Part of Speech Tagging in Luyia: A Bantu Macrolanguage

Kenneth Steimel
Indiana University
Bloomington, Indiana
ksteimel@iu.edu

Abstract

Luyia is a macrolanguage in central Kenya. The Luyia languages, like other Bantu languages, have a complex morphological system. This system can be leveraged to aid in part of speech tagging. Bag-of-characters taggers trained on a source Luyia language can be applied directly to another Luyia language with some degree of success. In addition, mixing data from the target language with data from the source language does produce more accurate predictive models compared to models trained on just the target language data when the training set size is small. However, for both of these tagging tasks, models involving the more distantly related language, Tiriki, are better at predicting part of speech tags for Wanga data. The models incorporating Bukusu data are not as successful despite the closer relationship between Bukusu and Wanga. Overlapping vocabulary between the Wanga and Tiriki corpora as well as a bias towards open class words help Tiriki outperform Bukusu.

1 Introduction

Luyia is a macrolanguage comprised of over 20 individual languages that form a dialect continuum. These languages have a very high degree of cognates. This provides a unique opportunity to examine the performance of Natural Language Processing tools on a cluster of very closely related Bantu languages. The structure of this paper is as follows: first, I provide some background on the Luyia languages, then, I evaluate the relative performance of different languages on two types of predictive part of speech tagging tasks. The first task involves direct training of an SVM part of speech tagging model on one language and then evaluation of this model on another language. The second task involves augmenting an SVM part of speech tagging model by training on a mixture of data from the target language and another variety of Luyia.

2 The Luyia languages

Oluluyia or Luyia[1] is a macro-language spoken in Kenya by approximately 5.3 million people (Simons and Fennig, 2017). The varieties of Luyia belong to JE30 and JE40 in the revised Guthrie Bantu classification scheme (Maho, 2009, 61-62). This macrolanguage consists of over 20 sub groupings. As a macrolanguage, Luyia consists of a number of linguistic groupings that are individual languages but that are treated as a single language in some contexts (Simons and Fennig, 2017). Older works like (Williams, 1973) refer to the linguistic groups that compose Luyia as dialects instead of unique languages. However, newer works on Luyia like Ebarb (2014) refer to Luyia as a dialect continuum with "geographically close varieties enjoy[ing] higher rates of mutual intelligibility" Ebarb (2014, 7). The map shown in figure 1 displays the geographic distribution of the Luyia languages.

The generalization that geographically closer languages have more in common is largely supported by the cognate probability table shown in 1. The Luyia languages are italicized in the table. The non-Luyia languages (Soga, Ganda, and Gusii) are closely related to the Luyia languages but are outside of

