

Preliminary Experiments on Unsupervised Word Discovery in Mboshi

Pierre Godard, Gilles Adda, Martine Adda-Decker, Alexandre Allauzen,
Laurent Besacier, Helene Bonneau-Maynard, Guy-Noël Kouarata, Kevin
Löser, Annie Rialland, François Yvon

► **To cite this version:**

Pierre Godard, Gilles Adda, Martine Adda-Decker, Alexandre Allauzen, Laurent Besacier, et al.. Preliminary Experiments on Unsupervised Word Discovery in Mboshi. Interspeech 2016, Sep 2016, San-Francisco, United States. Interspeech 2016 proceedings. <hal-01350119>

HAL Id: hal-01350119

<https://hal.archives-ouvertes.fr/hal-01350119>

Submitted on 29 Jul 2016

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.

Preliminary Experiments on Unsupervised Word Discovery in Mboshi

Pierre Godard^{1,2}, Gilles Adda¹, Martine Adda-Decker^{1,3}, Alexandre Allauzen^{1,2}, Laurent Besacier⁴,
Hélène Bonneau-Maynard^{1,2}, Guy-Noël Kouarata³, Kevin Löser^{1,2}, Annie Rialland³, François Yvon¹

¹LIMSI, CNRS, Université Paris-Saclay, France

²Université Paris-Sud, France

³LPP, CNRS-Paris 3/Sorbonne Nouvelle, France

⁴Laboratoire d'Informatique de Grenoble (LIG)/Univ. Grenoble Alpes, France

{godard, gadda, madda, allauzen, hbm, loser, yvon}@limsi.fr,

laurent.besacier@imag.fr, {guy.kouarata, annie.rialland}@univ-paris3.fr

Abstract

The necessity to document thousands of endangered languages encourages the collaboration between linguists and computer scientists in order to provide the documentary linguistics community with the support of automatic processing tools. The French-German ANR-DFG project *Breaking the Unwritten Language Barrier* (BULB) aims at developing such tools for three mostly unwritten African languages of the Bantu family. For one of them, Mboshi, a language originating from the “Cuvette” region of the Republic of Congo, we investigate unsupervised word discovery techniques from an unsegmented stream of phonemes. We compare different models and algorithms, both monolingual and bilingual, on a new corpus in Mboshi and French, and discuss various ways to represent the data with suitable granularity. An additional French-English corpus allows us to contrast the results obtained on Mboshi and to experiment with more data.

Index Terms: automatic alignment, automatic transcription, machine translation, Bantu languages, language documentation

1. Introduction

The project *Breaking the Unwritten Language Barrier* (BULB),¹ which brings together linguists and computer scientists, aims at supporting linguists in documenting unwritten (or hardly written) languages. In order to achieve this goal, we develop tools for language documentation by building upon technology and expertise from the area of natural language processing, most prominently automatic speech recognition and bitext alignment. As a development and test bed for this methodology, three less-resourced African languages from the Bantu family have been chosen: Basaa, Myene and Mboshi.

The BULB project methodology can be summarized into three main steps: (1) *Collection* of a large corpus of speech at a reasonable cost. For this, we use standard mobile devices and a dedicated software called LIG-AIKUMA [1].² After initial recording, the data is re-spoken by a reference speaker to enhance the signal quality, and orally translated into a target (well-documented) language (French in our case). (2) *Automatic transcription* of the Bantu languages at phoneme level and the French transcription at word level, followed by the *automatic alignment* between recognized Bantu phonemes and

French words. (3) *Tool development*: Implement tools that will assist linguists in their documentation work, taking into account their needs and existing technology’s capabilities. At this stage of the project (end of first year), we have focused on the data acquisition and have also begun to work on automatic transcription and alignment ([2, 3]).

Contributions. This paper is related to the second step of the project methodology. More precisely, we address the *automatic alignment*. This problem (also called *word discovery*) consists in automatically discovering lexical units (as well as their pronunciations) in an unknown (and unwritten) language without any supervision. While several algorithms were proposed in the past to address this problem (see details in section 2), they were only simulating unwritten language cases using well-resourced languages such as English or Spanish. One contribution of this paper is to benchmark several algorithms, both monolingual and bilingual, for a real endangered (and unwritten) language: Mboshi. For bilingual approaches, we also investigate which unit (word or lemma) on the source (written) language is more suitable to discover words in a target (unwritten) language. Finally, an additional French-English corpus allows us to contrast the results obtained on Mboshi and to experiment with larger amounts of data.

