are valuable in terms of patrimony and some encode cultural knowledge (about plants, or mushrooms, for example). These debates were recorded with a tablet in a home in Brazzaville as shown in the following picture (1):

Figure 1. Recordings of debates with a tablet in a home in Brazzaville

Each debate lasts between 24 and 80 minutes. The whole duration of the recorded debates was 25h18mn whose 20h22 were respoken and translated (see 5.2 and 5.3.). The speakers signed a consent form for being recorded and have their recordings made public but some of these debates, being culturally sensitive, might have to be checked by specialists (botanists, for example) for an appropriate avaibility.

c) Conjugations

To facilitate the process of grammar development, verb conjugations are of special interest. Verb morphology is complex in Mboshi as well as in Bantu languages in general. This complexity introduces additional challenges for automatic processing, particularly in word segmentation and mapping between Mboshi word/morphemes and their French counterparts in translation. Conjugations were added to the corpora as linguistic facilitators for the planned processings. We knew also that it was quite easy to make systematic recording of conjugations as we could build on the experience that speakers acquire in conjugations during their schooling, which is in French from the beginning in Congo-Brazzaville. The corpora was designed to cover a large part of the verbal conjugations. 50 verb based on the main root patterns (C (V), CV or CVC) were conjugated at 15 tenses/aspects. The subjects were also varied, to capture the agreements between the subjects and the verbs. 18 different subjects (pronouns or nouns with various class prefixes) were necessary to cover the spectrum of the agreement patterns.

These conjugations were recorded by one speaker, based on conjugation tables (lists of verbs, subjects, tenses). The total duration of this recording is 5h56mn. No respeaking or translation of these conjugated forms were done as the verbs and their conjugations were known.

d) Reading of Bible extracts

Extracts of Bible in Mboshi were read by 6 speakers. The total amount of these recordings is: 4h06mn. No respeaking or translation had to be done.

e) Comments on pictures

Guy-Noël Kouarata took 1500 pictures illustrating plants, artifacts, animals and everyday activities to be included later on in an Encyclopedia or to be archived as culturally sensitive. Figures (2) and (3) provide a sample of these pictures.

Figure 2. Mask for the kyebe-kyebe danse and ceremony

Figure 3. Traditional fishtrap

These pictures were commented by 2, 3 or 4 speakers. Each comment lasted between 20 seconds to 3 minutes. Currently, they are not respoken or translated but kept for further completion and processing. . .

5.2 Phase 2: Respeaking

The task of respeaking was performed by three different native speakers in a quiet room in Brazzaville. Speakers were instructed to repeat, eliminating hesitation pauses, speaking slowly, still naturally. The task was found more difficult than expected, with a strong tendency among the speakers to come back to their usual rate of speech, which was often quite fast. The respeaking was made for 20h22 mn of source data.

5.3 Phase 3: Translation

The translation in French was performed by Guy-Noël Kouarata in a quiet room. A rather literal translation was preferred in order to improve the possibility of matching automatically Mboshi words and their translation to French. The translation function of LIG-AIKUMA was used. The whole pipeline (oral audio corpus, respeaking, translation) was obtained for 20h22mn hours of source speech.

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

Table 1. Mboshi parallel corpora : current state.

5.4 Manual transcription

A small part of the debates was transcribed, based on the annotation conventions of the Mboshi-French dictionary (Beapami and al. 2000). The duration of the manually transcribed part is 1h 10 minutes.

6. First linguistic analysis of the corpus with the help of automatic treatments

Forced text-to-speech alignment was applied to the 5178 read sentences. LIMSI's STK speech processing toolkits for the ASR were used (Gauvain and Lamel, 2003; Lamel and Gauvain 2005). A variant dictionary was built, based on the manual transcripts and rules generating variants. With forced alignment and pronunciation variants in the dictionary, various types of vowel deletion or morpheme deletion at word junction (triggered by phonetic processes or specific to some morphemes, in particular, the connective ones) could be sorted out and better quantified and understood. (Cooper-Leavitt J. and al. 2017 a, b).

7. Conclusion

This article presents multimodal and parallel corpora recorded in Mboshi (Bantu C25). Their current state is summarized in Table 1. These corpora were designed to document the language itself, to contribute to the patrimony preservation of Mboshi people and to provide materials which could be processed by current language-based technologies. Parallel corpora collection with a smartphone/tablet and the friendly-user application AIKUMA can also have other purposes than linguistic ones, for example in music, as songs and lyrics can be stored in parallel. The final corpus which will contain more annotations, respeaking and oral translation will be made available to the research community at the end of the BULB project.

8. Acknowledgements

This work was funded by the French ANR and the German DFG under grant ANR-14-CE35-0002. It also received financial support from the EFL LABEX (10-LABX-0083).

