BULBasaa: A Bilingual Bàsàá-French Speech Corpus for the Evaluation of Language Documentation Tools

Fatima Hamlaoui$^{\diamond\ddagger}$, **Emmanuel-Moselly Makasso**‡, **Markus Müller***, **Jonas Engelmann**‡, **Gilles Adda**†, **Alex Waibel****, **Sebastian Stüker***

$^{\diamond}$University of Toronto, Toronto, Canada
‡Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin, Germany
*Karlsruhe Institute of Technology, Karlsruhe, Germany
†LIMSI, CNRS, Paris, France
*LTI, CMU, Pittsburgh, USA
contact: f.hamlaoui@utoronto.ca, m.mueller@kit.edu, sebastian.stueker@kit.edu

Abstract

Bàsàá is one of the three Bantu languages of BULB (*Breaking the Unwritten Language Barrier*), a project whose aim is to provide NLP-based tools to support linguists in documenting under-resourced and unwritten languages. To develop technologies such as automatic phone transcription or machine translation, a massive amount of speech data is needed. Approximately 50 hours of Bàsàá speech were thus collected and then carefully re-spoken and orally translated into French in a controlled environment by a few bilingual speakers. For a subset of ≈10 hours of the corpus, each utterance was additionally phonetically transcribed to establish a golden standard for the output of our NLP tools. The experiments described in this paper are meant to provide an automatic phonetic transcription using a set of derived phone-like units. As every language features a specific set of idiosyncrasies, automating the process of phonetic unit discovery in its entirety is a challenging task. Within BULB, we envision a workflow where linguists are able to refine the set of automatically discovered units and the system is then able to re-iterate on the data, providing a better approximation of the actual phone set.

Keywords: Bàsàá, Northwest Bantu, Computational linguistics, unsupervised phone discovery, under-resourced languages

1. Introduction

About two-thirds of the approximately 7,000 languages spoken today count less than 1,000 speakers and are in danger of disappearing (Simons and Fennig, 2017). In a world that is rapidly changing, some estimate that 70% to 90% of the languages spoken today will fall out of use by the end of this century. To provide communities with modern tools that will increase the vitality of their language and support its use in a variety of contexts, as well as to help linguists in their efforts to learn about human cognition though the study of language diversity, the BULB (Breaking the Unwritten Language Barrier) project has been developing Natural Language Processing-based tools to help and accelerate language documentation. To achieve this goal, a critical mass of speech data is necessary. In the present paper, we describe the corpus created for Bàsàá (Cameroon), one of the three under-resourced Bantu languages on which the project concentrates and outline a method for the unsupervised evaluation of the derived phone set. The other two languages part the of the project are Myene (Gabon) and Mboshi (Republic of Congo) (Godard et al., 2018). We first provide basic information about Bàsàá in Section 2. In Section 3., we describe the data collection process and provide a detailed overview of the corpus. In the remainder of the paper, we concentrate on the subset of our data that was used for evaluation. Data preparation steps are outlined in Section 4. and an example use of the *BULBasaa* corpus is shown in Section 5. Section 6. concludes the paper.

2. Bàsàá

Bàsàá (A43 in (Guthrie, 1948)) is spoken in the south of Cameroon, by approximately 300,000 speakers (Simons and Fennig, 2017). It is a two tone language (High and Low), that on the surface contrasts High, Low, Falling, Rising and downstepped High tones. On the segmental level, Bàsàá has a 7-vowel system (Figure 1), in which vowel length is contrastive.

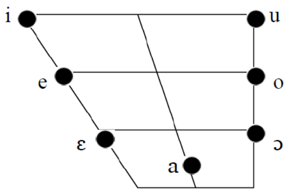


Figure 1: Bàsàá vowels (Makasso and Lee, 2015)

As discussed in detail by (Hyman, 2003), the phonology of Bàsàá is, in many ways, not very typical of Bantu languages. Both open and closed syllables are allowed, and surface syllable onsets are not required. Its consonant system is particularly complex (Table 1) and consonant oppositions depend on the position of the syllable within the prosodic stem. In a prosodic stem that contains up to three syllables (four consonants and three vowels), the total number of underlying consonant contrasts is only possible on the first consonant of the stem and progressively decreases as one reaches the end of the stem (Hyman, 2003; Makasso and Lee, 2015).

