Topic modelling to support English text selection for translation into South Africa's other official languages

Mazarura, Jocelyn
Department of Statistics, University of Pretoria
jocelyn.mazarura@up.ac.za

de Wet, Febe
Department of Electrical and Electronic Engineering, Stellenbosch University & School of Electrical, Electronic and Computer Engineering, North-West University
fdw@sun.ac.za

Abstract

Appropriate training data is a prerequisite for the development of natural language processing (NLP) techniques. Vast amounts of language data are typically required to develop NLP tools that perform at state-of-the-art level. Such abundant resources are currently only available in a few languages. The remaining languages have to find alternative ways to become "NLP-enabled". The aim of the study reported on here is to make more language data available to support NLP development in the official languages of South Africa. In this paper we present the idea of generating text data by means of translation. We also propose the use of topic modelling to identify text in a highly resourced source language that will yield meaningful translations in under-resourced target languages. More specifically, the paper describes how topic modelling was used to identify English Wikipedia articles that should be suitable for translation into South Africa's 10 other official languages.

Keywords: DHASA, topic modelling, text selection, translation, under-resourced languages

1 Introduction

The resources required to implement state-of-the art performance in natural language processing (NLP) applications like text generation, question answering, automatic speech recognition, etc. are currently only available in a few well-resourced languages. The other 6 000-odd languages that are spoken in the world today have to find alternative ways to prepare themselves for a digital future society in which the ability to process language and speech automatically will be a prerequisite for participation. Applications like search engines and chat bots are developed using text data while others, like automatic dictation systems, require both text and speech data to be implemented. The current study is specifically aimed at developing text resources that are suitable for language modelling in South Africa's official languages.

A number of text resources have been developed during previous projects and are available in the resource catalogue of the South African Centre for Digital Language Resources (SADiLaR[1]). However, the extent of the existing text corpora is not adequate to implement large vocabulary automatic speech recognition (ASR). Moreover, the most extensive text resource that is available in all the official languages, the NCHLT Text Corpora (Eiselen & Puttkammer 2014), was derived from documents published on the South African government website domain (*.gov.za). As a result, the vocabulary and language usage is domain specific. The aim of the current work is to develop new text resources that can be used in addition to or in combination with existing corpora. The new data should ideally represent a diversity of domains and should contribute to expanding existing lexica.

One obvious and potentially rich source of text data is the world wide web. While many languages have at least some presence on the web, there are also many which are not represented on the internet at all. However, the fact that a language is present on the web does not guarantee that the available data is suitable for NLP research. Web-text is also not by default of good quality. In addition, some internet texts are extremely domain specific. Especially for under-resourced languages, online texts tend to be restricted to religious publications or government documents.

A strategy that has been proposed to strengthen the internet presence of under-represented languages is to create text by, for instance, encouraging first language speakers to write Wikipedia articles (Pretorius & Wolff 2020). The approach has been successful in some instances and have resulted in the first ever contributions to Wikipedia in many languages. Despite the advantage of producing "native" text, this is a slow process and has, to date, not yielded the amount of text data required by the data-hungry deep learning methods that are currently widely used in the field of NLP.

This paper proposes large scale translation as an alternative to text generation[2]. Despite the differences between *real* and translated text, translations could bridge the gap between having almost no data available at all and having just enough data to bootstrap semi-supervised systems that can contribute to developing further resources. Proposing translation as a means to generate text in an under-resourced language also means deciding on what should be translated. Many of the obvious answers to this question, e.g. the 100 most popular Mandarin Wikipedia articles, would not necessarily yield meaningful text in other languages. The research reported on in this paper represents an attempt to find a systematic way of identifying text that is suitable for translation and that would yield meaningful text in the target languages. More specifically, we describe how English Wikipedia articles that should be suitable for translation into South Africa's 10 other official languages were identified.

