Unsupervised Word Segmentation: does tone matter ?

Pierre Godard, Kevin Löser, Alexandre Allauzen, Laurent Besacier, François Yvon

Unsupervised learning of word segmentation: does tone matter ?

Pierre Godard[1], Kevin Löser[1], Alexandre Allauzen[1], Laurent Besacier[2], François Yvon[1]

[1] LIMSI, CNRS, Université Paris-Saclay, France,
{godard,loser,allauzen,yvon}@limsi.fr
[2] Laboratoire d'Informatique de Grenoble (LIG), Université Grenoble Alpes, France,
laurent.besacier@imag.fr

Abstract. In this paper, we investigate the usefulness of tonal features for unsupervised word discovery, taking Mboshi, a low-resource tonal language from the Bantu family, as our main target language. In a preliminary step, we show that tone annotation improves the performance of *supervised learning* when using a simplified representation of the data. To leverage this information in an unsupervised setting, we then present a probabilistic model based on a hierarchical Pitman-Yor process that incorporates tonal representations in its backoff structure. We compare our model with a tone-agnostic baseline and analyze if and how tone helps unsupervised segmentation on our small dataset.

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 [2], and there are too few field linguists to document all of these endangered languages. Innovative speech data collection methodologies [4,5] as well as computational assistance [1,15] were recently proposed to help them in their documentation and description work. This paper follows similar objectives and focuses on the unsupervised discovery of words from an unsegmented sequence of symbols. While such material could be obtained by automatic phone recognition from speech, we investigate here oracle data (hand-annotated symbols from linguists) in a rather limited resource setting (only a few thousand sentences). In this context, our main question is to evaluate the usefulness of tone information in unsupervised word segmentation, and we use Mboshi, a mostly unwritten African language of the Bantu family, as our main test case.

The integration of tonal information in segmentation relies on Bayesian nonparametric (BNP) models, popularized in Natural Language Processing (NLP) by [8,10,9]. In this approach, (pseudo)-words or (pseudo)-morphs are generated by a bigram model over a non-finite inventory, through the use of a Dirichlet process (DP), or a more general Pitman-Yor process (PYP), which enables to discover units that follow a power-law distribution, a universal characteristic of language lexicons. These algorithms were originally designed as computational

models of language acquisition and were mainly applied to non-tonal languages, with the exception of [11], who investigated the use of adaptor grammars for unsupervised word segmentation of Mandarin Chinese. Tones were shown to have a small impact on segmentation accuracy and were reported to yield a small improvement for simple grammars and no improvement with more complex ones. Also worth mentioning is the work of [12], who studied the role of prosodic information (at a macroscopic level) for word segmentation. The approach was tested on English and Japanese: for both languages, it was shown that prosodic boundaries were actually helping word segmentation.

Contributions: we investigate in this paper whether and how tone annotation can help word segmentation in the case where the distribution of tones obey morphological and syntactical, as well as lexical, constraints. After briefly presenting some peculiarities of Mboshi, we first show (in Section 2) that tones help disambiguate word boundaries in a *supervised setting* - suggesting that this information could also help the unsupervised discovery of words. We then present in Section 3 a new hierarchical BNP model, which uses generalized back-off schemes to integrate tonal representations in word segmentation. Based on the experiments reported in Section 4, we conclude that, notwithstanding the inherent unstability of BNP models, tonal information can improve word discovery procedures.

2 A Preliminary Study: Supervised Word Segmentation

In order to assess the potential for tones to help discovering a tonal language's word units, we first conducted a supervised experiment making use of decision trees. We start this section by presenting a short sketch of the language targeted in this work: Mboshi.