In the context of Northwest Bantu languages, a group that displays more diversity than in the rest of the Bantu family (Bearth, 2003), Bàsàá is a rather well described language. Just like most of the languages of its family (and proba-

	Bilabial	Alveolar	Palatal	Velar	Lab. velar	Uvular	Glottal
Plosive	p	t		k	k^w g^w		
Affricate			ʧ ʤ				
Implosive	ɓ						
Prenasalized	^{m}b	nd	^{n}ʤ	^{ŋ}g			
Nasal	m	n	ɲ	ŋ	ŋ^w		
Tap/Flap		ɾ̥ ɾ					
Fricative	ɸ β	s		x ɣ		χ	h ɦ
Approximant	w		j				
Lateral approx.		l					

Table 1: Bàsàá consonants (Makasso and Lee, 2015)

bly beyond, of Africa), most linguistic studies are however based on a handful of speakers and on elicited data. This methodology has numerous advantages and has yielded significant results. Detailed descriptions of its phonology, its morphology and its syntax are available. Entire aspects of the language however remain inaccessible: those that can be better captured through the study of natural, uncontrolled speech. With *BULBasaa*, one of our longer term aims is also to provide the basis for a reference corpus of spoken Bàsàá.

A number of publications are available in Bàsàá (bibles, dictionaries, collections of idiomatic expressions, language learning methods). Some of them use an IPA-based transcription of the language, including a representation of tones, such as the *Dictionnaire Basaá-French* (Lemb and De Gastines, 1973). As already pointed out by (Lemb and De Gastines, 1973), and this is probably still true, many publications however use alphabets that follow the transcription conventions of European languages such as French and English, and also German, for the least recent ones, with the effect that a number of contrasts are not marked and tones are not represented. Instead, accents are often used to distinguish different sounds.

3. *BULBasaa*

3.1. Data Collection

As described in (Adda et al., 2016; Stüker et al., 2016), one central aspect of BULB is to use the resource-economic methodology developed by S. Bird and colleagues, (Bird et al., 2014) which consists in (i) collecting both elicited and natural speech, (ii) proceeding to the careful re-speaking of the data by selected native speakers, to ensure the best acoustic quality possible for the purpose of automatic phone transcription, and (iii) translating the data into a major language (here, French), to speed up the documentation process and provide a basis for machine translation.

Following this method, the *BULBasaa* corpus comprises two types of Bàsàá data. One half of the recordings (uncontrolled speech) was acquired from a local Bàsàá speaking radio station in the first phase of the project. The other half of the recordings (controlled speech) was made by one of the authors (E.-M. Makasso) in various locations in the Centre and Littoral regions (Yaoundé, Douala, Eseka, Edea, Messondo) shown on Figure 2, using the LIG-Aikuma application (Gauthier et al., 2016). The data collection process took place during several missions between the summer 2015 and the summer 2017. The recording equipment used consisted of four Android powered tablets (Samsung SM-T550/T800/T810).

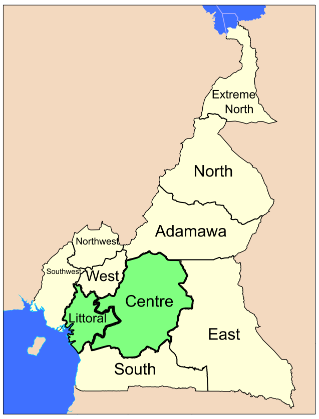
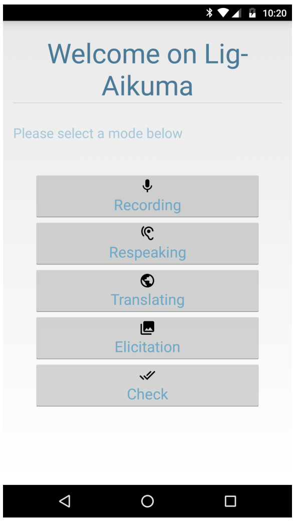


Figure 2: Provinces of Cameroon
(Wikipedia, 2006)

LIG-Aikuma is an improved version of AIKUMA, the Android application originally developed by Steven Bird and colleagues (Bird et al., 2014), which is also meant to facilitate data collection of the type of above-described parallel speech (original>re-speaking>translation). It offers four recording modes, shown in Figure 3.