The rest of this paper is organized as follows. In Section 2, we summarize the related works on unsupervised word discovery. Details on Mboshi language and data collection are given in Section 3. Section 4 presents the algorithms used while Section 5 is dedicated to experiments and results. Finally, Section 6 concludes and gives some perspectives.

2. Unsupervised word discovery

2.1. Previous work

The feasibility of automatically discovering lexical units (as well as their pronunciations) in an unknown (and unwritten) language without any supervision was examined in [4]. This goal was achieved by unsupervised aggregation of phonetic strings into word forms from a continuous flow of phonemes (or from a speech signal) using a monolingual algorithm based on cross-entropy. Applied to a speech translation task, this approach led to almost the same performance as the baseline approach, while being theoretically applicable to any unwritten language. A phone-based speech translation approach that made use of cross-lingual supervision was introduced in [5]. This bilingual approach works on a scenario in which a hu-

¹<http://www.bulb-project.org>

²<https://forge.imag.fr/frs/download.php/706/MainActivity.apk>

man translates the audio recordings of the unwritten language into a written language. Alignment models as used in machine translation [6, 7] were then learned on the resulting parallel corpus made of foreign phone sequences and their corresponding English translation. [8] combined this approach with monolingual techniques and also did contrastive comparisons. [9, 10] then continued to work on this task by enhancing the alignment model and examined the impact of the choice of the written language to which the phoneme sequence is aligned.

Working with a similar goal in mind, and using bilingual information in order to jointly learn the segmentation of a target string of characters (or phonemes) and its alignment to a source sequence of words, [11, 12] are building on Bayesian monolingual segmentation models introduced by [13] and further expanded in [14]. This trend of research has become increasingly active in the past years, moving from strategies using segmentation as a preprocessing to the alignment steps, to models aiming at jointly learning relevant segmentation and alignment. [15] reports performance improvements for the latter approach on a bilingual lexicon induction task, with the additional benefit of achieving high precision even on a very small corpus, which is of particular interest in the context of BULB.

2.2. Open issues and specificities of the BULB context

Many questions still need to be addressed. Implicit choices are usually made through the way data are specified and represented. Taking, for example, tones into account, prosodic markers, or even a partial bilingual dictionary, would require different kinds of input data, and the development of models able to take advantage of this additional information.

A second observation is that most attempts to learn segmentation and alignments need to inject some prior knowledge regarding the desired form of the linguistic units which should be extracted. This is because most machine learning schemes deployed in the literature tend to produce degenerated and trivial (over-segmented or conversely under-segmented) solutions. The additional constraints needed to control such phenomena are likely to greatly impact the nature of the units that are identified. Supporting the documentation of endangered languages within the framework of BULB should lead us to question the linguistic validity of those constraints and the results they produce. The Adaptor Grammar framework [16, 17], which enables the specification of high-level linguistic hypotheses appears to be of particular interest. Another important aspect of our endeavour is the noisy nature of the input produced by the phonemicization of the unwritten language. Processing a phoneme lattice instead of a (one-best) phonemic transcription, following [18], seems to be a promising strategy.

More generally, a careful inventory of priors derived from the linguistic knowledge at our disposal should be undertaken. This is especially true regarding cross-lingual priors we can postulate about French on the one hand, and Basaa, Myene and Mboshi on the other hand: without taking such priors into account, it is dubious that general purpose unsupervised learning techniques will succeed in delivering any usable linguistic information.

3. Data collection

3.1. Mboshi

Mboshi originates from the “Cuvette” region of the Republic of Congo and is also spoken in Brazzaville and in the diaspora. The number of Mboshi speakers is estimated at 150,000 (Congo

National Inst. of Statistics, 2009). A dictionary [19] is available and, just like Basaa and Myene, the language benefits from recent linguistic studies [20, 21]. During the last decades, Mboshi (Bantu C 25) has been studied to describe its grammar including morphological and phonological systems [20, 22, 23].