2.1 Mboshi language

Mboshi is a language spoken in Congo-Brazzavile, and it was one of the languages documented by the BULB (Breaking the Unwritten Language Barrier) project [1,15], using the LIG-AIKUMA tool [5]. Preliminary experiments for a small portion of it were reported in [7]. Mboshi is a two tone Bantu language whose words are typically composed of roots and affixes. Almost all Mboshi words include at least one prefix, while the presence of several prefixes and one suffix is also very common. While the language can be considered as rarely written, linguists have nonetheless defined a non-standard grapheme form of it, considered to be close to the language phonology. The suffix structure tends to be a single vowel (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 affricates and pre-nasalized plosives (coded with symbols *dz* and *mb*). Mboshi also makes use of short and long vowels (coded respectively as V and VV). With respect to tones, 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. Concerning grammatical tones, word root, prefix and suffix all bear specific tones which tend to be realized as such in their surface forms. Tonal modifications may 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.

2.2 Data and representations

Our study uses a corpus in Mboshi built both from translated reference sentences for oral language documentation [6] and from a Mboshi dictionary [3]. This corpus, comprising more than 9K sentences, is split in three parts that we call `S`, `M` and `L` and for which we give basic statistics in Table 1. Subset `S` can be considered as homogeneous, as it comes from a single source; subsets `M` and `L` come from two different sources and exhibit more lexical diversity.

name	#sent	#tokens	#types
`S`	1,174	6,238	1,664
`M`	4,904	27,990	7,271
`L`	9,256	52,433	11,440

Table 1. Corpus statistics for the Mboshi corpus (`letter+tone` representation).

In our transcriptions, Mboshi's tonal system consisting of a pair of high and low tones is simply represented using diacritics on vowels: acute accent for a high tone, and no accent for a low tone. Our approach consists in varying the representation of the input text and comparing the full transcription with diacritics (`letter+tone`) to i) the transcription `letter` where diacritics are removed, ii) the transcription `xV` where vowels are replaced by the symbol 'V', iii) the transcription `xLH` where high-toned vowels are replaced by the symbol 'H' and low-toned vowels are replaced by 'L', iv) the transcriptions `CV` (resp. `CLH`) where consonants in `xV` (resp. `xLH`) are replaced by a generic symbol, 'C' (see Table 2). We expect the systematic comparison of tonal (`letter+tone`, `xLH`, and `CLH`) with their non tonal counterpart (`letter`, `xV`, `CV`) to shed some light on the usefulness of this information.

[3] The distinction between high and low tones is phonological (see [14]).

name	representation
`letter`	wa ayɛɛ la midi
`letter+tone`	wa áyɛɛ la midí
`CV`	CV VCVV CV CVCV
`CLH`	CL HCLL CL CLCH
`xV`	wV VyVV lV mVdV
`xLH`	wL HyLL lL mLdH

Table 2. Various representations of the text.

2.3 Disambiguating Word Boundaries with Decision trees

For each representation of the text, we train a decision tree classifier[4] to predict a binary decision corresponding to the presence or absence of a word boundary after each character. This prediction is based on features encoding a fixed-length window of characters centered around the decision point; and consider in our experiments windows of varying size[5] We report the precision, recall and F-measure computed on ambiguous[6] word boundaries in Table 3 on the `S` corpus[7] where 10% of the number of sentences have been held out for testing purposes. For the pseudo-orthographic (`letter` and `letter+tone`) text representations, it seems that the tonal information is of little value to disambiguate the frontiers. However, as we simplify the text representation, despite a drop in F-measure, the contrast between a representation ignoring tones (`CV`) and one that captures them (`CLH`) becomes much sharper, suggesting that a tonal signal can be used to improve segmentation. This motivates the design of new models that could capture this signal directly on pseudo-orthographic representations and help improve the precision of unsupervised segmentation.

representation	P	R	F-measure
`letter`	0.92	0.93	0.92
`letter+tone`	0.91	0.89	0.90
`xV`	0.86	0.89	0.87
`xLH`	0.88	0.89	0.88
`CV`	0.70	0.61	0.65
`CLH`	0.78	0.72	0.75

Table 3. Precision, Recall and F-measure on word boundaries in various text representations of corpus `S` with a decision tree classifier (11-words window width).

