# Unsupervised Word Segmentation from Speech with Attention

Pierre Godard, Marcely Zanon Boito, Lucas Ondel, Alexandre Berard,
François Yvon, Aline Villavicencio, Laurent Besacier

# Unsupervised Word Segmentation from Speech with Attention

*Pierre Godard*[*1], *Marcely Zanon Boito*[⋆1], *Lucas Ondel*[◇], *Alexandre Berard*[⋆‡], *François Yvon*[*], *Aline Villavicencio*[†], *Laurent Besacier*[⋆]

[*]LIMSI, CNRS, Université Paris-Saclay, Orsay, France
[⋆]LIG, UGA, G-INP, CNRS, INRIA, Grenoble, France
[◇]BUT, Brno, Czech Republic
[†]CSEE, University of Essex, UK
[‡]CRISTAL, Université de Lille, France

contact: pierre.godard@limsi.fr, marcely.zanon-boito@univ-grenoble-alpes.fr
(1) Both first authors have contributed equally to this paper

## Abstract

We present a first attempt to perform attentional word segmentation directly from the speech signal, with the final goal to automatically identify lexical units in a low-resource, unwritten language (UL). Our methodology assumes a pairing between recordings in the UL with translations in a well-resourced language. It uses Acoustic Unit Discovery (AUD) to convert speech into a sequence of pseudo-phones that is segmented using neural soft-alignments produced by a neural machine translation model. Evaluation uses an actual Bantu UL, Mboshi; comparisons to monolingual and bilingual baselines illustrate the potential of attentional word segmentation for language documentation.

**Index Terms**: computational language documentation, encoder-decoder models, attentional models, unsupervised word segmentation.

## 1. Introduction

Speech technology often relies on minimal linguistic expertise and textual information to build acoustic and language models. However, for many languages of the world, text transcripts are limited or nonexistent; therefore, recent efforts have been devoted to *Zero Resource Settings* [1, 2, 3] where the aim is to build speech systems without textual or linguistic resources for e.g.: (1) unwritten languages [4, 5]; (2) models that mimic child language development [6]; (3) documentation of endangered languages by analyzing speech recordings using automatically discovered linguistic units (phones, morphs, words, etc) [7].

This paper focuses on unsupervised word segmentation from speech: the system must output time-stamps delimiting stretches of speech, associated with class labels, corresponding to real words in the language. This task is already considered in the Zero Resource Speech Challenge[1] in a fully unsupervised setting: systems must learn to segment from a collection of raw speech signals only. We investigate here a slightly more favorable case where speech utterances are multilingually grounded: they are aligned, at the sentence level, to a written translation in another language. Such a condition is realistic in language documentation, where it is common to collect speech in the language of interest and have it translated or glossed in another language [8]. In this context, we want to examine whether we can take advantage of the weak form of supervision available in these translations to help word segmentation from speech. Our hypothesis is that textual translations should help in segmenting speech into words in the unwritten language, even in the absence of (manually obtained) phonetic labels. As a first contribution in this direction, we recently proposed to leverage attentional encoder-decoder approaches for unsupervised word segmentation [9]. However, this was done from an unsegmented sequence of (manually obtained) true phone symbols, not from speech. It was shown that the approach proposed can compete with a Bayesian Non Parametric (BNP) baseline [10] on a small corpus in Mboshi language (5k sentences only).

In this paper, our **contribution** is to develop an attentional encoder-decoder word segmentation from speech (§2) that operates in two steps: (a) automatic Acoustic Unit Discovery (AUD), based on Bayesian models, to generate time-marked pseudo-phone symbols from the speech; (b) encoder-decoder word segmentation using these pseudo-phones.[2] Experiments with AUD outputs of increasing complexity (see §3) are presented for word boundary detection using the Mboshi corpus recently made available [13] (§4). Our best pipeline from speech has a word boundary F-measure of 50.0% while segmenting from true phone symbols leads to 61.0%.

## 2. Attentional Encoder-Decoder Approach for Word Discovery

For word segmentation, given a parallel corpus pairing sequences of pseudo-phone units in the unwritten language (UL) with sequences of words in the well-resourced language (WRL), we compute attention matrices as the result of training a standard Neural Machine Translation (NMT) system translating from the WRL into the UL. Then, these soft-alignment matrices are post-processed to derive word boundaries.