Mboshi is a tone language. There are 2 tones (high vs low tones) which may play lexical (*ibea* (to borrow) vs *iba* (to call)) as well as grammatical (*ybea* (which borrows) vs *yebea* (it borrows)) roles. The phonemic inventory is described using 25 consonants and 7 vowels. A specificity of the consonantal system (as compared to English) are for example the labiodental consonants (pf, bv) and the pre-nasalised consonants (mb, mbv, nd, ndz).

One particular aspect of Mboshi that can impact automatic word discovery is its agglutinative morphology. Words are basically formed as a sequence or more generally as a combination of morphological constituents, the latter undergoing a variety of phonological processes during word formation. For example, noun morphology can be quickly described as a combination of a root form to which a collection of affixes may be added. Simple nominal roots may be monosyllabic (*/-VV/* or */-CV/*), disyllabic (*/-CVV/*, */-VCV/* or */-CVCV/*) and trisyllabic (*/-CVCVCV/* or */-CVVCV/*) and may be augmented by prefixes only, whereas verbo-nominal roots also allow for suffix agglutination. There is a long tradition in describing Bantu languages with the help of a rich set of nominal class prefixes [24]. Whereas Bleek’s classification proposes 18 classes, the number of classes varies across languages and even within a language depending on the authors. Most recent work on Mboshi describes a system using 13-14 classes [22, 21]. The verbal morphology can be described by a prefix-root-extension-suffix pattern. Affixes allow for situating an action with respect to an acting agent, a time moment or duration, a place, etc.

A description of phonological processes in Mboshi, as well as studies of its tonal system have recently been carried out [25, 21] and a bilingual French-Mboshi dictionary is also being developed [26].

3.2. Corpus collection

Our objective was to collect large volumes of data from dozens of speakers in different speaking styles. All existing written documents including a 4,500-entry Mboshi dictionary [27], traditional tales and biblical texts in Mboshi were gathered. Furthermore, 1,200 reference sentences for oral language documentation [28] were translated and written in Mboshi by one of the authors (GNK) who is a native speaker of Mboshi. Only these 1,200 sentences translated in Mboshi (called *Bouquiaux* data set in the rest of the paper) were used in the experiments of this paper. However, this is worth mentioning that 50h of speech data in Mboshi language were collected recently with LIG-AIKUMA in Congo-Brazzaville [29]. This larger data set will be used in future studies.

4. Methods assessed

4.1. Algorithms

We contrast here a series of recent methods for performing the automatic segmentation of an input stream of symbols into meaningful lexical or sub-lexical units. *dpseg* [13, 30]³ implements a Bayesian non-parametric approach, where (pseudo)-

³<http://homepages.inf.ed.ac.uk/sgwater/resources.html>

morphs are generated by a bigram model over a non-finite inventory, through the use of a Dirichlet-Process (DP). Estimation is performed through Gibbs sampling. Another implementation of Goldwater et al’s proposal is `pgibbs`⁴ [31], where the DP is replaced by a more general Pitman-Yor Process (PYP); this implementation notably provides an effective parallelization of the sampling process through blocked sampling (our experiments use a 3-gram model). An extension of this model is proposed in [14], which replaces the base distribution of the PYP LM by another hierarchical PYP language model at the character level ; we use here the implementation of [18], denoted `lattice_lm`⁵ in Table 2. `pypshmm` [32] is another generalization of `dpseg`, introducing some morphotactics through word classes: in this model, sentences are produced by a non-parametric semi-Markov model, where both the number of states and the number of types are automatically adjusted based on the available data. Two hierarchical PYPs processes are also embedded in this architecture: one for controlling the number of classes (states) and one for controlling the number of words; as in [14], the base distribution is also a hierarchical PYP language model.

Having a French translation of the input at our disposal also allows us to contrast the above monolingual approaches with bilingual models. We consider here the Model 3P of [33], which generalizes the IBM alignment Model 3 of [6] to the case where the target side is an unsegmented character stream. We use the implementation (`pisa`⁶) of the authors.

