DATA-DRIVEN PART-OF-SPEECH TAGGING FOR THE GIKUYU LANGUAGE: DEVELOPMENT, CHALLENGES AND PROSPECTS

Gabriel Kamau

Department of Computer Science, Dedan Kimathi University of Technology, Nyeri Kenya

ABSTRACT

This paper presents the development of a data-driven Part-of-Speech (POS) tagger for Gikuyu, a Bantu language spoken in Kenya. Gikuyu, like many indigenous African languages, is under-resourced, with limited computational tools for linguistic processing. By employing a corpus sourced primarily from the Gikuyu Bible and leveraging a Memory-Based Tagging (MBT) approach, this study demonstrates the feasibility of creating a robust POS tagging system. The tagger achieved a precision of 90.44%, a recall of 88.34%, and an F-score of 91.35%. These results underscore its potential for applications in machine translation, speech recognition, and language preservation. The study highlights the challenges of working with under-resourced languages, including data collection and annotation, and provides recommendations for future work, including integration with broader NLP tasks.

KEYWORDS

Natural Language Processing, Part-of-Speech Tagging, Gikuyu Language, Data-Driven Approach, Low-Resource Languages

1. INTRODUCTION

Natural Language Processing (NLP) bridges the gap between human languages and computational systems, enabling applications such as machine translation, text mining, and sentiment analysis. A foundational task in NLP is Part-of-Speech (POS) tagging, which assigns grammatical labels—such as nouns, verbs, and adjectives—to words in a text. This task supports downstream applications like named entity recognition, syntax parsing, and speech synthesis Kumar et al. (2023)

The Gikuyu tribe forms the largest indigenous tribe in Kenya. According to the Kenya Population and Housing Census (2019), the Gikuyu language is spoken by approximately 8.1 million people in Kenya alone. The language however like other indigenous languages in Kenya faces challenges of digitization and language tools development. While significant advances have been made for European and Asiatic languages, African languages, including Gikuyu, remain underrepresented in NLP research. Kenya's official languages, English and Kiswahili, dominate written and spoken communication, relegating Gikuyu and other indigenous languages primarily to oral use. Consequently, the language risks extinction without research-based preservation efforts.

Table 1 shows the population distribution of the Kenyan indigenous languages based on 2019 Kenya Population and Housing Census (KPHC).

DIO:10.5121/ijnlc.2024.13602

Table 1: Kenya indigenous languages population distribution

SN	LANGUAGE	SPEAKERS
1	Gikuyu	8,148,668
2	Luhya	6,823,842
3	Kalenjin	6,358,113
4	Luo	5,066,966
5	Kamba	4,663,910
6	Kenyan Somalis	2,780,502

This work addresses the gap in computational resources for Gikuyu language by developing a data-driven POS tagging system. The objectives include collecting and annotating a corpus, building a tagging model, and evaluating its performance. The study aligns with Sustainable Development Goals (SDGs) 4.6 and 4.7, promoting linguistic diversity and global citizenship education.

2. LITERATURE REVIEW

Computational linguistics, a field at the intersection of computer science and linguistics, has seen tremendous advancements in recent years (Hellwig & Nehrdich,2021). Among its critical areas is Natural Language Processing (NLP), which enables machines to understand, interpret, and generate human language. A fundamental task within NLP is Part-of-Speech (PoS) Tagging, the process of assigning grammatical categories to words within a text, such as nouns, verbs, and adjectives (Joshi et al, 2020). While substantial progress has been made in developing PoS taggers for European and Asiatic languages, African languages continue to be underrepresented in computational linguistics research. This gap underscores the necessity for targeted efforts in creating tools and resources for these languages.