2 Background

Previous attempts to translate NLP tasks and data sets from highly resourced to under-resourced languages have encountered various challenges related to the differences between the source and target languages. For instance, during the development of the *African Wordnet*, it was found that many concepts in existing Wordnets were not lexicalised in South African languages (Griesel & Bosch 2020). This difference was overcome by using the SIL Comparative African Wordlist (SILCAWL) introduced by Snider & Roberts (2004) as a guideline to add new synsets to the African Wordnet. This list consists of a vast selection of words relating to various things, such as body parts, emotions and stages of life in English and French. The content of the entries in the list is – to a large extent – lexicalised in South Africa's official languages. Using the SILCAWL as a point of departure has the additional advantage that it does not perpetuate culturally and cognitively biased language resources.

Another example of an NLP resource that was created by translation is the *FLORES* dataset, a multilingual resource that was compiled to enable benchmarking in low-resource and multilingual machine translation (Goyal et al. 2022). The dataset consists of 3001 sentences translated into – at the time of writing – 200 languages, but this number is continuously expanding. Afikaans, Sepedi, SiSwati, Xitsonga, Setswana, isiXhosa and isiZulu are already included in the FLORES set. A translation into Sesotho will be released soon, followed by isiNdebele and Tshivenda.

The FLORES sentences were selected from 842 English Wikipedia articles. Articles were chosen to represent a wide range of topics including crime, disasters, entertainment, geography, health, nature, politics, science, sports, and travel. One might argue that more translations should be generated by expanding the existing list of articles with more articles on the same or similar topics. However, topics were assigned manually to the sentences in the FLORES collection which means that expanding on or repeating the process is not possible. Moreover, further inspection of the FLORES dataset revealed that the source domains, *Wikinews*, *Wikijunior* and *WikiVoyage*, contain many articles that may not always be relevant in the African context and which would probably give rise to the same lexicalistion issues that were encountered during the African Wordnet project.

The *No Language Left Behind Seed Data* (NLLB-Seed) was created in a similar manner as the FLORES data and comprises a set of professionallytranslated Wikipedia sentences (NLLB Team et al. 2022). However, in contrast to the FLORES data, the NLLB-Seed sentences were taken from Wikimedia's so-called "list of articles every Wikipedia should have". The list includes topics in different fields of knowledge and human activity, similar to those in the SILCAWL. The NLLB-Seed data currently consists of around six thousand sentences in 39 languages. Although the list could be a good point of departure for an extensive translation effort, it is not currently foreseen that the list will expand over time. In contrast, the topic modelling method proposed here could generate any number of sentences and could even be repeated if required.

The remainder of this paper is structured as follows. Section 3 provides an overview of topic modelling and Section 4 explains how models are evaluated. Section 5 describes how data was selected to create a Wikipedia topic model related to the SILCAWL. The results of the investigation are presented in Section 6, followed by a discussion in Section 7.

3 Topic modelling

Topic modelling is a text mining technique used to uncover latent topics in large collections of documents. Most topic models are unsupervised, thus making them powerful tools in addressing real-world problems as data is typically unlabelled in practice. Topic models have found use in many applications, including sentiment analysis (Rana et al. 2016), text classification, document categorisation, summarisation, etc. (Boyd-Graber et al. 2017).

Latent Dirichlet Allocation (LDA) is one of the most popular topic models (Blei et al. 2003). It is a three-level hierarchical Bayesian model, in which each document is modelled as a finite mixture over a set of latent topics, where a topic is defined as a distribution over the words of the vocabulary. LDA is a generative probabilistic model as it models the assumption that each document is formed through the following generative process:

1. Assuming that the corpus contains K topics, randomly choose K topic distributions, ϕ_k.
2. For each of the M documents in the corpus, randomly select a distribution over topics, θ_m.
3. Assume that each document contains n_m words. For each word, w_{mn}, in the m^{th} document.
    a. Randomly choose a topic, from the distribution over topics from Step 2, where z_{mn} denotes an indicator of the selected topic.
    b. Randomly choose a word from the selected topic based on its distribution from Step 1.

This generative process can be summarised graphically as in Figure 1.

LDA can be easily applied to any dataset using the gensim[3] package in Python. The main outputs of interest are typically the topics and topic distributions of each document. Gensim also has the capability to identify topic distributions in unseen documents, which is a feature that is useful for classifying new documents.