[4] We use `scikit-learn`'s implementation (`http://scikit-learn.org/stable/modules/tree.html`)
[5] With padding at the beginning and end of the sentence.
[6] We exclude word boundaries corresponding to the beginning and end of the sentence.
[7] Similar results are obtained for the larger corpora.

3 Non-parametric segmentation models with tone information

3.1 Pitman-Yor processes

PYPs are a class of stochastic processes used as models for sparse probability distributions with a countably infinite support and distributed according to a power-law, and therefore especially suited to model distributions arising in linguistic data [16,9]. A PYP is defined by some *base distribution* P_0, and two *concentration* ($\theta \in]-d, \infty[$) and *discount* ($d \in [0, 1[$) hyperparameters, and generates sparse versions of P_0, whose degree of sparsity is controlled by θ and d. Rather than explicitly defining $\mathrm{PYP}(P_0, \theta, d)$, we use the Chinese Restaurant Process (CRP) metaphor to recall its main properties.

We assume a restaurant with K tables, with n_k customers seated at table k ($k \in \{1, \ldots, K\}$), and $N = \sum_{k=1}^{K} n_k$ the total number of customers. Each table has a label l_k from the domain $\mathcal{D}(P_0)$ of P_0. The restaurant is initially empty, with no table. Customers enter one by one and:

- seat at table k table with probability proportional to $n_k - d$
- seat at an empty table with probability proportional to $\theta + K \cdot d$. l_{K+1} is then chosen by sampling from P_0 and K is incremented.

The table layout defines a distribution $P(l)$ over $\mathcal{D}(P_0)$, where $P(l)$ is the probability of obtaining l by uniformly picking a random customer and returning its table label. When $N \to \infty$, $P(l)$ converges to a sample of $\mathrm{PYP}(P_0, \theta, d)$.

Given a particular table setup, the probability of sampling a label x is computed as:

$$P(x|n_1 \ldots n_K, l_1 \ldots l_K) \propto \left(\sum_{k=1}^{K} \mathbf{1}_{l_k = x} \cdot (n_k - d) \right) + (\theta + Kd) \cdot P_0(x).$$

3.2 Sampling segmentations using a PYP-distributed unigram word model

In this work, we model a sentence as a concatenation of words drawn from a unigram distribution P_w generated by a $\mathrm{PYP}(P_{spl}, \theta, d)$ as in [10]. P_{spl} is the *spelling model*, for instance a n-gram model over character sequences (see below). We tokenize a corpus $s_1 \ldots s_n$ of n sentences by Gibbs-sampling every segmentation s_i conditioned on all other segmentations.[8] As explained above, sampling $x \sim P_w$ can be done by maintaining a CRP associated to P_w, such that for every token t in the current lexicon L there is a customer seated at a table labelled

[8] Following [13], we use a forward filtering-backward sampling (FFBS) algorithm to sample segmentations. As this method only approximates the posterior distribution, we also perform a Metropolis-Hastings correction step.

with the type of t. We also placed agnostic priors on the PYP hyperparameters: $\theta \sim \exp(1)$ and $d \sim Beta(1, 5)$ and resampled these hyperparameters as well as the CRP table layouts every 200 iterations.

3.3 A spelling model with tones

The spelling model defines a distribution over character strings that reflects how word forms should look like. Obvious candidate spelling models are n-gram models of character sequences:[9]

$$P_{spl}(w) = \prod_{i=1}^{l} P(c_i|c_{i-n+1} \dots c_{i-1}) \cdot P(\text{stop}|c_{l-n+1} \dots c_l)$$

In order to also integrate tone information in the spelling model P_{spl}, we define the set of *contexts* $\mathcal{K}$ as the set of sequences $\tau_1 \dots \tau_j \, c_{j+1} \dots c_k$ (with $k \in \{0, \dots, n-1\}$), where τ_i are *tones* symbols from the set {H(high tone), L(low tone), C(consonant)} and c_i are regular characters. A *context* is thus a sequence of length at most $n-1$ comprising a prefix of tones and a suffix of characters.

