Lexicon Discovery for Language Preservation using Unsupervised Word Segmentation with Pitman-Yor Language Models (FGNT-2015-01)

Oliver Walter, Reinhold Haeb-Umbach

Department of Communications Engineering
University of Paderborn, Germany
{walter,haeb}@nt.uni-paderborn.de

Jan Strunk, Nikolaus P. Himmelmann

Institut für Linguistik
Universität zu Köln, Germany
{jan.strunk,n.himmelmann}@uni-koeln.de

Abstract

In this paper we show that recently developed algorithms for unsupervised word segmentation can be a valuable tool for the documentation of endangered languages. We applied an unsupervised word segmentation algorithm based on a nested Pitman-Yor language model to two austronesian languages, Wooi and Waima'a. The algorithm was then modified and parameterized to cater the needs of linguists for high precision of lexical discovery: We obtained a lexicon precision of of 69.2% and 67.5% for Wooi and Waima'a, respectively, if single-letter words and words found less than three times were discarded. A comparison with an English word segmentation task showed comparable performance, verifying that the assumptions underlying the Pitman-Yor language model, the universality of Zipf's law and the power of n-gram structures, do also hold for languages as exotic as Wooi and Waima'a.

Index Terms: unsupervised vocabulary acquisition, unsupervised language acquisition, unsupervised word segmentation

1. Introduction

Linguists estimate that of the approximately 6,500 languages worldwide, about half of them will have vanished within less than 100 years [1]. Language documentation efforts, such as the US National Science Foundation's Documenting Endangered Languages Program [2], the Hans Rausing Endangered Languages Project [3] or the "Dokumentation bedrohter Sprachen" (DoBeS) project funded by the Volkswagen foundation [4], have created large multimedia corpora of many languages to document what may otherwise be lost forever. Their manual transcription which is required to conduct phonetic or linguistic research on them, is a significant challenge, if not the least from a financial point of view.

Thus the use of speech and language technologies has been studied for automatic transcription and segmentation [5, 6], and projects, such as the AUVIS project [7], have been initiated to bring phonetic and linguistic researchers and natural language processing (NLP) scientists and engineers together to develop tools according to the needs of field researchers. CLARIN [8], the Common Language Resources and Technology Infrastructure, provides a one-stop access to language resources and tools to discover, explore, exploit, annotate, analyse or combine digital language data in written, spoken, video or multimodal form, for researchers in the humanities or social sciences. Participating centers are offering access services to data, tools and expertise through this portal.

Many of the languages threatened by extinction are only spoken and not written. Tools for the discovery of the inventory of phonemes, the word list and a pronunciation dictionary, are for the most part not available to linguists todate. These tasks, however, share similarities with the research objectives of 'zero resource speech technologies', a field of research that has found much attention in recent years due to a large IARPA program in this field [9]. Lexicon discovery can be viewed as an unsupervised learning problem with an infinite, or at least unknown, number of items to be learned. Exactly for this task, nonparametric Bayesian methods have proven to be very effective.

Lee and Glass developed a latent variables model for acoustic unit discovery, based on a Dirichlet Process (DP) mixture [10]. With the DP, the number of units need not be specified in advance and can grow with the data. They showed that the discovered sub-word units were highly correlated with English phonemes and the resulting segmentation outperformed other state-of-the-art methods. A similar statistical model, however now for input of categorical nature, has been used for the unsupervised segmentation of character sequences into words [11]. The key idea in both cases is that the succession of labels or speech frames within a word or an acoustic unit is more predictable than at unit boundaries. We have extended the unsupervised word segmentation algorithm of [11] to cope with noisy, i.e., error-prone input as received by an ASR decoder, and were able to discover words and learn a language model and pronunciation lexicon from continuous speech input only [12].

In this paper we report on experiments with word discovery algorithms for two austronesian languages, Wooi [13] and Waima'a [14]. Wooi is spoken at the western tip of Yapen Island in Indonesia. Approximately 1600 people still speak Wooi. Waimaa is an endangered Austronesian language from Timor Lorosa'e (East Timor). The precise extent of the Waimaa speaking area and poulation remain to be determined.

On text corpora from these languages we perform unsupervised word segmentation employing Bayesian language modeling based on the Pitman-Yor process. We experiment with language model orders and investigate the usefulness of word length models. Further, we propose a method to tune lexical discovery results towards high precision to cater the needs of linguists. A comparison with an English corpus of similar size, the Wall Street Journal (WSJ) corpus, yields interesting insights.