4 Model evaluation

By virtue of being unsupervised, evaluating topic models is a challenging task. Although some authors use perplexity[4] or held-out likelihood as measures of topic model performance, it has been shown that such intrinsic measures of model performance do not correlate with human understanding of topics (Chang et al. 2009). In other words, models with better perplexity scores often produce less humanly interpretable topics, which defeats the exploratory goals of topic modelling. In light of this, coherence measures have become more popular. In this research, we make use of the well-known UMass coherence measure as it has been shown to align well with human evaluations of coherence (Mimno et al. 2011). Considering the top N words in a topic, the UMass coherence for a topic is calculated in gensim as in Equation 1.

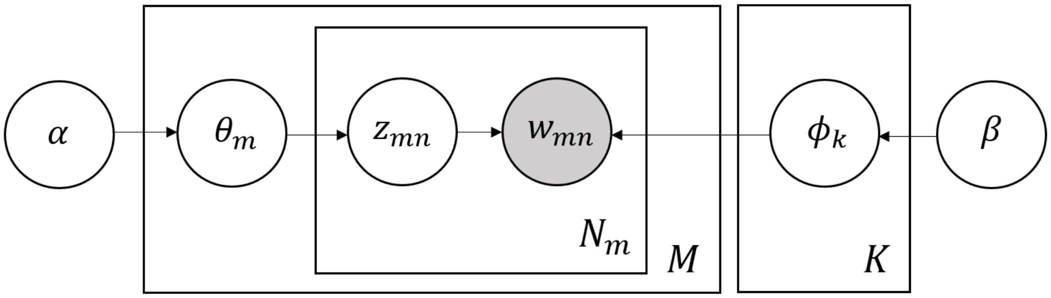

Figure 1: Graphical model for LDA. The shaded circles are used to represent observed variables, whilst unshaded circles represent unobserved variables. The arrows represent dependencies between variables and rectangles indicate replicated structures. α and β denote hyperparameters.

$$C_{\text{UMass}} = \frac{2}{N \cdot (N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \frac{P\left(w_i, w_j\right) + \varepsilon}{P\left(w_j\right)} \tag{1}$$

$P\left(w_j\right)$ denotes the probability of word w_j occurring in the training set of documents. $P\left(w_i, w_j\right)$ denotes the probability of words w_i and w_j occurring together in the training set and ε denotes a smoothing parameter. The overall coherence for a topic model is then taken as the average of all the topic coherences. In general, a higher coherence score is desired as it indicates better topic coherence.

5 Document selection process

The objective of this research is to create a collection of sentences that can be used for NLP research in under-resourced languages, in a manner that is both methodological and systematic. To this end, we propose using topic modelling as a selection method. This section provides the details of the selection process.

5.1 Topic model training

Given the aim of our research, we sought to create a data set that is generally relevant and not domain specific by selecting sentences from Wikipedia articles. To this end, we first created a training set, based on articles from various general topics. These general topics were taken from the SILCAWL comparative word list introduced in Section 2. The English words in the list were used as search terms in Wikipedia and the associated articles were retrieved from the web. After cleaning the data and removing any potentially offensive words, the resulting corpus contained 1 698 documents with an average of 1 132 words per document.

A topic model was then trained on this data using gensim. The hyperparameters were learned from the data, but the model was run for different numbers of topics, $K = \{100, 200, ..., 1000\}$. The average coherence score for each model is shown in Figure 2. It can be seen that the coherence does not increase significantly beyond $K = 800$, thus $K = 800$ was chosen to be the ideal number of topics for the training data.

Table 1 shows the top 25 topics from $K = 800$ arranged in order of decreasing coherence. It is evident from the table that the topic model was able to pick up various topics including, for example, poverty (topic 98), health (topic 224) and time (topic 551). Note that the choice of overall topic names are determined by the user and thus are subjective. The full set of topics can be found at https://github.com/jrmazarura/wikipedia-sentence-selection.

Table 1: Top 25 topics from training set.

