International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 10, No.1 (Apr-2021)

http://dx.doi.org/10.12785/ijcds/100146

# A Corpus Based Transformation-Based Learning for Hausa Text Parts of Speech Tagging

**Jamilu Awwalu[1], Saleh Elyakub Abdullahi[2] and Abraham Eseoghene Evwiekpaefe[3]**

*[1,3] Department of Computer Science, Nigerian Defence Academy, Kaduna, Nigeria*
*[2]Department, of Computer Science, Nile University of Nigeria, Abuja, Nigeria*

*Received 3 Apr. 2020, Revised 3 Feb. 2021, Accepted 26 Feb. 2021, Published 21 Apr. 2021*

**Abstract:** Parts of Speech tagging also known as POS tagging is a division under semantic analysis in Natural Language Processing. It has been an active research area for a very long time especially for languages such as; English, Arabic, Mandarin, Czech, Bahasa Melayu, Wolof, and Igbo. Hausa language belongs to the West Chadic languages, it is spoken in parts of several countries such as; Nigeria, Niger, Benin, Cameroun, Chad, Burkina Faso, Sudan, Congo, Ghana, and Togo. However, despite all these wide number of speakers, the Hausa language lacks Natural Language Processing (NLP) resources such as POS taggers. This limits NLP research such as Information Retrieval, Machine Translation, and Word Sense Disambiguation on Hausa language. Different Machine Learning (ML) approaches have yielded varying performance in POS tagging, thus indicating the critical role ML approach plays on performance of POS taggers. In this study, we create a Hausa language POS tagset, called Hausa tagset (HTS), apply Transformation-based Learning as a hybrid tagger, Hidden Markov Model (HMM) and N-Gram as probabilistic taggers to perform POS tagging of the Hausa language. Results on taggers testing based on precision, recall, and f1-measure from this study shows that TBL tagger scored 64%, 52%, and 53% outperforming the HMM tagger which scored 55%, 7%, and 5%. Comparing TBL with the N-gram taggers, the TBL and Unigram taggers achieved 53% f1-measure while the bigram and the Trigram taggers achieved 52%. On recall, the TBL achieved 6% more than the Unigram and 7% more than the Bigram and Trigram. In terms of precision, the TBL scored lowest compared to the N-gram taggers by scoring 64%, while the Unigram tagger achieved 70%, followed by the Bigram and Trigram both scoring 69%. Although the TBL tagger majorly outperformed other (i.e. HMM, Unigram, Bigram, Trigram) taggers on all evaluation metrics except for Unigram precision, both TBL and Unigram tagger achieved same level of f1-measure, and differ on precision, and recall with a balanced difference as TBL exceeded the Unigram tagger by 6% on recall while the Unigram tagger exceeded the TBL tagger by 6% also on precision.

**Keywords:** Parts of Speech Tagging, Hausa, Transformation Based Learning, Machine Learning, Hidden Markov Model, N-gram

## 1. Introduction

Machines are made to think and execute tasks like humans, they also aid humans in analyzing huge amount of data that is almost impossible for humans to analyze without the aid of computers. Artificial Intelligence (AI) has gained much attention with advancements made on different areas such as its role in supporting the building of intelligence machines. It has enhanced human lives in different aspects and improved the manufacturing and service industry over the past two decades [1]. Such areas include Computer Vision and Scene Recognition, Natural Language Processing (NLP), Neural Computing and Expert Systems [1]. Human Language Technology (HLT) consists of areas such as, Speech Recognition, Machine Translation, Text Generation, and Text Mining [2]. NLP is a field in computing that focusses on developing systems that allow communication between computers and people using their everyday language [3]. The idea of making computers process and understand human language is as old as the computers themselves, this idea has been called using different names such as language and speech processing, human language technology, natural language processing, computational linguistics, and speech recognition and synthesis [4].

NLP has been an active research area for several decades with advancements that can be grouped into series of activities. These activities include syntactic analysis, morphological and lexical analysis, discourse integration, pragmatic analysis, and semantic analysis [5]. Parts of Speech (POS) Tagging according to [6] is the assignment of a part of speech to each given word within a sentence. It works by assigning POS label to the different words contained in a text [7]. Parts of speech include verbs, nouns, adjectives, adverbs, conjunction pronouns and their sub-categories. Parts of speech

*E-mail address: awachi.jami@nda.edu.ng, saleh.abdullahi@nileuniversity.edu.ng, aeevwiekpaefe@nda.edu.ng*

tagging is a part of semantic analysis that has evolved with various machine learning approaches proposed over decades of heavy research to solve the problem of tagging rich-resource languages [8]. POS tagging requires large annotated corpora or linguistically motivated rules [9]. POS tagging approaches according to [10] are in three categories; statistical, rule based, and hybrid tagging. The rule based tagging uses set of rules that are applied along with the contextual information used to assign POS tags.

Transformation Based Learning (TBL) for POS tagging is built on Brill's tagger which according to [11] is an advanced method of the rule based approach. According to [12] TBL is a supervised Machine Learning algorithm that generates set of transformation rules, which correct classification mistakes of a baseline classifier.

Hausa is a West-Chadic language that is spoken mainly in northern Nigeria, Niger, Benin Ghana, Cameroun, Congo, Burkina Faso, Sudan, Chad, Togo and other parts of northern Africa. More than 85% of people who speak the 195 available Chadic languages speak Hausa [13]. This is the possible reason why international news corporations across Europe, America and Asia are broadcasting news in Hausa language. Also, Facebook added Hausa to its list of available languages users can view and use on its website. However, despite the wide number of speakers and advancements in Hausa Text Summarization by [14] and Text Normalization by [15] there is no Hausa language POS tagset and Parts of Speech (POS) Tagger implemented mainly on a Hausa POS tagset at the time of this research.

Different machine learning approaches exists for POS tagging using annotated corpus. These approaches could be supervised, semi-supervised, or unsupervised. Based on a selected approach, further modelling techniques are applied for accomplishing POS tagging. According to [16], computers analyze natural language in two basic approaches i.e. the corpus based approach, and the corpus driven approach. In the corpus-driven (CD), the corpus serves the empirical background from which lexicographers detect linguistic phenomena and extract data with no prior expectations or assumptions. The corpus-based (CB) approach uses an underlying corpus as an inventory of language data that is split into the training and testing data. This research explores the implementation of a Hausa tagset HMM, N-Gram, and Transformation Based Learning POS Tagger for Hausa language using the CB approach.

Although contributions have been made on Hausa NLP resources such as text normalization [15], summarization [14], verb conjugator by African Language Technology (AFLAT) named Verbix, no NLP resource such as Hausa POS Tagset for analyzing correspondences between surface form and lexical forms of words classes for Hausa language at the time of writing this report. Also, on Parts of Speech (POS) tagging, [17] contributed a parameter file for Hausa Language that can be used with the TreeTagger developed for English language by [18] which has been used for several languages by creating parameter files for Czech, Danish, Bulgarian, and English Languages. This TreeTragger was mainly developed for the English language, and is a rule based system on Decision Tree (DT). However, rule based linguistic systems as stated by [19] are limited by problems of over-specification, under-specification, and noise in rules. The following factors further complicate rule specification and limit the TreeTagger parameter file for Hausa POS tagging:

i. Hausa is a highly inflectional and morphologically-rich language.

ii. Lexical, syntactic, semantic, discourse, and pragmatic ambiguities are high in Hausa language.

This study proposes TBL for POS tagging because it combines probabilistic features and rule templates [20]. With the combination of rules and probability, this study aims to overcome limitations of rule or probabilistic based taggers such as HMM and N-Gram in POS tagging, and answer the following Research Questions (RQ):

**RQ1**: What is the possible set of TBL transformation rules for tagging Hausa POS.

**Approach**: Development of set of transformation rules that define POS tags based on the following (a) current, first previous, second previous, third previous words and tags. (b) The first next, second next, third next words and tags.

**RQ2**: What is the difference in terms of performance of generative and hybrid taggers in Hausa language text tagging.

**Approach**: (a) Implement TBL as a hybrid tagger, HMM and N-gram as generative taggers on Hausa language. (b) Evaluate and compare results to determine performance of the different tagging approaches.

**RQ3:** What is the impact of adopting corpus driven approach for TBL POS tagging of Hausa language

**Approach:** (a) Implement corpus driven tagging approach for Hausa language. (b) Assess performance of results from step a.

## 2. RELATED WORK

Transformation-Based Learning (TBL) has been successfully used in solving various Natural Language Processing (NLP) problems [21] such as Parts of Speech Tagging, sentence boundary disambiguation, and dialogue act tagging [22]. Transformation Based Learning (TBL) is built on Brill's tagger which according to [11] is an advanced method of the rule based approach. It repeatedly stacks rules on top of each other to improve accuracy and performance. TBL according to [12] is a supervised Machine Learning algorithm that generates set of transformation rules, which correct classification mistakes of a baseline classifier.

### A. *Classical Transformation Based Learning (TBL)*

The main functionality of the TBL is composed of a three steps process; the initial annotator, the templates, and the scoring function [20]. The implementation of classical TBL is shown in Fig. 1 while Table I shows related research based on the classical TBL implementation.