### 2.1. Neural Architecture

The NMT architecture, equations (1)-(4), are inspired by [14]. A bidirectional encoder reads the input sequence $x_1, ..., x_A$ and produces a sequence of encoder states $\mathbf{h} = h_1, ..., h_A \in \mathbb{R}^{2\times n}$, where $n$ is the chosen encoder cell size. At each time step $t$, the decoder uses its current state $s_{t-1}$ and an attention mechanism

[1]http://zerospeech.com/2017

[2]End-to-end speech processing can be performed with an encoder-decoder architecture for speech translation (e.g. [11, 12]); early attempts to train end-to-end from speech to translated text, in our language documentation scenario, were not viable due to limited data.

to compute a probability distribution $y_t$ over a target vocabulary of size $|V|$. It then generates the symbol $z_t$ having the highest probability, stopping upon generating a special end-of-sentence token. The decoder updates its internal representation $s_t$, using the ground-truth symbol $w_t$, instead of the generated symbol $z_t$, since in our alignment setting the reference translations are always available, even at test time. Our system is described by the following equations:

$$c_t = \text{attn}(\mathbf{h}, s_{t-1}) \quad (1)$$

$$y_t = \text{output}(s_{t-1} \oplus E(w_{t-1}) \oplus c_t) \quad (2)$$

$$z_t = \arg\max y_t \quad (3)$$

$$s_t = \text{LSTM}(s_{t-1}, E(w_t) \oplus c_t), \quad (4)$$

where $\oplus$ is the concatenation operator. $s_0$ is initialized with the last state of the encoder (after a non-linear transformation), $z_0 =$ <BOS> (special token), and $E \in \mathbb{R}^{|V| \times n}$ is the target embedding matrix. The output function uses a maxout layer, followed by a linear projection to $\mathbb{R}^{|V|}$, as in [14].

The attention mechanism is defined as:

$$e_{t,i} = v^T \tanh\left(W_1 h_i + W_2 s_{t-1} + b_2\right) \quad (5)$$

$$\alpha_{t,i} = \text{softmax}(e_{t,i}) \quad (6)$$

$$c_t = \text{attn}(\mathbf{h}, s_{t-1}) = \sum_{i=1}^{A} \alpha_{t,i} h_i \quad (7)$$

where $v$, $W_1$, $W_2$, and $b_2$ are learned jointly with the other model parameters. At each time step ($t$) a score $e_{t,i}$ is computed for each encoder state $h_i$, using the current decoder state $s_{t-1}$. These scores are then normalized using a softmax function, thus giving a probability distribution over the input sequence $\sum_{i=1}^{A} \alpha_{t,i} = 1$ and $\forall t, i, 0 \leq \alpha_{t,i} \leq 1$. The context vector $c_t$ used by the decoder is a weighted sum of the encoder states. This can be understood as a summary of the useful information in the input sequence for the generation of the next output symbol $z_t$. Likewise, the weights $\alpha_{t,i}$ can be viewed as defining a soft-alignment between the input $x_i$ and output $z_t$.

### 2.2. Word Segmentations from Attention

The main aspects of our approach are detailed below.

**Reverse Architecture:** in NMT, the soft alignments probabilities are normalized for each *target symbol* $t$ (i.e. $\forall t, \sum_i \alpha_{i,t} = 1$, with $i$ indexing the source symbols). However, there is no similar constraint for the source symbols, as discussed by [5]. Rather than enforcing additional constraints on the alignments, as in the latter reference, we propose to reverse the architecture and to translate from WRL words into UL symbols, following [9]. This "reverse" architecture notably prevents the attention model from ignoring some UL symbols. As experiments with actual phone sequences have shown that the best results were obtained with this WRL-to-UL translation [9], we will use this reverse architecture throughout.

**Alignment Smoothing:** to deal with the length discrepancy between UL (pseudo-phones) and WRL (words), we implemented the alignment smoothing procedure proposed by [5]. It consists of first adding temperature to the *softmax* function (we use T=10 for all experiments) used by the attention mechanism; and then post-processing the resulting soft-alignment probability matrices, averaging each score with the scores of the two neighboring words. Even if boosting many-to-one alignments should not hold in the case of the reverse architecture, we keep it for our experiments given the gains reported by [9], even in the reverse case.

