*Proc. of the International Conference on Computing and Advances in Information Technology (ICCAIT 2023)*
*21-23 November 2023, Ahmadu Bello University, Zaria, Nigeria*

# A Kanuri Part of Speech Tagger

Aminu Tukur
*Department of Mathematics and Computer Science, Borno State University*
*Maiduguri, Nigeria*
tukuraminu@bosu.edu.ng

Adamu jibrin
*Department of Mathematics and Computer Science, Borno State University*
*Maiduguri, Nigeria*
adamujibrin@bosu.edu.ng

Ummi Inuwa
*Department of Languages and Literary Studies, Borno State University*
*Maiduguri, Nigeria*
ummiinuwa@gmail.com

***Abstract*— Part-of-Speech (POS) tagging is a technique for assigning each word of a text with a fitting part of speech tag. POS tagging can be used in TTS (Text to Speech), Information Retrieval, Shallow Parsing, Information Extraction and Linguistic Research for corpora and also as an intermediate step for higher level Natural Language processing (NLP) tasks. Despite the recent effort of researchers to develop POS Tag tools of Nigerian Languages like Hausa, Yoruba and Igbo, Some Languages are yet to be explored in the research community. Consequently, this research work proposes a technique for POS tagging of Kanuri Language (sentences). Lacked Tagged Corpus using Part of Speech Make it difficult to develop such tool. Therefore, we will collect a corpus of Kanuri based text from Kanuri Literature and News Article from Borno State Ministry of Information and Culture, Borno Radio Television Corporation (BRTV) and Borno State Emirate Council, Maiduguri. We will find the words frequency, rank, range and probability using the concept of Hidden Markov Model (HMM). Furthermore, we will build a corpus of Kanuri based by tagging each word with its appropriate part of speech, which serve as the training corpus for our proposed Part of Speech Tagging Model. POS is a Pre-requisite for a Major NLP Task. Unfortunately, for Kanuri language, research towards tools and techniques for these pre-requisite activities is still in the initial stages, successful implementation will serve as the Foundation of Kanuri NLP and advance NLP task Like Speech Recognition.**

***Keywords*— *Natural Language Processing (NLP), Hidden Markov Model (HMM), Text to Speech (TTS), Part of speech (POS), Kanuri.***

## I. INTRODUCTION

The advancement of Information and Communication Technology (ICT) has led to the massive amount of user-generated content on social media, blogs, and websites, with recent exponential growth of such content. Natural Language Processing (NLP) has emerged as one of the most competitive study fields in Artificial Intelligence (AI). NLP is a branch of computer science that aims to define concepts and find tactics, and develop software that can read, write, speak, and interpret natural human languages to encourage communication between people and computers.

NLP enables computers to describe in plain language how people use language. A large number of NLP applications, word sense disambiguation, information retrieval, voice recognition, and machine learning, use Part of Speech (POS) tagging as part of the application processing. POS tagger is an essential tool, which assigns part of speech tags for each word of the sentence [2]. Basic POS taggers typically categorize the words into nouns, verbs, adjectives, etc., but more sophisticated taggers can also add other categories like numbers or gender. POS tagging is used as a pre-processing step by a wide range of language processing systems, greatly enhancing their recall and precision [3].

Kanuri is a Language widely spoken across the Northeast part of Nigeria (mostly in Borno and Yobe States), southeastern Niger, western Chad, northern Cameroon, parts of Libya, and the Darfur region of Sudan. The language is associated with the Kanem and Bornu empires that dominated the Lake Chad region for thousands of years. An estimated 7.8 million people speak Kanuri as a first language; the majority of them live in Nigeria. The basic expression order of Kanuri sentences is “subject–object–verb”. It is typologically rare all together having postpositions and post-nominal modifiers – for example, "Fatima’s pot" would be expressed *as nje Fatima-be*, "pot Fatima-of" [4].

Until recently, Kanuri has been written using the basic “Ajami” Arabic script, mainly in religious, court contexts and letters, for at least four hundred years. More recently, it is written in a modified Latin script. The Gospel of John published in 1965 was produced in Roman and Arabic script. A standardized Romanized orthography was developed by the Kanuri Research Unit and the Kanuri Language Board (known as the Standard Kanuri Orthography in Nigeria). The elaboration, based on the dialect of Maiduguri, was approved by the Orthography Committee of the Kanuri Language Board, under the Chairmanship and supervision of Abba Sadiq, Waziri of Borno. It was officially approved by the Kanuri Language Board in Maiduguri, Nigeria, in 1975 [5]; [6]. In spite all of the effort there is still scarcity of linguistic resources for Kanuri. Therefore, developing linguistic resources for under-resourced languages like Kanuri is an added advantage.

