A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments

P. Godard[*], G. Adda[*], M. Adda-Decker[†], J. Benjumea[‡], L. Besacier [⋆], J. Cooper-Leavitt[*], G-N. Kouarata[†], L. Lamel[*], H. Maynard[*], M. Müller[◇], A. Rialland[†], S. Stüker[◇], F. Yvon[*], M. Zanon-Boito [⋆]

[*]LIMSI, CNRS, Université Paris-Saclay, Orsay, France
[†]LPP, CNRS, Paris, France
[‡]ENS, EHESS, PSL Research University, CNRS, INRIA, Paris, France
[⋆]LIG, UGA, G-INP, CNRS, INRIA, Grenoble, France
[◇]KIT, Karlsruhe, Germany
contact: pierre.godard@limsi.fr, laurent.besacier@univ-grenoble-alpes.fr

Abstract

Most speech and language technologies are trained with massive amounts of speech and text information. However, most of the world languages do not have such resources or stable orthography. Systems constructed under these almost zero resource conditions are not only promising for speech technology but also for computational language documentation. The goal of computational language documentation is to help field linguists to (semi-)automatically analyze and annotate audio recordings of endangered and unwritten languages. Example tasks are automatic phoneme discovery or lexicon discovery from the speech signal. This paper presents a speech corpus collected during a realistic language documentation process. It is made up of 5k speech utterances in Mboshi (Bantu C25) aligned to French text translations. Speech transcriptions are also made available: they correspond to a non-standard graphemic form close to the language phonology. We present how the data was collected, cleaned and processed and we illustrate its use through a zero-resource task: spoken term discovery. The dataset is made available to the community for reproducible computational language documentation experiments and their evaluation.

Keywords: language documentation, field linguistics, spoken term discovery, word segmentation, zero resource technologies, unwritten languages.

1. Introduction

Many languages will face extinction in the coming decades. Half of the 7,000 languages spoken worldwide are expected to disappear by the end of this century (Austin and Sallabank, 2011), and there are too few field linguists to document all of these endangered languages. Innovative speech data collection methodologies (Bird et al., 2014; Blachon et al., 2016) as well as computational assistance (Adda et al., 2016; Stücker et al., 2016) were recently proposed to help them in their documentation and description work.

As more and more researches are related to computational language documentation (Duong et al., 2016; Franke et al., 2016a; Godard et al., 2016; Anastasopoulos and Chiang, 2017), there is a need of realistic corpora to fuel reproducible and replicable language studies at the phonetic, lexical and syntactic levels.

Our work follows this objective and presents a speech dataset collected following a real language documentation scenario. It is multilingual (Mboshi speech aligned to French text) and contains linguists' transcriptions in Mboshi (in the form of a non-standard graphemic form close to the language phonology). The corpus is also enriched with automatic forced-alignment between speech and transcriptions. The dataset is made available to the research community[1]. This dataset is part of a larger data collection conducted on Mboshi language and presented in a companion paper (Adda et al., 2018).

Expected impact of this work is the evaluation of efficient and reproducible computational language documentation approaches in order to face fast and inflexible extinction of languages.

This extended abstract is organized as follows: after presenting the language of interest (Mboshi) in section 2., we describe the data collection and processing in sections 3. and 4. respectively. Section 5. illustrates its first use for an unsupervised word discovery task. Our spoken term detection pipeline is also presented and evaluated in this section. Finally, section 6. concludes this work and gives some perspectives

2. Language description

Mboshi (Bantu C25) is a typical Bantu language spoken in Congo-Brazzavile. It is one of the languages documented by the BULB (Breaking the Unwritten Language Barrier) project (Adda et al., 2016; Stücker et al., 2016).

Phonetics, phonology and transcription Mboshi 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. It has a set of prenasalized consonants (mb, nd, ndz, ng, mbv) which are common in Bantu languages (Embanga Aborobongui, 2013; Kouarata, 2014).

While the language can be considered as rarely written, linguists have nonetheless defined a non-standard graphemic form for it, considered to be close to the language phonol-

[1]It will be made available for free from the "Evaluations and Language resources Distribution Agency" but its current version is online on: `https://github.com/besacier/mboshi-french-parallel-corpus`

ogy. Affricates and pre-nasalized plosives were coded using multiple symbols (e.g. two symbols for dz, three for mbv). Long and short vowels were coded respectively as V and as VV.
Mboshi displays a complex set of phonological rules. The deletion of a vowel before another vowel in particular, common in Bantu languages, occurs at 40% of word junctions (Rialland et al., 2015). This tends to obscure word segmentation and introduces an additional challenge for automatic processing.