Topic Number	Coherence score	Topic words
780	-0.521	specie animal mammal group year early include large ago small
686	-0.565	millipede centipede specie order leg pair segment group body large
792	-0.581	bamboo culm flower shoot specie plant seed year grow time
162	-0.613	scorpion tarantula specie spider female leg male include large pressure
98	-0.661	poverty poor live income child people world population day social
164	-0.663	world theory concept term thing sense view understand form refer
628	-0.672	emotion theory experience individual emotional person social action behavior people
748	-0.675	band tour album rock record music member year group play
636	-0.694	dew form water surface small find night device include leave
155	-0.695	faisal saudi saud king country establish religious issue family royal
670	-0.713	leisure boy time activity sport family include male work increase
541	-0.726	duck tribe subfamily make family include breed water generally word
218	-0.727	cave form rock water include early find sea large deep
309	-0.736	giraffe neck long animal protein depend find high form percent
574	-0.741	chameleon catfish specie family tongue large sound signal include body
18	-0.744	kiln hut heat fire dry temperature wood build design type
771	-0.751	lice louse host body head bird specie egg live feed
551	-0.754	date year calendar start ad name reference early event begin
315	-0.755	boar specie wild butterfly area male year population large plant
704	-0.759	marshe water marsh plant pool form high occur habitat type
306	-0.770	perfection fly perfect number concept man art century body great
463	-0.771	material cut blade metal shape wood tool design type small
185	-0.782	pregnancy miscarriage week woman risk fetus birth increase age term
224	-0.785	health disease include medical condition care treatment factor study risk
334	-0.787	cattle zebu breed animal milk african originate carry region period

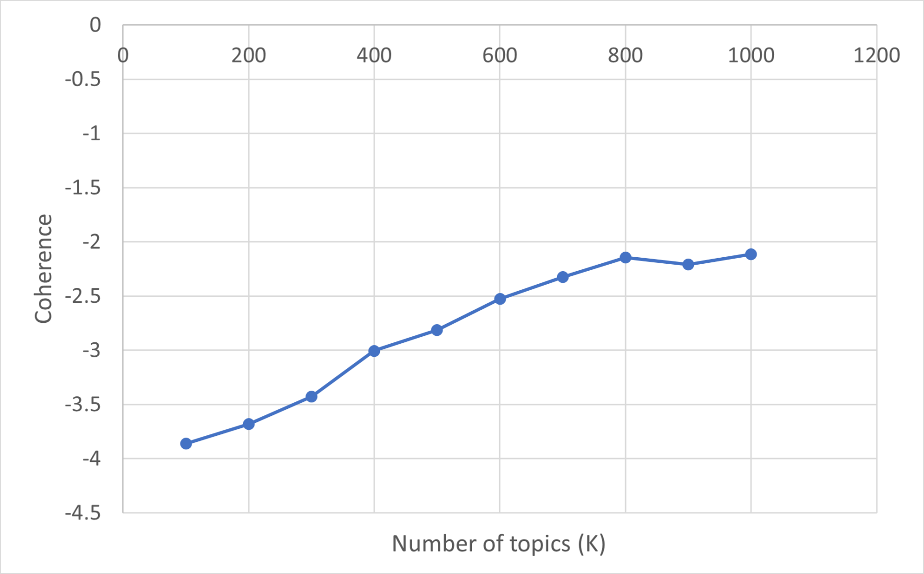

Figure 2: Average coherence score for $K = \{100, 200, ..., 1000\}$.

5.2 Document selection using trained topic model

The next step in the document selection process was to use the trained topic model to select articles from Wikipedia and then select sentences from them. For each article in Wikipedia, the model was used to determine the topic distribution and only articles satisfying the following conditions were retained:

1. The dominant topic of the article had to be amongst the 500 topics with the highest coherence scores from the trained model.
2. The dominant topic needed to make up at least a quarter of the article's topic distribution.

The 500 topic cutoff used in #1 was selected by assessing the coherence scores. Based on inspection, the interpretability of the topics appeared to noticeably deteriorate as the coherence scores decreased below -2. The first 500 topics had coherence scores above this value. In addition, the 500-topic cutoff also ensured that a broad variety of topics were covered. Choosing The 25% cutoff used in #2 not only ensured that the topic made up a significant proportion of the article's topic, but it was also necessary to ensure that only *some* of the articles from Wikipedia were selected.