The paper is organized as follows. In the next Section we give a summary of unsupervised word segmentation employing a Pitman-Yor language model. The datasets are described in Section 3 and word segmentation experiments are described in sections 4 and 5, before concluding the paper with Section 6.

2. Unsupervised Discovery of Lexical Units

The task we are facing is as follows: we are given a corpus of unsegmented character sequences (or any other input of categorical nature, such as letters or phonemes) and wish to segment the input into words. This will give us a word list for that corpus, a first step towards a lexical analysis of the language.

When doing so, we are faced with a chicken-and-egg problem: we need a word list and a language model (i.e., probabilities of words and word sequences) to do the segmentation. However, to estimate the language model, we need to have a segmentation. Before coming back to this problem we first describe a language model which is able to cope with an unknown number of words.

2.1. Nested Pitman-Yor Language Model

The statistical model we are going to use to solve the task must be able to assign a probability to unknown words based on their spelling and handle an *a priori* unknown number of already discovered words (e.g. be nonparametric). It must also capture the only two constraints we impose:

i) The occurrence of the words follow a power law distribution

ii) n-gram structures apply for both, the word, and the character level

The first assumption, known by the name Zipf's law, has shown to be reasonable for most natural, and artificial languages [15], and we presume it will also hold for the two exotic languages we consider here, Wooi and Waima'a. The second one has been shown to be effective in modeling language data and relies on the fact, that predictability is the fundamental requirement to differentiate anything structured from noise.

A model which meets all of these requirements is the Nested Pitman-Yor Language Model (NPYLM), first introduced in [11]. This model incorporates two hierarchical Pitman-Yor language models (HPYLM), one for the discovered words and one for the characters [16].

The HPYLM is based on the Pitman-Yor (PY) process, which is a generalized Dirichlet process governed by two parameters – a discount parameter d and a strength parameter θ – and a base distribution G_0, which is defined over a probability space X of tokens (words or characters). The drawing process for the PY process may be explained through a Chinese-restaurant analogy: at any time the process has a number of "tables", which can grow infinitely large, and each of these tables has a symbol from X associated with it. A new draw (a "customer") is either "seated" at an existing table and assigned the symbol (word) associated with it, or assigned to a new table. When a new table is selected, the symbol assigned to it is drawn from G_0. The overall process results in draws from a distribution G:

$$G \sim \mathrm{PY}\left(d, \theta, G_0\right). \tag{1}$$

The parameter θ controls how similar this drawn distribution is to the base distribution which itself can be seen as a mean distribution of the drawn ones. Draws from the distribution G obey the power law and hence incorporate the prior knowledge of Zipf's law.

The n-gram structure is captured by embedding the above in a hierarchical structure. At the n-gram level a separate PY process is instantiated for *every* $(n-1)$-word context. The *base* distribution for each PY process at the n-gram level is a $(n-2)$-context specific PY process. One can view the entire structure as a tree, where the root node gives the unigram probability, its children the bigram probability and so forth. This can be interpreted as smoothing since a certain amount of the probability mass is moved to the shorter context. The degree of smoothing is controlled by the discount parameter d. So instead of one, there are several distributions, one for each context:

$$G\left(\mathbf{u}\right) \sim \mathrm{PY}\left(d, \theta, G\left(\pi\left(\mathbf{u}\right)\right)\right). \tag{2}$$

The notation $\pi\left(\mathbf{u}\right)$ describes the shorter context, e.g. if $\mathbf{u} = (w_{i-1}, \ldots, w_{i-n+1})$ then $\pi\left(\mathbf{u}\right) = (w_{i-1}, \ldots, w_{i-n+2})$. To cope with yet unknown words, another tree is built, this time for the characters, which is used to calculate the likelihood of the character sequence forming the word to serve as the base distribution for the word model, resulting in the aforementioned NPYLM.