2.1. Swahili POS Tagger

Swahili, a widely spoken language in Eastern and Central Africa, has benefited from data-driven PoS tagging efforts. De Pauw et al. (2006) used a data-driven approach for Kiswahili, utilizing the Helsinki Corpus of 3.65 million words. They compared multiple taggers, including Maximum Entropy Modelling (MXPOST) and Support Vector Machines (SVM). MXPOST outperformed others, achieving high accuracy. To ensure unbiased training, the corpus was randomized and divided into training (80%), validation (10%), and blind test (10%) sets. The taggers achieved remarkable accuracy on the blind test set, with MXPOST standing out for its use of maximum entropy modelling, incorporating lexical, contextual, and morphological features.

This project highlights the value of combining existing annotated corpora and advanced algorithms but demonstrates the limitations of corpus availability for Gikuyu.

2.2. Kamba POS Tagger

The Kamba language, predominantly spoken in Machakos, Makueni, and Kitui regions of Kenya, represents a significant example of under-resourced languages. Kituku et al. (2015) developed a POS tagger for Kikamba, using a Memory-Based Tagger (MBT). The project processed a corpus of approximately 30,000 words collected from online sources and documents in Kamba. After cleaning and formatting the corpus, manual annotation of ten PoS categories (e.g., adjectives, nouns, verbs, and punctuation) was completed using Microsoft Excel. The formatted dataset was then processed on a Linux-based MBT system.

The results demonstrated a precision of 83%, a recall of 72%, and an F-score of 75%, with overall accuracy reaching 90.68%. However, the tagger's effectiveness was influenced by the quality of annotations and the limited size of the corpus.

The system's language dependence however, limits its applicability to Gikuyu due to linguistic variations between the two languages.

2.3. Kipsigis POS Tagger

Kipsigis, a member of the Kalenjin linguistic family, represents another under-resourced language. The Kipsigis tagger, developed using MBT, processed a smaller corpus of 14,000 words. Following methods similar to the Kamba project, the corpus was manually cleaned, annotated, and formatted. The tagger achieved a precision of 88.375%, a recall of 88.25%, and an F-score of 88.625%, with overall accuracy reaching 94.46%.

The results were promising but highlighted the limitations of relying solely on traditional MBT methods. The study underscores the challenges of limited training data and the necessity of customized linguistic tools for under-resourced languages (Kituku et al., 2015).

2.4. Setswana POS Tagger

M Setswana, a Bantu language spoken in Botswana, South Africa, Zimbabwe, and Namibia, presents unique challenges due to its disjunctive orthography. Malema et al. (2017) employed a rule-based approach supported by a dictionary and a morphological analyser for Setswana POS tagging, focusing on relative and possessive structures. Despite achieving an identification rate of 82%, the approach struggled with longer sentences due to limited morphological analysis tools.

Additionally, while this effort marked a step forward for Setswana NLP, the approach was limited by its reliance on rule-based systems, which are generally less adaptable than data-driven models.

2.5. Research Gaps

While various approaches have been applied to Bantu languages, existing tools are language-specific and lack generalizability to Gikuyu. This study fills the gap by developing the first POS tagger tailored to Gikuyu, leveraging both linguistic insights and data-driven techniques.

3. METHODOLOGY

The methodology for developing the Gikuyu Part-of-Speech (POS) tagger involved a structured approach that included data collection, annotation, training, and evaluation. A data-driven approach using a Memory-Based Tagger (MBT) was selected due to its adaptability and performance for under-resourced languages (Liang & Huang, 2023). Figure 1 summarizes the POS tagger architecture.

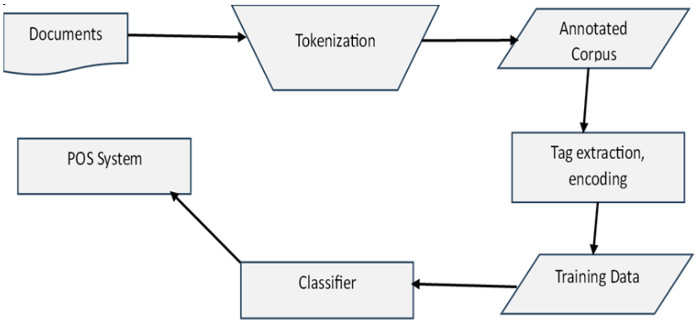


Figure 1: POS tagger Architecture

3.1. Data Collection

The primary corpus was extracted from the Gikuyu Bible, a comprehensive and structured text that provided a consistent linguistic framework. Other supplementary sources included online Gikuyu texts, cultural stories, and educational materials. Web scraping tools such as Google Scholar and Monster Crawler were employed to collect additional data, ensuring diversity and richness in the corpus.