4.2. Data overview

Corpus	repr.	#tokens	#types	snt len	wrđ len
Bouquiaux fr	wd	7245	1502	6.2	
Bouquiaux fr	le	7826	1107	6.7	
Bouquiaux mb	ph	6177	1460	5.3	6.8
<hr/>					
Ted fr 0.5K	wd	6820	1733	13.6	
Ted fr 0.5K	le	7174	1325	14.3	
Ted en 0.5K	ph	6122	1492	12.2	6.5
<hr/>					
Ted fr 1K	wd	13456	2705	13.5	
Ted fr 1K	le	14227	1972	14.2	
Ted en 1K	ph	12123	2247	12.1	6.6
<hr/>					
Ted fr 2K	wd	26250	4413	13.1	
Ted fr 2K	le	27808	3119	13.9	
Ted en 2K	ph	23719	3561	11.8	6.7
<hr/>					
Ted fr 5K	wd	64201	7929	12.8	
Ted fr 5K	le	68123	5239	13.6	
Ted en 5K	ph	57731	6305	11.5	7.0
<hr/>					
Ted fr 10K	wd	129958	12268	13.0	
Ted fr 10K	le	138185	7728	13.8	
Ted en 10K	ph	116551	9538	11.6	7.2

Table 1: Corpus statistics. ‘snt len’ gives the average sentence length (in words). ‘wrđ len’ reports the average word length (in phones).

Our primary data source is the Mboshi-French parallel corpus derived from the 1,200 Bouquiaux reference sentences (see 3), for which we vary the representations both on the Mboshi (mb) and the French (fr) side. For the Mboshi, we compare a graphemic and a phonemic representation. The phoneme

⁴<https://github.com/neubig/pgibbs>

⁵http://www.phontron.com/lattice_lm/

⁶<https://code.google.com/archive/p/pisa/>

sequence in Mboshi is considered as almost perfect (no ASR) for the moment. For this, since Mboshi has very little written material and no official agreement exists on writing conventions, a basic grapheme-phoneme perl script was developed to generate a pronunciation dictionary (the used writing conventions are very close to the oral forms). For French, we use both word (wd) and lemmatized (le) forms,⁷ with the hope that the latter will yield less sparse models. We reiterate the same experiments and contrast with French-English data derived from the TedTalk corpus⁸, where we remove word separators on the English side, and also optionally replace the orthography with phonemic strings (as found in a dictionary). The use of a well-resourced language pair allowed us to experiment with corpora of increasing sizes (up to 10K sentences). Basic statistics regarding these datasets are given in Table 1.

5. Experiments and discussion

5.1. Protocol

For all methods evaluated in this study, we attempted to optimize their parameters and hyperparameters on the smallest (0.5K sentences) extract of TedTalk using `hyperopt` [34]. The search algorithm was run several hundreds of times for each method, and the optimal parameters were then frozen to carry out the experiments on the Mboshi-French parallel corpus and the French-English corpora of larger sizes.

5.2. Results

The results in terms of word boundaries prediction are summarized in table 2 for the Mboshi-French parallel corpus and a TedTalk corpus of similar size (0.5K). To assess the possible impact of having more data, figures 1 and 2 show the evolution of respectively the F₁-measure and precision computed on the TED datasets of increasing sizes using a phonetic representation.

The first observation is that our results in Mboshi are much worse than in English, even though the data size are comparable. This highlights the need to collect and process actual under-resourced languages in order to get a fair approximation of the performance of Natural Language Processing techniques – this discrepancy might be due to the fact that we use impoverished representations in Mboshi, where important phonetic information (eg. tones) regarding word structure has been removed; this might also be due to the construction of the Mboshi-French corpus itself, showing less lexical diversity than the TedTalk corpus.⁹

Another general remark is that the non-parametric LMs improve their performance with more samples (see figures 1 and 2). The simpler `dpseg`, which only implements a 2-gram model, obtains very stable and satisfactory performance across the board; `pgibbs` benefits from faster implementations, yet, even after 10K sentences, its segmentation accuracy is still slightly smaller than `dpseg`’s. It is notable that `dpseg`, `pgibbs` and `lattice_lm` display a strong tendency to over-

⁷Lemmatization is performed with the TreeTagger.

⁸<https://wit3.fbk.eu/>

⁹We examined the distribution of words across both corpora, plotting their frequencies for increasing Zipfian ranks. It appears that the French word distribution inside Bouquiaux is similar to its Mboshi counterpart, while differing significantly from the French word distribution observed for the TedTalk corpus. On the latter corpus the distribution of French tokens presents a strong similarity to its English counterpart, which led us to attribute the performance discrepancy to the construction of both corpora rather than to their linguistic differences.