The model and its underlying sufficient statistics Σ, also called "seating arrangement" in the Chinese-restaurant analogy, are used to calculate the predictive probability of a word w given its context $\mathbf{u}$:

$$\begin{aligned}\Pr\left(w|\mathbf{u}\right) &= \frac{c_{\mathbf{u}w\cdot} - d_{|\mathbf{u}|}t_{\mathbf{u}w}}{\theta_{|\mathbf{u}|} + c_{\mathbf{u}\cdot\cdot}} \\ &\quad + \frac{\theta_{|\mathbf{u}|} + d_{|\mathbf{u}|}t_{\mathbf{u}\cdot}}{\theta_{|\mathbf{u}|} + c_{\mathbf{u}\cdot\cdot}} \Pr\left(w|\pi\left(\mathbf{u}\right)\right) \\ &=: \Pr_{\text{local}}\left(w|\mathbf{u}\right) + \Pr\left(\text{FB}|\mathbf{u}\right)\Pr\left(w|\pi\left(\mathbf{u}\right)\right)\end{aligned} \tag{3}$$

with the obvious definition for $\Pr_{\text{local}}\left(w|\mathbf{u}\right)$ and $\Pr\left(\text{FB}|\mathbf{u}\right)$.

Again, a Chinese-restaurant analogy may be used to interpret eq. (3). $c_{\mathbf{u}w\cdot}$ describes the counts of word w in the context $\mathbf{u}$. $t_{\mathbf{u}w}$ describes how many "tables" are occupied by this word. $d_{|\mathbf{u}|}$ and $\theta_{|\mathbf{u}|}$ are the discount and strength parameters of the Pitman-Yor process at the context length $|\mathbf{u}|$, which are shared by all contexts of the same length. The $\cdot$ indicates a marginalization, so $c_{\mathbf{u}\cdot\cdot} = \sum_w c_{\mathbf{u}w\cdot}$ is the count of all words in the context $\mathbf{u}$. In the generative perspective $\Pr\left(\text{FB}|\mathbf{u}\right)$ is the probability for assigning a new table for a new draw. The parameters $d_{|\mathbf{u}|}$, $\theta_{|\mathbf{u}|}$ are sampled after each iteration as is explained in [16].

With the help of eq. (3) we are able to assign a probability to a given word sequence.

2.2. Unsupervised Word Segmentation

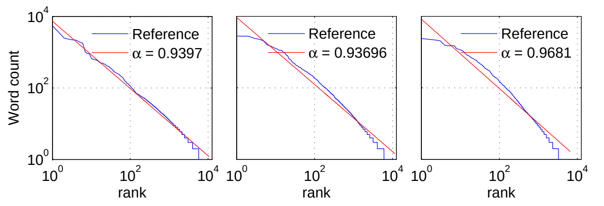
Given the unsegmented character sequence at the input, a segmentation into words can be carried out by an iterative process, which alternates between drawing a segmentation for a sentence using the estimated NPYLM to calculate the segmentation likelihood and reestimating the language model parameters given the drawn segmentation. Starting from a random initial segmentation and the language model estimated from it, we choose a sentence of the corpus at random, remove its contribution to the word and letter counts and estimate the language model on the remaining corpus. Then this language model is employed to draw a new segmentation for that sentence. This can be efficiently accomplished with the *Forward-Filtering Backward-Sampling* algorithm described in [11]. This process is repeated until a stable segmentation is found.

It has been shown that good segmentation results on both, character sequences [11, 17] and phoneme sequences [17] can be achieved with the help of the NPYLM.

3. Datasets

We carried out unsupervised word segmentation experiments on three corpora:

- The first corpus contains the text prompts of the WSJ-CAM0 training data, consisting of 5612 sentences with 95442 running words and a lexicon size of 10658 words.



(a) WSJ (b) Wooi (c) Waima'a

Figure 1: Word count as a function of rank

The results on this corpus serve as reference results to set the performance on the austronesian corpora into comparison to a well-known database.

- The second corpus contains text transcriptions of 743 recording sessions of the Wooi language. The sessions are separated into 37028 intonation units, with 123848 running words and a lexicon size of 11512 words.
- The third corpus contains 391 recording sessions of the Waima'a language. The sessions are separated into 19888 intonation units, with 88751 running words and a lexicon size of 6535 words.

The three copora are comparable in terms of the number of words in the running text and the lexicon size. While the WSJ corpus contains text from newspaper articles, the Wooi and Waima'a corpus contain spontaneous speech.