We provide an overview of the type of data collected in the next Section. Our uncontrolled speech recordings consisted of radio shows of approximately 45 minutes, including musical opening and interludes. The files thus required some cleaning before re-speaking and translation could be achieved. Additionally, to facilitate their use on an early version of LIG-Aikuma, the files were segmented into small chunks of ≈1.5 minutes before being uploaded onto the tablets.

Re-speaking of the data was mostly done by one female speaker (in her early thirties) in Berlin (Germany), where the BULB team of linguists working on Bàsàá was based, and by one male speaker (in his mid-thirties) in Yaoundé. As to the oral translation into French, a consequent part of it was also done by one male speaker (in his early forties) in Berlin and by two speakers in Yaoundé, one female (in her forties) and one male (in his early thirties). Note that although the process of orally translating natural discourse data by bilingual speakers is probably much faster than in



Welcome on Lig-Aikuma

Please select a mode below

Figure 3: Screenshot LIG Aikuma (version July 2016)

more traditional methods of language documentation, it is also a cognitively demanding task for which we found it preferable to hire more than one speaker.

To have a baseline for comparison for the output of our ASR tools, approximately 10 hours of uncontrolled speech recordings were manually phonetically transcribed (against financial compensation) by a team of Bàsàá speaking linguistics students in Yaoundé.

3.2. Data Description

Table 2 provides an overview of the make-up of the *BULBasaa* corpus.

Data type	Amount
Original recordings (uncontrolled speech)	22,62 hours (1058 files)
Re-speaking (uncontrolled speech)	23,73 hours (1058 files)
Translation (uncontrolled speech)	33,55 hours (1142 files)
Picture-based elicitation	8,3 hours (2419 files)
Unprocessed bilingual controlled speech	32,10 hours (164 files)

Table 2: Overview of BULBasaa recordings

As already mentioned above, a significant part of the original Bàsàá speech data consists in radio recordings, where speakers discuss a variety of topics (every-day-life-related, local, social, cultural, historical etc). We also collected some more uncontrolled speech on the form of personal stories, tales and a traditional ceremony.

In addition, for our controlled data, we had one of our female speakers (from the Cenre region) record the entire *Dictionnaire Basaá-French* (Lemb and De Gastines, 1973), a total of 6,000 words together with their translation and example sentences to illustrate the various meanings of each word. The dictionary is a large part of the bilingual controlled speech recordings that we still need to process.

We also had a male speaker record sections of the *Enquête et description des langues orales* linguistic questionnaire (Bouquiaux and Thomas, 1976) dealing with the human being as a physical (anatomy) and as a social entity (family relations).

Finally, we used a subset of the pictures made by our colleague Guy-Noël Kouarata, a member of the BULB team working on Mboshi (Bantu C25), during one of his field trips in Republic of Congo, to elicit data parallel to the ones he collected (Godard et al., 2018).

4. Data Preparation

4.1. Pre-processing

The manual phonetic transcriptions were checked for consistency and occasional faulty symbols were removed or replaced by their correct Bàsàá counterpart. Some symbols were mapped so that, when compounded with diacritics representing tones, they correspond to existing symbols in UTF-8 encoding. The tone markers were checked as well, removing some erroneous sequences, and keeping only the 5 diacritics representing the different surface tones present in the language (see section 2). As French borrowings were transcribed using French orthography, we semi-automatically added a symbol to mark these words.

4.2. Forced Alignment

As the provided phonetic transcriptions did not contain any timing information and were not segmented based on individual recordings, we first split them to create segments for each recording. As this is a labor intensive process, we only prepared a total of 56 min / 814 speech segments of the re-spoken data in this manner. The alignment of the phones to the audio was performed using a speech recognition system. With a very limited amount of available data, we could not train an ASR system on Bàsàá data only, but instead used a system trained on multiple source languages (French, German, Italian, Russian, Turkish). The system was trained using data from the Euronews Corpus (Gretter, 2014). This corpus contains 70h of TV broadcast news per language. Trained jointly on this data, the system featured a multilingual phone set. It was adapted to Bàsàá by first manually mapping the phone set used for the phonetic transcription to the phones of the multilingual recognizer. Next, one iteration of Viterbi training was performed to adapt the acoustic model to the new language. Using this system, we force aligned the transcripts to the audio. Those alignments were used to measure against in the following experiments. As we did not have much data for adaptation, the alignments were not perfect and do not represent a gold standard.