**Hard Segmentation Generation:** once the soft-alignment matrices $\alpha$ are obtained for all utterances in the corpus, a word segmentation is inferred as follows. We first transform soft-alignments into hard-alignments by aligning each UL symbol $w_t$ with the word $x_i$ such that: $i = \arg\max_{i'} \alpha_{t,i'}$. The source sequence is then segmented according to these hard-alignments: if two consecutive symbols are aligned with the same WRL word, they are considered to belong to the same UL word.

## 3. Acoustic Unit Discovery (AUD)

Our AUD systems are based on the Bayesian non-parametric Hidden Markov Model (HMM) of [15]. This model is topologically equivalent to a phone-loop where each acoustic unit is represented by a left-to-right HMM. To cope with the unknown number of units needed to properly describe the speech, the model assumes a potentially infinite number of symbols. However, the prior over the weight of the acoustic units (a Dirichlet Process [16]) will act as a sparsity regularizer, leading to a model which explains the data with a relatively small unit set.

We implemented two variants of this original model. The first one, referred to as *HMM*, approximates the Dirichlet Process prior by a simpler symmetric Dirichlet prior, as proposed by [17]. This approximation, while retaining the sparsity constraint, avoids the complication of dealing with the variational treatment of the stick breaking process in Bayesian non-parametric models. The second variant, denoted Structured Variational AutoEncoder (*SVAE*) AUD, is based on the work of [18] and embeds the HMM model into a Variational Auto-Encoder (VAE) [19] where the posterior distribution of the HMM and the VAE parameters are trained jointly using Stochastic Variational Bayes [20, 18]. To initialize the model, the prior distribution over the HMM parameters (mixture weights, means and covariance matrices) was set to the posterior distribution of the phone-loop trained in a supervised fashion (Baum-Welch training) on the TIMIT data set. This procedure can be seen as a cross-lingual knowledge transfer as the AUD training on the UL language is essentially adapting the English phone set distribution to the Mboshi corpus. Finally, both models were trained using two features sets: the well-known *MFCC* + $\Delta$ + $\Delta\Delta$ features and the Multilingual BottleNeck (*MBN*) features [21]. Note that the MBN features were not trained on any Mboshi data, and only use languages as listed in [21]).

## 4. Word Segmentation Experiments

### 4.1. Corpus, Baselines and Metric

We used the Mboshi5k corpus [13] in all our experiments.[3] 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 [7]. This speech dataset was collected following a real language documentation scenario, using Lig_Aikuma,[4] a mobile app specifically dedicated to fieldwork language documentation, which works both on Android powered smartphones and tablets [8]. The corpus is multilingual (5,130 Mboshi speech utterances aligned to French text) and contains linguists' transcriptions in Mboshi in the form of a non-standard graphemic

[3]The dataset is documented in [13] and available at https://github.com/besacier/mboshi-french-parallel-corpus

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

| | |
|---|---|
| Mboshi | wáá ngá iwé léekundá ngá sá oyoá lendúma saa m ótéma |
| French | si je meurs enterrez-moi dans la forêt oyoa avec une guitare sur la poitrine |

Figure 1: *A tokenized and lowercased sentence pair example in our Mboshi-French corpus.*

| language | split | #sent | #tokens | #types |
|---|---|---|---|---|
| Mboshi | train | 4,616 | 27,563 | 6,196 |
| | dev | 514 | 2,993 | 1,146 |
| French | train | 4,616 | 38,843 | 4,927 |
| | dev | 514 | 4,283 | 1,175 |

Table 1: *Corpus statistics for the Mboshi corpus*

form close to the language phonology. Correct word segmentation of the Mboshi transcripts was also provided by the linguists and a forced-alignment between speech and transcripts was computed to obtain time-stamps-delimited word tokens for evaluation. The corpus is split in two parts (*train* and *dev*) for which we give basic statistics in Table 1. We also include an example of a sentence pair from our corpus in Figure 1.