	Phonemes						Graphemes					
	Mboshi			Ted 0.5K			Mboshi			Ted 0.5K		
	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
dpseg	61.0	75.5	67.5	74.6	91.5	82.2	55.8	78.2	65.1	63.9	94.4	76.2
pgibbs	60.2	78.3	68.1	67.3	91.1	77.4	54.9	93.0	69.0	55.0	92.8	69.1
lattice1m	45.3	77.6	57.2	58.4	96.3	72.8	45.5	77.8	57.5	52.9	95.5	68.1
pypshmm	51.0	66.3	57.7	76.9	81.5	79.2	59.4	68.1	63.5	71.2	79.0	74.9
pisa (wd)	27.3	59.8	37.5	35.8	80.1	49.5	25.8	57.1	35.5	23.9	82.3	37.0
pisa (le)	32.2	24.7	27.9	41.4	69.4	51.8	30.2	23.9	26.7	26.1	75.3	38.8

Table 2: Precision (P), recall (R) and F₁-measure for the prediction of word boundaries.

	Mboshi(1460)			Ted 0.5K (1492)		
	#types	P	R	#types	P	R
dpseg	753	44.8	23.1	1044	48.0	33.6
pgibbs	763	43.8	22.9	1056	39.7	28.1
lattice1m	726	26.7	13.3	726	33.8	20.9
pypshmm	809	30.0	16.6	1272	41.8	35.6
pisa (wd)	1954	13.9	18.6	1890	15.0	19.0
pisa (le)	2255	11.4	17.6	2081	15.5	21.6

Table 3: Number of types and precision (P) and recall (R) of the resulting vocabularies. The number of types of the reference is indicated in the first line.

segment their input, achieving very high recall (> 90% on the phonetic English data), and a less satisfactory precision. The recent PYP semi-Markov model (pypshmm) achieves performances that are in the same ballpark, but shows a milder tendency to oversegment the input.

The difference between input representations on the Mboshi side are small; yet it seems that better results are obtained for all methods when using a phonetic representation; pypshmm differs from the other methods in that respect: this might be because it tends to produce longer segments than the other monolingual models, which proves beneficial on graphemic forms, that are typically longer than phonetic representations.

In comparison, the bilingual method achieves disappointing results: even on the larger dataset, its performance is almost 20 points worse than the best monolingual approach. Using lemmas instead of word forms here proves to be beneficial for the TedTalk, with a gain for all corpora sizes, while this leads to a decrease in performance on the Mboshi-French parallel corpus.

Table 3 gathers statistics at the lexicon level and allows us to compare the automatically extracted lexicons with the reference. From the number of types, we confirm that dpseg, pgibbs, lattice1m and pypshmm tend to oversegment, resulting in a compact lexicon, while pisa exhibits the inverse tendency. Whereas lattice1m yields poor recall values, dpseg and pgibbs achieve a more balanced trade off between precision and recall.

6. Conclusions

This paper presented preliminary experiments on unsupervised word discovery in Mboshi, a very under-resourced African language from the Bantu family. Several mono-lingual and bi-lingual automatic segmentation methods were assessed. Among them, dpseg, which only implements a 2-gram model, obtained stable and interesting performance. However, this approach (as well as pgibbs and lattice1m) tends to heavily oversegment the input, achieving very high recall and a

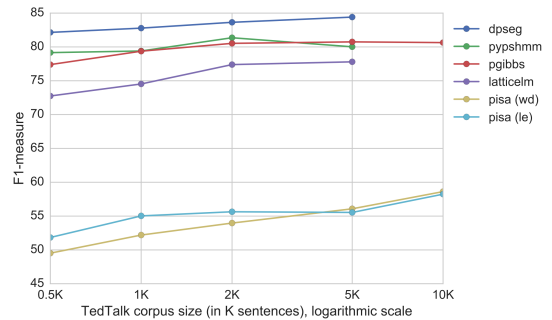


Figure 1: Evolution of the F₁-measure on TED Talk corpora of different sizes (phonemic representation)