Morphology Mboshi words are typically composed of roots and affixes, and include almost always at least one prefix, while the presence of several prefixes and one suffix is also very common. The suffix structure tends to be a single vowel V (e.g. -a or -i) whereas the prefix structure may be both CV and V. Most common syllable structures are V and CV, although CCV may arise due to the presence of affricates and pre-nasalized plosives mentioned above.
The noun class prefix system, another typical feature of Bantu languages, has an unusual rule of deletion targeting the consonant of prefixes[2]. The structure of the verbs, also characteristic of Bantu languages, follows: Subject Marker — Tense/Mood Marker — Root-derivative Extensions — Final Vowel. A verb can be very short or quite long, depending of the markers involved.

Prosody Mboshi prosodic system involves two tones and an intonational organisation without downdrift (Rialland and Aborobongui, 2016). The high tone is coded using an acute accent on the vowel while low tone vowel has no special marker. Word root, prefix and suffix all bear specific tones which tend to be realized as such in their surface forms.[3] Tonal modifications may also arise from vowel deletion at word boundaries.
A productive combination of tonal contours in words can also take place due to the preceding and appended affixes. These tone combinations play an important grammatical role particularly in the differentiation of tenses. However, in Mboshi, the tones of the roots are not modified due to conjugations, unlike in many other Bantu languages.

3. Data Collection

We have recently introduced Lig_Aikuma[4], a mobile app specifically dedicated to fieldwork language documentation, which works both on android powered smartphones and tablets (Blachon et al., 2016). It relies on an initial smartphone application developed by (Bird et al., 2014) for the purpose of language documentation with an aim of long-term interpretability. We extended the initial app with the concept of retranslation (to produce oral translations of the initial recorded material) and speech elicitation from text or images (to collect speech utterances aligned to text or images). In that way, human annotation labels can be replaced by weaker signals in the form of parallel multimodal side information (images or text in another language). Lig_Aikuma also implements the concept of *respeaking* initially introduced in (Woodbury, 2003). It 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 analyze later on (analysis by a linguist or by a machine). 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).
Focusing on Mboshi data, our corpus was built both from translated reference sentences for oral language documentation (Bouquiaux and Thomas, 1976) and from a Mboshi dictionary (Beapami et al., 2000). Speech elicitation from text was performed by three speakers in Congo-Brazzaville and led to more than 5K spoken utterances. The corpus is split in two parts (*train* and *dev*) for which we give basic statistics in Table 1. We shuffled the corpus prior to splitting in order to have comparable distributions in terms of speakers and origin[5]. There is no text overlap for Mboshi transcriptions between the two parts.

language	split	#sent	#tokens	#types
Mboshi	train	4,617	27,568	6,197
	dev	514	2,993	1,146
French	train	4,617	38,848	4,928
	dev	514	4,283	1,175

Table 1: Corpus statistics for the Mboshi corpus

4. Data Processing

4.1. Cleaning, Pre-/Post-Processing

All the characters of the Mboshi transcription have been checked, in order to avoid multiple encodings of the same character. Some characters have also been transcoded so that a character with a diacritic corresponds effectively to the UTF-8 composition of the individual character with the diacritic. Incorrect sequences of tones (for instance tone on a consonant) have been corrected. It was also decided to remove the elision symbol in Mboshi.
On the French side, the translations were double-checked semi-automatically (using linux *aspell* command followed by a manual process – 3.3% of initial sentences were corrected this way). The French translations were finally tokenized (using *tokenizer.perl* from the Moses toolkit[6]) and lowercased.

4.2. Forced Alignment

As the linguists' transcriptions did not contain any timing information, the creation of timed alignments was necessary. We used a speech recognition system to force align the phonetic transcriptions to the audio. With the dataset being very limited in size, it was not possible to train an ASR system entirely on Mboshi data. Therefore, we used an already

[2]A prefix consonant drops if the root begins with a consonant (Rialland et al., 2015).

[3]The distinction between high and low tones is phonological (see (Rialland and Aborobongui, 2016)).

[4]http://lig-aikuma.imag.fr

[5]Either (Bouquiaux and Thomas, 1976) or (Beapami et al., 2000).

[6]http://www.statmt.org/moses/

trained multilingual system and adapted it to Mboshi using the given data. This system was then used to generate the forced-alignments that will serve as a gold standard for the evaluations presented in the next section.

5. Illustration: Unsupervised Word Discovery from Speech

In this section, we illustrate how our corpus can be used to evaluate an unsupervised discovery of word units from raw speech. This task corresponds to $task2$ of the *Zero Resource Speech Challenge 2017*[7]. We present below the evaluation metrics used as well as a monolingual baseline system which does not take advantage yet of the French translations aligned to the speech utterances (bilingual approaches may be also evaluated with this dataset but we leave that for future work).