9. Bibliographical References

Adda, G., Stücker, S., Adda-Decker, M., Ambouroue, O., Besacier, L., Blachon, D., Bonneau-Maynard, H., Godard, P, Hamlaoui F., Idiatov D., Kouarata G.-N., Lamel, L, Makasso, E.-M., Rialland, A., Van de Velde, M., Yvon F., and Zerbian S. (2016). Breaking the unwritten language barrier: The Bulb project. *In Proceedings of SLTU (Spoken Language Technologies for Under-Resourced Languages)*, Yogyakarta, Indonesia

Amboulou, C. (1998). *Le Mbochi : langue bantoue du Congo Brazzaville (zone C, groupe C20)*. Ph.D. thesis, INALCO, Paris.

Beapami, R. P., Chatfield R., Kouarata, G.-N. and Embengue Waldschmidt, A. (2000). *Dictionnaire Mbochi-Français*. SIL-Congo Publishers, Brazzaville.

Bird, S., Hanke, F. R., Adams, O. and Lee, H. (2014). Aikuma: A mobile app for collaborative language documentation. *Workshop on the Use of Computational Methods in the Study of Endangered Languages*, Baltimore, USA. pp 1-5

Bouquiaux, L. and Thomas J. (1976). *Enquête et description des langues à tradition orale*. SELAF, Paris, France

Cooper-Leavitt J., Lori L., Rialland A., Adda-Decker M., Adda G. (2017 a). Developing an Embosi (Bantu C25) speech variant dictionary to model vowel elision and morpheme deletion. In *Conference of Interspeech 2017 Proceedings.*

Cooper-Leavitt J., Lori L., Rialland A., Adda-Decker M., Adda G. (2017 b). Corpus base linguistic exploration via forced alignments with a "light-weight" ASR tool. In *Workshop on Language Technology*, Poznan, Poland.

Embanga Aborobongui M. (2013). *Les processus segmentaux et tonals en mbondzi (variété de la langue Embosi. C25)*. Ph.D thesis, Paris 3 University, Paris.

Fontaney, L. (1989). Mboshi: steps toward a Grammar: Part I. *Pholia* 3, pp. 87-169

Fontaney, L. (1989). Mboshi: steps toward a Grammar: Part II. *Pholia* 4, pp. 71-131

Gauthier, E., Blachon D., Besacier L., Kouarata, G.-N., Adda-Decker, M. and al. (2016). LIG-AIKUMA: a Mobile App to Collect Parallel Speech for Under-Resourced Language Studies. *Interspeech 2016* (short demo paper), Sept 2016, San-Francisco, Interspeech 2016 Proceedings.

Gauvain, J.-L. and Lamel L., 2003. Large vocabulary speech recognition based on statistical methods. In Chou and Wang J. (eds.), *Pattern Recognition in Speech and Language Processing*. CRC Press, pp. 149–189.

Godard, P., Adda, G., Adda-Decker, M., Benjumea, J., A., Besacier, Cooper-Leavitt J., L., Kouarata, G.-N., Lamel L., Maynard, H., Rialland A., Stüker, S., Yvon, F. and Zanon-Boito M. (2018). A Very Low Resource language Speech Corpus for Computational language Documentation Experiments, In *LREC 2018* (in press), Japan

Godard, P., Adda, G., Adda-Decker, M., Allauzen, A., Besacier, L., Bonneau-Maynard, H., Kouarata, G.-N., Löser, K. Rialland A., and Yvon, F. Preliminary Experiments on Unsupervised Word Discovery in Mboshi. In *Interspeech 2016*, San Francisco, California, USA

Kouarata, G.-N. (2014) *Variations de forme dans la langue Mbochi (Bantu C25)*, Ph.D. thesis, Lyon 2 University, Lyon.

Lamel, L. and Gauvain J.-L. (2005). *Speech recognition*. In R. Mitkoc (ed.). The Oxford Handbook of Computational Linguistics, Oxford University Press, Oxford, pp. 305–32

Ndongo Ibara Y. P. (2012). *Embosi-English Dictionary*. Peter Lang, Frankfurt

Rialland, A., Embanga Aborobongui M., Adda-Decker, M. and Lori Lamel, (2015). Dropping of the class-prefix consonant, vowel elision and automatic phonological mining in embosi (Bantu C 25). In Kramer, R., Zsiga E., and O. Tlale Boyer (eds.), *Selected Proceedings of the 44th Annual Conference on African Linguis*tics. Cascadilla Press, Somerville, USA, pp 221-230

Rialland A. and Embanga Aborobongui M. (2016) How intonations interact with tones in Embosi (Bantu C25), a two-tone language without downdrift. In L. Downing and A. Rialland (eds.), *Intonation in African tone languages.* Mouton de Gruyter, Berlin. pp. 195-222

Stücker, S., Adda; G., Adda-Decker, M., Ambouroue, O., Besacier, L., Blachon, D., Bonneau-Maynard, H., Godard, P, Hamlaoui F., Idiatov D., Kouarata G.-N., Lamel, L, Makasso, E.-M., Rialland, A., Van de Velde, M., Yvon F., and Zerbian S. (2016). Innovative technologies for under-resourced language documentation: The Bulb project. In *Proceedings of CCURL (Collaboration and Computing for UnderResourced Languages: toward an Alliance for Digital Language Diversity)*, Portoroz Slovenia.

.