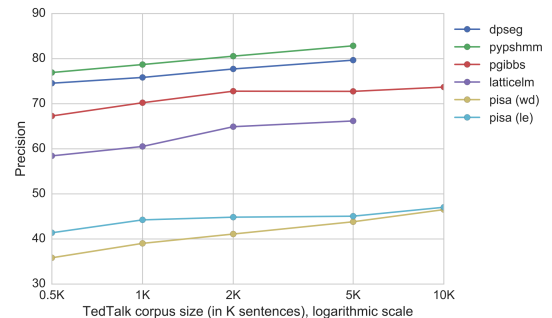


Figure 2: Evolution of the precision on TED Talk corpora of different sizes (phonemic representation)

lower precision. Interestingly, the PYP semi-Markov model (pypshmm) obtained comparable performance with a milder tendency to oversegment the input. Further investigations are needed to confirm the encouraging performance obtained by dpseg and pypshmm since we only worked on a small dataset for which the Mboshi phoneme sequence was considered as perfect.

As a consequence, our data set will be extended soon in order to benchmark word discovery approaches on (a) increased amount of data, and (b) true ASR outputs (word hypothesis for French and phoneme hypothesis for Mboshi). As other mid-term perspectives we intend to work with more contrasting languages and to take into account tones or prosodic markers.

7. Acknowledgements

This work was partly funded by the French ANR and the German DFG under grant ANR-14-CE35-0002.