[1] Also spelled Luhya

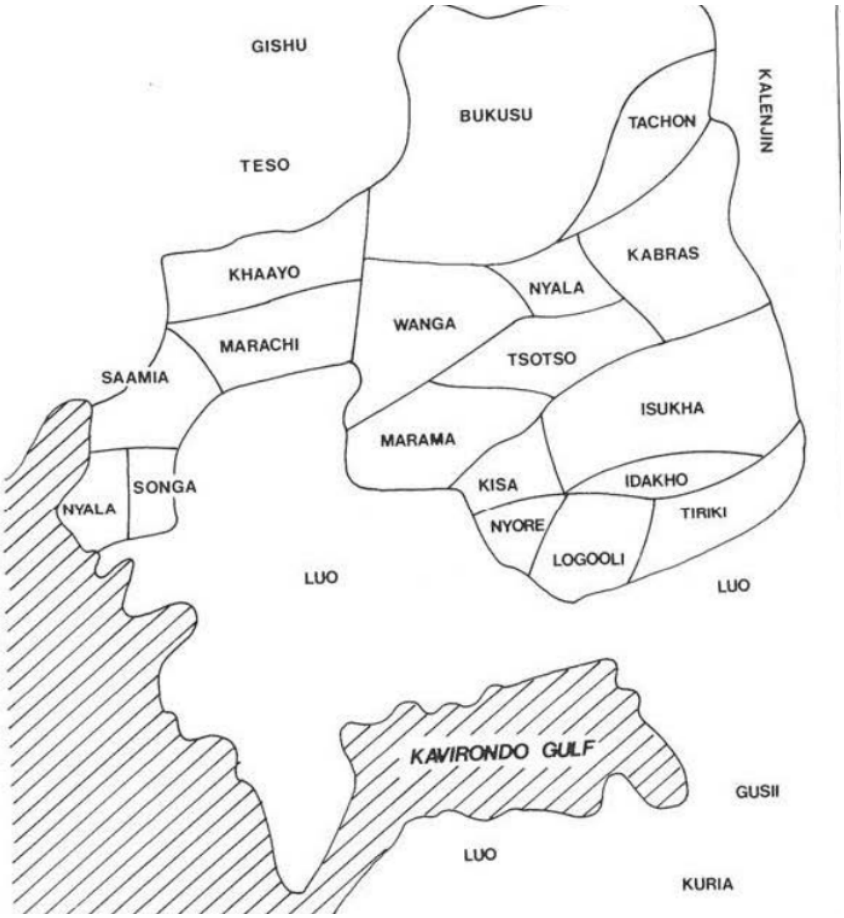

Figure 1: A map of the Luyia languages (Hinnebusch et al., 1981)

	Ganda	Soga	Saamia	Wanga	Bukusu	Idakho	Logooli
Soga	83.0						
Saamia	54.0	59.0					
Wanga	54.0	59.0	88.0				
Bukusu	52.0	60.0	77.0	81.0			
Idakho[3]	48.0	51.0	71.0	71.0	68.0		
Logooli	51.0	55.0	75.0	78.0	70.0	80.0	
Gusii	40.0	42.5	48.5	47.5	42.5	44.0	47.0

Table 1: Cognate percentages using 200-word lists (Hinnebusch et al., 1981, 184)

the macro-language[2]. The language pairs with cognate probabilities higher than 75% are geographically adjacent. This threshold of 75% is estimated to indicate sufficient mutual intelligibility to consider languages sharing such percentages to be dialects of a single language (Ladefoged et al., 1972) as cited in (Hinnebusch et al., 1981, 184).

The languages being examined in this study are Wanga (lwg), Bukusu (bxk), and Tiriki (ida). Bukusu is classified as JE31c while Wanga is JE32a in the New Guthrie Classification (Maho, 2009, 61). However, both belong to the Masaba-Luhya group (Maho, 2009). Tiriki (JE413), is part of a more distant branch of Luhya, the Logooli-Kuria group (Maho, 2009, 62).

3 Related Work

Most previous approaches to Bantu part of speech tagging have been based upon two-level finite state morphologies (Hurskainen, 1992; Pretorius and Bosch, 2009). However, these methods are expensive in terms of person hours and monetary cost because they rely on highly trained grammar writers. While Pretorius and Bosch (2009) use an existing morphology to expedite morphology creation in a related language, there are other disadvantages to finite-state morphologies for POS labels; systems like these are also difficult to expand. Most significantly for this particular setting, these require fairly extensive knowledge about the language being modeled. As documentation of the Luyia languages is still in progress, development of a finite state grammar with sufficient coverage does not seem to be feasible at this time. An alternative course is to develop a corpus and then use probabilistic methods to do the analysis.

Corpus creation is also a laborious process, however, there have been attempts to mitigate the expense. For example, (Yarowsky and Ngai, 2001) used cross-linguistic projection to apply existing resources for English to bilingual text corpora and project the analysis of the English text onto the second language using statistically derived word alignments. While this method is successful, what if reliable parallel texts are not available?

Hana et al. (2004) uses a morphological analyzer for their target language (Russian) and then used transition probabilities from a related language (Czech) using a Hidden Markov Model (HMM). This approach is similar in that I am using information from a related language to bolster taggers without doing projection. However, I propose the possibility of using robust machine learning techniques, rather than relying on an expensive morphological analyzer. Morphologically rich languages can have a large number of morphological variants for a single word resulting in high rates of Out of Vocabulary (OOV) terms. Statistical taggers perform worse on tokens that are not found in the training data. A common way to make taggers more robust to OOV terms is to use components of words as features during training rather than whole words. Such methods include using induced and hand labeled suffixes as features (Hasan and Ng, 2009), initial and final characters of words combined with the tags of surrounding words (Màrquez et al., 2000) and morphologically segmented words (Tachbelie et al., 2011). However, unlike these approaches, I use a simple character n-gram model instead of building a segmenter. This is similar

[2]These languages are not the subject of this paper. They are only included as a point of reference for these probabilities

[3]While Hinnebusch et al. (1981) does not include figures for Tiriki, Idakho is a good reference for what the expected values would be for Tiriki. Both languages share the same ISO-639-3 code (ida) and are considered by some to be part of a single Idakho-Isukha-Tiriki variety (Simons and Fennig, 2017).