Our neural (*attentional*) word segmentation is compared with two baselines: a naive bilingual baseline (*proportional*) that segments the source according to the target as if the alignment matrix between symbols (AUD symbols in Mboshi and graphemes in French) was diagonal;[5] a monolingual baseline [10] which implements a Bayesian non-parametric approach, where (pseudo)-words are generated by a bigram model over a non-finite inventory, through the use of a Dirichlet process (refered to as *dpseg*). We evaluate with the *Boundary* metric from the *Zero Resource Challenge 2017* [22, 3]. It measures the quality of a word segmentation and the discovered boundaries with respect to a gold segmentation (P, R and F-score are computed).

### 4.2. Details of the NMT system

We use the LIG-CRIStAL NMT system.[6] Our models are trained using the Adam algorithm, with a learning rate of $0.001$ and a batch size ($N$) of 32. We minimize the cross-entropy loss between the output probability distribution $p_t = softmax(y_t)$ and a reference translation $w_t$. Our models use global attention and bidirectional layer in the encoder; encoder and decoder have 1 layer each, with a cell size of 64. Dropout is applied with a rate equal to 0.5. For NMT training, we split the 5,130 sentences into training and development, with about 10% of the corpus for the latter. However, the soft-alignment matrices are obtained from both train and dev sets after forced-decoding and segmentation is evaluated on all 5,130 utterances.

### 4.3. Results

Unsupervised word segmentation results obtained from speech with different AUD configurations as well as from true phones (upper-bound performance corresponding to a *topline*) are reported in Table 2, using the *Boundary* metric. We trained 5 different NMT models changing the train/dev split[7] and either (i) averaging the scores over the 5 runs (columns *att. (biling.)* in Table 2) or (ii) averaging the obtained soft-alignment matrices (columns *att. average* in Table 2). The latter slightly boosts boundary detection performance. For all AUD configurations, our method outperforms two baselines (*dpseg* and *proportional*), as well as a pure speech-based baseline using segmental DTW [23], which only achieves a F-score of 19.3 on our data. While competitive with true phones, the results of the monolingual method (*dpseg*) are heavily degraded on discovered (noisier) units, as also reported by [2]. Conversely, our method is much more robust to noise and seems better suited for real-world scenarios. While straightforward, the bilingual baseline (*proportional*) is rather strong compared to its monolingual counterpart (*dpseg*). This suggests that multilingual grounding provides a useful signal for word segmentation from speech.

Regarding AUD specifically, we observe that the best F-score for word boundary detection was obtained with MBN features and the SVAE model. The results of our *attentional* segmentation are the best results reported so far on this corpus. This confirms that we can effectively take advantage of the weak supervision available in the translations in order to help word segmentation from speech.

### 4.4. Discussion

The NMT system requires a sequence of unsegmented symbols (the phones) and their aligned sentence translations in order to provide segmentation. Therefore, the AUD method chosen to encode the speech input has an impact on the quality of the final segmentation. Our best word segmentation results (see Table 2) are obtained using the SVAE model (this holds for both Bayesian and neural segmentation approaches). One natural explanation would be to posit that phone boundaries (and consequently word boundaries) are more accurately detected by the SVAE model than the HMM model. [24] show that this is true in terms of precision for phone boundaries, and in term of normalized mutual information, but that the recall on these boundaries is lower than its HMM counterpart. This indicates that the SVAE model extract more consistent pseudo-phone units, although it misses some boundaries, than the HMM model, and we confirm here the result of [24] showing that this is beneficial for the word segmentation task.

Another additional explanation might be that shorter sequences of symbols are easier to segment. For instance, even if the attention helps the system to better deal with long sentences, it is still prone to performance degradation when faced with very long sequences of symbols [14]. Table 3 (left side) shows how the different AUD approaches are encoding the UL sentences. We observe that the HMM model uses more symbols to represent an utterance, while the SVAE model offers a more concise representation.

Table 3 also reports information regarding the generated segmentation using a single attention model (right side), showing that our best model (MBN SVAE) results in segmentations that relate closely to the *topline* in terms of number of tokens per sentence. This best model also achieved a vocabulary size close to the *topline*, of 14,837 types compared to 13,878. This

[5] Blank spaces on the French side are then used to segment the Mboshi input.

[6] See https://github.com/eske/seq2seq.