3.2. Corpus Characteristics

The corpus consisted of approximately 10,000 words, divided into sentences for easy tokenization. The text was chosen to represent a variety of linguistic forms and grammatical structures in Gikuyu, including nouns, verbs, adjectives, and adverbs.

3.3. Manual Annotation

Manual annotation was conducted using Microsoft Excel, where each word was assigned a POS tag in a parallel column. This process involved:

i. Tokenizing sentences into individual words.
ii. Assigning tags based on predefined categories, such as nouns (NOUN), verbs (VERB), and conjunctions (CONJ).
iii. Ensuring consistency in tagging rules to improve the model's reliability.

Table 2. shows the comprehensive Tags Format adopted.

Table 2: Tags Format

Noun	NOUN	
Pronoun	PRONOUN	
Verb	VERB	
Adverb	ADVERB	
Adjective	ADJECTIVE	
Preposition	PREPOSITION	
Conjunction	CONJUCTION	
Comma	COMMA	
Colon	COLON	
Semi colon	SEMCLN	
Full stop	F-STOP	
Quotation Marks	QUOTES	
Question Mark	QUESM	
Apostrophe	APOSTR	
Exclamation mark	EXCLMK	
Numerals	NUME	

3.4. Tagset Design

The tags used were carefully chosen to align with common linguistic categories while addressing the unique needs of the Gikuyu language. The tagset included:

i. NOUN: Denoting objects, names, or places.
ii. VERB: Representing actions or states.
iii. ADJECTIVE: Qualifying nouns.
iv. PREPOSITION, CONJUNCTION, PRONOUN, etc.: Functional categories.
v. Punctuation markers: COMMA, F-STOP, APOSTR, etc.

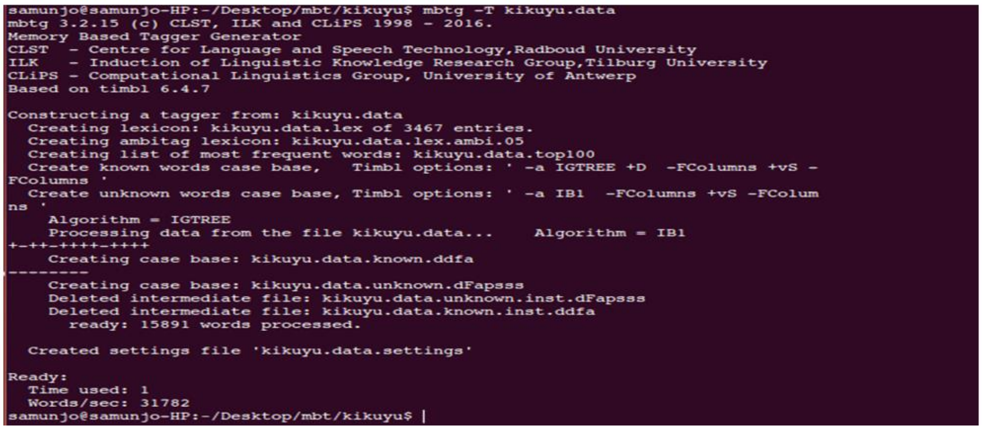
This tagset was inspired by established frameworks from related Bantu language projects, such as Kiswahili and Kamba (Kituku et al., 2015; De Pauw et al., 2006).

3.5. Training Process

A Memory-Based Tagger (MBT) was selected for its effectiveness in handling small datasets, characteristic of under-resourced languages. MBT uses a machine learning algorithm that relies on lexical, contextual, and orthographic features of the input words. The annotated corpus was converted into a text file format compatible with MBT. The conversion included:

i. Replacing spaces with tab delimiters.
ii. Adding sentence delimiters (<utt>).