Language	Number of Sentences	Number of Words
Wanga	1,294	7,337
Bukusu	1,107	7,940
Tiriki	1,393	9,319

Table 2: Total corpus size for the Luyia corpora

to the MaxEnt tagger in Gambäck et al. (2009) but with all possible character n-grams rather than n-grams anchored to the beginning or end of the word and with fewer contextual features. This is not unfounded as De Pauw et al. (2012) found that contextual information was not necessary to achieve good tagging accuracy for machine learning in Bantu languages.

In this light I, like Tachbelie et al. (2011), am interested in how much data is required to obtain satisfactory prediction accuracies (85%). I experiment with using a corpus from one language directly to tag the target language as well as mixing data from both a source and target language to tag the target language.

4 Corpora

The Luyia corpora used for this analysis are extracted from a Fieldworks database (International, 2018) manually annotated by language documentation researchers (Green et al., 2018). For this analysis, the corpus was filtered to include only sentences where all words were labeled with POS tags. The size of the filtered Luyia corpora are displayed in table 4. All three corpora are roughly the same size with Tiriki being slightly larger than the other two by about 1,000 words.

The Swahili corpus used for comparison with the Luyia corpora differs significantly in terms of annotation methods, composition and size. The Helsinki Corpus of Swahili is a large, automatically annotated corpus using a two-level finite state morphology. This corpus consists of parliamentary proceedings and newspaper articles in contrast to the personal narratives that comprise the Wanga corpus. In addition, the full Helsinki Corpus consists of 25 million words and is much larger than the Wanga corpus. Due to my limited access to the Helsinki Corpus, in this study I use a much smaller sampling taken from 70 files consisting of 1,000 sentences each. These files were collected by searching for the top 10 most common words in the corpus[4] and collecting a number of JSON files[5] for each using the web interface to the annotated corpus.

5 Methods

The following sections describe methods for tagging Wanga data using a modified version of the Helsinki Corpus of Swahili tagset (Hurskainen and Department of World Cultures, University of Helsinki, 2016)[6]. All experiments used Wanga as the target language. The test set was kept the same for all experiments. The training dataset was slightly over half of the filtered Wanga corpus at 591 sentences and 3274 words. In this study, I seek to determine how well one variety of Luyia can be used to tag another variety of Luyia directly and how effectively one variety can augment a small training corpus from the target language. In essence, how much data in the target language is required? The first set of experiments uses no data from the target language while the other experiments use various amounts of target language data mixed with data from another variety of Luyia. Table 3 displays the size of the datasets at different training splits. A Support Vector Machine (SVM) is a geometric supervised learning method that takes labeled data and

[4]Determined using the unannotated version of the Helsinki corpus which can be downloaded in its entirety.

[5]The number of JSON files downloaded for each seed word differed based upon how independent the words are. For example, one of the most common words was Mhesimiwa which means 'honorable', a title that precedes many names in the parliamentary proceedings portion of the corpus. The files from this search term contain a sampling of the corpus heavily biased towards proper nouns. A small number of files were used for **Mhesimiwa** while a larger number were used for more grammatically neutral terms like **na**, meaning 'and'.

[6]All punctuation was reduced to a single 'PUNCT' tag instead of having separate parts of speech for each punctuation mark

Training Proportion	Wanga Words	Tiriki Words	Bukusu Words	Swahili Words	Sentences
0.05	378	379	545	1404	59
0.1	758	785	1054	2973	118
0.2	1431	1648	2057	5779	236
0.5	3686	4247	4800	14914	590

Table 3: Number of words and sentences for various training dataset sizes

learns a hyperplane to separate one class from another [7].

The features extracted for each word were the following:

- Character unigrams, bigrams, trigrams and 4-grams internal to the word
- The first 3 characters of the preceeding word or the whole word if less than 3 characters [8]
- The first 3 characters of the following word or the whole word if less than 3 characters [9]

The set of all attested features were collected from the training data for each experiment and then filled in the counts of each feature for each word in a kind of "count-hot" encoding. Only features found in the training set were used to extract features from the test set. Two sets of SVM models were employed in this study. First a Luyia language by itself is used to train a model and then evaluate on the test portion of the Wanga dataset. Next, a combination of the source Luyia language and the Wanga training set are used to train a model and then evaluate on that same Wanga test dataset.