5. Unsupervised Phone Discovery

One of the first steps in documenting a new language is to establish its phone inventory. This is a difficult and time-consuming process. We aim at supporting linguists during this step by inferring a set of phone-like units automatically. Given that unknown languages potentially feature a number of idiosyncrasies, we want to provide an iterative approach, where linguists can make adaptions to the set of discovered phone-like units and the system is then able to incrementally derive a better set of discovered units. Our approach of unsupervised phone discovery (UPD) consists of three steps: (i) segmenting recordings into phone-like units, (ii) extracting articulatory features for each segment, and (iii) clustering segments based on the extracted features. In Sections 5.1. to 5.3., we describe each step of this pipeline. The experiments and the numbers reported were carried out using the part of the data for which the forced alignments were created.

5.1. Segmentation

A deep bi-directional long short-term memory network (DBLSTM) based approach was used to segment the recordings (Franke et al., 2016). The network was trained multilingually using the same data set as the multilingual ASR system in Section 4.2.: Data from 5 languages (English, French, German, Italian, Turkish) from the Euronews Corpus (Gretter, 2014).

5.2. Articulatory Feature Extraction

We trained detectors for 7 different types of articulatory features (AFs, see Table 5.2.) using data from French, German and Turkish from the same dataset (Müller et al., 2017). In order to create training data, we trained ASR systems for each language and used those systems to generate phonetic alignments. We then mapped the phones to AFs. As the pronunciation dictionaries were automatically built using MaryTTS (Schröder and Trouvain, 2003), we used the AF definitions embedded within MaryTTS' language definition files to establish the phone to AF mapping.

Type	# Classes	Description
cplace	8	Place of articulation
ctype	6	Type of articulation
cvox	2	Voiced
vfront	3	Tongue x position
vheight	3	Tongue y position
vlng	4	Type of vowel
vrnd	2	Lips rounded

Table 3: Overview of AF types used

The AF detectors were DBLSTM based and are trained using one-hot targets. Previous studies have shown, that using only the inner third of each segment does increase the classification performance as the articulators in the vocal tract have already reached their final position for producing the current sound and lesser co-articulation artifacts are being encountered. Figure 4 shows AFs extracted over an utterance. Co-articulation artifacts can be seen as the articulators transition from one position to the next, whereas the recognition at the center of each phone remains stable.

5.3. Clustering

As final step in our pipeline, we clustered the segments into a set of phone-like units using the extracted AFs. To determine AFs per segment, we extracted and averaged them for the inner third. A single AF vector for each segment was obtained this way. We used k-Means for clustering of the segments. This method required the number of classes as parameter. Here, we determined the class count supervised, based on the reference transcripts.

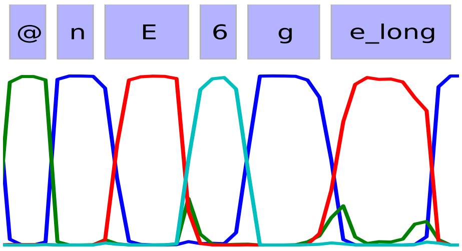


Figure 4: Example of extracted AFs

In order to evaluate the result of the clustering, we used an unsupervised evaluation method based on a TTS system (Baljekar et al., 2015). The inferred units were used to train a TTS system on the data set. This TTS system was then used to synthesize Bàsàá speech. To asses the performance of the TTS systems, the mean cepstral distortion (MCD) score (Mashimo et al., 2001) was used. It is a measure of the distortion between the synthesized and the real speech. The lower the quality of the TTS system is, the higher the distortion and the MCD score are.

The TTS was trained and tested on the full data set. Training and testing on the same data was possible, as we only want to assess the quality of the discovered units, and not the TTS system itself. As baseline, we computed the MCD score of a TTS system trained on the reference transcriptions. Table 4 shows the scores for clusterings. While the score for the automatically derived units increases, the clustered units could be used to synthesize Bàsàá speech.