Fig. 1 displays the word counts on the three corpora as a function of the rank of the word in a list ordered according to word frequency. In the chosen log-log scale adherence to Zipf's law can be easily checked. If for the probability of the n-th most probable word w_n it holds that

$$\Pr(w_n) \propto \frac{1}{n^\alpha}, \tag{4}$$

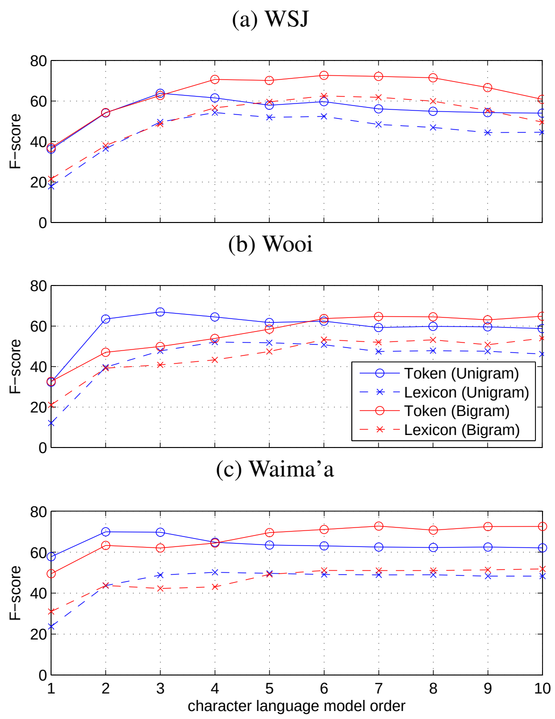
the distribution of word probabilities will form a straight line with slope $-\alpha$. The WSJ corpus shows a straight line, while the curves for the Wooi and Waima'a corpus slightly deviate from a line for the lower ranks. For the higher ranks the lines follow the power law property. The maximum likelihood estimates of α are 0.9397 for the WSJ corpus, 0.93696 for the Wooi corpus and 0.9681 for the Waima'a corpus. These results show that the assumption of a power law property by the Pitman-Yor language model holds for English as well as for exotic languages like Wooi and Waima'a.

4. Word Segmentation Experiments

In a first experiment we applied the word segmentation algorithm to all three corpora and evaluated the quality of the segmentation result for different orders of the character language model and the word language model. As performance measures we used the token F-score and the lexicon F-score.

The F-score F is the harmonic mean of the precision P and the recall R. The precision is defined as the ratio of the number of correctly found words N_{correct} and the total number of found words N_{found}, while recall is defined as the ratio of the number of correctly found words N_{correct} and the number of words to be found $N_{\text{reference}}$:

$$F = 2\frac{P \cdot R}{P + R}, \; P = \frac{N_{\text{correct}}}{N_{\text{found}}}, \; R = \frac{N_{\text{correct}}}{N_{\text{reference}}}. \tag{5}$$



(a) WSJ

(b) Wooi

(c) Waima'a

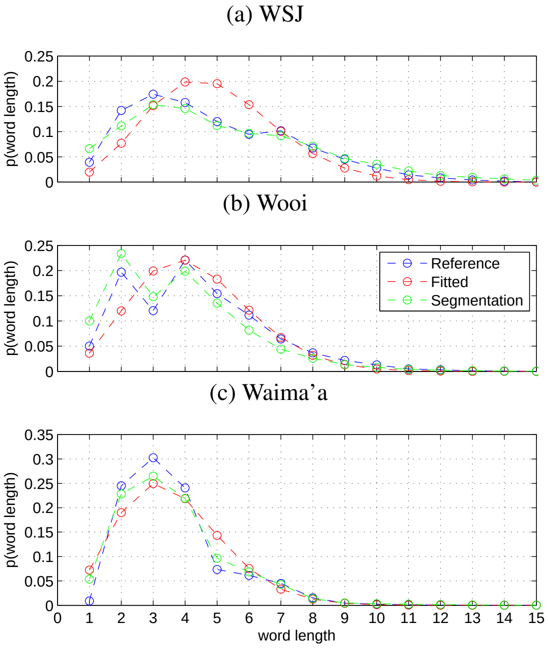
Figure 2: Token and lexicon F-score as function of model order

The token F-score is determined by aligning the segmented sequence with the reference sequence, and choosing the alignment with the lowest edit distance, allowing deletions, insertions, substitutions and matches. Matches are then counted as correctly found tokens. For the lexicon F-score a word in the learned lexicon is counted as correctly found if it is present in the reference lexicon.