Figure 2 shows a sample training run while Figure 3 shows an example from the training corpus:



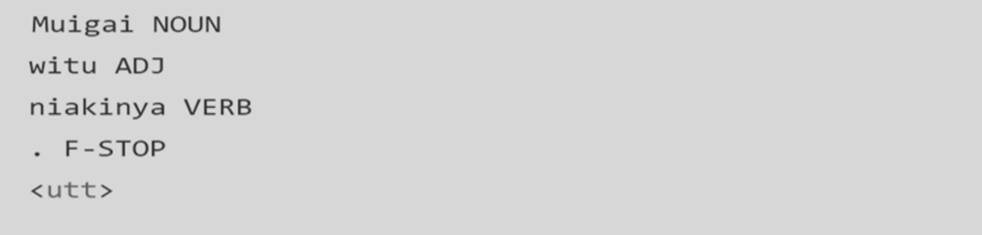
```
samunjo@samunjo-HP:~/Desktop/mbt/kikuyu$ mbtg -T kikuyu.data
mbtg 3.2.15 (c) CLST, ILK and CLiPS 1998 - 2016.
Memory Based Tagger Generator
CLST  - Centre for Language and Speech Technology,Radboud University
ILK   - Induction of Linguistic Knowledge Research Group,Tilburg University
CLiPS - Computational Linguistics Group, University of Antwerp
Based on timbl 6.4.7

Constructing a tagger from: kikuyu.data
  Creating lexicon: kikuyu.data.lex of 3467 entries.
  Creating ambitag lexicon: kikuyu.data.lex.ambi.05
  Creating list of most frequent words: kikuyu.data.top100
  Create known words case base,   Timbl options: ' -a IGTREE +D  -FColumns +vS -
FColumns '
  Create unknown words case base, Timbl options: ' -a IB1  -FColumns +vS -FColum
ns '
    Algorithm = IGTREE
    Processing data from the file kikuyu.data...    Algorithm = IB1
+-++-++++-++++
    Creating case base: kikuyu.data.known.ddfa
--------
    Creating case base: kikuyu.data.unknown.dFapsss
    Deleted intermediate file: kikuyu.data.unknown.inst.dFapsss
    Deleted intermediate file: kikuyu.data.known.inst.ddfa
      ready: 15891 words processed.

  Created settings file 'kikuyu.data.settings'

Ready:
  Time used: 1
  Words/sec: 31782
samunjo@samunjo-HP:~/Desktop/mbt/kikuyu$ |
```

Figure 2: A sample training run



```
Muigai NOUN
witu ADJ
niakinya VERB
. F-STOP
<utt>
```

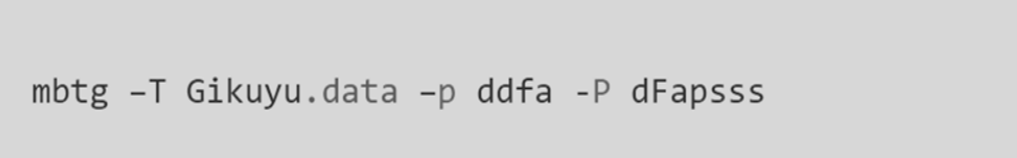
Figure 3: Sample annotated corpus

3.6. Training Configuration

The training process used default configurations in MBT to optimize the balance between complexity and computational efficiency:

i. ddfa: Focused on two disambiguated tags to the left and one ambiguous tag to the right for known words.
ii. dFapsss: Applied additional orthographic features for unknown words, including the first and last three letters.

The command-line script executed the training process as shown in Figure 4



```
mbtg -T Gikuyu.data -p ddfa -P dFapsss
```

Figure 4: command-line script

During training, the model generated configuration files for known and unknown word patterns. These were used to predict POS tags for new input data.