# of units	MCD Score
Baseline	5.15
UPD Clustering	5.64

Table 4: Comparison of MCD Scores

6. Conclusion

We have presented the *BULBasaa* corpus, a Bàsàá-French bilingual corpus made up of both controlled (elicited) and uncontrolled (natural) speech. As Bàsàá was an under-resourced language, but its basic grammatical properties are fairly well documented, this dataset enables the evaluation of methods aimed at language documentation. From a computational linguistic perspective, our goal in the near future is to focus on improving the reference alignments, as well as to increase the performance of our setup for unsupervised phone discovery. From a linguistic perspective, one of our goals is to prepare the corpus so as to allow for the exploration of both well and under-studied grammatical aspects of the language, in particular through the study of natural speech, as well to allow for the study of aspects of a variety of French (the variety spoken by our Bàsàá speakers) that has not received much attention so far. We plan to share these language resources through the ELDA agency after the end of the project.

7. Acknowledgments

This work was realized in the framework of the ANR-DFG project BULB (ANR-14-CE35-002). The research reported here was in part conducted at the 2017 Frederick Jelinek Memorial Summer Workshop on Speech and Language Technologies, hosted at Carnegie Mellon University and sponsored by Johns Hopkins University with unrestricted gifts from Amazon, Apple, Facebook, Google, and Microsoft. Many thanks go to the rest of the BULB@ZAS team, Carole Ngo Sohna, Chantale Ngo Ndindjock, Rodolphe Maah, Paul-Roger Bassong, and Henry Salfner, as well as to Ute Andersohn, Joerg Dreyer, Elke Dresler, Annerose Schroeder, Wolfgang Kreher and Manfred Krifka for their administrative and practical support at ZAS.

8. Bibliographical References

Adda, G., Stüker, 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.

Baljekar, P., Sitaram, S., Muthukumar, P. K., and Black, A. W. (2015). Using articulatory features and inferred phonological segments in zero resource speech processing. In *Sixteenth Annual Conference of the International Speech Communication Association.*

Bearth, T., (2003). *The Bantu languages*, chapter Syntax, pages 121–142. Routledge.

Bird, S., Hanke, F. R., Adams, O., and Lee, H. (2014). Aikuma: A mobile app for collaborative language documentation. *ACL 2014*, page 1.

Luc Bouquiaux et al., editors. (1976). *Enquête et description des langues à tradition orale.* SELAF, Paris.

Franke, J., Müller, M., Hamlaoui, F., Stüker, S., and Waibel, A. (2016). Phoneme Boundary Detection using Deep Bidirectional LSTMs. In *Speech Communication; 12. ITG Symposium; Proceedings of.* VDE.

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

Godard, P., Adda, G., Adda-Decker, M., Benjumea, J., Besacier, L., Cooper-Leavitt, J., Kouarata, G.-N., Lamel, L., Maynard, H., Müller, M., 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.

Gretter, R. (2014). Euronews: A Multilingual Benchmark for ASR and LID. In *Fifteenth Annual Conference of the International Speech Communication Association.*

Guthrie, M. (1948). *The classification of the Bantu languages.* Oxford University Press for the International African Institute.

Hyman, L., (2003). *The Bantu languages*, chapter Bàsàa (A43), pages 257–282. Routledge.

Lemb, P. and De Gastines, F. (1973). *Dictionnaire basaá-français.* Collège Libermann.

Makasso, E.-M. and Lee, S. J. (2015). Basaá. *Journal of the International Phonetic Association*, 45(1):71–79.

Mashimo, M., Toda, T., Shikano, K., and Campbell, N. (2001). Evaluation of cross-language voice conversion based on gmm and straight.

Müller, M., Franke, J., Stüker, S., and Waibel, A. (2017). Improving Phoneme Set Discovery for Documenting Unwritten Languages. *Elektronische Sprachsignalverarbeitung (ESSV) 2017.*

Schröder, M. and Trouvain, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6(4):365–377.

Gary F. Simons et al., editors. (2017). *Ethnologue: Languages of the World, Twentieth edition.* SIL International, Dallas, Texas. Online version: `http://www.ethnologue.com`.

Stüker, 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 Under-Resourced Languages : toward an Alliance for Digital Language Diversity)*, Portorož Slovenia.

Wikipedia. (2006). Departments of cameroon — Wikipedia, the free encyclopedia. [Online; accessed 28-September-2017].