In recent times, there is significant interest in developing POS tagger of Nigerian Languages which include POS of Hausa [7]; Igbo [8], and Yoruba [9]. Regardless of all the effort of NLP of Nigerian Language, some Languages like “Kanuri” were not explored.

Nevertheless, given the availability and increasing number of Kanuri internet users and the exponential growth of Kanuri online content. Analysis and Processing of such text becomes necessary in the field of Natural Language Processing. Unfortunately, there are some pre-requisites to Conducting NLP research which include POS tagging. Although there are numerous existing POS tagging techniques. Unfortunately, none was trained in the Kanuri text. Therefore, there is a need to develop a standard POS Corpus of Kanuri and tools that recognize a Kanuri Text (Such Corpus).

Hence in this research, we proposed a statistical approach for Kanuri POS tagging toward the realization of Kanuri Natural Language Processing (kanuriNLP). The aim of this research is to develop a Kanuri part of speech Corpus and Part of Speech tagging tool.

The following paper is organized as follows: Section II, Present Literature review (Natural Language Processing, Part of Speech Tagging), Section III presents the Methodology, which comprises the Proposed model, the implementation of the proposed approaches using worked example, and also the proposed experiment with the required steps. Finally, Section IV Conclusion along with the scope for future work.

## II. LITERATURE REVIEW

In this Section we review literature on the Natural Language Processing and different Part of Speech Tagging (POS) techniques.

### *A. NATURAL LANGUAGE PROCESSING*

Today, there is a lot of discussion and study on the subject of natural language processing. It is one of the oldest areas of machine learning study and is utilized in important domains like text processing, speech recognition, and machine translation. Significant advances in computation and Artificial Intelligence have made thanks to natural language processing [10]. Furthermore, the different text and speech processing algorithms used in natural language processing, and also difference between the various algorithms used in NLP was highlighted. NLP as the branch of artificial intelligent (AI) which has become an interesting area of study in computational statistics and text determining. Emphasis on how NLP encompasses approaches that uses computer to analyze and determine similarity and translate between languages, which tempt the interest of researchers.

Subsequently, [11] discussed the essential element of NLP and how researchers have extracted data and encode natural language information for analytical processing starting from how NLP is been originated from linguistic rules. They also discuss how the medical NLP is been evolved including vectorization, word embedding, classifications and also how is used to automate speech recognition and also the role of machine learning and neural network in improving NLP was highlighted.

### *B. PART OF SPEECH TAGGING*

Part-of-Speech (POS) tagging is a technique for assigning each word of a text with a fitting part of speech tag. The significance of part-of-speech (also known as POS, word c-lasses, morphological classes, or lexical tags) for language processing is the large amount of information they give about a word and its neighbor. POS tagging can be used in different Languages to aid TTS (Text to Speech) conversion, and even improve linguistic research information retrieval, shallow parsing, information extraction, as an intermediate step for advanced level NLP tasks such as parsing, semantics, translation, and many more [12]; [13]. POS tagging is an important step in natural language processing, literally a pre-requisite of Advance NLP of any Language. Corpora that have been POS-tagged are very useful not just for linguistic research, e.g. finding instances or frequencies of particular constructions but for further computational processing, such as syntactic parsing, speech recognition, stemming, word sense disambiguation [14].

POS tagging is the first task in every NLP application. The complexity of computing POS tag lies in the number of computational steps an algorithm uses for determining POS tag for a given sentence [15]. Usually, POS tagging follows two approaches; Rule based and corpus base. Rule-based taggers try to assign a tag to each word using a set of handwritten rules of that Particular Language. These rules could indicate, for example, that a word following a determiner, and an adjective must be a noun. Of course, this means that the set of rules must be properly written and checked by human experts as shown in [12]. With the advent and availability of ICT resources and online content of low resources Languages different POS model were developed such as a POS tagger for both Kannada and Hindi by [16], employing Machine Learning (ML) and Deep Learning (DL) approach. Other languages include [17] the development of a Part-of-Speech tagger Indigenous language of Aotearoa, also known as New Zealand and Annotating Norwegian language varieties on Twitter for part-of-speech by [18]. Though, some used the corpus-based approach for the training of data or knowledge resource for POS Tagging and NER [19], instead of the Machine Learning or Statistical taggers.