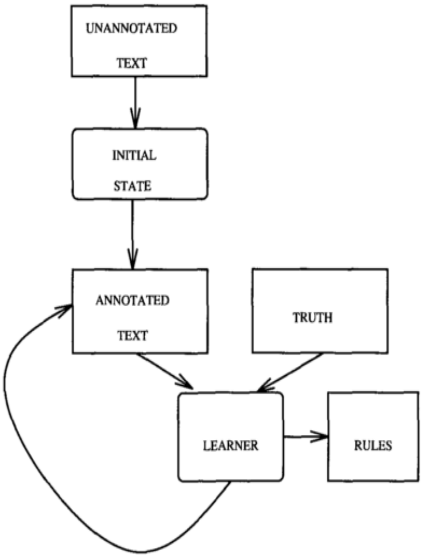


Figure 1. Transformation Based Learning. [20]

TABLE I. RELATED RESEARCHES BASED ON CLASSICAL TBL

| Technique | Description | Authors |
|---|---|---|
| Posting Act Tagging Using TBL | Extended traditional approaches used in dialogue act tagging and POS tagging by incorporating regular expressions into rule templates used by TBL. | [22] |
| POS-Tagger for English-Vietnamese Bilingual Corpus | Used TBL method and SUSANNE training corpus to train English POS-tagger. | [23] |

### B. *Modified Transformation Based Learning (TBL)*

The basic implementation of the TBL algorithm has been modified to fit specific requirements or to improve its performance for particular tasks. These modifications are summarized in Table II.

TABLE II. MODIFIED VERSIONS OF TBL

| Technique | Description |
|---|---|
| TBL in the Fast Lane by [21] | Due to the long time taken in training TBL on large datasets, the study by [21], proposed an approach of minimizing the training time of TBL without affecting its performance. Findings on the performance of the proposed approach reveal significant reduction in training duration while the same performance as the standard TBL is maintained. |
| Robust TBL using Ripple Down Rules for POS Tagging by [24] | Based on stored POS tagging rules, [24] proposed the modification of TBL where incremental knowledge is acquired. New rules are added only in order to correct errors discovered in the existing ones. This approach according to [24] allows the control of rules interaction in a systematic way. Results from conducted experiments on 13 languages reveals improvement in terms of speed when training the tagger and competitive accuracy was obtaining when compared with state-of-the-art morphological and POS taggers. |
| Entropy Guided TBL by [12] | The Entropy Guided TBL proposed by [12] is a combination of TBL with the characteristics of Decision Tree's (DT) feature selection. Based on this combination, POS tagging transformation rules are generated. These rules eliminate the need for a domain expert to build TBL transformation templates, and the rules are more efficient that those generated by DT only. Result from evaluation on three linguistics tasks in English and Portuguese shows the proposed method i.e. Entropy Guided TBL performs better than TBL (with hand-written rules) and DT. |
| Relational TBL by [25] | Motivated by the similarity between human activity recognition tasks and tagging process in NLP, [25] developed a relational TBL tagging system. This was based on the principles of inductive programming logic so as to make the proposed approach able to cope with background theory and relational representations. |
| Adaptive TBL for Improving Dictionary Tagging by [26] | The study by [26] proposed the modification of TBL in POS tagging by employing an adaptive approach which allows users produce dictionary of high quality that is parsed into lexicographic units such as pronunciations, headwords, translations, and parts of speech. This is done using little amount of data used during training. Results from conducted experiments on two dictionaries reveals a rise in tokens or individual words tagging accuracy from 83% to 93%, and from 91% to 94%, also on contiguous phrases from 64% to 90%, and 83% to 93%. |

## 3. METHODOLOGY

This section discusses our methodology as illustrated in Fig. 2 based on data collection and corpus formation, Transformation Based Learning Tagging, Hidden Markov Model Tagging, N-Gram Model Tagging, Words categorization using Experts Advisory Group on Language Engineering Standards (EAGLES) guideline on morpho-syntactic annotation, and Performance evaluation methods.

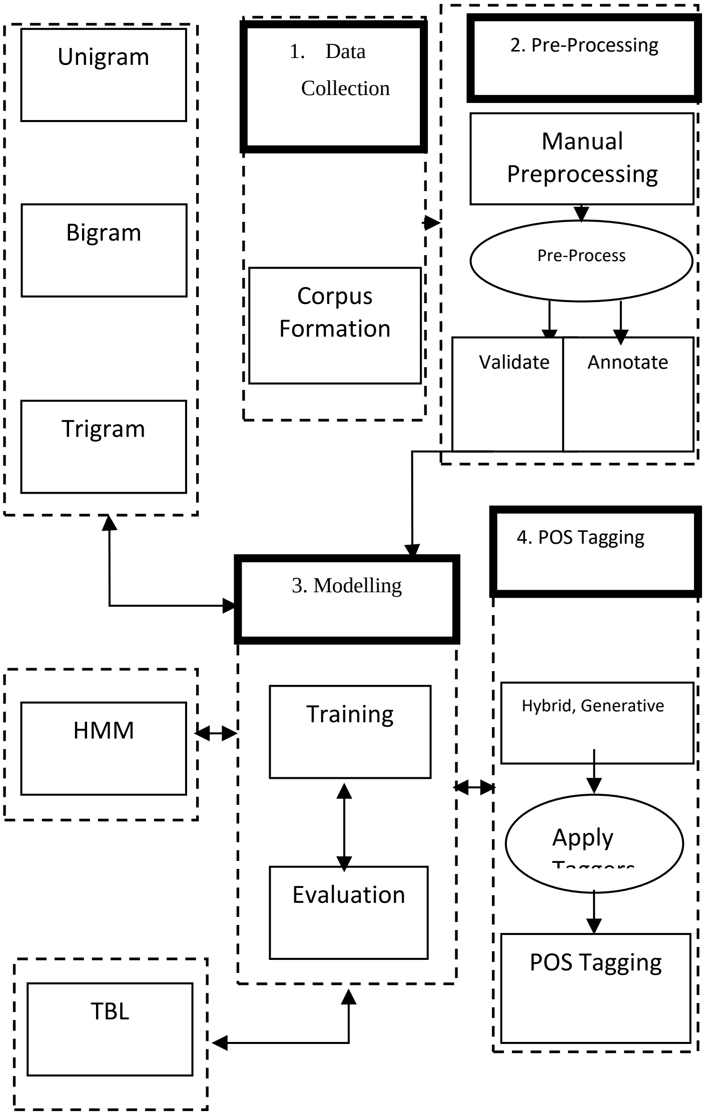


Figure 2. Workflow of Methodology

### A. *Cross Industry Standard Protocol for Data Mining (CRISP-DM)*

The CRISP-DM as a formal methodology is well developed and has been applied in several knowledge discovery researches [27]. We adopt it for the implementation of our research design as shown in Fig. 3. Although the data mining aspects of CRISP-DM addresses the process of pattern discovery in data entities, in our POS tagging work our interest is to find and identify POS classes in text which is linked to data mining on four out of the six steps of CRSIP-DM methodology.

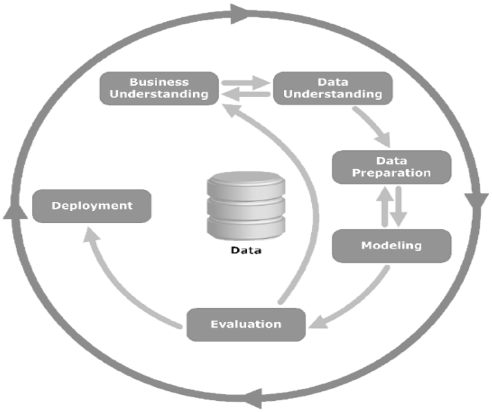


Figure 3. CRISP-DM Stages [28]

They are: data understanding, data preparation, modelling, and evaluation. Therefore, the use of the CRISP-DM methodology would aid in providing a structural approach to POS tagging. The description of our four selected processes from figure 3 is presented below:

- Data understanding: This phase marks the initial process of defining the data to be used for this study, and then proceed with activities that enable us to make use of the collected data. The data definition for this study is written Hausa text that is based on interviews and news narration from newspapers.
- Data preparation: In this phase, we carry out data pre-processing activities. These activities are related to cleansing and integrity check needed for the construction of the final corpus from the initial raw data collected. Our aim at this point is to check for data quality and any associated problems with the data in order to have a first insight into the collected data. While outliers can be considered anomalies or noise and thus castoff in data mining, they would remain in our corpus as they are not our focus and they also belong to classes of POS. The outcome of this phase is the Hausa Corpus, which is used to create Hausa language POS tagset (HTS). Some of the activities are (i) Data transformation. Our activities in this phase involve syntactic modifications applied to the collected data. Stemming is one of the activities where we would transform words into normalized forms. (ii) Data exploration and visualisation. In this stage, we focus on data distribution using relevant graphical tools in order visualise the structure of the collected and transformed data. The process of data quality check is shown in Fig. 4, while the different pre-processing activities are shown in Fig. 5.