After this step, the final collection of sentences was compiled by randomly selecting an equal number of articles from each topic and then randomly selecting one sentence from each article. For example, to create a collection of 5 000 sentences, 10 articles belonging to each of the 500 topics were randomly selected and one sentence was randomly selected from each of the articles. Only sentences with at least 15 words made up of 3 characters or more were selected. This was done to avoid retrieving sentences that were very short and did not contain much information.

6 Results

The final selection of sentences can be found at https://github.com/jrmazarura/wikipedia-sentence-selection. Table 2 shows a snippet of the sentences selected by the proposed method. From the table, it is evident that many topics do not appear to be related to the South African or even African context as desired. In selecting the articles with dominant topics belonging to our top 500 topics, it can be seen that many of the sentences are about seemingly unrelated topics.

For example, consider Topic 218 in Table 1. This topic appears to be related to ocean caves. The trained topic model was used to retrieve related articles from Wikipedia and the titles of the articles were recorded (the list of titles is also available on GitHub). Some of the Wikipedia articles retrieved have titles such as *Cave*, *Stalactite* and *Stalagmite*. However, there are much more which seem less important in our context, such as *Velebit caves* (caves in Croatia), *Abrakurrie Cave* (caves in Australia) and *Ogof Agen Allwedd* (caves in Wales) amongst many others. The topic model was able to select 482 cave-related articles, but unfortunately, most of them are about specific caves from all over the world. Ultimately, this means that sampling random sentences from these randomly selected articles will most likely produce sentences that are unrelated to South Africa. The following are examples of some of the cave-related sentences that were selected:

"The Bull Thistle Cave Archaeological Site is an archaeological site on the National Register of Historic Places, located in Tazewell County, Virginia."

"The Caves of Hotton are speleothem caves located in Wallonia near Hotton in Belgium, which were discovered in 1958 and are around 5 or 6 km long and 70 metres deep."

This observation is likely due to the topic model trying to find topic distributions for over five million articles over only 800 topics. It is reasonable to assume that such a large corpus is likely to contain much more than 800 topics. In addition, topics such as the cave topic that was considered as an example, are likely to have many related subtopics, e.g. *African caves* or *ice caves*, etc. If there is only one topic related to caves, then all articles will be grouped into one topic creating a collection of many cave-related articles covering a broad scope.

7 Discussion & future work

Our first attempt at using topic modelling to identify Wikipedia articles that would be suitable for translation from English into the 10 other South African languages yielded an extremely diverse set of sentences, many of which contain words that are probably not lexicalised in any of the target languages. Training the topic models on the SIL-CAWL was clearly not adequate to avoid this challenge.

In the next phase of the research we will aim to find pruning criteria that could be used to "focus" the topic models on content that is more relevant to the South African context. We would also like to establish a relationship between the selected text and the resulting translation in the target language to ensure that both relevant and translatable text has been chosen. Such a measure will also allow a comparison between the proposed topic model approach to text selection and a fully random baseline.

Notes

[1] `https://repo.sadilar.org/`

[2] In this paper *text generation* only refers to text produced by humans. Automatic text generation will not be considered because the technique itself relies on the availability of text data, a resource that is not available in the circumstances relevant to this study.

[3] `https://radimrehurek.com/gensim/auto_examples/index.html`

[4] Perplexitiy measures a model's ability to predict new data. A low perplexity score is assumed to indicate good model performance.