4. Evaluation

4.1. Cross-Validation

A 10-fold cross-validation technique was applied to ensure robust performance assessment. This method involves splitting the dataset into 10 equal subsets, training the model on 9 subsets, and testing it on the remaining subset. The process repeats until every subset has been used for testing.

4.2. Evaluation Metrics

The performance of the Gikuyu Part-of-Speech (POS) tagger was evaluated using four standard metrics: precision, recall, F-score, and accuracy (Gupta & Sign, 2024). These metrics are widely adopted in NLP for assessing classification tasks and sequence labeling systems like POS tagging.

4.2.1. Precision

Precision measures the proportion of correct POS tag predictions out of all tags predicted by the model. It is expressed as shown in Formula 1:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \qquad \mathit{Formula}\ 1$$

In the case of the Gikuyu POS tagger, precision evaluates how accurately the model assigns a specific POS tag to a word without including incorrect classifications. For this study, the tagger achieved an average precision of 90.44%, demonstrating its reliability in correctly predicting tags for the given input.

4.2.2. Recall

Recall measures the proportion of correctly predicted POS tags out of all actual instances of that tag in the dataset. It is calculated as shown in Formula 2:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \qquad \mathit{Formula}\ 2$$

The tagger's average recall score was 88.34%, reflecting its effectiveness in identifying true instances of POS tags in the Gikuyu dataset. High recall is particularly important for minimizing the omission of valid POS tags.

4.2.3. F-Score

The F-Score is the harmonic mean of precision and recall, providing a balanced measure that considers both false positives and false negatives. It is given by Formula 3.

$$\text{F} - \text{Score} = 2\ X\ \frac{\text{Precision X Recall}}{\text{Precision} + \text{Recall}} \qquad \mathit{Formula}\ 3$$

The Gikuyu tagger achieved an F-Score of 91.35%, which highlights its balanced performance across precision and recall. This metric is particularly relevant for under-resourced languages like Gikuyu, where misclassifications can have a significant impact on usability

4.2.4. Accuracy

Accuracy measures the proportion of all correctly classified tags (both known and unknown) out of the total predictions made. It is calculated as shown in Formula 4.

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \qquad \mathit{Formula}\ 4$$

Table 3 shows the evaluation results.

Table 3: Results

Fold	**Precision**	**Recall**	**F-score**
Fold1	0.8722	0.85384	0.8862
Fold2	0.9357	0.9328	0.9340
Fold3	0.8739	0.9381	0.9327
Fold4	0.9202	0.8523	0.9156
Fold5	0.8797	0.8489	0.8858
Fold6	0.9313	0.9308	0.9307
Fold7	0.9313	0.8793	0.9307
Folf8	0.8699	0.8563	0.8809
Fold9	0.9203	0.8671	0.9205
Fold10	0.9099	0.8746	0.9177
AVARAGE	**90.44 %**	**88.34 %**	**91.35 %**

For the Gikuyu tagger, accuracy was computed separately for known and unknown words:

i. Known Word Accuracy: 93.35%
ii. Unknown Word Accuracy: 88.35 %
iii. Overall Accuracy: 90.69%

Table 4 shows the accuracy confusion matrix.

Table 4 Accuracy confusion matrix

Fold	**Accuracy (Known)**	**Accuracy (Unknown)**	**Accuracy (Overall)**
Fold1	0.9250	0.8952	0.9256
Fold2	0.9329	0.9152	0.8989
Fold3	0.9403	0.8634	0.9369
Fold4	0.9434	0.8752	0.9032
Fold5	0.9269	0.8523	0.8869
Fold6	0.9389	0.8838	0.8724
Fold7	0.9323	0.8852	0.9245
Folf8	0.9137	0.8653	0.9035
Fold9	0.9420	0.8762	0.9074
Fold10	0.9397	0.8865	0.9100
AVARAGE	**93.35 %**	**88.35%**	**90.69 %**

The disparity between known and unknown word accuracy reflects the challenges inherent in handling novel inputs, a common issue in low-resource language processing. Improving the system's ability to generalize to unknown words remains a key area for future work.