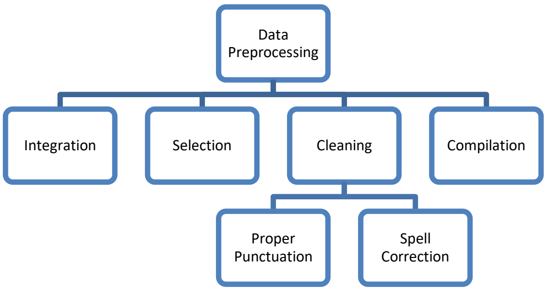


Figure 4. Preprocessing Activities on Hausa Corpus

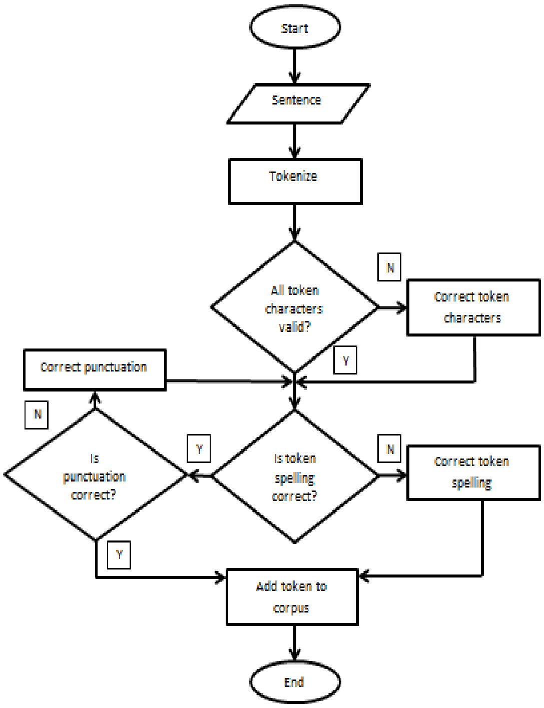


Figure 5. Process of data quality checking for corpus formation

The details of the different manual preprocessing activities as shown in figure 4 are: (a) Integration: To integrate the collected data from different news pages and sources to form one corpus. (b) Selection: To select the contents from the corpus one sentence at a time for inspection and cleaning. (c) Cleaning: correcting spelling, punctuation, and invalid characters. (d) Compilation: Compiling all cleaned words to form a preprocessed corpus.

- Modelling. In this phase, we identify all entities and the relationship between them. The modelling generally uses algorithms. It is historically rooted in mathematics, numerical analysis, and statistics. However, for complex data, there are different techniques that are used such as statistical, nearest neighbour, classification and information or context based approaches. In this study we use TBL, HMM, and G-Gram to model our POS tagger. This process of modelling is the training stage for our tagger as shown in figure 4. The modelling would be based on the TBL, HMM, and N-gram implementation from Natural Language Toolkit (NLTK).
- Evaluation: In this phase, we focus on evaluating the modelling process from the previous phase i.e. Data Modelling as shown in Fig. 6. The evaluation process is the testing phase where we evaluate the performance of the taggers on the HTS. This process continues hand in hand with the modelling from phase iii until a reliable and satisfactory performance measure is achieved. SCIKIT learn machine learning library would be used for the evaluation of the tagging process by the TBL, HMM, and N-gram taggers.

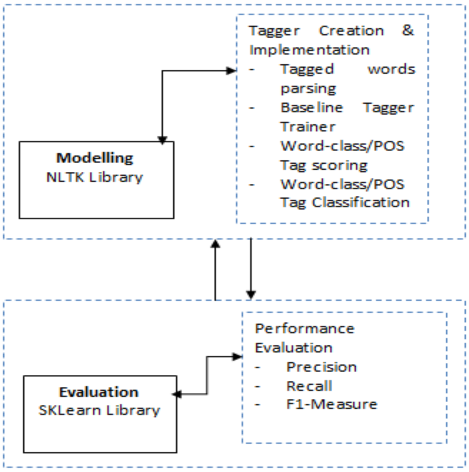

Modelling
NLTK Library
Tagger Creation & Implementation
- Tagged words parsing
- Base line Tagger Trainer
- Word-class/POS Tag scoring
- Word-class/POS Tag Classification
Evaluation
SKLearn Library
Performance Evaluation
- Precision
- Recall
- F1-Measure

Figure 6. Modelling and Evaluation

### B. *Data Collection and Corpus Formation*

Data to be used for corpus construction in this study is grouped into two parts in order to have a well triangulated and balance corpus.

- First part is collected from news media which would come in digital format. Therefore, requesting a digital news transcript would save the time of converting the paper records to digital or soft copy.
- Second part, Hausa literary documents would be obtained and transcribed.

The pre-processing includes tasks of formatting to standard Hausa text issues like abbreviated or slang words. The outcome of this task would provide us with a clean text available for POS tag labelling as the next step in the framework.

### *C. HTS Tagset Design*

According to [29] for most researchers that are developing taggers and part-of-speech tagsets, the traditional categorization of grammar: noun, adjective, verb, preposition, adverb, pronoun, and interjection and conjunction is not enough. There may be need to tag other grammatical terms, such as morphological subcategories which includes person for verbs and number for nouns and tense, and/or syntactic

The EAGLES tagset design guidelines provide a framework that is flexible which in theory comprises everything the tagset designer would desire to mark up, and at the same time without restricting the tagset designer's choice. It also discourages "reinventing the wheel" while promoting consistency and linguistic resources reusability for different languages [30].

In categorizing words according to EAGLES standard, the following constraints as stated by [31] are recognized in the word categories description by the means of morphosyntactic tags:

- Obligatory: the main parts of speech such as noun, conjunction, and verb belong here, as obligatorily specified.
- Recommended: These are the grammatical classes that fall under standard conventional grammatical descriptions such as number, gender, and person.
- Special extensions: Divided into two, It contains (a) Language specific attributes or values, and (b) Generic attributes or values. However, In practice, generic and language-specific features cannot be clearly distinguished [31].

Considering the above constraints, we categorized words from the collected corpus into different word classes or POS. The Penn TreeBank POS Tagset design is adopted for this study. It is a standard Tagset that is being widely used for tagging process. The Penn TreeBank contains 36 different word tags as shown in Table 3. However, the adoption of the TreeBank was not for all the tags it contains, it is for tags that are available in the HTS. Other tags introduced in the study as shown in Table 4 were carefully included based on the defined standards of Tagset development by [29].

TABLE III. LIST OF POS TAGS IN THE PENN TREEBANK PROJECT

| **Tag** | **Tag** | **Tag** | **Tag** |
|---|---|---|---|
| CC | MD | RBR | LS |
| CD | NN | RBS | VBP |
| DT | NNS | RP | WP |
| EX | NNP | SYM | RB |
| FW | NNPS | TO | VBZ |
| IN | PDT | UH | WP$ |
| JJ | POS | VB | VBN |
| JJR | PRP | VBD | WDT |
| JJS | PRP$ | VBG | WRB |

TABLE IV. INTRODUCED WORD TAG

| **Tag** |
|---|
| NM |

Following the TreeBank design, the corpus was annotated as part of the pre-processing activities for the implementation of this study. The summary of HTS design is presented in Table 5.

TABLE V. HTS DESIGN SUMMARY

| **S/No.** | **Design Aspect** | **Status Description** |
|---|---|---|
| 1 | Purpose | Public purpose: to create a tagset to serve as a resource for research in Hausa NLP. |
| 2 | Size | 3,000 words |
| 3 | Language | Hausa – Kananci dialect |
| 4 | NLP Context | POS |
| 5 | Availability | i. Freely available at https://github.com/HausaNLPResearch/Hausa-POS-Tagset-HTS-<br>ii. Downloadable |
| 6 | Data Materials | Written words |
| 7 | Materials Genre | i. Discussion<br>ii. Narrative |
| 8 | Annotation | Entire corpus tagged for POS |

### *D. Tagger Modelling*

The algorithms for modelling the taggers are Transformation Based Learning, Hidden Markov Model, and N-gram Models.

- Transformation Based Learning

This section discusses how this study practically applies TBL for POS tagging on HTS. Some of the benefits involved in using TBL are (i) Small set of simple rules are learnt that can be enough for tagging. (ii) Development as well as debugging is very easy in TBL because the learned rules are easy to understand. (iii) Complexity in tagging is reduced because in TBL there is interlacing of machine learned and human-generated rules. (iv)Transformation-based tagger is fast.

There are different approaches to implementing TBL which includes statistical techniques or simply annotation of all words with a particular or random POS tag. In this study we take the statistical approach which employs n-gram taggers in annotating the initial input i.e. unannotated text. The statistical approach uses maximum likelihood estimate (MLE) on n-gram taggers (i.e., Bi-gram, and n-gram). The appropriate tag in this approach is determined by applying Equation (1) to the targeted text and as shown in Equation (2) computation for implementation.

$$P\left(t_i \mid w_i\right) \tag{1}$$

$$P\left(t_i \mid w_i\right) \approx \frac{C(t,w)}{C(w)} \tag{2}$$

As shown in the above equations, the number of times a word *(w)* appears in the training set is denoted as *C(w)*, while *C(ti, wi)* denotes number of times word w appears with tag t. Equation (3) shows the Bigram computation process of MLE as detailed in Equation (4) where *C(ti-1, wi)* is count of times word *wi* appears with *tagi-1*.