[7] The difference between best and worst configurations varied from 0.5% to 1.3% for AUD, and 1.6% for true phones.

| AUD feat. | AUD model | dpseg (monoling.) | | | proportional baseline (biling.) | | | attentional (biling.)* | | | att. average (biling.)+ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** | **F** |
| **MFCC** | **HMM** | 27.9 | 80.2 | 41.3 | 42.6 | 49.9 | 46.0 | 51.6 | 44.9 | 48.0 | 55.5 | 43.7 | 48.9 |
| **MFCC** | **SVAE** | 29.8 | 69.1 | 41.7 | 42.2 | 51.9 | 46.6 | 52.7 | 45.0 | 48.5 | 55.7 | 44.1 | 49.2 |
| **MBN** | **HMM** | 27.8 | 72.6 | 40.2 | 42.5 | 48.1 | 45.2 | 50.8 | 44.5 | 47.4 | 54.1 | 42.9 | 47.8 |
| **MBN** | **SVAE** | 30.0 | 72.9 | 42.5 | 42.5 | 51.6 | 46.6 | 57.2 | 43.0 | 49.1 | 60.6 | 42.5 | 50.0 |
| **true phones** | | 53.8 | 83.5 | 65.4 | 44.5 | 62.6 | 52.0 | 60.5 | 59.9 | 60.3 | 62.8 | 59.3 | 61.0 |

Table 2: *Precision, Recall and F-measure on word boundaries over the Mboshi5k corpus, using different AUD to extract pseudo-phones from speech. True phones toplines are also provided. *averaged scores over 5 different runs; +averaged 5 attention matrices*

| | Phones per Sentence | | | Tokens per Sentence | | |
|---|---|---|---|---|---|---|
| | avg | max | min | avg | max | min |
| **true phones** | 21.8 | 60 | 4 | 6.0 | 21 | 1 |
| **MFCC HMM** | 37.0 | 95 | 11 | 3.6 | 22 | 1 |
| **MFCC SVAE** | 26.3 | 73 | 7 | 7.6 | 26 | 1 |
| **MBN HMM** | 32.1 | 93 | 12 | 5.0 | 14 | 1 |
| **MBN SVAE** | 23.4 | 71 | 7 | 5.4 | 21 | 1 |

Table 3: *AUD methods differ in their ability to encode speech utterances (left side); which impacts the final segmentation of the attentional model (right side).*

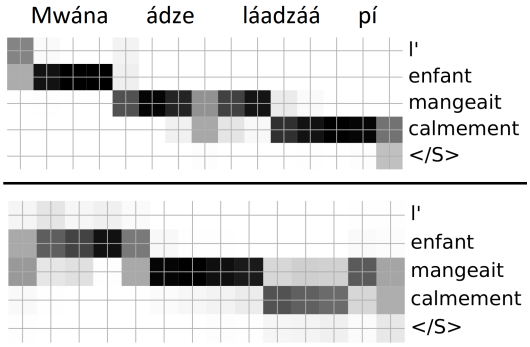


Figure 2: *NMT output alignment for true phones (top) and AUD using MBN SVAE (bottom). For illustration purposes, we give the transcription of the audio in Mboshi.*

is another clue as to why the MBN SVAE is performing best on our task.

Analyzing the averaged attention model's results in Table 2, we can see an increase in performance of about 0.8% in all cases. This improvement also holds for tokens and types scores (not reported here). However, while the *topline* achieves 34.3% of vocabulary (types) retrieval, our best AUD setup achieves 13.5% only. This illustrates the difficulty of the word discovery task – a task already challenging with true phones – in noisy setups. The large difference between true phones and pseudo-phones for type's retrieval could be explained by the fact that a single change in the pseudo-phone sequence representing two speech segments of a same word will have the consequence to split the word cluster in two parts (in two types). A deeper analysis of the word clusters obtained is probably necessary to better understand how AUD from speech affects the word discovery task, and to come up with ways to better cluster speech segments in relevant types.

The attention-based segmentation technique remains much more robust for word boundary detection than our monolingual (Bayesian) approach. Figure 2 shows an example of a (good quality) soft alignment (attention) matrix produced in our best setup (MBN SVAE).