Table 2: Selected sentences from Wikipedia

	Selected sentences
1.	The plateau originated in the Gondwanan breakup and is one of the five major submerged parts of Zealandia, a largely submerged continent.
2.	As such, it has become the site of the small Tabor Mountain Ski Resort, which is one of Prince George's two local ski hills, the other being the small Hart Highlands Alpine Park on the north side of the city.
3.	The stratovolcano lies above the regional Liquine-Ofqui Fault zone, and the ice-covered massif towers over the south portion of Pumalín Park.
4.	The sedimentary rock was more fragile than the metamorphic rock formed by the contact of the magma and the surrounding sedimentary rock.
5.	Examination by a severe weather team from the Bureau of Meteorology examined the damage in the Bucca and Kolan region and recorded it as an 'F4' on the Fujita scale.
6.	Formed in 1922 as the Westby Ski Club, the all-volunteer club held the first ski jumping tournament southeast of Westby, near Bloomingdale, Wisconsin in 1923.
7.	It is often considered to be one of the most spoiled of the Munros, due to the Glenshee Ski Centre which covers the eastern slope of the mountain.
8.	The de Lalande crater is named after the French astronomer Marie-Jeanne de Lalande (1768-1832), illegitimate daughter of astronomer Joseph Jerome de Lalande (1732-1807).
9.	The hill is 522 metres (1712 feet) high and is the highest point of the relatively low-lying county of Renfrewshire and indeed the entire Clyde Muirshiel Regional Park of which it is a part, having a considerable Topographic isolation.
10.	Additionally, a volcano was hypothesized to exist in the Nova Iguaçu area, in Rio de Janeiro, and was called the Nova Iguaçu Volcano.
11.	There is no precise definition of surrounding base, but Denali, Mount Kilimanjaro and Nanga Parbat are possible candidates for the tallest mountain on land by this measure.
12.	The Dease Plateau is a sub-plateau of the larger Yukon Plateau, and is located in far northern British Columbia, Canada, northwest from the Deadwood River to and beyond the Yukon-British Columbia boundary.
13.	The ruins on its top derive from the year 1449, when Oberhohenberg Castle together with the town of Hohenberg, at the foot of the mountain, were destroyed in a local feud.
14.	Enoggera Hill is a small mountain of the Taylor Range in Australia in the Brisbane suburb of Enoggera, in Queensland, with a peak of 273 meters (896 feet) above sea level.
15.	SnoCountry Mountain Reports was the first and is now the largest snow conditions reporting service in the world.

Acknowledgements

The localisation of the Mozilla Common Voice platform in South Africa is supported by the German Federal Ministry of Economic Cooperation and Development (BMZ), represented by the GIZ project FAIR Forward - Artificial Intelligence for All.

References

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent dirichlet allocation', *Journal of machine Learning research* **3**(Jan), 993–1022.

Boyd-Graber, J., Hu, Y., Mimno, D. et al. (2017), 'Applications of topic models', *Foundations and Trends® in Information Retrieval* **11**(2-3), 143–296.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. & Blei, D. (2009), 'Reading tea leaves: How humans interpret topic models', *Advances in neural information processing systems* **22**.

Eiselen, R. & Puttkammer, M. (2014), Developing text resources for ten South African languages, *in* 'Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)', European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 3698–3703.
URL: *http://www.lrec-conf.org/proceedings/lrec2014/pdf/1151_Paper.pdf*

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzman, F. & Fan, A. (2022), 'The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation', *Transactions of the Association for Computational Linguistics* **10**, 522–538.

Griesel, M. & Bosch, S. (2020), Navigating challenges of multilingual resource development for under-resourced languages: The case of the African wordnet project, *in* 'Proceedings of the first workshop on Resources for African Indigenous Languages', pp. 45–50.

Mimno, D., Wallach, H., Talley, E., Leenders, M. & McCallum, A. (2011), Optimizing semantic coherence in topic models, *in* 'Proceedings of the 2011 conference on empirical methods in natural language processing', pp. 262–272.

NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H. & Wang, J. (2022), 'No language left behind: Scaling human-centered machine translation'.

Pretorius, L. & Wolff, F. (2020), *Wikipedia as a Transformative Multilingual Knowledge Resource*, Cambridge University Press, p. 232–246.

Rana, T. A., Cheah, Y.-N. & Letchmunan, S. (2016), 'Topic modeling in sentiment analysis: A systematic review.', *Journal of ICT Research & Applications* **10**(1).

Snider, K. & Roberts, J. (2004), 'SIL comparative African wordlist (SILCAWL)', *Journal of West African Languages* **31**(2), 73–122.