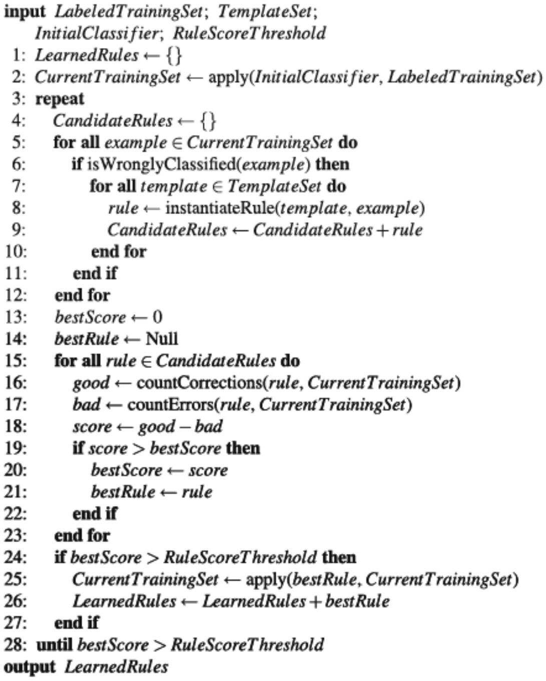
$$P\left(t \mid t,w\right) \approx \frac{C(tt,w)}{C(t,w)} \tag{3}$$

$$t = \arg\max P\left(t \mid t,w\right) \tag{4}$$

Due to out of vocabulary (OOV) occurrence in statistical n-gram models, which limits it performance, we improvise by implementing back-off. The back off implementation is done where when it fails in determining the initial annotation for a word, we back-off to bigram, then trigram.

The temporary Tagset is the tagged data from the Initial State Annotator (ISA). This temporary Tagset is then compared with the manually annotated Tagset in order to check the correctness of the output of tagged data by comparing it with the output of the initial annotation carried out by the initial state annotator. Then, the annotated text becomes the input to the Learner where a list of transformations in a particular order are learned, then applied to the generated output of ISA to make it as close to truth as possible.

The complete process of our TBL training as shown in Fig. 7 is based on the following steps: (i) Iteratively instantiating transformation rule templates in the training dataset (ii) Using the number of positive and negative counts to score the rules. (iii) Choosing the rule with the highest score then applying it to the training dataset. (iv) Ending the learning process when the rule with the highest score fails to meet the thresholds.



```
input LabeledTrainingSet; TemplateSet;
    InitialClassifier; RuleScoreThreshold
LearnedRules ← {}
CurrentTrainingSet ← apply(InitialClassifier, LabeledTrainingSet)
repeat
    CandidateRules ← {}
    for all example ∈ CurrentTrainingSet do
        if isWronglyClassified(example) then
            for all template ∈ TemplateSet do
                rule ← instantiateRule(template, example)
                CandidateRules ← CandidateRules + rule
            end for
        end if
    end for
    bestScore ← 0
    bestRule ← Null
    for all rule ∈ CandidateRules do
        good ← countCorrections(rule, CurrentTrainingSet)
        bad ← countErrors(rule, CurrentTrainingSet)
        score ← good − bad
        if score > bestScore then
            bestScore ← score
            bestRule ← rule
        end if
    end for
    if bestScore > RuleScoreThreshold then
        CurrentTrainingSet ← apply(bestRule, CurrentTrainingSet)
        LearnedRules ← LearnedRules + bestRule
    end if
until bestScore > RuleScoreThreshold
output LearnedRules
```

Figure 7. Pseudocode for learning transformations [12]

- Hidden Markov Model (HMM)
  As a probabilistic tagger, the HMM tagger is basically made of two components where A shows probabilities of transitions while B shows the observation likelihood. Component A: This contains the probability of the word tag transition as shown in Equation (5), this is the probability of a word tag occurring considering the immediate previous tag.

$$P\left(t_i \mid t_{i-1}\right) \tag{5}$$

$$P\left(t_i \mid t_{i-1}\right) = \frac{C\left(t_i \mid t_{i-1}\right)}{C\left(t_{i-1}\right)} \tag{6}$$

This is computed as shown in Equation (6) by counting the times a tag occurs first in labelled corpus and how frequently it is followed by the second tag. An example in Equation (7) calculates the Maximum Likelihood Estimate (MLE) of a VB tag following an MD tag.

$$P(VB \mid MD) = \frac{C(MD,VB)}{C(MD)} = \frac{251}{301} = 0.83 \tag{7}$$

As illustrated in Equation (7), the MD occurred 301 times followed by the VB tag 251 times, hence the MLE of 0.83.

Component B: This contains the probability of given a tag associating with a particular word as shown in Equation (8). An example in Equation (9) calculates the MLE an MD associating with the word will.

$$P(w_i \mid t_i) = \frac{C(t_i \mid w_i)}{C(t_i)} \qquad (8)$$

As illustrated Equation (9) where it shows the MLE to the question of "If we were going to generate a VB, how likely is it that this verb is samu?"

$$P(samu \mid VB) = \frac{C(VB, samu)}{C(VB)} = \frac{277}{1267} = 021 \qquad (9)$$

- N-Gram Taggers

  The N-gram is based on statistical tagging algorithm. The modelling for the N-gram taggers in this study is based on three taggers i.e. Unigram, Bigram, and Trigram taggers as shown in Fig. 8.

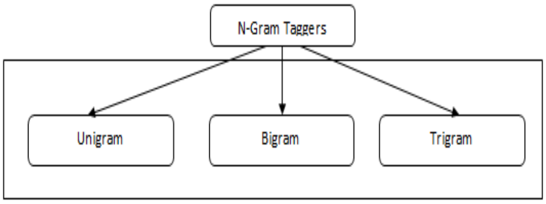


Figure 8. N-gram taggers

In the Unigram tagging i.e. n = 1where n is the number of token; we assign the most likely tag using statistical approach. An example is assigning the noun tag 'NN' to any occurrence of the word 'dama = right', since 'dama' is used as a NN such 'dama = right' more often than it is used as an adjective such as 'dama = many' or verb such as 'dama = mix' . Before the N-gram tagger can be used in tagging process, it must go through training on a corpus in order to determine the tags which are mostly associated to each word. The default tagger is used as the back-off in assigning the tag 'NN' to words it encounters, this means that the back-off tagger is left with the responsibility of tagging unknown words to the Unigram Tagger in the dataset for training as 'NN' that is. The Bigram i.e. n = 2 and the Trigram i.e. n = 3 works in the same manner as the Unigram in terms of backing-off, the only difference for the Bigram and Trigram taggers is that they take the word context into consideration when tagging a particular word. In their (Bigram and Trigram) training, a frequency distribution is created. This frequency distribution describes the frequencies with which, every word is assigned a tag in dissimilar contexts. For the bigram, the context is made of two words i.e. the word which is to be assigned a tag and the assigned tag to the word before it. For Trigram, the context is comprises three words which are; the assigned tags to the two preceding words, and the word which is to be tagged.

When tagging, the tagger the n-gram taggers use frequency distribution to tag words by assigning the tag with the maximum frequency given the context to each word. When an unknown token is encountered i.e. a context for which no data has been learnt, the trigram tagger backs off to the bigram tagger, if same encounter occurs in bigram, the bigram tagger backs off to the unigram tagger. The general concept of N-gram computation is shown in Equation (10) and described in Table 6

$$P(w_n \mid w_1^{n-1}) \approx P(w_n \mid w_{n-N+1}^{n-1}) \qquad (10)$$

TABLE VI. TABULAR DESCRIPTION OF EQUATION 10

| Character | Description |
|---|---|
| *C* | Count of word occurrence |
| *P* | Probability or likelihood of given Word or 'W' to a class value |
| *W* | Word |
| *N* | Number of words |

## 4. Experiments and Results

In this section, the steps and details involved in the experiments conducted and results obtained are discussed.

### A. Experimental Setup

The setup of experiments conducted in this study is discussed.

- Data Description and Corpus Formation

  The data for HausaTagset (HTS) Corpus was collected from different Hausa online newspapers. The description of the data is presented in Table 7. One might argue that the bigger the training size, the better model. However this is far from correct as proven by [32].

TABLE VII. HTS CORPUS

| Title | Hausa Tagset (HTS) Corpus V1.0 |
|---|---|
| Tokens | 3093 |
| Format | Text Only |
| Genre | Interviews and News Narration |

- Tagset Implementation

  The Penn TreeBank POS Tagset was adopted for this study. It is a standard Tagset that is being widely used for tagging process. The Penn TreeBank contains 36 different word tags as previously shown in Table 3. However, the adoption of the TreeBank was not for all tags it contains, it was for tags that are available in the HTS Corpus. The complete tags for all tokens numbering 38 are shown in Table 8.