Consequently, In Nigeria, a lot of researchers have shifted their dynamic toward NLP of Nigerian Languages which has gained a wider audience. A process has been applied to improve a part-of-speech (POS) tagged corpus for the Igbo Language using an inter-annotation agreement (IAA) exercise to iteratively revise the tagset used in the creation of the initial tagged corpus, with the aim performance [20]. The result shows a significant improvement in accuracy increasing from 88% to 94%. Although, the tagged corpus is re-usable for other dialects of the language. A Machine Learning Method was presented Another Approached was presented, using POS tagging of an Igbo corpus that can achieve a good performance with the small dataset available for the language [21]. The Table below shows the summary of POS taggers of Nigerian Languages.

Table1: POS Tagging of Nigerian Languages

| S/no | Title | Reference |
|---|---|---|
| 1. | Part-of-speech Tag set and Corpus Development for Igbo,An African | [8] |
| 2. | Part-of-Speech tagging of Yoruba Standard, Language of Niger-Congo family | [9] |
| 3. | AcomparativestudyofhiddenMarkovmodelandconditionalrandomfieldsonaYorùbapart-of-speech Tagging task | [22] |
| 4. | Parts-of-Speech Tagging of Hausa-Based Texts Using Hidden Markov Model. | [7] |

From the Table 1 above, we observed that there is POS Taggers of different Languages of Nigeria precisely; Hausa, Yoruba and Igbo. The work mentioned can serve as the Foundation of our work.

## III. METHODOLOGY

This Section presents the general research methodology that will be used in conducting this study, it is shown as a sequence of activities that will be carried out in phases and steps.

### *A. Proposed Model*

This research proposed a model for the assignment of Part-of-Speech to Kanuri text, usually from sentences. We will use the collected Tag corpus for the training of the proposed model. The Model is shown below:

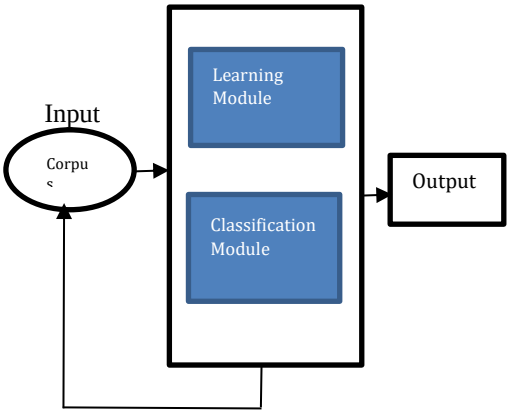


**Fig 1**: Architecture of the Proposed Model.

From Fig 3.1, we propose a model that will recognize Kanuri-based text. The model has the following component: The Pre-Processing stage which include the corpus and the implementation stage which include the "Sentence Collection", "Sentence Analyzer" and "POS Tagging module".

The explanation of the component of the proposed model is shown in bellow section below:

### *B. Description of the Proposed Model*

The description of each component from Fig 3.1 is discussed in this section.

**Pre-Processing stage**

The pre-processing stage is the phase in which we build the corpus.

i. **Corpus:** Will consist of writing from Kanuri Literature and News Sources, the part of Speech will be use from (Jarrett, 2007) "A dictionary of Manga, a Kanuri Language of Eastern Niger and NE Nigeria Dictionary". A combination of both Writing and their related part of speech. It will serve as a training set for our model.

ii. **Input:** 25% of the Corpus serves as the input serve for the analysis.

iii. **Hidden Markov Model (HMM):**

The **Hidden Markov Model (HMM)** is a generative sequence model or classifier that use probabilistic approach that maps a sequence of observations to a sequence of labels. It is a model where the states represent labels (e.g. words, letters, etc.) and the transitions represent the probability of jumping between the states [23]. It follows what is called the **Markov process** meaning that the probability of being at some point in a state is independent on the previous history (*memoryless*). The HMM is suitable for Low dataset. HMM uses the following equation:

$$P(q_j|q_i) = \frac{N(i,j)}{\sum_{j'} N(i,j')} \quad \text{equ 1.}$$

Where N(i,j) is the number of transitions from state qi to state qj.

HMM model in this instant consist of two phases which include "The Sentence Analyzer module" and "Tagging Module".

**i. The Sentence Analyzer**

The learning phase of the module based on Hidden Markov Model (HMM), the module will learn from the "kanuri" corpus that we will build. After the learning is complete it can assign a tag to an unseen word.