5.1 Direct Source to Wanga

For this set of experiments features were extracted and then those features were used to train on the different training dataset sizes described in table 3 The model created was then used to predict part of speech labels for the Wanga training set.

5.2 Wanga and Source Mixture to Wanga

After establishing the baseline of using Wanga character n-grams alone to tag Wanga, the Wanga data was combined in a 2 to 1 ratio with the source language data. Thus, four training sets were created which were composed of 5, 10, 20 and 50% of the Wanga Corpus. Half the number of sentences used in the Wanga Corpus were extracted from the source language corpus and mixed in with the Wanga data. For example, with the Wanga dataset containing 59 sentences, 30 randomly sampled Tiriki sentences were mixed in. This was done because I wanted to ensure that the Wanga data was not overwhelmed by the source language data[10]. The features discussed in section 5 were extracted from these combined training sets and the SVM models were trained. The resulting models were used to tag the test set.

6 Results

First, I discuss results for training a model one of the source languages and then directly applying this model to predict parts of speech for the Wanga test data. Then, language mixture models are discussed.

[7] The particular implementation used for this work (Pedregosa et al., 2011) employed 1 versus 1 for multiclass-classification. One versus rest classification is also common

[8] These character sequences were appended to "BEG_" making these features distinct. The three initial characters of the preceding word are used because Bantu languages (including Luyia) make heavy use of prefixes for inflectional markers. By using initial characters, I aim to make use of these inflectional markers without adding significantly to the dimensionality of the training vectors.

[9] These character sequences were appended to "AFT_" making these features distinct.

[10] However, some trials with larger and smaller mixing ratios were also conducted. These yielded worse results.

Training Set Size	Swahili	Tiriki	Bukusu	Wanga
0.05	26.91	**57.76**	54.51	76.36
0.1	27.46	**60.84**	58.67	81.89
0.2	29.61	**62.47**	56.22	86.10
0.5	25.56	**63.07**	62.49	91.06

Table 4: Accuracy of training on one language

Training Set Size	Swahili	Tiriki	Bukusu	Wanga
0.05	76.45	**78.40**	**77.30**	76.36
0.1	80.58	**82.66**	79.94	81.89
0.2	86.76	86.49	85.88	86.10
0.5	90.34	90.04	90.20	91.06

Table 5: Accuracy of language mixture models

6.1 Direct Source to Target Tagging

Table 4 displays the accuracies obtained by applying the models trained on data from one language to label parts of speech for the Wanga test set. Swahili is provided as a baseline while Wanga itself is provided as an upper bound. The training set size proportion is with reference to the Wanga corpus.

Surprisingly, Tiriki, the more distantly related variety, is more effective at tagging Wanga than Bukusu. The accuracy of the taggers trained on Tiriki are higher across the board. In addition, the accuracy of the tagger trained on Tiriki consistently rises as the amount of training data increases. For training on Bukusu, the tagger benefits from having access to more data overall. However, the drop at 0.2 may indicate that the machine learner is overfitting on Bukusu data and is not able to generalize to predicting on the Wanga test data.

Though the tagger trained on Tiriki performs much worse than the upper bound, the highest dataset size for Tiriki is approaching the accuracy obtained for the smallest Wanga training set. A new corpus for a Luyia language could get preliminary POS tags by training on a large portion of another Luyia language. Future research will have to investigate if the trends observed here, where the more distantly related of two varieties is most effective, generalizes to other source-target pairs.

6.2 Language Mixture Tagging

Now we turn to results obtained by training on a mixture of the Wanga training dataset and the source language dataset. Table 5 displays the accuracies of predictions using combination models of different sizes. The Wanga values listed are the prediction accuracies from training on Wanga alone. They repeat the Wanga data in table 4. The entries in bold represent cases where the accuracy was considerably higher than the accuracy of training on Wanga alone.

Once again, Tiriki performs better than Bukusu despite the fact that Bukusu is more closely related to Wanga. The combination model that uses Bukusu is only considerably higher than the baseline for the lowest training set size. In addition, the combination model that incorporates Tiriki data outperforms the baseline by a wider margin in this case. The Tiriki combination model also performs much better than the baseline for the 0.1 training set size.

Overall, augmenting Wanga taggers with data from another variety of Luyia is beneficial at smaller training set sizes. However, this augmentation has no effect or even a slightly negative effect on performance for larger training set sizes.