TABLE VIII. AVAILABLE TAGS IN HTS

| Tag | Tag | Tag | Tag | Tag |
|---|---|---|---|---|
| PRP | JJ | “ | WDT | WP |
| VB | . | VBZ | ( | RBR |
| NN | VBD | ; | ) | NPS |
| CC | NNS | PRP$ | NNPS | JJR |
| IN | VBG | EX | PDT | WRB |
| NP | POS | ? | ! | NM |
| , | RB | : | - | UH |
| RP | VBN | DT | | |

- Corpus Driven Approach

  With the goal of building a generic Hausa text POS tagger that has very little dependence on training corpus domain, the implementation would not require a huge volume of manually tagged or untagged Hausa corpus. The corpus driven approach was selected as the approach of tagging implementation. The corpus-driven (CD) approach is an approach where the corpus serve the empirical basis which lexicographers mine and detect linguistic phenomena with no prior expectations and assumptions. Claims or conclusions are exclusively drawn on the basis of observations from the corpus. On the other hand, the corpus-based (CB) approach makes use of corpus as an underlying inventory of language data from which the right components are fetched out to come up with intuitive knowledge. This helps in verifying expectations, allowing the quantification of linguistic phenomena, finding proof for current theories. It is an approach where corpus is interrogated and data is used to confirm linguistic pre-set explanations and assumptions. It acts, therefore, as additional supporting material.

## *B. Tagging Experimentation*

The tagging implementation was done in three levels. (i)Pre-processing (ii) Baseline tagger training (iii)POS tagging

The output from the first two levels serves as input to the third level where the actual POS tagging is conducted. The pre-processing level takes input in the form of text file, while the baseline tagger trainer takes in the annotated text and begins the process of training the tagger by splitting the text input into training and testing portions. The tagger accepts two inputs, one is pre-processed text and the other is trained output from the baseline model trainer. Fig. 9 shows the different steps involved in implementation of the tagger.

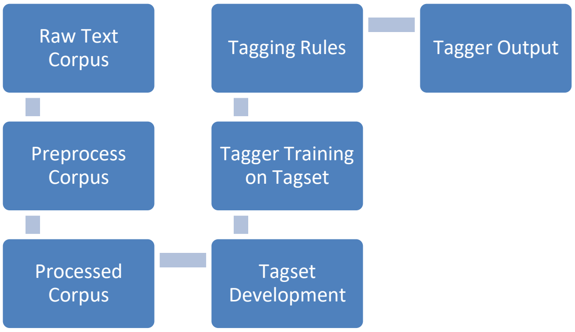


Figure 9. Tagging Experimentation

- Baseline Tagger Training

  The training of the baseline tagger was implemented in two stages. (i) The holdout approach was used in training the baseline tagger with 70:30 split. (ii) The implementation of the TBL tagging process based on the description of approach to answer research question 1 which is based on the classical TBL algorithm. (iii) Rules generated: Several rules were generated from the baseline tagger training. Some of the rules are shown in table 10 (iv) The baseline tagger accuracy was calculated from the test set which is 30% of the annotated HTS Corpus. This was determined by retrieving part of the manually annotated corpus containing different word classes. The reason for this was to have a very strong test set that can be certified with respect to the class it belongs to i.e. all possible word classes from the complete corpus. The accuracy of the tagger was calculated in terms of precision, recall, and f- measure as described in the subsequent sections. The result from the baseline tagger on POS tagging is obtained following the description and procedures in Fig 9. And a cross-section of generated rules shown in Table 9.

TABLE IX. GENERATED RULES

| Number | Rules |
|---|---|
| Rule 1 | IN->PRP if Pos:VB@[-1] & Pos:VB@[1] |
| Rule 2 | VB->VBD if Word:Na@[-1] & Word:aiki@[1] |
| Rule 3 | NN->VB if Pos:VBD@[-1] & Pos:IN@[1] |
| Rule 4 | POS->PRP if Word:za@[-1] |
| Rule 5 | DT->UH if Pos:VBG@[1] |
| Rule 6 | PDT->NN if Pos:EX@[1] |
| Rule 7 | VBN->IN if Pos:,@[1] |
| Rule 8 | DT->PDT if Pos:IN@[2] |
| Rule 9 | PDT->VBZ if Pos:POS@[-1] |
| Rule 10 | POS->PRP if Pos:VBZ@[1] |
| Rule 11 | RP->DT if Pos:?@[2] |
| Rule 12 | VB->VBD if Pos:VBZ@[2] |
| Rule 13 | VB->VBG if Pos:PDT@[2] |
| Rule 14 | CC->IN if Pos:RP@[-3,-2,-1] |
| Rule 15 | NN->VB if Pos:.@[-3,-2,-1] |
| Rule 16 | CC->DT if Pos:PRP@[-1] & Pos:NM@[1] |
| Rule 17 | CC->POS if Pos:VBG@[-1] & Pos:PRP@[1] |
| Rule 18 | IN->VB if Pos:PRP@[-1] & Pos:.@[1] |
| Rule 19 | NN->NP if Pos:IN@[-1] & Pos:IN@[1] |
| Rule 20 | NN->VB if Pos:VB@[-1] & Pos:DT@[1] |

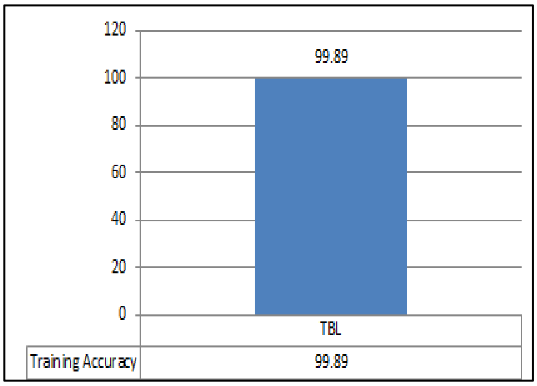


Figure 10. Baseline Tagger Accuracy

- Tagging Implementation

The baseline tagger experimented and evaluated from the previous section is applied on the corpus and tasked with tagging their respective POS classes. As the baseline tagger for our POS tagging implementation, we use the approach that, for all words as unannotated text, the words are passed through the initial-state annotator by assigning the output of our manually created annotator as n-gram tagger. Secondly, the result are compared to the truth (manually tagged corpus). The tagging outcome is shown in Fig. 11 while the expected outcome is shown in Fig. 12

```
[[('Malam', 'NN'), ('Musa', 'NN'), ('Ibrahim', 'NP'),
('shi', 'DT'), ('ne', 'DT'), ('Sarkin', 'NN'), ('garin',
'NN'), ('na', 'POS'), ('Dan', 'NP'), ('Bami', 'NN'),
('shima', 'NN'), ('ya', 'PRP'), ('tabbatar', 'VBD'),
('da', 'CC'), ('aukuwar', 'NN'), ('lamarin', 'NN'),
('harma', 'NN'), ('ya', 'PRP'), ('yi', 'VB'), ('kira',
'VB'), ('ga', 'IN'), ('matasa', 'NN'), ('da', 'CC'), ('su',
'PRP'), ('guji', 'NN'), ('daukar', 'VB'), ('doka', 'NN'),
('a', 'IN'), ('hannunsu/NN', 'NN'), ('tare', 'JJ'), ('da',
'CC'), ('jan', 'NN'), ('hankalin', 'NN'), ('iyaye', 'NN'),
('dasu/CC', 'NN'), ('kula', 'NN'), ('da', 'CC'),
('tarbiyar', 'NN'), ('yaransu', 'NN'), ('tare', 'JJ'),
('da', 'CC'), ('yi', 'VB'), ('masu', 'NN'), ('addu'a',
'NN'), ('a', 'IN'), ('dukkan', 'NN'), ('lokutta', 'NN'),
('.', '.')], [('Shiko', 'NN'), ('D.P.O', 'NN'), ('na',
'POS'), ('wannan', 'PRP'), ('shiyar', 'NN'),
('dake/CC', 'NN'), ('da', 'CC'), ('ofishi', 'NN'), ('a',
'IN'), ('garin/NN', 'DT'), ('Hunkuyi', 'NN')]
```

Figure 11. TBL outcome

Malam/NN Musa/NP Ibrahim/NP shi/PRP ne/DT Sarkin/NN garin/NN na/CC Dan/NN Bami/NN shima/PRP ya/PRP tabbatar/VBD da/CC aukuwar/VBD lamarin/VB harma/CC ya/PRP yi/VB kira/VB ga/CC matasa/NNS da/CC su/PRP guji/VBD daukar/VB doka/NN a/IN hannunsu/NN/POS tare/CC da/CC jan/VB hankalin/NN iyaye/NNS dasu/CC/PRP kula/VB da/CC tarbiyar/NN yaransu/NNS tare/CC da/CC yi/VB masu/PRP addu'a/NN a/IN dukkan/JJ lokutta/NNS ./. Shiko/CC D.P.O/NN na/CC wannan/PRP shiyar/NN dake/CC/VBG da/CC ofishi/NN a/IN garin/NN/DT Hunkuyi/NN

Figure 12. Expected Outcome

## C. *Performance Evaluation*

The performance of the TBL tagger implementation based on the baseline tagger is discussed in this section and shown in Fig. 13 and appendix 'A'.