For every non-empty context $\kappa \in \mathcal{K}$, we further assume that the conditional distribution $P(c|\kappa)$ recursively arises from a PYP with base distribution $P(c|\beta(\kappa))$, where β is a *backoff function*. The base distribution of the unigram distribution (κ is empty) is the uniform distribution. In this setting, base distributions of PYPs themselves arise from PYPs, giving rise to a *hierarchical PYP* whose structure is defined by the backoff function β. To enrich the model with awareness of tone patterns, we designed the following backoff scheme, where characters are first replaced by tones (rightwards), then dropped (rightwards), as follows:

$$\begin{aligned} c_1 \dots c_{n-1} &\xrightarrow{\beta} \tau_1 c_2 \dots c_{n-1} \xrightarrow{\beta} \tau_1 \tau_2 c_3 \dots c_{n-1} \\ &\xrightarrow{\beta} \dots \xrightarrow{\beta} \tau_1 \dots \tau_{n-1} \xrightarrow{\beta} \tau_2 \dots \tau_{n-1} \\ &\xrightarrow{\beta} \tau_3 \dots \tau_{n-1} \xrightarrow{\beta} \tau_{n-1} \xrightarrow{\beta} \emptyset. \end{aligned}$$

On a Mboshi example, backoff will thus unfold as follows: `ámid` $\xrightarrow{\beta}$ `Hmid` $\xrightarrow{\beta}$ `HCid` $\xrightarrow{\beta}$ `HCLd` $\xrightarrow{\beta}$ HCLC $\xrightarrow{\beta}$ CLC $\xrightarrow{\beta}$ LC $\xrightarrow{\beta}$ C $\xrightarrow{\beta}$ $\emptyset$. This model is referred to as MULTI in our experiments. We also evaluated an alternative backoff scheme β', where only one single tone is remembered:

$$\begin{aligned} c_1 \dots c_{n-1} &\xrightarrow{\beta'} \tau_1 c_2 \dots c_{n-1} \xrightarrow{\beta'} \tau_2 c_3 \dots c_{n-1} \\ &\xrightarrow{\beta'} \dots \xrightarrow{\beta'} \tau_{n-2} c_{n-1} \xrightarrow{\beta'} \tau_{n-1} \xrightarrow{\beta'} \emptyset \end{aligned}$$

[9] where we add initial padding symbols as needed.

This is illustrated on the same Mboshi example: ámid $\xrightarrow{\beta'}$ Hmid $\xrightarrow{\beta'}$ Cid $\xrightarrow{\beta'}$ Ld $\xrightarrow{\beta'}$ C $\xrightarrow{\beta'}$ $\emptyset$. This model is referred to as LAST in our experiments.

We compare our tone models MULTI and LAST to a baseline PYP n-gram spelling model — referred to later on in our experiments as BASE — that is unable to distinguish between high and low tones. In this baseline, the backoff scheme β_{BASE} is simply defined by:

$$c_1 \dots c_{n-1} \xrightarrow{\beta_{\text{BASE}}} c_2 \dots c_{n-1} \xrightarrow{\beta_{\text{BASE}}} c_3 \dots c_{n-1}$$
$$\xrightarrow{\beta_{\text{BASE}}} \dots \xrightarrow{\beta_{\text{BASE}}} c_{n-2}c_{n-1} \xrightarrow{\beta_{\text{BASE}}} c_{n-1} \xrightarrow{\beta_{\text{BASE}}} \emptyset$$

It corresponds on the previously used example to: ámid $\xrightarrow{\beta_{\text{BASE}}}$ mid $\xrightarrow{\beta_{\text{BASE}}}$ id $\xrightarrow{\beta_{\text{BASE}}}$ d $\xrightarrow{\beta_{\text{BASE}}} \emptyset$.

4 Experiments

Several setups are considered with varying corpus sizes (S, M, and L) and text representations (letter+tone, letter, etc.). We report precision, recall and F-measure on word boundaries (BP, BR, BF) and word types (LP, LR, LF) with respect to S used as a test set for all corpora sizes. Additionally, we make the spelling model's Markov order vary between 1 and 6 and also conduct experiments with the BASE model on all the text representations discussed in Section 2.2. Each configuration (144 in total) is ran 5 times resulting in a total of 720 measures. Table 4 gives the results of the best thirty runs in terms of F-measure on word boundaries (BF).