5.1. Evaluation Metrics

Word discovery is evaluated using the *Boundary*, *Token* and *Type* metrics from the *Zero Resource Challenge 2017*, extensively described in (Ludusan et al., 2014) and (Dunbar et al., 2017). They measure the quality of a word segmentation and the discovered boundaries with respect to the gold corpus. For these metrics, the precision, recall and F-score are computed; the *Token* recall is defined as the probability for a gold word token to belong to a cluster of discovered tokens (averaging over all the gold tokens), while the *Token* precision represents the probability that a discovered token will match a gold token. The F-score is the harmonic mean between these two scores. Similar definitions are applied to the *Type* and *Boundary* metrics.

5.2. Unsupervised Word Discovery Pipeline

Our baseline for word discovery from speech involves the cascading of the following two modules:

- unsupervised phone discovery (UPD) from speech: find pseudo-phone units from the spoken input,
- unsupervised word discovery (UWD): find lexical units from the sequence of pseudo-phone units found in the previous step.

Unsupervised phone discovery from speech (UPD)
In order to discover a set of phone-like units, we use the KIT system which is a three step process. First, phoneme boundaries are found using BLSTM as described in (Franke et al., 2016b). For each speech segment, articulatory features are extracted (see (Müller et al., 2017a) for more details). Finally, segments are clustered based on these articulatory features (Müller et al., 2017b). So, the number of clusters (pseudo phones) can be controlled during this process.

Unsupervised word discovery (UWD)
To perform unsupervised word discovery, we use $dpseg$ (Goldwater et al., 2006; Goldwater et al., 2009)[8]. It implements a Bayesian non-parametric approach, where

method	P	R	F
gold FA phones + dpseg	68.2	82.6	74.7
(Jansen and Van Durme, 2011)	27.3	12.0	16.6
UPD+dpseg pipeline (5 units)	22.5	39.0	28.5
UPD+dpseg pipeline (30 units)	21.1	51.1	29.9
UPD+dpseg pipeline (60 units)	21.0	52.4	30.0

Table 2: Precision, Recall and F-measure on word boundaries. Pipeline compared with an unsupervised system based on (Jansen and Van Durme, 2011).

(pseudo)-morphs are generated by a bigram model over a non-finite inventory, through the use of a Dirichlet-Process. Estimation is performed through Gibbs sampling.
Godard et al. (2016) compare this method to more complex models on a smaller Mboshi corpus, and demonstrate that it produces stable and competitive results, although it tends to oversegment the input, achieving very high recall and a lower precision.

5.3. Results

Word discovery results are given in Tables 2, 3 and 4 for *Boundary*, *Token* and *Type* metrics respectively[9].
We compare our results to a pure speech based baseline which does pair-matching using locally sensitive hashing applied to PLP features and then groups pairs using graph clustering (Jansen and Van Durme, 2011). The parameters[10] of this baseline spoken term discovery system are the same as the ones used for the *Zero Resource Challenge 2017* (Dunbar et al., 2017).
A topline where $dpseg$ is applied to the gold forced alignments (phone boundaries are considered to be perfect) is also evaluated.
For the pipeline, we experience with different granularities of the UPD system (5, 30 and 60 units obtained after the clustering step).
We note that the baseline provided by the system of (Jansen and Van Durme, 2011) has a low coverage (few matches). Given that our proposed pipeline provides an exhaustive parse of the speech signals, it is guaranteed to have full coverage and, thus, shows better performance according to all metrics. The quality of segmentation, in terms of tokens and boundaries is, however, still low for all systems. Increasing the number of pseudo phone units improves *Boundary* recall but reduces precision (and overall, there is not much difference in term of *Boundary* F-measure).

6. Conclusion

We have presented a speech corpus in Mboshi made available to the research community for reproducible computational language documentation experiments. As an illus-

[7] `http://zerospeech.com`
[8] `http://homepages.inf.ed.ac.uk/sgwater/resources.html`

[9] Results may change in the final version because we are currently investigating improvement of our forced alignments used to generate the gold standards for the metrics. Moreoever, only one run per condition is reported here. In the final paper, we also plan to report an average of multiple runs.
[10] Notably the DTW threshhold – which controls the number of spoken clusters found – is set to 0.90 in our experiment

method	P	R	F
gold FA phones + dpseg	34.3	41.3	37.5
(Jansen and Van Durme, 2011)	2.3	0.9	1.3
UPD+dpseg pipeline (5 units)	1.7	3.3	2.2
UPD+dpseg pipeline (30 units)	1.6	4.7	2.4
UPD+dpseg pipeline (60 units)	1.5	4.8	2.4

Table 3: Precision, Recall and F-measure on word tokens. Pipeline compared with an unsupervised system based on (Jansen and Van Durme, 2011).

method	P	R	F
gold FA phones + dpseg	21.4	28.2	24.3
(Jansen and Van Durme, 2011)	3.1	1.7	2.2
UPD+dpseg pipeline (5 units)	1.9	3.7	2.5
UPD+dpseg pipeline (30 units)	1.6	2.9	2.1
UPD+dpseg pipeline (60 units)	1.6	2.8	2.0

Table 4: Precision, Recall and F-measure on word types. Pipeline compared with an unsupervised system based on (Jansen and Van Durme, 2011).

tration, we presented the first unsupervised word discovery (UWD) experiments applied to a truly unwritten language (Mboshi).
Future works include enrichment of our dataset to evaluate also a bilingual lexicon discovery task (possible with encoder-decoder approaches, as shown in (Zanon Boito et al., 2017)).