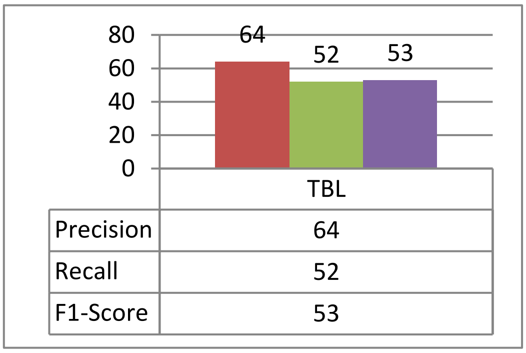
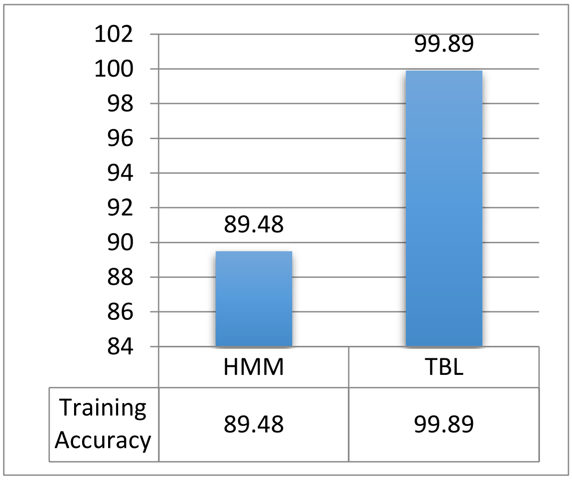


Figure 13. TBL Training Accuracy

### *D. Comparison of TBL with HMM and N-Gram Taggers*

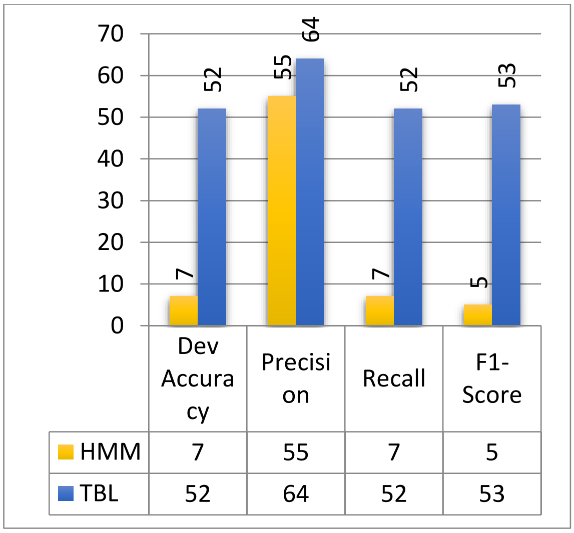
The tagger from this study based on TBL has shown good result from the evaluation conducted as elaborated in the previous section. To verify that, the accuracy of the TBL tagger is compared to that of the Hidden Markov tagger (HMM) and n-gram taggers on same corpus (i.e. HTS). The result of this comparison can be used to further validate the performance of the TBL tagger on the HTS corpus. The taggers from both taggers were compared in this section in order to have a comparative view of performance in tagging the HTS.

- Comparison of TBL with HMM tagger: The results were obtained separately and put together as discussed and shown in Fig. 14, 15, and appendix 'A'.



| | HMM | TBL |
|---|---|---|
| Training Accuracy | 89.48 | 99.89 |

Figure 14. TBL and HMM Training Accuracy

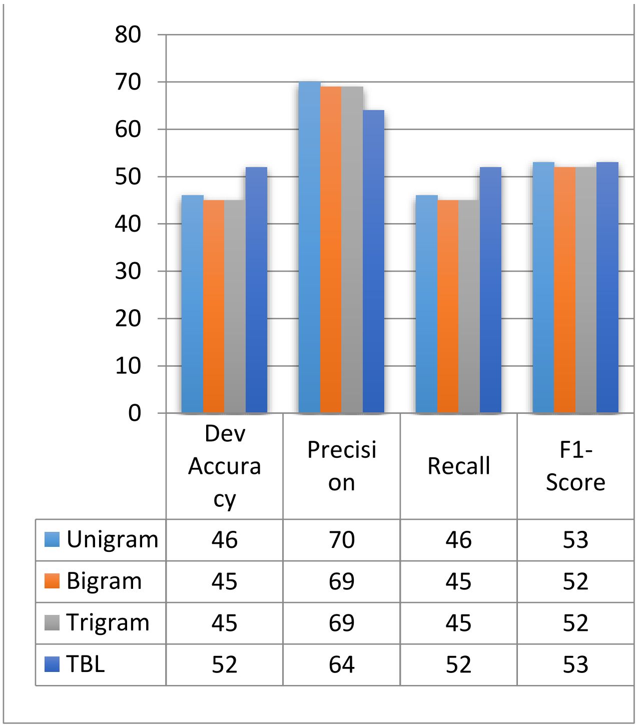


| | Dev Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| ■ HMM | 7 | 55 | 7 | 5 |
| ■ TBL | 52 | 64 | 52 | 53 |

Figure 15. TBL and HMM Implementation Accuracy

From the graph presented in Fig. 14 it can be seen that both TBL and HMM taggers achieved an impressively high training accuracy level i.e. 89.48% for HMM and 99.89% for the TBL tagger hence the TBL tagger achieving 10% more accuracy than the HMM tagger. On the development accuracy presented in figure 15 the TBL tagger consistently performed more than the HMM tagger. Unlike its performance in training, the HMM performed poorly on the development by achieving very low results in accuracy, recall, and the f1-measure except for precision where it achieved 55% while TBL achieved 64% making a difference of 9%. The HMM tagger confusion matrix shown in appendix 'A' contains information about actual and predicted tagging done by the HMM tagger. The performance of the tagger here is evaluated using the data in the matrix and the entries in the confusion matrix have the same interpretation process as explained in the discussion section of this study.

- Comparison of TBL with N-gram taggers: The taggers compared with the TBL tagger in this section are three n-gram taggers. The TBL tagger in this context is a hybrid tagger, while the n-gram taggers are generative taggers. The result of tagging performance on the development (i.e. the tagging implementation) is shown in Fig. 16 while their accuracy in terms of training accuracy on the HTS is presented in Fig. 17



| | Dev Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| ■ Unigram | 46 | 70 | 46 | 53 |
| ■ Bigram | 45 | 69 | 45 | 52 |
| ■ Trigram | 45 | 69 | 45 | 52 |
| ■ TBL | 52 | 64 | 52 | 53 |

Figure 16. TBL and N-Gram Implementation Accuracy

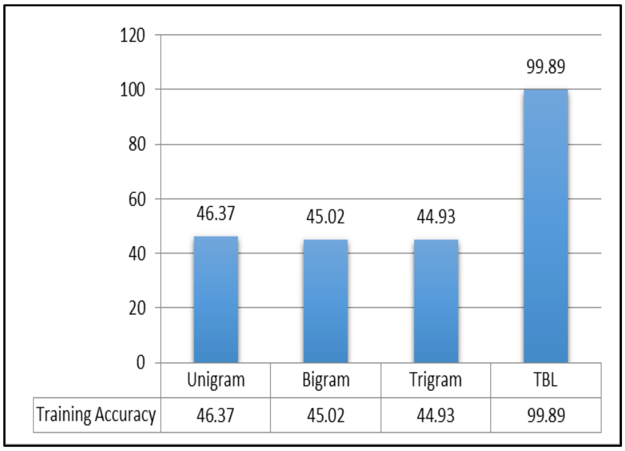


Figure 17. Training Accuracy of TBL and N-gram Taggers

From the graph presented in Fig. 17 it can be seen that all taggers achieved almost same level of f1-measure with the TBL and Unigram taggers achieving 53% while the bigram and the Trigram taggers achieved 52% leaving only a difference of 1% f1-measure. However in terms of recall and development accuracy, the TBL achieved 6% more than the Unigram and 7% more than the Bigram and Trigram. In terms of precision, the TBL achieved the lowest by scoring 64%, while the Unigram tagger achieved the highest score of 70%, followed by the Bigram and Trigram both scoring 69%.

## 5. Discussion

In this section, the results obtained from the tagging experimentation in the previous sections are discussed.

### *A. Experimental Results*

The different taggers results and strategies are presented and discussed in this section. For the TBL tagger, the supervised approach was used with the default back-off of unigram, bigram, and trigram.

The results from the tagging process produces both open and closed words classes. This includes words such as preposition as a closed word class in and noun as an open word class in the Hausa language. This did not affect performance of the tagging process because that was how the tagger was trained based on the open and closed word classes forming the different POS tags of the Hausa language as discussed in the previous section.

The results from the implementation of this tagger for parts of speech tagging were presented in the previous section i.e. 4c where a test of 30% of the HTS was used to evaluate the accuracy of the tagger. The test set of the HTS consist of all word classes available in the training. This was in order to (I) ensure a balanced training and test set was used, (II) eliminate tagger biasness on training or testing.

It was observed that the TBL tagger achieved a high accuracy level on the baseline tagger than the HMM and all n-gram taggers. This is not strange as the TBL tagger is a hybrid in the form of generative and discriminative taggers. Whereas the HMM and n-gram taggers are generative taggers, the TBL tagger boosts upon their weaknesses in the tagging process, thus performing better achieving better accuracy levels both in training and development (i.e. implementation).

Although the TBL tagger outperformed the other (i.e. HMM, Unigram, Bigram, Trigram) taggers, there was not much difference between it and the Unigram tagger. Just as they (i.e. TBL and Unigram tagger) achieved same level of f1-measure, and differ on precision, recall, and development accuracy, the difference is balanced as TBL exceeded the Unigram tagger by 6% on recall and development accuracy, while the Unigram tagger exceeded the TBL tagger by 6% also on precision.

