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

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

▶ To cite this version:


HAL Id: hal-01710043
https://hal.archives-ouvertes.fr/hal-01710043
Submitted on 15 Feb 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Parallel Corpora in Mboshi (Bantu C25, Congo-Brazzaville)

Annie Riall1, Martine Adda-Decker1, Guy-Noël Kouarata1, Gilles Adda2, Laurent Besacier3, Lori Lamel2, Elodie Gauthier3, Pierre Godard2, Jamison Cooper-Leavitt2

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
 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 Mboshi, mainly translations of the Bible. Mboshi has linguistic resources, that is, grammatical texts in Mbochi, mainly translations of the Bible.

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 Mboshi, mainly translations of the Bible.

Mboshi has linguistic resources, that is, grammatical studies (Fontaney, 1988, 1989; Amboulo, 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, ë, a, ô, o, u) with an opposition between long and short vowels. Its consonantal system includes the following phonemes: p, t, k, b, d, ñ, l, r, m, n, mb, nd, ndz, ng, mbv, f, s, ɣ, s, pf, bv, ts, dz, w, j. We can notice the absence of /g/, the presence of a voiced bilabial fricative /ɲ/ and a set of prenasalised 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/vW, Cj/vWV).

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 + ášri → okondzáseri
c11.chief + 3sg.say.REC.
«The chief said.»

(2) o-yúlu + álámbi → o-yúyáámbi
c11.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+kúsù → akúsu «tortoises»

(5) ba+áši → 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 [midimíga] «he was loved»
SM3d5- 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.
Figure 1. Recordings of debates with a tablet in a home in Brazzaville

Figure 2. Mask for the kyebe danse and ceremony
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.

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

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


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


