ISSN (Print): 2476-8316
ISSN (Online): 2635-3490
Dutse Journal of Pure and Applied Sciences (DUJOPAS), *Vol. 6 No. 2 June 2020*

Parts-of-Speech Tagging of Hausa-Based Texts Using Hidden Markov Model

***Aminu Tukur[1], Kabir Umar[2], Anas Sa'idu Muhammad[3]**

[1,2]Faculty of Computer Science and Information Technology,
Bayero University, Kano.
tukuraminu85@gmail.com

[3]Department of Nigerian Langauges,
Bayero University, Kano.

Email: ukabir.se@buk.edu.ng

Abstract

In the area of Computational Linguistics, Hausa-Based text as an aspect of Natural Language Processing (NLP) is a virgin research area with work done on stemming of Hausa words. However, there are areas that are left untouched like sentiment analysis, lemmatization and Part-of-Speech (POS) tagging. This paper presents a technique for POS tagging of Hausa sentences using the Hidden Markov Model. We collected a corpus of Hausa-Based texts from Freedom Radio and AfriHausa, we train the model using text collected annotated with the eight basic POS with the addition of processes in the form of a number and tense maker as independent POS tagged sets. The Result obtains from our training and testing of Hausa Corpus includes the Average of 76.795% accuracy, with Adjective Achieving the highest accuracy of 100% and Conjunction having the lowest Accuracy of 50%.

Keywords: Lemmatization, Natural Language processing, Sentiment Analysis, Stemming, Part-of-Speech (POS) Tagging

INTRODUCTION

Hausa is a Chadic language and is the second most spoken language with approximately 40 million native speakers and about 18 million second-language speakers all located in 13 different countries in Africa. The first writing system of the Hausa language, called *Ajami*, was based on the Arabic script. Presently, the Romanized orthography, called *Boko*, is used, since its introduction by the British at the beginning of the twentieth century (Caron, 2015; Newman & Newman, 2001). The standard Hausa alphabet consists of 28 characters with 23 from the English alphabet excluding q, v and x with the addition of ɓ, ɗ, ƙ, and 'y (Newman, Yaro, Newman, & Dresel, 1979). The Hausa language consists of a large amount of

**Author for Correspondence*

inflectional and derivational morphology (Newman, 2000; Schuh, 2019). Given the increasing number of Hausa internet users and the exponential growth of Hausa online content, exploration in this language has gained the attention of many researchers in the last decade. Recently, researchers have developed an interest in performing and developing an algorithm that analyses Hausa text (Bimba, Idris, Khamis, & Mohd Noor, 2015).

A valid example is one of the tools designed to solve the problem of information retrieval of the Hausa-based text is situated on a model that will stem Hausa text (Bimba, *et al* 2015). It was designed on the criteria of an algorithm that reduces all words with the same root or stem to a common form, usually by stripping each word of its derivational and inflectional affixes. For illustrative purposes, words like *maroowaci* can be stem to *roowa*, similar to English words like connected, connecting, will be truncated to connect. In another array of research, a spelling corrector of Hausa text was designed and the characteristics of the language alphabet were considered (Salifou & Naroua, 2014). A study was also conducted to develop a model to automatically summarize Hausa-Based text on feature extraction using Naïve Model (Bashir, Rozaimee, & Isa, 2017). Virtually, it can be argued that in the area of Natural Language Processing (NLP) of Hausa, a lot has been covered (Bashir, *et al* 2017; Bimba, *et al* 2015; Maitamaa, *et al* 2014; Salifou & Naroua, 2014). Nevertheless, there are areas that are left untouched and are at the infancy stage like sentiment analysis. However, the basic approach that is adopted in conducting sentiment analysis includes the data preprocessing stages which consist of tokenization, stemming, part-of-speech (POS) tagging, feature extraction and stop word elimination. What makes the analysis of other languages, like English, Indo-European, and other Asian languages to be successful is the availability of a tool or rather resources called “Stanford NLP” to perform the analysis mentioned earlier. Therefore, in this paper, we are presenting a method of POS Tagging of the Hausa-Based text using the Hidden Markov Model (HMM). Basically, the fundamental aim is to perform Expert verification.