As observed from the confusion matrices (see appendix A) of the different taggers (i.e. TBL, HMM, Unigram, Bigram, Trigram) experimented in this study, the word tag assignment by the different taggers take varying forms especially for the wrongly assigned tags.

The TBL tagger error was mostly on the NN tag where it assigned the NN for majority of the words wrongly. However, the overall error rate on all tags was not far with the highest tagging error found on VB tag as 7.4% and a cumulative of 9.5% which are both higher than the actual tagging accuracy on VB i.e. 6.5% but less than the highest achieved tag accuracy found on PRP tag as 9.6%.

Unlike the TBL tagger, the HMM tagger errors was mostly on the NP tags. The tagger assigned the NP tagger for all word tags. The error rate of word tags such as VB tagged as NP was higher (i.e. 15.5%) than the actual accurate tagging of the NP as NP which was 4.7%. Whereas the TBL tagger achieved its highest accurate tagging on PRP tag, the HMM PRP tag accuracy of 0.7% was very much below the tagging error of PRP on PRP. The cumulative tagging error of the PRP tag was 15.4% with the wrong tagging as NP contributing the bulk of the error i.e. 14.6%. Similarly the CC word tag cumulative error rate of 12.1% was by far over the accuracy of 0.5%. The bulk of the CC tagging error i.e. 11.9% was also contributed by the tagging of CC as NP just like PRP and VB tags.

The Unigram tagging error was more spread compared to both HMM and TBL but more concentrated on first 11 word classes. But more particularly, it wrongly tagged 60% of words classes to the NONE class. Also, it is in this NONE class that the highest word tag error was recorded which is 7.7% on reference to the VB tag. Similar tagging errors were observed on the bigram and trigram taggers where the Bigram tagger recorded exact figures on the confusion matrix as the unigram tagger. For the Trigram tagger, besides having similar performance behaviour with the Unigram and Bigram taggers in terms of sparse tagging error, it's major wrongly tagging on the NONE tag was also 60% but on varying tags not as the case for exact similarity on the NONE tagging error by the Unigram and Bigram taggers.

### B. Answers to Research Questions

In this section, the answers to questions of this research are provided and discussed. This is based on the outcome of experiments conducted in the previous section.

***RQ1:*** *What is the basic set of TBL transformation rules for tagging Hausa POS.*

The basic set of transformation ruled for TBL POS tagging as obtained from implementation of TBL on the HTS tagset are 39 rules. Although the number of rules may vary according to input parameters and tagset, in any case these rules can be reflective and common across tagsets. The generated rules from this study are listed in appendix b. These rules are not exhaustive for Hausa language POS tagging, which is why they are termed basic rules.

***RQ2:*** *What are the differences or similarities in terms of performance of generative and hybrid taggers in Hausa language tagging.*

The comparison between the three different taggers as observed from this study reveals that (a) on tagger training, both hybrid and generative taggers achieved an impressively high training accuracy level with hybrid tagger achieving highest training accuracy of 99.89% followed by the generative taggers with 89.48% for HMM and 46% for N-Gram. (b) On development i.e. the implementation of trained tagger on the HTS, accuracy performance was evaluated on precision, recall, and f1-measure. In terms of precision, the n-gram taggers achieved highest of 70%, followed by the hybrid tagger achieving 64%, then the HMM tagger with 55%. The n-gram taggers persist by achieving high f1-measure of 53% which is same as the f1-measure for the hybrid tagger but far higher than the HMM tagger which was 5%. However in terms of recall, the hybrid tagger achieved the highest score of 52% followed by the n-gram taggers achieving 46%, then the HMM tagger achieving 7%. This means that the hybrid tagger has the highest ratio of correct tagging on actual tags from the tagset, while the n-gram taggers have the highest ratio of correct tagging on total tagset, and both the hybrid and n-gram taggers achieving same weighted average of the precision and recall accuracy levels. This shows that the HMM tagger least performed compared to the other two types of taggers, while the hybrid and n-gram taggers have balanced accuracy.

***RQ3:*** *What is the impact of adopting corpus driven approach for TBL POS tagging of Hausa language*.

There have been arguments on corpus driven vs corpus based approaches to POS tagging. With this study implementation of POS tagging based on a data driven approach, results from the approach appears to be promising due to satisfactory state of the art accuracy level achieved as shown in the results section. Therefore, it can be concluded that the corpus driven approach impact on TBL POS tagging has a positive impact.

## 6. Conclusion

This study presented various research directions in the fields of POS tagging of Hausa language and Hausa language NLP resources ranging from the general introduction and its applicability, justification of the relevance of this study, the review of literature on POS tagging approaches and language resources, the formulation of the research methodology, and output were presented. Also the performance evaluation, and implementation and results evaluation based on defined methodology and standards.

From the conducted experiments, It was observed that the TBL tagger achieved a higher accuracy level on the baseline tagger than the HMM and all n-gram taggers. This is not strange as the TBL tagger is a hybrid in the form of generative and discriminative taggers. Whereas the HMM and n-gram taggers are generative taggers, the TBL tagger boosts upon their weaknesses in the tagging process, thus performing better and achieving better accuracy levels both in training and development (i.e. implementation). Although the TBL tagger outperformed the HMM, Unigram, Bigram, Trigram taggers, there was not much difference between it and the Unigram tagger. Just as they (i.e. TBL and Unigram tagger) achieved same level of f1-measure, and differ on precision, recall, and development accuracy, the difference is balanced as TBL exceeded the Unigram tagger by 6% on recall and development accuracy, while the Unigram tagger exceeded the TBL tagger by 6% also on precision.

In summary, this study presents a number of resources, tools, and developed a standard tagset for Hausa language called the HTS downloadable from https://github.com/HausaNLPResearch/Hausa-POS-Tagset-HTS-. The TBL tagger as a hybrid tagger was implemented using the. Results from the tagging experiments shows improvements on TBL accuracy levels compared to accuracy of that of n-gram and HMM taggers from this study. As such it is realized that not only the learning algorithm used in training a tagger matters in tagging accuracy, but also the composition of taggers model also affects accuracy of POS taggers.

## REFERENCES

[1] A. Pannu, “Artificial Intelligence and its Application in Different Areas,” *Int. J. Eng. Innov. Technol.*, vol. 4, no. 10, pp. 79–84, 2015.

[2] P. Patheja, A. Waoo, and R. Garg, “Part of Speech Tagging,” in *International Conference on Computer and Automation Engineering, 4th (ICCAE 2012)*, 2012, pp. 281–286.

[3] R. J. Mooney, “CS 343: Artificial Intelligence Natural Language Processing,” 2010.

[4] D. Jurafsky and J. H. Martin, “Speech and Language Processing,” in *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd ed., Pearson Prentice Hall, 2008, p. 17.

[5] A. Chopra, A. Prashar, and C. Sain, “Natural Language Processing,” *Int. J. Technol. Enhanc. Emerg. Eng.*, vol. 1, no. 4, pp. 131–134, 2013.

[6] Robin, “World of Computing. Articles on Natural language Processing,” *Parts of Speech Tagging*, 2009. [Online]. Available: http://language.worldofcomputing.net/pos-tagging/parts-of-speech-tagging.html. [Accessed: 28-Feb-2018].

[7] S. L. Pandian and T. V Geetha, “Morpheme based Language Model for Tamil Part-of-Speech Tagging,” *Polibits*, vol. 38, pp. 19–25, 2008.

[8] M. Albared, N. Omar, M. Juzaidin, and A. Aziz, “Arabic Part Of Speech Disambiguation : A Survey,” *Int. Rev. Comput. Softw.*, vol. 4, no. 5, 2009.

[9] S. Dandapat, “Part-of-Speech Tagging for Bengali,” MS Thesis, Dept. of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, India, 2009.

[10] D. Kumawat and V. Jain, “POS Tagging Approaches: A Comparison,” *Int. J. Comput. Appl.*, vol. 118, no. 6, pp. 975–8887, 2015.

[11] M. Alex and L. Q. Zakaria, “Kadazan Part of Speech Tagging using Transformation-Based Approach,” in *Procedia Technology*, 2013, vol. 11, no. Iceei, pp. 621–627.

[12] C. N. dos Santos and R. L. Milidiú, “Entropy Guided Transformation Learning,” in *Algorithms and Applications*, 2012, pp. 9–21.

[13] E. Halvor and T. Rolf, “Language families,” in *Linguistics for Students of Asian and African Languages*, 2005, p. 25.

[14] M. Bashir, A. Rozaimee, and W. M. Wan Isa, “Automatic Hausa Language Text Summarization,” *World Appl. Sci. J.*, vol. 35, no. 9, p. 7, 2017.

[15] J. Z. Maitama, U. Haruna, A. Y. Gambo, B. A. Thomas, N. B. Idris, A. Y. Gital, and A. I. Abubakar, “Text normalization algorithm for facebook chats in Hausa language,” in *The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M)*, 2014, pp. 1–4.

[16] B. Hladka, “Czech Language Tagging,” Institute of Formal and Applied Linguistics, Charles University, 2000.