4.2.5. Cross-Validation with K-Fold Testing

To ensure robust evaluation, a 10-fold cross-validation technique was employed. This approach involves dividing the dataset into 10 equal parts, using nine parts for training and one for testing iteratively, until all folds have been tested. This method minimizes overfitting and provides a reliable estimate of model performance (Musabeyezu et al. 2023). Results across the 10 folds were averaged to produce the final evaluation metrics.

4.3. Comparison to Related Work

The evaluation metrics of the Gikuyu tagger were benchmarked against POS tagging systems for other African languages:

i. Kamba: Kituku et al. (2015) reported an F-Score of 75% and overall accuracy of 90.68%.
ii. Kiswahili: De Pauw et al. (2006) achieved state-of-the-art accuracy using MXPOST with a large annotated corpus.
iii. Setswana: Malema et al. (2017) achieved 82% identification accuracy using a rule-based approach.

Compared to these systems, the Gikuyu tagger demonstrates a competitive performance, particularly given the constraints of limited corpus size and manual annotation.

5. RESULTS AND DISCUSSION

The Gikuyu Part-of-Speech (POS) tagger demonstrated strong performance across evaluation metrics, indicating its reliability and utility for processing the Gikuyu language. The key results are summarized below:

1. Precision, Recall, and F-Score:

    i. Precision: 90.44%
    ii. Recall: 88.34%
    iii. F-Score: 91.35%

These metrics indicate that the tagger accurately identified POS tags while minimizing false positives and negatives. The F-Score, which balances precision and recall, confirms the robustness of the tagger in both recognizing and predicting POS categories.

2. Accuracy:

    i. Known Word Accuracy: 93%
    ii. Unknown Word Accuracy: 88%
    iii. Overall Accuracy: 91%

The higher accuracy for known words compared to unknown ones highlights a common challenge in NLP, particularly for under-resourced languages. The system's ability to handle unknown words with 88% accuracy is noteworthy, given the limited corpus size and lack of advanced external resources.

3. Cross-Validation Results:

Using a 10-fold cross-validation approach, the tagger consistently performed well across all folds, with precision and recall varying slightly depending on the fold but maintaining an overall F-Score above 91%.

The Gikuyu tagger demonstrates strong performance despite the challenges of a smaller corpus and limited resources, highlighting the effectiveness of the MBT approach for under-resourced languages as shown in table 5.

Table 5: Comparison of Results with Related Works

Language	Approach	Corpus Size	Precision	Recall	F-Score	Accuracy
Gikuyu	MBT(Data-Driven)	10000 Words	90.44%	88.34%	91.35%	91%
Kamba	MBT(Data-Driven)	30000 Words	83%	72%	75%	90.68%
Kiswahili	MXPOST(Data-Driven)	3.65 Million Words	>95%	N/A	N/A	>95%
Setswana	Rule-Based	Limited	N/A	N/A	82%	82%

5.1. Discussion

The results of this study illustrate the feasibility and effectiveness of applying a data-driven approach to POS tagging for under-resourced languages. Several factors contributed to the tagger's success:

1. Use of Memory-Based Tagging (MBT)

MBT leverages lexical, contextual, and orthographic features, making it suitable for handling limited linguistic resources (Abdullahi & Traore, 2023) . The tagger's robust performance for known words demonstrates its capability to generalize within the boundaries of its training data. However, its slightly lower performance for unknown words suggests a need for additional techniques, such as morphological analysis or external lexicons, to improve predictions.

2. Impact of Manual Annotation

Manual annotation of the corpus ensured high-quality training data, which is crucial for achieving reliable predictions. The tags applied to the dataset adhered to a consistent schema, supporting accurate model training and evaluation. However, manual annotation is labor-intensive and may limit scalability.

5.2. Challenges Specific to Gikuyu

Morphological Complexity: Gikuyu features complex morphology, including inflections and agglutination, which pose challenges for POS tagging. The lack of a morphological analyzer limited the system's ability to generalize for unknown word forms.