RELATED WORK

In this section, we review literature on general concept of POS tagging and the applicability of different POS tagging technique on Hausa-Based text.

Part of Speech (POS) Tagging

Part-of-Speech (POS) tagging is a technique for assigning each word of a text with a fitting POS tag set. The significance of POS (also known as word classes, morphological classes, or lexical labels) for language processing is a large amount of information they give about a word and its neighbor (Bernard, 1991). POS tagging can be used in Text-to-Speech (TTS), information retrieval, shallow parsing, information extraction, linguistic research for corpora and also as an intermediate step for higher-level NLP tasks such as parsing, semantics, translation, and many more (Ankita & Abdul, 2018; Hasan, UzZaman, & Khan, 2007). POS tagging remains an important step in natural language processing, as corpora that have been POS-tagged are very useful not just for linguistic research, in case of finding concordances or frequencies of particular constructions, but for further computational processing, such as syntactic parsing, speech recognition, stemming, word sense disambiguation (Hana, Feldman, & Brew, 2006). POS tagging is usually the first task in most NLP applications. The complexity of computing POS tag lies in the number of computational steps an algorithm uses for determining the POS tag for a given sentence (Ankita & Abdul, 2018). Other tasks that use NLP application include identifying the entity, for instance, Named Entity Recognition (NER), which is the basic operation in NLP. NER is done with the help of POS tags (Ankita & Abdul, 2018).

Usually, POS tagging follows two approaches; Rule-Based and Corpus-Based. Rule-Based taggers are used in assigning a tag to each word using a set of handwritten rules. The set of rules must be properly written and checked by human experts, the rules could specify, for instance, that a word following a determiner and an adjective must be a noun (Hasan, UzZaman, & Khan, 2007). While the corpus-based approach, also known as feature engineering (Li, Graca, & Taskar, 2012) uses the training data or knowledge resource for POS tagging and NER (Marquez, 1999). In addition to this, morphological tagging is used in assigning POS, case, number, gender and other morphological information to each word in a corpus. Despite the importance of morphological tagging, there are many languages that lack annotated resources of this kind, mainly due to the lack of training corpora which are usually required for applying standard statistical taggers. Many words in a language like English are associated with more than one POS, and correct interpretation of a word depends on it being assigned the correct category (Kupiec, 1992). Disambiguated category information for the words in a document is useful for text analysis in information retrieval systems, and also for stress assignment in TTS applications. This rested to the fact that many language models for speech recognition systems are based on a POS model of word sequences.

Consequently, tagging programs typically use a dictionary containing the set of possible categories of each word in the vocabulary can assume. In the process, the set of 3300 rules used a context of up to four words and same criteria was used as an aid in tagging 77% of the corpus the Brown Corpus correctly (Francis & KuEera, 1982; Keera & Francis, 1967). Although, some specific problems must be addressed when applying taggers to other languages than English. Researches have shown that it is challenging to develop a POS tagger for other languages like the Middle Eastern language such as Hebrew and Arabic given the high morphological complexity and a high degree of ambiguity in Semitic texts. Besides the problems derived from the richer morphology of the particular language, there is a more general problem consisting of the lack of large, manually annotated corpora for training (Ratnaparkhi, 1996). Most of the recent corpus-based POS taggers in the literature are either statistically based, that uses HMM (Ankita & Abdul, 2018; Jelinek, *et al* 1994; Kupiec, 1992; Merialdo, 1994; Weischedel, Meteer, Schwartz, Ramshaw, & Palmucci, 1993; Yousif, 2019).