**ii. Classification Module**

This is the "POS Tagging" classification phase. The module will assign Part-of-speech to Kanuri word of the sentences as an output "Tag set"

### *C. Experiment Planning*

#### *I. Data set*

A corpus of Kanuri from "Literature" will be used and the POS tag from "A dictionary of Manga, a Kanuri Language of Eastern Niger and NE Nigeria Dictionary" by Kevin Jarrett, will be used in the tagging. After the tagging of the POS is complete, it will serve as our training set. For testing we are going to use 25% of the data set.

#### *II. ii.Performance Metrics.*

**Accuracy:** Simply measure that the Algorithm makes the correct classification.

We will evaluate the accuracy of our classification using:

$$\text{ACCURACY OF SYSTEM} = \frac{Number\ of\ Correct\ Tag\ (Verified\ by\ Dictionary)}{Total\ No.\ of\ Tag} X\ 100$$

eqn (3.1)

$$\text{ACCURACY OF Expert} = \frac{Number\ of\ Correct\ Tag(verified\ by\ Expert)}{Total\ No.\ of\ Tag} X\ 100$$

eqn (3.2)

The Accuracy is the most adopted form of evaluating the efficiency of POS tagging as shown in Section.

#### *III. Evaluation Strategy.*

For the training of the Kanuri Corpus, Train/Test split validation is carried out and expert verification of some of the testing as well. The corpus data is separated into 2 subsets, we are going to categorize it as (25/75) %, Where 25% will serve as testing set and the 75% will serve as the Training set for Kanuri POS Tagging tool.

#### *IV. Experimental Design*

In our research we performed two different experiments which are normally divided into test and training set as shown in section 3.3.3. We will verify our result by an expert from the

field of Kanuri Linguistics as shown in 3.5. Verification by expert will help in reducing the bias.

### *D. Expert Analysis*

Since it is the first research for POS of Kanuri, we are going to Perform Expert verification of our techniques in other to verify the efficiency of the automatic tagging, so that our tool can be used for future Analysis of KanuriNLP.

For the Analysis a sentence will be given; in which we will split the sentence into words, each is assign a POS from our proposed model, and another column will be provided for the verification, whether the word is correctly tag or wrongly tag by the proposed model.

The table of the Analysis will include the following Headings: S/n, Words. POS Tag, Verification. The Sample of the analysis is shown in table 3.1 below using the Example "*Modu Kiri Chezi"*

TABLE 1: QUESTIONNAIRE SAMPLE FOR THE EXPERT.

| S/n | Words | POS Tag Automatically By the Computer system | Verification by Expert |
|---|---|---|---|
| 1. | Modu | NN | ✓ |
| 2. | Kiri | VB | X |
| 3. | Chezi | VB | ✓ |

From table 3.1 above the symbol " ✓ " simply means the model has correctly Tag the Words and the symbol "X" means the model Has Wrongly tag the Sentence. We will use the eight basic part of speech NN: Noun, VB: Verb, ADJ: Adjective, ADV: Adverb, PRON: Pronoun, PREP: Preposition, CONJ: Conjunction, TM: Tense maker, NUMB: Number. While other Symbols are maintained to be in the form of symbols.

## IV. CONCLUSION

In our paper, we proposed a technique of developing POS tagging of Kanuri based text using Hidden Markov Model. For our experiment we will create Corpus of Kanuri and perform the POS tagging on a corpus of News Article collected from Borno Radio Television. The work is on progress; with the successful implementation, we plan to extend it to cover order aspect of Kanuri NLP.

Developing a POS tag Corpora of Kanuri is an important tool for studying the languages and improving language learning and teaching. They provide a wealth of information about how languages work and can be used to create effective teaching materials and activities. Developing a Kanuri POS tagging tool can be used as a foundation for Kanuri Language Processing "KanuriNLP" which leads to Text to Speech (TTS), Information Retrieval, Shallow Parsing, Information Extraction and Linguistic Research for corpora and also as an intermediate step for higher level Natural Language processing (NLP) tasks.

## REFERENCES


[1] Zakari, R. Y. (2021). A Systematic Literature Review of Hausa Natural Language Processing. July, 4–11. https://doi.org/10.24203/ijcit.v10i4.86

[2] Kanakaraddi, S. G., & Nandyal, S. S. (2018). Survey on Parts of Speech Tagger Techniques. Proceedings of the 2018 International Conference on Current Trends towards Converging Technologies, ICCTCT 2018, 1–6. https://doi.org/10.1109/ICCTCT.2018.8550884.