7 Analysis

Tiriki outperforming Bukusu on the combination tagging task is somewhat unsurprising: if Bukusu and Wanga are extremely similar varieties, the Bukusu data may contribute nothing new. However, if Bukusu and Wanga are so similar, Bukusu would be expected to perform better on the direct tagging task. The fact that Tiriki is still performing better on this task is very surprising.

Training Set Size	Swahili	Tiriki	Bukusu	Wanga
0.05	25.33	47.75	41.62	58.17
0.1	25.85	49.77	46.54	64.24
0.2	28.47	51.31	43.21	68.12
0.5	22.74	49.75	51.08	72.68

Table 6: Out of Vocabulary accuracy for source-Wanga

Training Set Size	Swahili	Tiriki	Bukusu	Wanga
0.05	58.84	63.47	62.38	58.17
0.1	63.31	66.51	62.60	64.24
0.2	70.13	70.66	70.78	68.12
0.5	70.73	74.66	72.72	72.68

Table 7: Out of Vocabulary accuracy for language mixture tagger

Performance on out of vocabulary (OOV) terms is higher for Tiriki than Bukusu. In tables 6 and 7 the accuracy of the Tiriki tagger is higher than or approximately the same as the Bukusu accuracies[11].

One possibility is that the Wanga and Tiriki corpora themselves have more in common, even if the varieties themselves do not. This does seem to be the case. Table 8 displays the percentage of words in each of the source corpora that are within the specified Levenshtein distance of a word in the Wanga corpus.

The Tiriki corpus has a three percent higher rate of absolute matches compared to Bukusu. While Bukusu has a higher percentage of words within a distance of 2, the large number of exact matches between Tiriki and Wanga is likely to blame for the increased performance of Tiriki relative to Bukusu[12].

Upon further analysis of the words that Tiriki gets correct but Bukusu does not (using the 0.05 training set size), it appears that the Tiriki classifier has a beneficial bias towards nouns and verbs. The Bukusu tagger tends to incorrectly predict closed classes like pronouns and conjunctions when the Tiriki classifier correctly predicted noun and verb tags. Most mistaken conjunction tags in Bukusu are for short words: all of these mistaken conjunctions and pronouns are less than the median word length of 9. The Tiriki classifier is more heavily biased towards predicting open class parts of speech than the Bukusu classifier.

Some of the Tiriki corpus's advantages originate from sources of ambiguity. Infinitive verbs like **okhushina** share some aspects of both verbs and nouns: they are part of the noun class system (class 15) and can trigger agreement with nominal modifiers like demonstratives, numerals and adjectives. However, they are semantically verbal and can take verbal morphemes like passive markers and causative markers. The Bukusu tagger consistently labels these as verbs while the Tiriki tagger tends to label them as nouns. The Wanga corpus uses the noun label more often for these infinitives resulting in higher performance for Tiriki.

Only a small handful of the cases where Tiriki outperforms Wanga appear to be due to errors in the gold standard labels. However, a more thorough investigation of the errors should be conducted using judgements from a native speaker of Wanga.

[11] An exception to this is in the application of the largest non-mixed models. Bukusu outperforms Tiriki by 1.33% for this scenario

[12] Though how this should be interpreted in light of the higher out of vocabulary scores for Tiriki is still unclear

Levenshtein Distance Threshold	Tiriki	Bukusu
0	31.82%	28.46%
1	55.67%	57.15%
2	78.89%	78.19%

Table 8: Percentage of source corpus within specified distance of any word in Wanga corpus

One last source of Tiriki's higher performance relative to Bukusu is due to borrowed words. The words that Tiriki predicts correctly but Bukusu misses are frequently loan words from English or Swahili. Words like **i-proverb** and **i-nursery** are borrowed from English with the addition of an augment prefix from Wanga. Swahili borrowed words are pervasive including coordinating conjunctions like **lakini**, meaning 'but' and numerals like **saba**[13]. The character n-gram model used in this work can rely on morphological clues to determine parts of speech for Luyia words. However, for borrowed words, these morphological queues are likely not strong enough. The Tiriki tagger's training data likely had tokens similar to or identical to these borrowed words.