Unlike what was observed in the supervised setting (Section 2), there seems to be some benefit in keeping the tonal information in the representation (letter+tone vs. letter): 22 out of the 30 best runs use tone information and top 9 configurations actually use the letter+tone representation. We can also observe that large values of n (beyond 3) do not help much and that $n = 3$ seems to be a good compromise. The benefit of our tone models (MULTI and LAST) is less conclusive when compared to the BASE model, which also obtained very good performance.

Also note that we did not observe a clear trend when increasing the corpus size. This might be due to the more heterogeneous nature of our Mboshi M, and L subsets mentioned in section 2.2. Of particular interest are the results obtained with the xLH representation and the BASE model: notwithstanding the replacement of 14 symbols (Mboshi has a 7 vowels inventory, each prone to carry a low or high tone) by only 2 symbols encoding the tonal information, the performance compares to the very best result on our segmentation task.

To serve as another baseline, we also ran dpseg [8,10] [10] for all the possible representations and corpus sizes (18 configurations). It implements a Non-parametric Bayesian approach, where (pseudo)-words are generated by a bigram model over a non-finite inventory, through the use of a Dirichlet-Process. We used

[10] http://homepages.inf.ed.ac.uk/sgwater/resources.html

corpus size	model	Markov order	representation	BP	BR	BF	LP	LR	LF
M	BASE	3	letter+tone	89.07	72.89	80.17	58.28	62.80	60.46
S	LAST	1	letter+tone	77.59	81.50	79.50	54.13	45.67	49.54
M	LAST	3	letter+tone	87.20	72.93	79.43	56.53	60.64	58.51
M	BASE	3	letter+tone	86.53	71.52	78.31	55.93	60.34	58.05
L	BASE	3	letter+tone	89.17	69.23	77.95	55.80	62.44	58.93
S	MULTI	3	letter+tone	88.58	69.08	77.62	52.12	54.03	53.05
M	LAST	3	letter+tone	83.82	71.48	77.16	52.94	57.87	55.30
M	LAST	6	letter+tone	76.29	78.04	77.16	47.98	46.45	47.21
S	LAST	1	letter+tone	76.86	77.47	77.16	54.45	47.42	50.69
L	BASE	2	letter	81.51	73.20	77.13	55.25	55.21	55.23
S	LAST	6	letter+tone	80.25	74.23	77.12	53.16	48.56	50.75
M	BASE	3	letter	88.85	68.13	77.12	57.36	63.46	60.26
M	BASE	3	letter	87.71	68.33	76.81	57.44	63.12	60.15
M	LAST	5	letter+tone	84.25	70.56	76.80	50.75	57.09	53.73
M	BASE	3	letter	89.35	67.24	76.73	56.64	63.39	59.83
M	LAST	3	letter+tone	85.66	69.23	76.58	52.92	58.29	55.48
M	BASE	4	letter	94.24	64.32	76.46	54.49	65.37	59.44
L	BASE	3	letter	89.10	66.96	76.46	54.05	62.78	58.09
L	BASE	4	letter+tone	90.90	65.88	76.39	52.28	62.08	56.76
M	BASE	3	xLH	87.96	67.26	76.23	52.88	62.44	57.26
S	LAST	1	letter+tone	72.64	79.94	76.11	48.96	42.61	45.57
S	LAST	1	letter+tone	76.38	75.67	76.02	51.83	45.07	48.22
M	BASE	3	letter	88.49	66.53	75.96	55.61	62.17	58.71
M	BASE	2	xLH	81.70	70.64	75.77	53.19	52.49	52.84
M	BASE	6	letter	74.72	76.62	75.66	48.74	47.31	48.01
L	BASE	2	xLH	81.24	70.62	75.55	50.59	53.15	51.84
M	LAST	4	letter+tone	86.32	67.16	75.54	50.11	57.21	53.42
M	BASE	2	xLH	81.89	69.91	75.42	52.28	52.24	52.26
M	BASE	3	letter+tone	86.54	66.65	75.30	52.70	57.57	55.03
L	BASE	3	letter+tone	86.66	66.57	75.30	51.76	58.35	54.86