Nevertheless, the rule-based is also used in some cases such as Drill's Transformation Based Learner [TBL] (Brill, 1992). However, the Maximum Entropy (MaxEnt) combines the advantages of all these methods. It uses a rich feature representation, like TBL and SDT, and generates a tag probability distribution for each word, like Decision Tree and Markov Model techniques, even though it is not suitable for small data set. However, some of the problems were overcome by using the two simpler tokenization models based on a model of word or word-segmentation and implement them using standard HMMs (Bar-Haim, Silam, & Winter, 2008). Several factors could affect the taggers' results, which include training data, vocabulary, syntactical and grammatical styles, and the method used in the classification of tags. Many methods are used to examine the performance of POS tagging. However, the simplified scheme of evaluating the correctness of the classification tag is accurate. In addition, it is essential to highlight some factors that could affect the accuracy of tagging like the type of vocabulary, tag set size, test corpus size, and the method used to compute the production. The accuracy is the number of correct answers divided by the total number of responses (both correct and incorrect).

The phrase recognizers by (Cutting, Kupiec, Pederson, & Penelope, 1992) also provide input to a system, which recognizes nominal arguments of verbs, specifically, Subject, Object, and Predicative arguments. However, the "SOPA" does not rely on other information (such as parity or voice) specific to the particular verbs involved. Some commonly used statistics for part of speech tagging are: how often a certain word was tagged in a certain way; how often two tags appeared in sequence or how often three tags appeared in sequence. These look a lot like the statistics a Markov Model would use. However, in the maximum entropy framework, it is possible to easily define and incorporate much more complex statistics, not restricted to n-gram sequences (Toutanova & Manning, 2000). Therefore, there is a need for a study that considers features that require Hausa language morphological analysis. As a result, this study focuses on developing a model to automatically Tag Hausa text using POS tag. For long, it has was argued that:

> Computers can't understand human language. They can't analyze and derive meaning from human language. So, there should be a way where a computer can understand, analyze and generate meaning from the human language. The way is known as Natural Language Processing. Some of the application of NLP is Automatic Summarization, Machine Translation, Speech Processing, Information Extraction, Question Answering System, Opinion Mining and Topic Segmentation etc. Most of these applications use POS tagging and NER (Hasan, UzZaman , & Khan, 2007).

Applicability of Existing

POS Tagging Technique to Hausa-Based Text

There is numerous POS Tagging techniques, however, none of which is trained on low resource language like Hausa. In this section, we are going to explain and show the grammatical structure of the Hausa Language.

Adjectives

Adjectives share their morphology with nouns, and some can even function as nouns, for instance, *tsoho* 'old' (adj.) and 'old man' (n.). However, they have some specific properties that distinguish them from nouns (Jabir, 2015). These are: (i) morphology: there are adjectives derived from nouns of quality that have a specific morphology and cannot function as nouns, for example, *zazzafa* 'very hot' (*zaafi*: 'heat'); (ii) Syntax: they function mainly as nominal modifiers or predicators; (iii) Their gender and number features are assigned by the noun they qualify and are not lexical properties.

Syntax

Adjectives appear in three different constructions: (i) predicative; (ii) pre-nominal attributive; (iii) post-nominal attributive. Let us illustrate these three constructions with a simple adjective, fari (m.); fara (f.); farare (pl.) all are referred to as 'white' in English. The pre-nominal attributive structure is <Adj - POSL N>. The post-nominal attributive function uses the structure <N Adj> without POSL as shown in the example below:

mace kyakkyawa; meaning "beautiful lady" in English.

Mace is noun and kyakkyawa is adjectives. Likewise, the adjectives can also appear before the noun, as in the case of:

kyakkyawar mace meaning "beautiful lady" in English.

Key: Adj: Adjective; POSL: possessive Link; M: masculine; F: feminine; Pl: Plural.

There are single words in Hausa that doesn't have a single word meaning. For instance:

gangara, meaning: Sloping ground (downward).

In practice, there is a need to develop a model that recognized the syntax of Hausa grammatical structure as designed in this study.