7. Acknowledgements

This work was partly funded by the French ANR and the German DFG under grant ANR-14-CE35-0002.
Some experiments reported of this paper were performed during the Jelinek Summer Workshop on Speech and Language Technology (JSALT) 2017 in CMU, Pittsburgh, which was supported by JHU and CMU (via grants from Google, Microsoft, Amazon, Facebook, and Apple).

8. 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.

Adda, G., Rialland, A., Kouarata, G.-N., Adda-Decker, M., , Besacier, L., Lamel, L., Gauthier, E., , Godard, P., and Coope-Leavitt, J. (2018). Parallel Corpora in Mboshi (Bantu C25, Congo-Brazzaville). In *Submitted to LREC 2018*, Japan.

Anastasopoulos, A. and Chiang, D. (2017). A case study on using speech-to-translation alignments for language documentation. pages 170–178. Association for Computational Linguistics.

Peter K. Austin et al., editors. (2011). *The Cambridge Handbook of Endangered Languages*. Cambridge University Press.

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

Blachon, D., Gauthier, E., Besacier, L., Kouarata, G.-N., Adda-Decker, M., and Rialland, A. (2016). Parallel speech collection for under-resourced language studies using the LIG-Aikuma mobile device app. In *Proceedings of SLTU (Spoken Language Technologies for Under-Resourced Languages)*, Yogyakarta, Indonesia, May.

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

Dunbar, E., Nga Cao, X., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., Anguera, X., and Dupoux, E. (2017). The zero resource speech challenge 2017. In *Automatic Speech Recognition and Understanding (ASRU), 2017 IEEE Workshop on*. IEEE.

Duong, L., Anastasopoulos, A., Chiang, D., Bird, S., and Cohn, T. (2016). An attentional model for speech translation without transcription. In *Proceedings of NAACL-HLT*, pages 949–959.

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.

Franke, J., Mueller, M., Hamlaoui, F., Stueker, S., and Waibel, A. (2016a). Phoneme boundary detection using deep bidirectional lstms. In *Speech Communication; 12. ITG Symposium; Proceedings of*, pages 1–5. VDE.

Franke, J., Müller, M., Stüker, S., and Waibel, A. (2016b). Phoneme boundary detection using deep bidirectional lstms. In *Speech Communication; 12. ITG Symposium; Proceedings of*. VDE.

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

Goldwater, S., Griffiths, T. L., and Johnson, M. (2006). Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 673–680, Sydney, Australia, July. Association for Computational Linguistics.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Jansen, A. and Van Durme, B. (2011). Efficient spoken term discovery using randomized algorithms. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 401–406. IEEE.

Kouarata, G.-N. (2014). *Variations de formes dans la langue mbochi (Bantu C25)*. Ph.D. thesis, Université Lumière Lyon 2.

Ludusan, B., Versteegh, M., Jansen, A., Gravier, G., Cao, X.-N., Johnson, M., and Dupoux, E. (2014). Bridging the gap between speech technology and natural language processing: an evaluation toolbox for term discovery systems. In *Proceedings of LREC*.

Müller, M., Franke, J., Stüker, S., and Waibel, A. (2017a). Improving phoneme set discovery for documenting unwritten languages. *Elektronische Sprachsignalverarbeitung (ESSV) 2017*.

Müller, M., Franke, J., Stüker, S., and Waibel, A. (2017b). Towards phoneme inventory discovery for documentation of unwritten languages. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE.

Rialland, A. and Aborobongui, M. E. (2016). How intonations interact with tones in Embosi (Bantu C25), a two-tone language without downdrift. In *Intonation in African Tone Languages*, volume 24. De Gruyter, Berlin, Boston.

Rialland, A., Embanga Aborobongui, G. M., Adda-Decker, M., and Lamel, L. (2015). Dropping of the class-prefix consonant, vowel elision and automatic phonological mining in Embosi. In *Proceedings of the 44th ACAL meeting*, pages 221–230, Somerville. Cascadilla.

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

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

Zanon Boito, M., Berard, A., Villavicencio, A., and Besacier, L. (2017). Unwritten languages demand attention too! word discovery with encoder-decoder models. In *Automatic Speech Recognition and Understanding (ASRU), 2017 IEEE Workshop on*. IEEE.