[3] Anoop, K. K. A. R. R. V. S. (2020). Parts-of-Speech tagging for Malayalam using deep learning techniques. International Journal of Information Technology. https://doi.org/10.1007/s41870-020-00491-z

[4] Jarrett, K. (2007). A dictionary of Manga.

[5] Pulse (2021-11-15). "Kanuri: Bornu's most dominant culture". Pulse Nigeria. Retrieved 21-Jan-2023

[6] Cyffer, N., Hutchison J, P., (1990). Dictionary of the Kanuri language. ISBN 90-6765-412-4

[7] Tukur, A., Umar, K., & Muhammad, A. S. I. Parts-of-Speech Tagging of Hausa-Based Texts Using Hidden Markov Model. Dutse Journal of Pure and applied sciences (DUJOPAS) 6(2). June 2020.

[8] Onyenwe, I. E., Uchechukwu, C., & Hepple, M. (2014). Part-of-speech tagset and corpus development for igbo, an african. In Proceedings of LAW VIII-The 8th Linguistic Annotation Workshop (pp. 93-98). Association for Computational Linguistics and Dublin City University.

[9] Adedjouma Sèmiyou, A., Aoga, J. O., & Igue, M. A. (2012). Part-of-speech tagging of yoruba standard, language of niger-congo family. Research Journal of Computer and Information Technology Sciences, 1, 2-5.

[10] Chithra, P., & Henila, M. (2019). International Journal of Computer Sciences and Engineering Open Access. International Journal of Computer Sciences and Engineering, 6(10), 628–632. https://doi.org/10.26438/ijcse/v6i1.161167

[11] Chen, P. H. (2020). Essential Elements of Natural Language Processing: What the Radiologist Should Know. Academic Radiology, 27(1), 6–12. https://doi.org/10.1016/j.acra.2019.08.010

[12] Hasan, F. M., UzZaman , N., & Khan, M. (2007). Comparison of Different POS Tagging Techniques (n-gram, HMM and Brill's tagger) for Bangla. Advance and innovation in Systems, Computer Science and engineering, 121-126.

[13] Ankita, & A. N. (2018). Part-of-speech Tagging and Named Entity Recognition Using Improved Hidden Markov Model and Bloom Filter. International Conference on Computing, Power and Communication Technologies (GUCON) (pp. 28-29). Greater Noida, UP, India: IEEE.

[14] Hana, J., Feldman, A., & Brew, C. (2006). Tagging Portuguese with a spanish tagger using cognates. ACM digital library (pp. 33-40). Trento, Italy: Association of Computational Linguistics.

[15] Dalai, T., Mishra, T. K., & Sa, P. K. (2023). Part-of-Speech Tagging of Odia Language Using Statistical and Deep Learning Based Approaches. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(6), 1-24.

[16] Advaith, V., Shivkumar, A., & Lakshmi, B. S. (2022, July). Parts of Speech Tagging for Kannada and Hindi Languages using ML and DL models. In 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) (pp. 1-5). IEEE.

[17] Finn, A., Jones, P. L., Mahelona, K., Duncan, S., & Leoni, G. (2022, May). Developing a Part-Of-Speech tagger for te reo Māori. In Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages (pp. 93-98).

[18] Mæhlum, P., Kåsen, A., Touileb, S., & Barnes, J. (2022). Annotating Norwegian language varieties on Twitter for part-of-speech. arXiv preprint arXiv:2210.06150.

[19] Márquez, L. (1999). Part-of-speech Tagging: A Machine Learning Approach based on Decision Trees. Universitat Politècnica de Catalunya.

[20] Onyenwe, I., Hepple, M., Uchechukwu, C., & Ezeani, I. (2015). Use of transformation-based learning in annotation pipeline of igbo, an african language. In Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (pp. 24-33). Association for Computational Linguistics.

[21] Onyenwe, I. E., Hepple, M., Chinedu, U., & Ezeani, I. (2019). Toward an effective Igbo part-of-speech tagger. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18(4), 1-26.

[22] Ayogu, I. I., Adetunmbi, A. O., Ojokoh, B. A., & Oluwadare, S. A. (2017, October). A comparative study of hidden Markov model and conditional random fields on a Yorùba part-of-speech tagging task. In 2017 International Conference on Computing Networking and Informatics (ICCNI) (pp. 1-6). IEEE.

[23] Mor, B., Garhwal, S., & Kumar, A. (2021). A systematic review of hidden Markov models and their applications. Archives of computational methods in engineering, 28, 1429-1448.