Adopted Algorithms

Hidden Markov Model (HMM) is the most popular algorithm for POS tagging, it analyzes the context of in which a word appears in a sentence and it is friendly for small dataset. After which assigns each word to its appropriate POS. HMM assigns POS tags by searching for the most expected tag for each word in a sentence. But, the HMM-based tagger does not find a tag for each word individually. On the contrary, it finds a tag sequence for a sentence as a whole. The model usually provides the conditional probability of a word given the preceding word, when the relation of the conditional probability is applied:

$$(P(Wn \mid \text{Wn-1}) = \frac{P(Wn-1,Wn)}{P(Wn-1)} \quad \ldots\ldots\ldots\ldots.2.1$$

The probability P() of a Word W n given the preceding token Wn-1 is equal to the probability of their bigram (Wn−1,Wn), divided by the probability of the proceeding Word. Example, *Musa yaa kashe kura* P(*Musa/yaa*), P(*ya/kashe)*, P(*kashe/kura*). The bigram model studies the pattern of the sentence and assign a probability value, it assigns probability in this sequence; what is the probability of word "b" following "a" and word "c" following word "b" and continuous……

EXPERIMENT PLANNING

Data set

A corpus of Hausa from "Freedom Radio" and "AfriHausa" will be used and the POS tag from a Hausa-English dictionary and English-Hausa vocabulary by Rev G.P. Bergery (1934) and grammar book written by M.K.M. Galadanci (1976) that gives detailed syntactic features of Hausa language will be used in the tagging process. After the manual tagging of the POS, it will serve as our training set for the model.

Performance Metrics

Accuracy: Simply measure that the Algorithm makes the correct classification. We will evaluate the accuracy of our classification using:

ACCURACY OF SYSTEM= $\frac{Number\ of\ System\ Correct}{Total\ No.\ of\ Tag}\ X\ 100$……………..3.1

ACCURACY OF Expert= $\frac{Expert}{Total\ No.\ of\ Tag}\ X\ 100$…………………………3.2

Accuracy is the most adopted form of evaluating the efficiency of POS tagging.

Evaluation Strategy

We are going to perform Train/Test split validation. The corpus data is separated into 2 subsets, we are going to categorize it as (25/75) %, Where 25% will serve as a testing set and the 75% will serve as the Training set.

Experimental Design

In our paper, two different experiments are performed, which are normally divided into test and training set. We will verify our result by an expert from the field of Hausa Linguistics as shown in section 3.3.5. Verification by an expert will help in reducing the bias.

Expert Analysis

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 NLP of Hausa. For the Analysis a sentence will be given; in which we will split the sentence into words, each word is assigned 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 1.1 below using the Example *Musa ya kashe kare:*

Table 1.1: Questionnaire Sample for the Expert.

S/n	**Words**	**POS Tag by our system**	**Verification by Expert**
1.	*Musa*	NN	✓
2.	*Ya*	VB	X
3.	*Tafi*	VB	✓

From table 1.1 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 have used 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. However, we are going to analyze it using the sample from Figure 1.1 below:

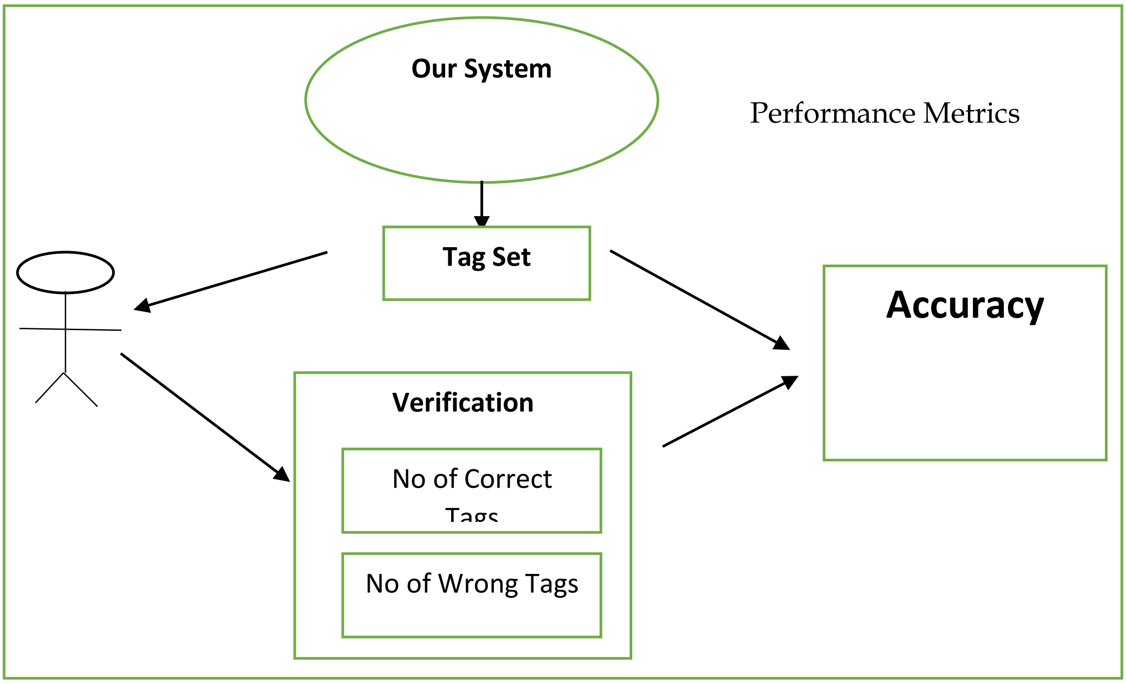

Figure 1.1: Expert Analysis Strategy

RESULT AND DISCUSSION

We have tag 3000 words, from our corpus using the eight basic parts of speech with the addition of Tenses maker and Number with their equivalent POs tags from the dictionary and the grammar book. we perform Split Validation technique. The model from equation 2.1 was used in the analysis. 25% percent of the corpus was used for testing while 75% was used for training of the model.

Result of Expert Analysis

We have used a sample of sentences that consists of 187 words for verification by an expert of Hausa Linguist. The 187 words served as an input for the system. The table consist of the following heading: “POS”, “Total No. of Tags”, “System Tag Correct”, “Expert Verified”, “System Accuracy”, and “Expert Accuracy”. The “Total No of tags” i.e. The POS tag, as automatically tagged by our system. The “System Tag correct”-the number “POS Tag” verified to be correct. “Expert Verified”- the Total number of POS Tag Verified which include both the Correct and Wrong POS Tag.

Table 1.2: Expert Verification

S/N	POS	Total No. of Tags	System Tag Correct	Expert Verified	System Accuracy %	Expert Accuracy %
1	NN	70	59	70	84.2	100
2	VB	25	17	25	68	100
3	PRON	13	10	13	76.9	100
4	ADV	-	-	-	-	-
5	ADJ	8	0	8	100	100
6	PREP	46	39	46	84.78	100
7	CONJ	4	2	4	50	100
8	TM	14	9	14	64.8	100
9	NUM	7	6	7	85.7	100
Total		**187**	**150**	**187**	**76.795**	**100%**

The Automatic POS Tagging tool of Hausa Sentence has achieved the Total Accuracy of 76.75%. The Accuracy of each POS Tag is represented using a line graph which is shown in Figure 1.2 below:

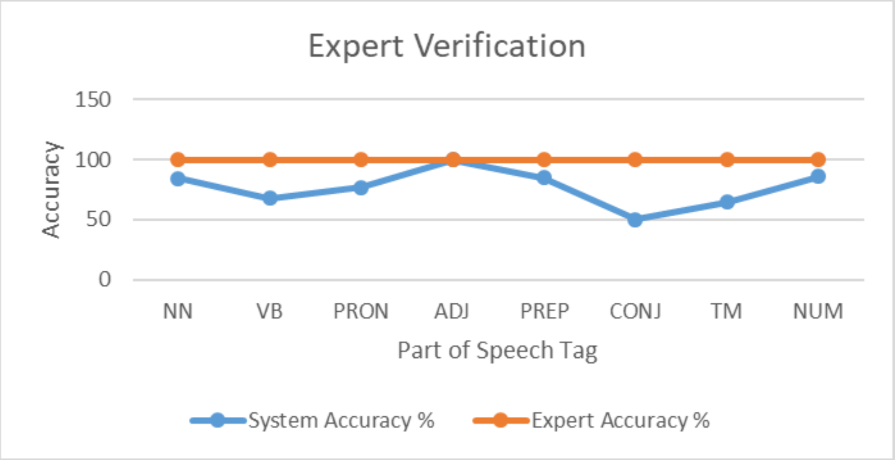

Figure 1.2: Accuracy of POS Tags verified by an Expert.