Table 4. Top 30 best F-measures on word boundaries (BF) for all models, corpora sizes, spelling model's Markov order, text representations and 5 runs for each (144 configurations * 5 runs were evaluated in total). A `dpseg` baseline was also run for all possible representations and corpus sizes (18 configurations) and best configuration lead to BF equals to 70.40%.

the same hyper-parameters as [7], which were tuned on a larger English corpus and then successfully applied to the segmentation of Mboshi. The best configuration lead to a F-measure on word boundaries (BF) equal to 70.40% which is significantly below the performance of the top-30 configurations reported in table 4.

Table 5 and 6 brings complementary view on our models' behavior. They compare different models/representations configurations for which F-measures on word boundaries are averaged over spelling model's Markov order and runs. Results in Table 5 are reported for the `S` corpus. They confirm the benefit of keeping tonal information in the representation. As expected, impoverished representations yield worse performance than the `letter+tone` or `letter` representations. However, they still convey enough signal to reliably detect word boundaries.This

result is encouraging for future experiments on true speech input, where coarse grain pseudo-phones or pseudo-syllable units could be extracted.

All these results should be taken with some care, given the large standard deviations of results inside each model/ representation combination. The standard deviation is even more important for the results of table 6 (`L` corpus). One explanation might be the heterogeneous nature of the `L` corpus, but it does not explain all the variability of our runs. The benefit of tone representation is also less clear for the `L` corpus than for the `S` corpus: as more data are considered in training, the usefulness of complex backoff schemes and rich information actually decreases.

In summary, it seems that our models do not take full advantage of the strong signal reflected in the last two lines of Table 3. This might be because these models cannot learn tonal regularities at the grammatical level and are by design limited to learn lexically-based tonal regularities. Yet, tones in Mboshi, as in most Bantu languages, play as much a grammatical role as a lexical one. Related is the current limitation of our models to a unigram word model embedding the more structured spelling model. Bigram dependencies at the word level were consistently shown to improve segmentation [10]. This will be an upcoming extension of the models presented in this work.

model	representation	BF	average	std dev	min	max
LAST	`letter+tone`		72.71	3.17	64.97	79.50
BASE	`letter+tone`		69.62	1.82	66.16	73.03
MULTI	`letter+tone`		69.48	3.93	61.46	77.62
BASE	`letter`		69.26	1.98	65.01	73.33
BASE	`xLH`		66.42	3.96	54.43	73.93
BASE	`xV`		63.44	4.48	48.91	71.32
BASE	`CV`		56.14	1.87	52.68	63.47
BASE	`CLH`		50.34	4.19	44.79	60.45

Table 5. Comparing different models/representations configurations: F-measures on word boundaries averaged over spelling model's Markov order and runs - `S` corpus

5 Conclusion

In a preliminary study, we showed that when learning a segmentation classifier on a simplified representation of a Mboshi corpus where all characters were collapsed to two 'vowel' and 'consonant' categories, supplying that classifier with tones provided an increase in performance and even led to a competitive segmentation accuracy despite the considerable simplification of the data. This suggested that segmentation could benefit from sensitivity to tonal cues, and we tried to leverage the latter in an unsupervised setting by introducing hierarchical n-gram spelling models that incorporate tone-conditional distributions

model	representation	BF average	std dev	min	max
BASE	`letter`	69.86	4.22	57.23	77.13
BASE	`xLH`	66.16	6.46	54.45	75.55
LAST	`letter+tone`	65.80	5.94	42.45	73.95
BASE	`letter+tone`	65.40	8.99	43.82	77.95
BASE	`xV`	63.98	7.16	44.12	73.85
MULTI	`letter+tone`	60.94	7.51	46.94	72.86
BASE	`CLH`	55.25	6.25	32.99	63.44
BASE	`CV`	49.90	8.34	23.61	64.85