Limited Corpus: The small size of the Gikuyu corpus, drawn mainly from the Bible, restricted the linguistic diversity of the training data. A larger, more varied corpus could improve the tagger's robustness and its application to diverse contexts.

Handling Unknown Words: A significant drop in accuracy for unknown words (88% compared to 93% for known words) emphasizes the need for enhanced generalization techniques, such as leveraging morphological patterns or external dictionaries.

5.3. Broader Implications

The development of a Gikuyu POS tagger has several potential applications:

Language Preservation: By digitizing linguistic structures, the tagger supports efforts to preserve Gikuyu, aligning with UNESCO's initiatives on linguistic diversity and multilingualism.

Machine Translation and Speech Processing: POS tagging is a foundational task for many NLP applications, including translation and speech recognition systems.

Educational Tools: The tagger could be integrated into educational software to teach Gikuyu in schools, especially in rural areas.

6. Conclusion and Future Work

This study successfully developed the first data-driven POS tagger for Gikuyu, achieving high accuracy and computational efficiency. The tagger can support NLP tasks such as machine translation, speech synthesis, and named entity recognition.

Future work could focus on:

1. Developing spellchecking capabilities.
2. Expanding the corpus to improve generalizability.
3. Adapting the tagger for other low-resource languages.
4. Exploring its application in syntax and semantic analysis tasks.

References

[1] Hellwig, O., & Nehrdich, S. (2021). Part-of-Speech Tagging for Low-Resource Languages: Tackling the Data Sparsity Problem. Journal of Language Modelling, 9(1), 41–66.

[2] Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. Transactions of the Association for Computational Linguistics, 8, 451–468.

[3] Kumar, S., Jyothi, P., & Bhattacharyya, P. (2024). Part-of-speech tagging for extremely low-resource Indian languages. Findings of the Association for Computational Linguistics: ACL 2024. Retrieved from ACL Anthology

[4] National Bureau of Statistics. (2019). 2019 Kenya population and housing census: Volume I - Population by county and sub-county. Retrieved from https://www.knbs.or.ke

[5] De Pauw, G., Schryver, G.-M., & Wagacha, P. W. (2006). Data-Driven Part-of-Speech Tagging of Kiswahili. Journal of African Languages and Linguistics.

[6] Kituku, B., Wagacha, P., & De Pauw, G. (2015). Kamba Part of Speech Tagger Using Memory-Based Approach. International Journal on Natural Language Computing, 4(2).

[7] Malema, G., Okgetheng, B., & Motlhanka, M. (2017). Setswana Part of Speech Tagging.

[8] Liang, Z., & Huang, J. (2023). Zero-resource cross-lingual part-of-speech tagging using unsupervised transfer methods. arXiv preprint arXiv:2401.05727. Retrieved from arXivar5iv

[9] Gupta, R., & Singh, A. (2023). Zero-resource cross-lingual part-of-speech tagging using HMM models and alignment-based corpus generation. arXiv preprint arXiv:2401.05727. Retrieved from arXiv

[10] Musabeyezu, T., Niyomutabazi, E., Chimhenga, E., Gotosa, K., Mizha, P., Agbolo, A., ... & Klakow, D. (2023). MasakhaPOS: Part-of-Speech Tagging for Typologically Diverse African Languages. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics. Toronto, Canada. Retrieved from ACL Anthology ACL Anthology

[11] Abdullahi, M., & Traore, S. (2023). Leveraging multilingual pre-trained models for improved African language POS tagging. Proceedings of the Conference on Natural Language Processing and Learning. Retrieved from ACL Anthology

[12] Daelemans, W., Zavrel, J., Van den Bosch, A., & Van der Sloot, K. (2007). Tilburg Memory-Based Learner Reference Guide.

AUTHOR

Gabriel Kamau, He is a lecturer in the Department of Computer Science, School of Computer Science and Information Technology of Dedan Kimathi University of Technology in Kenya. He has a teaching experience in undergraduate programmes spanning for over fifteen years