## 5. Related Work

Word segmentation in a monolingual setup was previously investigated from text input [10] and from speech [23, 25, 26, 27]. Word discovery experiments from text input on Mboshi were reported in [28]. Bilingual setups (cross-lingual supervision) for word segmentation were discussed by [29, 30, 31, 9], but applied to speech transcripts (true phones). Looking at NMT from speech, the research by [11, 12] are recent examples of approaches to end-to-end spoken language translation, but using much larger data conditions than ours.

Among the most relevant to our approach are the works of [5] on speech-to-translation alignment using attentional NMT and of [32] for language documentation. However, the former does not address word segmentation and is not applied to a language documentation scenario, while the latter does not provide a full coverage of the speech corpus analyzed.

## 6. Conclusions

Different from these related works and inspired by [9], this paper presented word segmentation from speech, in a bilingual setup and for a real language documentation scenario (Mboshi). The proposed approach first performs AUD to generate pseudo-phones from speech, and then uses these units in an encoder-decoder NMT for word segmentation. Our method leads to promising results for word segmentation from speech, outperforming three baselines in noisy (pseudo-phones) setups and finally delivering the best results reported so far for the Mboshi5k corpus. Future work includes investigating sources of weak supervision and minimal viable corpus sizes.

## 7. Acknowledgements

This work was partly funded by French ANR and German DFG under grant ANR-14-CE35-0002 (BULB project). This work was started at JSALT 2017 in CMU, Pittsburgh, and was supported by JHU and CMU (via grants from Google, Microsoft, Amazon, Facebook, Apple). It used the Extreme Science and Engineering Discovery Environment (NSF grant number OCI-1053575 and NSF award number ACI-1445606).

# 8. References


[1] J. Glass, "Towards unsupervised speech processing," in *Proc. IEEE-ISSPA*, 2012, pp. 1–4.

[2] A. Jansen, E. Dupoux, S. Goldwater, M. Johnson, S. Khudanpur, K. Church, N. Feldman, H. Hermansky, F. Metze, R. Rose, M. Seltzer, P. Clark, I. McGraw, B. Varadarajan, E. Bennett, B. Borschinger, J. Chiu, E. Dunbar, A. Fourtassi, D. Harwath, C.-y. Lee, K. Levin, A. Norouzian, V. Peddinti, R. Richardson, T. Schatz, and S. Thomas, "A summary of the 2012 JH CLSP Workshop on zero resource speech technologies and models of early language acquisition," in *Proc. ICASSP*, 2013.

[3] E. Dunbar, X. Nga-Cao, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier, X. Anguera, and E. Dupoux, "The zero resource speech challenge 2017," in *Proc. Automatic Speech Recognition and Understanding (IEEE ASRU)*, 2017.

[4] L. Besacier, B. Zhou, and Y. Gao, "Towards speech translation of non-written languages," in *Spoken Language Technology Workshop, 2006. IEEE*, 2006, pp. 222–225.

[5] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, "An attentional model for speech translation without transcription," in *Proc. NAACL-HLT*, San Diego, CA, 2016, pp. 949–959.

[6] E. Dupoux, "Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner," *Cognition*, vol. 173, pp. 43–59, 2018.

[7] G. Adda, S. Stüker, M. Adda-Decker, O. Ambouroue, L. Besacier, D. Blachon, H. Bonneau-Maynard, P. Godard, F. Hamlaoui, D. Idiatov, G.-N. Kouarata, L. Lamel, E.-M. Makasso, A. Rialland, M. V. de Velde, F. Yvon, and S. Zerbian, "Breaking the unwritten language barrier: The BULB project," *Procedia Computer Science*, vol. 81, pp. 8–14, 2016.

[8] 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," *Procedia Computer Science*, vol. 81, pp. 61–66, 2016.

[9] M. Zanon Boito, A. Berard, A. Villavicencio, and L. Besacier, "Unwritten languages demand attention too! Word discovery with encoder-decoder models," in *Proc. Automatic Speech Recognition and Understanding (IEEE ASRU)*, 2017.

[10] 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.

[11] A. Bérard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," in *NIPS workshop on End-to-end Learning for Speech and Audio Processing*, 2016.

[12] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly transcribe foreign speech," *arXiv preprint arXiv:1703.08581*, 2017.