Fig. 2 shows the results for the token and lexicon F-score as a function of the character language model order when using a unigram and a bigram word language model. For the unigram word language model, the F-scores reach their peak at around a spelling model order of 3. For the bigram word language model, the F-scores increase with increasing spelling model order, only for the WSJ corpus the F-score decreases for character language model orders greater than 8. Increasing the character language model order further makes no sense since the number of words of length larger than 8 characters is very small.

The bigram word language model performs better than the unigram word language model for the WSJ corpus (72.6 % token and 62.5 % lexicon F-score) and the Waima'a corpus (72.73 % token, 51.82 % lexicon) corpus. For the Wooi corpus, the peak token F-score for the unigram is at 66.99 % and for the bigram at 64.85 %. For the lexicon F-score the bigram word language model performs best, achieving an F-score of 54.01 %. Based on these results we chose a character language model order of 7 and a word model order of 2, since this setup gave good results across the three corpora.

In the next set of experiments we experimented with explicit word length modeling. It has been observed that an n-gram model at the character level results in inadequately low probabilities assigned to long words, because the model has a largely exponential distribution over length [18]. To correct this, a Poisson distribution of the word length is imposed by dividing the likelihood of a character sequence of length k,


(a) WSJ

(b) Wooi

(c) Waima'a

Figure 3: Word length distribution for reference and result

$P(c_1, \cdots, c_k)$ at the base distribution, by the probability of a word length k according to the language model and multiplying it with the probability of a word length k according to the Poisson distribution [11]:

$$P(c_1, \cdots, c_k) \leftarrow P(c_1, \cdots, c_k) \frac{P_p(k; \lambda)}{P(k)}, \tag{6}$$

where the Poisson distribution is given by

$$P_p(k; \lambda) = e^{-\lambda} \frac{\lambda^k}{k!}. \tag{7}$$

The Poisson parameter λ is estimated after each iteration of the word segmentation algorithm on the segmentation result.

Fig. 3 shows the distribution of the word length in the reference corpus and in the segmentation result with length modeling activated. It is striking how well the distribution of the word lengths of the segmentation result follows the distributions in the reference corpus. Even the dip at word length 3 in the Wooi corpus is reflected in the automatically found segmentation! The figure also shows a Poisson distribution fitted to the word length distribution measured on the reference text, which resulted in a mean value of 4.9, 4.3 and 3.6 characters per word for WSJ, Wooi and Waima'a, respectively.

Obviously the Poisson correction has little influence on the segmentation. This can be seen by the fact, that the distribution measured on the segmentation result is quite different from a Poisson distribution. Indeed, the explicit word length correction of eq. (6) did not improve the segmentation performance.

5. Optimization of Lexicon F-Score

To be a valuable tool for linguists, the lexicon discovered by the word segmentation algorithm should have both high precision and high recall. However, a word segmentation algorithm based on n-gram probabilities cannot be expected to perform well for words which occur very rarely in the corpus. Errors will therefore concentrate on infrequent words.

To analyse this effect, we carried out word segmentation with a bigram word model and a 7-gram character model as described and, in a postprocessing step, kept only word tokens that occurred at least a given count in the segmentation result. Table 1 shows the effect of discarding infrequent words on the token and lexicon F-score.

Table 1: F-scores of the different setups

	WSJ		Wooi		Waima'a	
Count	Token	Lex.	Token	Lex.	Token	Lex.
1	69.6	61.8	64.8	52.0	73.6	51.0
2	71.6	69.7	65.6	59.8	74.8	63.9
3	72.3	76.4	66.2	64.4	75.2	68.1
3/2	76.9	76.5	67.8	64.3	76.8	68.1

If only those words with a count of at least 1, 2 or 3 were kept in the segmentation result and the reference, F-scores at lexicon level rose significantly. Since infrequent words do not occur often in the segmentation, the token F-score is not greatly influenced. Discarding words with higher counts did not improve the results significantly but discarded too many words. In the table, the row with the first entry being 3/2 means, that only those words with a count of at least 3 and additionally with a length of at least 2 characters were kept. As can be seen, discarding single-character words did slightly improve the segmentation result. Discarding longer words decreased the scores. Table 2 shows the resulting precisions as well as the number of found words N_f and remaining words N_r for the last setup.