The representation of Figure 1.2 using Histogram is shown in Figure 1.3 below:

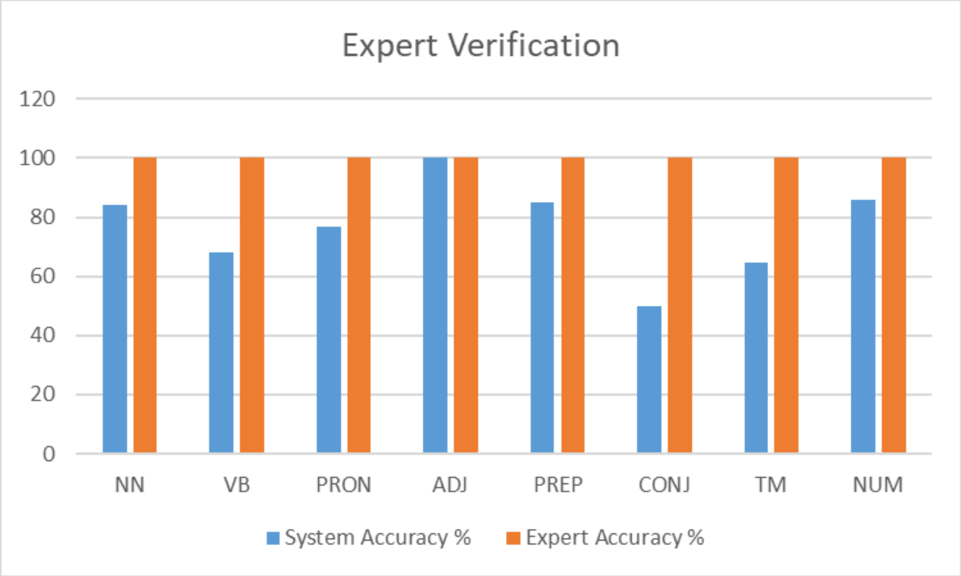

Figure 1.3: Expert Verification

From Figure 4.1 we can see that ADJ has performed well thereby achieving the accuracy of 100% as verified by an expert the remaining POS tags; NN, VB, ADJ, PREP, CONJ, TM and NUMB Achieved the following Accuracy consecutively: 84.2, 68, 76.9, 100, 84.78, 50, 64.8, 85.7. The POS tag with the lowest accuracy is the CONJ, which achieved 50%.

CONCLUSION

In our Research, a novel POS model of tagging Hausa sentences using HMM was presented. We have built a corpus of over 3000 words for the experiment. The corpus was manually tagged using the eight-part of speech as used by Bergery (1934) and M.K.M. Galadanci (1976) with the addition of Tense Marker (TM) and number to build the corpus. We Evaluate the Accuracy of Each POS as verified by an expert in the field of Hausa linguistics, ADJ has the highest 100%, while CON has low accuracy of 50%, the remaining tags perform within the range of 0.7-0.8 approximately. However, lack of enough data set and incorrect tagging has contributed to the low performance of some POS Tags Accuracies. The future work will be to increase the dataset for training in other to increase the Accuracy. Another work can include the breaking down of the POS tags into its Smaller Tag (entities) like the six categories of Verb.

REFERENCES

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)* (p. 28-29). Greater Noida, UP, India: IEEE.

Bar-Haim, R., Silam, K., & Winter, Y. (2008). Part-of-speech tagging of Modern Hebrew text. *Natural Language Engineering, 14*(2), 223-51.