[13] P. Godard, G. Adda, M. Adda-Decker, J. Benjumea, L. Besacier, J. Cooper-Leavitt, G. Kouarata, L. Lamel, H. Maynard, M. Müller, A. Rialland, S. Stüker, F. Yvon, and M. Z. Boito, "A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments," in *Proc. LREC*, Miyazaki, Japan, 2018.

[14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, San Diego, CA, 2015.

[15] L. Ondel, L. Burget, and J. Černockỳ, "Variational inference for acoustic unit discovery," *Procedia Computer Science*, vol. 81, pp. 80–86, 2016.

[16] Y. W. Teh and M. I. Jordan, "Hierarchical Bayesian nonparametric models with applications," in *Bayesian Nonparametrics: Principles and Practice*, N. Hjort, C. Holmes, P. Müller, and S. Walker, Eds. Cambridge University Press, 2010.

[17] K. Kurihara, M. Welling, and Y. W. Teh, "Collapsed variational Dirichlet process mixture models," in *Proc. IJCAI*. Morgan Kaufmann Publishers, 2007, pp. 2796–2801.

[18] M. Johnson, "Composing graphical models with neural networks for structured representations and fast inference," in *Advances in Neural Information Processing Systems*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 2946–2954.

[19] D. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. ICLR*, Banff, Australia, 2014.

[20] M. Hoffman, "Stochastic variational inference," *Journal of Machine Learning Research*, vol. 14, pp. 1303–1347, 2013.

[21] F. Grézl and M. Karafiát, "Adapting multilingual neural network hierarchy to a new language," in *Proc. SLTU*, 2014, pp. 39–45.

[22] B. Ludusan, M. Versteegh, A. Jansen, G. Gravier, X.-N. Cao, M. Johnson, and E. Dupoux, "Bridging the gap between speech technology and natural language processing: an evaluation toolbox for term discovery systems," in *Proc. LREC*, 2014.

[23] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *Proc. Automatic Speech Recognition and Understanding (IEEE ASRU)*, 2011, pp. 401–406.

[24] L. Ondel, P. Godard, L. Besacier, E. Larsen, M. Hasegawa-Johnson, O. Scharenborg, E. Dupoux, L. Burget, F. Yvon, and S. Khudanpur, "Bayesian Models for Unit Discovery on a Very Low Resource Language," in *Proc. ICASSP*, Calgary, Alberta, Canada, 2018.

[25] C.-y. Lee, T. J. O'Donnell, and J. Glass, "Unsupervised lexicon discovery from acoustic input," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 389–403, 2015.

[26] C. Bartels, W. Wang, V. Mitra, C. Richey, A. Kathol, D. Vergyri, H. Bratt, and C. Hung, "Toward human-assisted lexical unit discovery without text resources," in *Spoken Language Technology Workshop (IEEE SLT)*, 2016, pp. 64–70.

[27] M. Elsner, S. Goldwater, N. Feldman, and F. Wood, "A joint learning model of word segmentation, lexical acquisition, and phonetic variability," in *Proc. EMNLP*. Association for Computational Linguistics, 2013, pp. 42–54.

[28] P. Godard, G. Adda, M. Adda-Decker, A. Allauzen, L. Besacier, H. Bonneau-Maynard, G.-N. Kouarata, K. Löser, A. Rialland, and F. Yvon, "Preliminary experiments on unsupervised word discovery in mboshi," in *Proc. Interspeech*, 2016.

[29] S. Stüker, "Towards human translations guided language discovery for ASR systems," in *Proc. SLTU*, Hanoi, Vietnam, May 2008.

[30] S. Stüker, L. Besacier, and A. Waibel, "Human Translations Guided Language Discovery for ASR Systems," in *Proc. Interspeech*. Brighton (UK): Eurasip, 2009, pp. 1–4.

[31] F. Stahlberg, T. Schlippe, S. Vogel, and T. Schultz, "Word segmentation through cross-lingual word-to-phoneme alignment." in *Spoken Language Technology Workshop (IEEE SLT)*, 2012, pp. 85–90.

[32] A. Anastasopoulos, S. Bansal, D. Chiang, S. Goldwater, and A. Lopez, "Spoken term discovery for language documentation using translations," in *Proc. Workshop on Speech-Centric Natural Language Processing*, 2017, pp. 53–58.