Table 2: Precisions and remaining words of 3/2 setup

	WSJ		Wooi		Waima'a	
	Token	Lex.	Token	Lex.	Token	Lex.
P	81.1	75.7	64.7	69.2	78.1	67.5
N_r	75204	4037	119225	4066	79663	2364
N_f	92561	11070	139718	8641	87927	5783

With such a parameterization the presented word segmentation can serve as a handwork saving preprocessing tool for linguists: while the algorithm takes care of all frequent words and thus most of the text, linguists can concentrate on infrequent words to complete the lexicon.

6. Conclusions

In this paper we have shown that unsupervised word segmentation based on a nested Pitman-Yor language model can be a valuable tool for linguistic analysis, adding to the set of natural language processing tools to discover, annotate, analyse or combine digital language data. On the austronesian languages Wooi and Waima'a we have achieved segmentation results that are comparable to those obtained on a well-known English corpus, demonstrating the universality of unsupervised learning approaches to language processing. We have also shown that significantly higher F-scores are obtained on frequent compared to infrequent words, leading to the conclusion, that the proposed algorithm can be used as a preprocessing tool, which saves, however, not frees linguists from manual effort.

7. References

[1] A. Schenk, “Da fehlen einem die Worte,” *Die ZEIT, (in German)*, no. 38, p. 71f, Sept. 12, 2013.

[2] Documenting endangered languages (DEL). [Online]. Available: http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=12816

[3] Hans Rausing endangered languages project. [Online]. Available: http://www.hrelp.org/

[4] Dokumentation bedrohter Sprachen (DoBeS). [Online]. Available: http://dobes.mpi.nl/

[5] O. Schreer and D. Schneider, “Supporting linguistic research using generic automatic audio/video analysis,” in *Potentials of Language Documentation: Methods, Analyses and Utilization*, F. Seifart, G. Haig, N. Himmelmann, J. D., A. Margetts, and P. Trilsbeek, Eds. University of Hawaii Press, 2012.

[6] C. DiCanio, H. Nam, D. H. Whalen, H. Timothy Bunnell, J. D. Amith, and R. C. Garca, “Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment,” *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2235–2246, 2013.

[7] AUVIS project. [Online]. Available: https://tla.mpi.nl/projects_info/auvis/

[8] CLARIN. [Online]. Available: http://www.clarin.eu/

[9] M. Harper. (2014) IARPA babel program. [Online]. Available: http://www.iarpa.gov/index.php/research-programs/babel

[10] C.-Y. Lee and J. Glass, “A nonparametric Bayesian approach to acoustic model discovery,” in *Proc. of 50th Annual Meeting of the ACL.* Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 40–49. [Online]. Available: http://dl.acm.org/citation.cfm?id=2390524.2390531

[11] D. Mochihashi, T. Yamada, and N. Ueda, “Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling,” in *Proc. of the Joint Conf. of Annual Meeting of the ACL and Internat. Conf. on Natural Language Processing of the AFNLP*, 2009.

[12] J. Heymann, O. Walter, R. Haeb-Umbach, and B. Raj, “Iterative Bayesian word segmentation for unsupervised vocabulary discovery from phoneme lattices,” in *submitted to*, May 2014.

[13] Kirihio, J. Karter, V. Unterladstetter, A. Arilaha, F. Morigerowski, A. Loch, Y. Sawaki, and N. P. Himmelmann. (2009-2015) DoBeS Wooi documentation. DoBeS Archive. MPI Nijmegen. [Online]. Available: http://www.mpi.nl/DOBES/

[14] Belo, M. C.A, J. Bowden, J. Hajek, N. P. Himmelmann, and A. V. Tilman. (2002-2006) DoBeS Waima’a Documentation. DoBeS Archive. MPI Nijmegen. [Online]. Available: http://www.mpi.nl/DOBES/

[15] C. Manning, D. Christopher, and H. Schütze, *Foundations of statistical natural language processing*. MIT press, 1999.

[16] Y. W. Teh, “A hierarchical Bayesian language model based on Pitman-Yor processes,” in *Proc. of the 21st Int. Conference on Computational Linguistics and the 44th annual meeting of the ACL*, 2006.

[17] O. Walter, R. Haeb-Umbach, S. Chaudhuri, and B. Raj, “Unsupervised word discovery from phonetic input using nested Pitman-Yor language modeling,” in *Proc. IEEE International Conference on Robotics and Automation*, Karlsruhe, 2013.

[18] M. Nagata, “Automatic extraction of new words from japanese texts using generalized forward-backward search,” in *Empirical methods in natural language processing*, 1996, pp. 48–59.