Table 6. Comparing different models/representations configurations: F-measures on word boundaries averaged over spelling model's Markov order and runs - `L` corpus

in their hierarchy. These models were compared to a standard Pitman-Yor n-gram spelling model for numerous settings. While we were able to observe some benefit in keeping the tonal information in the representation (`letter+tone` vs. `letter`), our proposed spelling model with tones was less successful in capturing tonal signal in unsupervised word segmentation. For this reason, in future work, we hope to exploit tone not purely within the spelling model, but also on the grammatical level beyond simple unigram word sequence models.

Acknowledgments

This work was partly funded by the French ANR and the German DFG under grant ANR-14-CE35-0002. We warmly thank Martine Adda-Decker and Annie Rialland (from LPP-CNRS) for the sketch on the Mboshi language, as well as Gilles Adda (from LIMSI-CNRS), for many meaningful conversations and contributions.

References

1. 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., Zerbian, S.: Breaking the unwritten language barrier: The Bulb project. In: Proceedings of SLTU (Spoken Language Technologies for Under-Resourced Languages). Yogyakarta, Indonesia (2016)
2. Austin, P.K., Sallabank, J. (eds.): The Cambridge Handbook of Endangered Languages. Cambridge University Press (2011)
3. Beapami, R.P., Chatfield, R., Kouarata, G., Embengue-Waldschmidt, A.: Dictionnaire Mbochi-Français. SIL-Congo Publishers, Congo (Brazzaville) (2000)
4. Bird, S., Hanke, F.R., Adams, O., Lee, H.: Aikuma: A mobile app for collaborative language documentation. ACL 2014 p. 1 (2014)
5. Blachon, D., Gauthier, E., Besacier, L., Kouarata, G.N., Adda-Decker, M., Rialland, A.: 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 2016)

6. Bouquiaux, L., Thomas, J.M.C. (eds.): Enquête et description des langues à tradition orale. SELAF, Paris (1976)
7. Godard, P., Adda, G., Adda-Decker, M., Allauzen, A., Besacier, L., Bonneau-Maynard, H., Kouarata, G.N., Löser, K., Rialland, A., Yvon, F.: Preliminary experiments on unsupervised word discovery in mboshi. In: Interspeech 2016 (2016)
8. Goldwater, S., Griffiths, T.L., Johnson, M.: 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. pp. 673–680. Association for Computational Linguistics, Sydney, Australia (July 2006), http://www.aclweb.org/anthology/P06-1085
9. Goldwater, S., Griffiths, T.L., Johnson, M.: Interpolating between types and tokens by estimating power-law generators. In: Advances in Neural Information Processing Systems 18. pp. 459–466. MIT Press, Cambridge, MA (2006)
10. Goldwater, S., Griffiths, T.L., Johnson, M.: A Bayesian framework for word segmentation: Exploring the effects of context. Cognition 112(1), 21–54 (2009)
11. Johnson, M., Demuth, K.: Unsupervised phonemic Chinese word segmentation using adaptor grammars. In: 23rd International Conference on Computational Linguistics (COLING) (2010)
12. Ludusan, B., Synnaeve, G., Dupoux, E.: Prosodic boundary information helps unsupervised word segmentation. In: Annual Conference of the North American Chapter of the ACL. pp. 953–963. Denver, Colorado, USA (2015)
13. Mochihashi, D., Yamada, T., Ueda, N.: Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1. pp. 100–108. Association for Computational Linguistics (2009)
14. Rialland, A., Aborobongui, M.E.: How intonations interact with tones in Embosi (Bantu C25), a two-tone language without downdrift. In: Intonation in African Tone Languages, vol. 24. De Gruyter, Berlin, Boston (2016)
15. 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., Zerbian, S.: 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 (2016)
16. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. Journal of the American Statistical Association 101(476), 1566–1581 (2006)