8 Future Work

I intend to implement this research for the other two other pairings given the three Luyia languages for which data is available at the moment. The findings discussed herein are for using Tiriki and Bukusu to tag Wanga. What differences emerge when using Bukusu and Wanga to tag Tiriki or Wanga and Tiriki to tag Bukusu? In addition, certain parts of speech benefit more with the mixed language models described in section 5.2. What kind of accuracy can be obtained by combining finite-state and machine learning approaches using machine learning for open class parts of speech and

9 Conclusion

Wanga, a Luyia language can be tagged more effectively by using data from other varieties of Luyia. However, this effectiveness has a few stipulations. The variety that is effective is not necessarily the variety that is closest to Wanga. In these experiments, Tiriki, the more distantly related language fared better than Bukusu in both the direct tagging tasks and the combination tasks. A number of factors led to this advantage including vocabulary overlap between the Wanga and Tiriki corpora and a bias towards predicting open class parts of speech using the Tiriki tagger. In addition, while the introduction of data in the combination models can help, this effect is limited to cases where the training set size is small. As there are no annotated corpora at all for the 20 Luyia languages not used in this document, these findings could be used to create annotated corpora. This could be done either by using all available annotated data from another Luyia variety11 to annotate the new variety or by annotating a very small corpus from the new variety and then creating a mixture model.

10 Acknowledgements

I would like to thank the organizers of VarDial 2018 for their work in creating a welcoming venue to discuss this research. I would particularly like to thank Marcos Zampieri and Preslav Nakov. In addition, this work is the result of valuable feedback from anonymous reviewers. I greatly appreciate the feedback they gave. Lastly, I would like to thank Michael Marlo for engaging me with Luyia as an undergraduate and providing me with the corpora that made this research possible.

References

Guy De Pauw, Gilles-Maurice de Schryver, and Janneke van de Loo. 2012. Resource-light Bantu part-of-speech tagging. In *Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL 8-AFLAT 2012)*, pages 85–92. European Language Resources Association.

Kristopher J Ebarb. 2014. *Tone and variation in Idakho and other Luhya varieties*. Ph.D. thesis, Indiana University.

Björn Gambäck, Fredrik Olsson, Atelach Alemu Argaw, and Lars Asker. 2009. Methods for Amharic part-of-speech tagging. In *Proceedings of the First Workshop on Language Technologies for African Languages*, pages 104–111. Association for Computational Linguistics.

Christopher R. Green, Michael J. K. Diercks, and Michael R. Marlo. 2018. A grammar sketch of Wanga.

[13] Both of these terms were actually originally borrowed from Arabic

Jiri Hana, Anna Feldman, and Chris Brew. 2004. A resource-light approach to russian morphology: Tagging Russian using Czech resources. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Kazi Saidul Hasan and Vincent Ng. 2009. Weakly supervised part-of-speech tagging for morphologically-rich, resource-scarce languages. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 363–371, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thomas J Hinnebusch, Derek Nurse, and Martin Joel Mould. 1981. *Studies in the classification of Eastern Bantu languages*, volume 3. Buske.

Arvi Hurskainen and Department of World Cultures, University of Helsinki. 2016. Helsinki Corpus of Swahili 2.0 Annotated Version.

Arvi Hurskainen. 1992. A two-level computer formalism for the analysis of Bantu morphology: an application to Swahili. *Nordic Journal of African Studies*, 1(1):87–122.

SIL International. 2018. Fieldworks language explorer. `https://software.sil.org/fieldworks/`, March. Release 8.3.12.

Peter Ladefoged, Ruth Glick, and Clive Criper. 1972. *Language in Uganda*. Oxford University Press.

Jouni Filip Maho. 2009. Nugl online: The online version of the new updated Guthrie list, a referential classification of the Bantu languages. *Online file: http://goto. glocalnet. net/mahopapers/nuglonline. pdf*.

Lluís Màrquez, Lluis Padro, and Horacio Rodriguez. 2000. A machine learning approach to POS tagging. *Machine Learning*, 39(1):59–91.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Laurette Pretorius and Sonja Bosch. 2009. Exploiting cross-linguistic similarities in Zulu and Xhosa computational morphology. In *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages – AfLaT 2009*, pages 96–103, Athens, Greece, March. Association for Computational Linguistics.

Gary Simons and Charles Fennig. 2017. Ethnologue: Langauges of the world, twentieth edition.

Martha Yifiru Tachbelie, Solomon Teferra Abate, and Laurent Besacier. 2011. Part-of-speech tagging for under-resourced and morphologically rich languages–the case of Amharic. *HLTD (2011)*, pages 50–55.

Ralph Williams. 1973. A lexico-statistical look at Oluluyia. In *Fourth Annual Conference on African Linguistics*.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.