Bashir, M., Rozaimee, A., & Isa, W. W. (2017). Automatic Hausa language text summarization based on feature extraction using Naive Bayes Model. *World Applied Science Journal, 35*(9), 2074-2080.

Bernard. (1991). Tagging English text with a probabilistic model. *IEEE Internation Conference on Acoustic, Speech and signal processing. 20.* Toronto, Canada: IEEE Internation Conference on Acoustic, Speech and signal processing.

Bimba, A., Idris, N., Khamis, N., & Mohd Noor, N. F. (2015). Stemming Hausa text: Using affix-stripping rules and reference look-up. *Springer*.

Brill, R. (1992). A simple rule based part of speech tagger. *In proceeding of the third conference on Applied natural Language Processings* (p. 152-155). Association of Computational Linguistics.

Caron, B. (2015). *Hausa Grammatical Sketch.*

Cutting, D., Kupiec, J., Pederson, J., & Penelope, S. (1992). A practical part of speech tagger. *3rd Conference On applied Natural Language Processing.*, (p. 133-140).

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.

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.

Jabir, S. A. (2015). Comparative study of English and Hausa nominal phrases. Sokoto.

Jelinek, F., Lafferty, J., Magerman, D. M., Mercer, R. L., Ratnaparkhi, A., & Roukos, S. (1994). Decision tree parsing using a Hidden Derivational Model. *In Proceedings of the Human Language Technology Workshop*, (p. 272-277). Plainsboro, New Jersy.

Kupiec, J. (1992). Robust part-of-speech tagging using a hidden Markov model. *Computer Speech & Language, 6*(3), 225-242.

Li, S., Graca, J. V., & Taskar, B. (2012). Wiki-ly supervised part-of-speech tagging. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural* (p. 1389-1398). Jejju Kasland, Korea: Association for Computational Linguistics.

Maitamaa, J. Z., Haruna, U., Gambo, Y., Thomas, B. A., Binti, N., yau, I. A., & Abubakar, A. I. (2014). *In the 5th Interenation* , (p. 1-4).

Marquez, L. (1999). *Part of speech tagging: A machine learning approach based on decision trees.* Department of Languages, University of Catalunya.

Marquez, L., Rodriguez, H., & Carmona, J. (1997). Improving POS tagging using Machine-Learning Techniques.

Merialdo, B. (1994). Tagging English text with a probabilistic model. *Computational Linguistics, 20*(2), 155-172.

Newman, P. (2000). *The Hausa language: An encyclopedic reference grammar.* New Haven: Yale University Press.

Newman, P. (2007). *A Hausa-English dictionary.* New Haven: Yale University Press.

Newman, P., Newman, R. M., Yaro, I., & Dresel, L. (1979). *Modern Hausa-English Dictionary.* Ibadan: University Press PLC.

Newman, R., & Newman, P. (2001). The Hausa lexicographic tradition. Lexikos. *Lexikos, 11*(1), 263-286.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech-tagging. *In Conference on Empirical Methods in Natural Language Processing.*

Salifou, L., & Naroua, H. (2014). Design of A spell corrector for hausa language. *international Journal of Computational Linguistics(IJCL), 5(2)*, pp. 14-26.

Schuh, R. G. (2019, 19 07). A Hausa story and Hausa verb morphology UCLA.

Toutanova, K., & Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy Part-of-speech Tagger. *In proceeding of the 2000 joint SIGDAT conference on Empirical Method in Natural Language Processing and very large corpora: Held in conjunction with the 38th annual meeting of the Association for Computational Linguistics. 13*, pp. 63-70. Association of Computational Linguistics.

Weischedel, R., Meteer, M., Schwartz, R., Ramshaw, L., & Palmucci, J. (1993). Coping with ambiguity and unknown words through probabilistic models. *Computation Linguistics, 19*(2), 361-382.

Yousif, J.H. (2019). Hidden Markov Model tagger for applications based Arabic text: A review. *International Journal of Computation and Applied Sciences IJOCAAS, 7*(1).