8. References

- [1] D. Blachon, E. Gauthier, L. Besacier, G.-N. Kouarata, M. Adda-Decker, and A. Rialland, "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.
- [2] G. Adda, S. Stücker, M. Adda-Decker, O. Ambourou, L. Besacier, D. Blachon, H. Bonneau-Maynard, P. Godard, F. Hamlaoui, D. Idiatov, G.-N. Kouarata, L. Lamel, E.-M. Makasso, A. Rialland, M. Van de Velde, F. Yvon, and S. Zerbian, "Breaking the unwritten kanguage barrier: The Bulb project," in *Proceedings of SLTU (Spoken Language Technologies for Under-Resourced Languages)*, Yogyakarta, Indonesia, 2016.
- [3] S. Stücker, G. Adda, M. Adda-Decker, O. Ambourou, L. Besacier, D. Blachon, H. Bonneau-Maynard, P. Godard, F. Hamlaoui, D. Idiatov, G.-N. Kouarata, L. Lamel, E.-M. Makasso, A. Rialland, M. Van de Velde, F. Yvon, and S. Zerbian, "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.
- [4] L. Besacier, B. Zhou, and Y. Gao, "Towards speech translation of non written languages." in *Proc. SLT*, M. Gilbert and H. Ney, Eds. IEEE, 2006, pp. 222–225.
- [5] S. Stücker, "Towards human translations guided language discovery for asr systems," in *Proceedings of the First International Workshop on Spoken Languages Technologies for Under-resourced languages (SLTU)*, Hanoi, Vietnam, May 2008.
- [6] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.
- [7] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Comput. Linguist.*, vol. 29, no. 1, pp. 19–51, 2003.
- [8] S. Stücker, L. Besacier, and A. Waibel, "Human Translations Guided Language Discovery for ASR Systems," in *10th International Conference on Speech Science and Speech Technology (InterSpeech 2009)*. Brighton (UK): Eurasip, 2009, pp. 1–4.
- [9] F. Stahlberg, T. Schlippe, S. Vogel, and T. Schultz, "Word segmentation through cross-lingual word-to-phoneme alignment." in *Proc. SLT*. IEEE, 2012, pp. 85–90.
- [10] —, "Pronunciation extraction from phoneme sequences through cross-lingual word-to-phoneme alignment," in *The 1st International Conference on Statistical Language and Speech Processing*, 2013, sLSP 2013.
- [11] J. Xu, J. Gao, K. Toutanova, and H. Ney, "Bayesian semi-supervised Chinese word segmentation for statistical machine translation," in *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Manchester, UK, 2008, pp. 1017–1024.
- [12] T. Nguyen, S. Vogel, and N. A. Smith, "Nonparametric word segmentation for machine translation," in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, Stroudsburg, PA, USA, 2010, pp. 815–823.
- [13] S. Goldwater, T. L. Griffiths, and M. Johnson, "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*, Sydney, Australia, 2006, pp. 673–680.
- [14] D. Mochihashi, T. Yamada, and N. Ueda, "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*, 2009, pp. 100–108.
- [15] O. Adams, G. Neubig, T. Cohn, and S. Bird, "Inducing Bilingual Lexicons from Small Quantities of Sentence-Aligned Phonemic Transcriptions," in *12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, 2015.
- [16] M. Johnson, T. L. Griffiths, and S. Goldwater, "Adaptor grammars: a framework for specifying compositional nonparametric bayesian models," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007, pp. 641–648.
- [17] M. Johnson, "Unsupervised word segmentation for Sesotho using adaptor grammars," in *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*. Columbus, Ohio: Association for Computational Linguistics, June 2008, pp. 20–27.
- [18] G. Neubig, M. Mimura, S. Mori, and T. Kawahara, "Learning a language model from continuous speech." in *Proc. INTER-SPEECH*, 2010, pp. 1053–1056.
- [19] R. P. Beapami, R. Chatfield, G. Kouarata, and A. Waldschmidt, *Dictionnaire Mbochi - Français*. Brazzaville: SIL-Congo, 2000.
- [20] C. Amboulou, "Le Mbochi: langue bantoue du Congo Brazzaville (Zone C, groupe C20)," Ph.D. dissertation, INALCO, Paris, 1998.
- [21] G. M. Embanga Aborobongui, "Processus segmentaux et tons en Mbondzi – (variété de la langue embosi c25)," Ph.D. dissertation, Université Paris 3 Sorbonne Nouvelle, 2013.
- [22] P. L. Bedrosian, "The Mboshi noun class system," *Journal of West African Languages*, vol. 6, pp. 27–47, 1998.
- [23] L. Fontaney, "Mboshi: Steps Towards a Grammar," *Pholia*, vol. 3, 4, pp. 87–167, 71–132, 1988, 1989.
- [24] W. Bleek, *De nominum generibus linguarum Africae Australis*. Bonn, 1851.
- [25] J.-M. Beltzung, A. Rialland, and M. Embanga Aborobongui, "Les relatives possessives en mbochi (C25)," *Zas Papers in Linguistics*, vol. 52, pp. 7–31, 2010.
- [26] G.-N. Kouarata, "Variations de formes dans la langue mbochi (bantu c25)," Ph.D. dissertation, Université Lumière Lyon 2, 2014.
- [27] B. Roch Paulin, R. Chatfield, G. Kouarata, and A. Embengue-Waldschmidt, *Dictionnaire Mbochi-Français*. Congo (Brazzaville): SIL-Congo Publishers, 2000.
- [28] T. Bouquiaux and J. Thomas, *Enquête et description des langues à tradition orale*. Paris: SELAF, 1976.
- [29] D. Blachon, E. Gauthier, L. Besacier, G.-N. Kouarata, M. Adda-Decker, and A. Rialland, "Parallel speech collection for under-resourced language studies using the lig-aikuma mobile device app," 2016, submitted to SLTU 2016.
- [30] S. Goldwater, T. L. Griffiths, and M. Johnson, "A Bayesian framework for word segmentation: Exploring the effects of context," *Cognition*, vol. 112, no. 1, pp. 21–54, 2009.
- [31] G. Neubig, "Simple, correct parallelization for blocked Gibbs sampling," Nara Institute of Science and Technology, Tech. Rep., 2014.
- [32] K. Löser and A. Allauzen, "Une méthode non-supervisée pour la segmentation morphologique et l'apprentissage morphotactique à l'aide de processus de Pitman-Yor," in *Actes de la conférence sur le Traitement Automatique des Langues Naturelles (TALN'16)*, Paris, 2016.
- [33] F. Stahlberg, T. Schlippe, S. Vogel, and T. Schultz, "Word segmentation and pronunciation extraction from phoneme sequences through cross-lingual word-to-phoneme alignment," *Computer Speech & Language*, 2014.
- [34] J. S. Bergstra, D. Yamins, and D. Tax, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *Proceedings of the 30th International Conference on Machine Learning*, ser. ICML-13, S. Dasgupta and D. Mcallester, Eds., vol. 28, no. 1, 2013, pp. 115–123.