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


To cite this version:
P. Godard, G Adda, M Adda-Decker, J Benjumea, Laurent Besacier, et al.. A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments. Language Resources and Evaluation Conference (LREC), May 2018, Miyazaki, Japan. hal-01807093

HAL Id: hal-01807093
https://hal.archives-ouvertes.fr/hal-01807093
Submitted on 4 Jun 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.
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

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 and some even lack a stable orthography. Building systems under these almost zero resource conditions is 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, 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 detail 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; Stüker 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. To our knowledge, very few corpora are available for computational analysis of endangered languages.

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. This dataset is part of a larger data collection conducted on Mboshi language and presented in a companion paper (Rialland 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 paper 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-Brazzaville. It is one of the languages documented by the BULB (Breaking the Unwritten Language Barrier) project (Adda et al., 2016; Stüker et al., 2016).

Phonetics, phonology and transcription

Mboshi has a seven vowel system (i, e, r, a, 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, p, mb, nd, ndz, ng, mbv, f, s, χ, 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).

version is online on: https://github.com/besacier/
mboshi-french-parallel-corpus
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 phonology. Affricates and prenasalized 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 almost always include at least one prefix, while the presence of several prefixes and one suffix is also very common. The suffix structure tends to consist of a single vowel V (e.g. -a or -i) whereas the prefix structures are V and CV. Most common syllable structures are V and CV, although CCV may arise due to the presence of affricates and prenasalized plosives mentioned above.

The noun class prefix system, another typical feature of Bantu languages, has an unusual rule of deletion targeting the consonant of prefix(es). 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 organization 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. 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\(^3\) 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 retranslating (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 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. There is no text overlap for Mboshi transcriptions between the two parts.

### 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 effectively corresponds 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\(^4\)) and lowercased. We propose an example of a sentence pair from our corpus in Figure 1.

#### 4.2. Forced Alignment

As the linguists’ transcriptions did not contain any timing information, the creation of timed alignments was necessary. The word and phoneme level alignments were

\[^3\]A prefix consonant drops if the root begins with a consonant (Rialland et al., 2015).

\[^4\]Either (Bouquiaux and Thomas, 1976) or (Beapami et al., 2000).
produced with what Cooper-Leavitt et al. refer to as ‘A light-weight ASR tool’ (Cooper-Leavitt et al., 2017a). The alignment tool is an ASR system that is used in a forced-alignment mode. That is, it is used to associate words in the provided orthographic level transcription with the corresponding audio segments making use of a pronunciation lexicon which represents each word with one or more pronunciations (phonemic forms). The word-position-independent GMM-HMM monophone models are trained using the STK tools at LIMSI (Lamel and Gauvain, 2015). A set of 68 phonemes are used to represent the pronunciation dictionary, with multiple symbols for each vowel representing different tones (Cooper-Leavitt et al., 2017b) and a symbol for silence. The acoustic model is estimated iteratively, with 5 rounds of segmentation and parameter estimation, and the model resulting from the last iteration was used to resegment the data. Since vowel elision and morpheme deletion are known to be characteristic of the Mboshi language, variants explicitly allowing their presence or absence are included in the pronunciation lexicon. Details of how this was implemented can be found in (Cooper-Leavitt et al., 2017a).

5. Illustration: Unsupervised Word Discovery from Speech

In this section, we illustrate how our corpus can be used to evaluate unsupervised discovery of word units from raw speech. This task corresponds to Track 2 in the Zero Resource Speech Challenge 2017 (http://zerospeech.com). 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 Speech 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). It implements a Bayesian non-parametric approach, where (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. 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 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). For each granularity, we perform 3 runs and report the averaged results.

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

\footnote{http://zerospeech.com}

\footnote{http://homepages.inf.ed.ac.uk/sgwater/resources.html}

Notably the DTW threshold – which controls the number of spoken clusters found – is set to 0.90 in our experiment
parse of the speech signals, it is guaranteed to have full coverage and, thus, shows better performance according to the Boundary metric. The quality of segmentation, in terms of tokens and types is, however, still low for all systems. Increasing the number of pseudo phone units improves Boundary recall but reduces precision. For Token and Type metrics, a coarser granularity provides slightly better results.

6. Conclusion

We have presented a speech corpus in Mboshi made available to the research community for reproducible computational language documentation experiments. As an illustration, we presented the first unsupervised word discovery (UWD) experiments applied to a truly unwritten language (Mboshi).

The results obtained with our pipeline on pseudo-phones are still far behind those obtained with gold labels, which indicates that there is room for improvement for the UPD module of our pipeline. The UWD module should, in future work, make use of the bilingual data available to improve the quality of the segmentation.

Future work also includes enriching our dataset with some alignments at the word level, in order to evaluate a bilingual lexicon discovery task. This is possible with encoder-decoder approaches, as shown in (Zanon Boito et al., 2017). As we distribute this corpus, we hope that this will help the community to strengthen its effort to improve the technologies currently available to support linguists in documenting endangered languages.

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


Dunbar, E., Nga Cao, X., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., Anguera, X., and Dupoux, E.

<table>
<thead>
<tr>
<th>method</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>gold FA phones + dpseg</td>
<td>53.8</td>
<td>83.5</td>
<td>65.4</td>
</tr>
<tr>
<td>Jansen and Van Durme, 2011</td>
<td>31.9</td>
<td>13.8</td>
<td>19.3</td>
</tr>
<tr>
<td>UPD+dpseg pipeline (5 units)</td>
<td>27.4</td>
<td>46.5</td>
<td>34.5</td>
</tr>
<tr>
<td>UPD+dpseg pipeline (30 units)</td>
<td>24.6</td>
<td>58.9</td>
<td>34.7</td>
</tr>
<tr>
<td>UPD+dpseg pipeline (60 units)</td>
<td>24.4</td>
<td>60.2</td>
<td>34.8</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>method</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>gold FA phones + dpseg</td>
<td>20.8</td>
<td>35.2</td>
<td>26.2</td>
</tr>
<tr>
<td>Jansen and Van Durme, 2011</td>
<td>4.5</td>
<td>1.6</td>
<td>2.4</td>
</tr>
<tr>
<td>UPD+dpseg pipeline (5 units)</td>
<td>2.1</td>
<td>4.4</td>
<td>2.9</td>
</tr>
<tr>
<td>UPD+dpseg pipeline (30 units)</td>
<td>1.4</td>
<td>4.6</td>
<td>2.2</td>
</tr>
<tr>
<td>UPD+dpseg pipeline (60 units)</td>
<td>1.4</td>
<td>4.7</td>
<td>2.1</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>method</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>gold FA phones + dpseg</td>
<td>7.5</td>
<td>13.9</td>
<td>9.7</td>
</tr>
<tr>
<td>Jansen and Van Durme, 2011</td>
<td>4.6</td>
<td>4.8</td>
<td>4.7</td>
</tr>
<tr>
<td>UPD+dpseg pipeline (5 units)</td>
<td>2.5</td>
<td>6.7</td>
<td>3.6</td>
</tr>
<tr>
<td>UPD+dpseg pipeline (30 units)</td>
<td>2.0</td>
<td>4.5</td>
<td>2.8</td>
</tr>
<tr>
<td>UPD+dpseg pipeline (60 units)</td>
<td>2.0</td>
<td>4.4</td>
<td>2.7</td>
</tr>
</tbody>
</table>

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