[17] A. Zeldes, “amir-zeldes/hausa.” 2019.

[18] H. Schmid, “Probabilistic Part-of-Speech Tagging Using Decision Trees,” in *Proceedings of the International Conference on New Methods in Language Processing*, 1994.

[19] S. Corston-Oliver and M. Gamon, “Combining decision trees and transformation-based learning to correct transferred linguistic representations,” in *Ninth Machine Translation Summit 2003*, 2003, pp. 55–62.

[20] E. Brill, “Transformation-Based Error-Driven Learning and Natural Language Processing : A Case Study in Part of Speech Tagging,” *Comput. Linguist.*, vol. 21, no. 94, pp. 543–566, 1995.

[21] G. Ngai and R. Florian, “Transformation-Based Learning in the Fast Lane,” in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2001, p. 8.

[22] T. Wu, F. M. Khan, T. A. Fisher, L. A. Shuler, and M. Pottenger, “Posting Act Tagging Using Transformation- Based Learning,” in *Foundations of Data Mining and knowledge Discovery*, 2005, pp. 2–13.

[23] D. Dien and H. Kiem, “POS-Tagger for English-Vietnamese Bilingual Corpus,” in *HLT-NAACL-PARALLEL ’03 Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond*, 2003, pp. 88–95.

[24] D. Q. Nguyen, D. Q. Nguyen, D. D. Pham, and S. B. Pham, “A robust transformation-based learning approach using ripple down rules for part-of-speech tagging,” *AI Commun.*, vol. 29, no. 3, pp. 409–422, Apr. 2016.

[25] N. Landwehr, B. Gutmann, T. Thon, L. De Raedt, and M. Philipose, “Relational Transformation-based Tagging for Activity Recognition,” *IOS Press*, pp. 1–19, 2008.

[26] B. Karagol-ayan, D. Doermann, and A. Weinberg, “Adaptive Transformation-based Learning for Improving Dictionary Tagging,” in *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006, pp. 257–264.

[27] Z. Il-agure and H. N. Itani, “Link Mining Process,” *Int. J. Data Min. Knowl. Manag. Process*, vol. 7, no. 3, pp. 45–51, 2017.

[28] P. Chapman, R. Kerber, J. Clinton, T. Khabaza, T. Reinartz, and R. Wirth, “The CRISP-DM process model,” *Cris. Consort.*, vol. 310, no. C, p. 91, 1999.

[29] E. S. Atwell, “Development of tag sets for part-of-speech tagging,” in *Corpus Linguistics: An International Hanbook*, 2008, pp. 501–526.

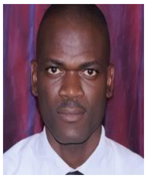
[30] A. Hardie, “Developing a tagset for automated part-of-speech tagging in Urdu,” *UCREL Tech. Pap.*, vol. 16, pp. 1–11, 2003.

[31] G. Leech and A. Wilson, *EAGLES - Recommendations for the Morphosyntactic Annotation of Corpora*. EAG–TCWG–MAC/R, 1996.

[32] E. Kjellqvist, “Part of Speech Tagging as an Application to Key Word Extraction Models for Swedish Company Web Pages,” 2005.

**Jamilu Awwalu** received Bachelor’s Degree in Business Computing and Information Technology from the University of Wales in 2010, M.Sc. degree in Information Technology (Artificial Intelligence) at University Kebangsaan Malaysia in 2015, and is currently a PhD candidate at the Nigerian Defence Academy Kaduna. His major research interests are NLP, computer vision, and machine learning.

**Saleh Elyakub Abdullahi** is a Professor of Computer Science at Nile University of Nigeria. He received PhD in Computer Science from Univ. of London in 1996, M.Sc Computer Science from the University of Lagos, B.Sc. in Computer Science from Ahmadu Bello University in 1987. His research interest includes algorithms, and simulation



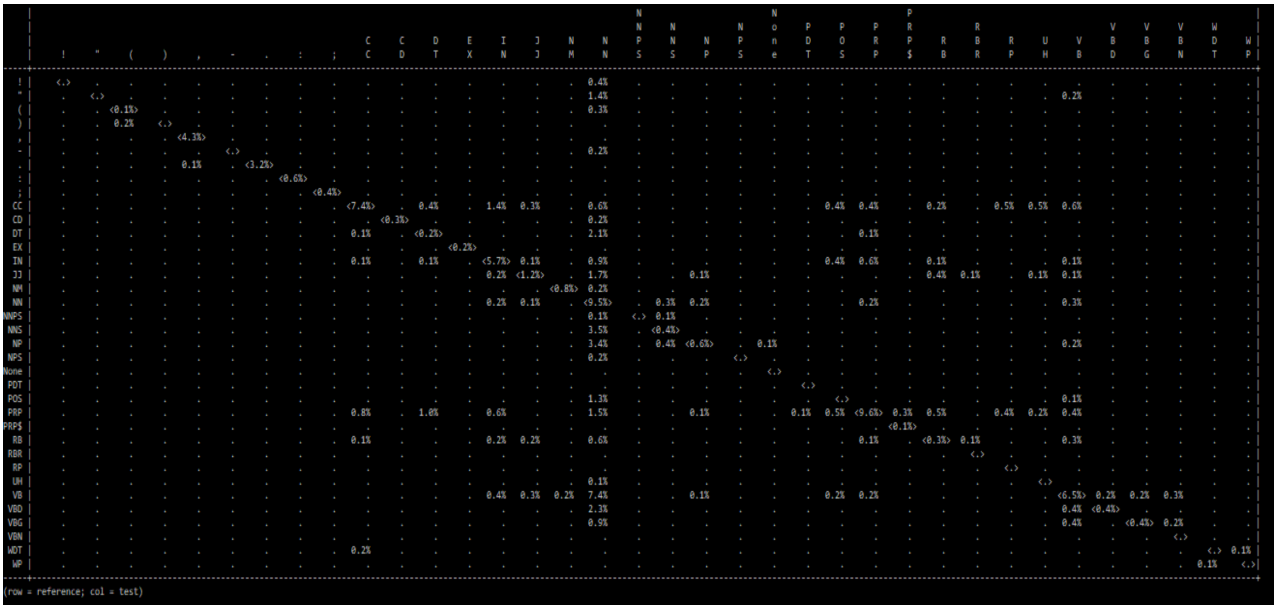
**Abraham Eseoghene Evwiekpaefe** is a lecturer in the Department of Computer Science at Nigerian Defence Academy. He received his Ph.D. in Computer Science from University of Benin, Nigeria**.** His research interest include Computational Mathematics, Security Challenges, Cloud Computing, E-commerce, and Information Systems

## APPENDICES

## Appendix A

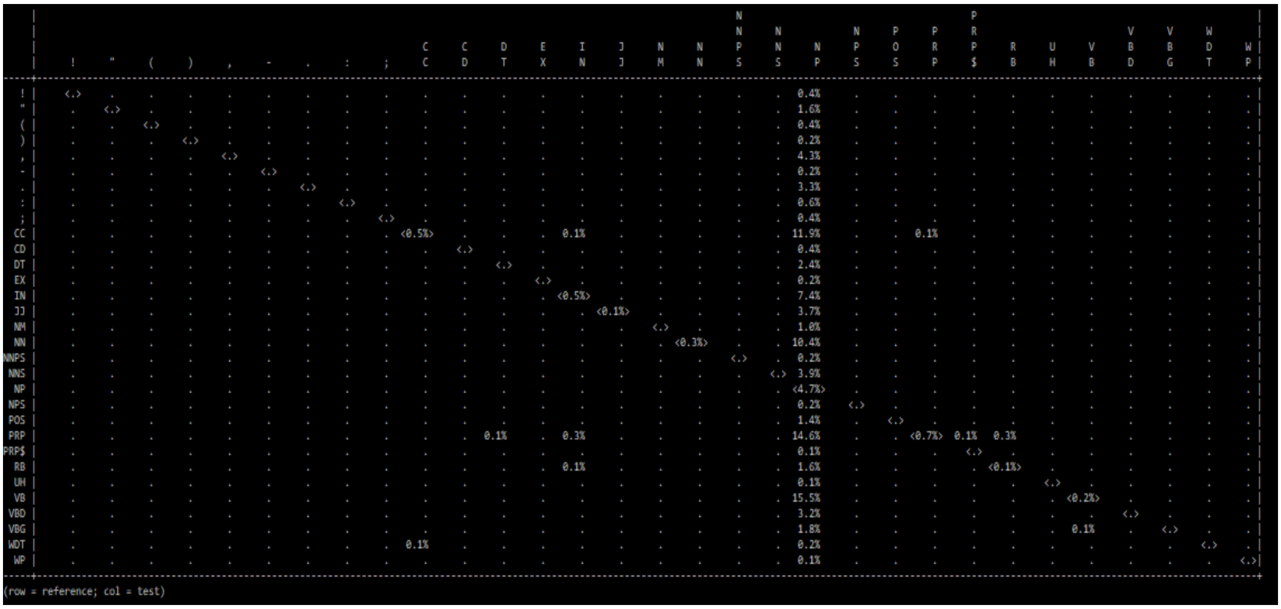
## Confusion Matrices

*1) Confusion Matrix for Transformation Based Learning (TBL) tagger*



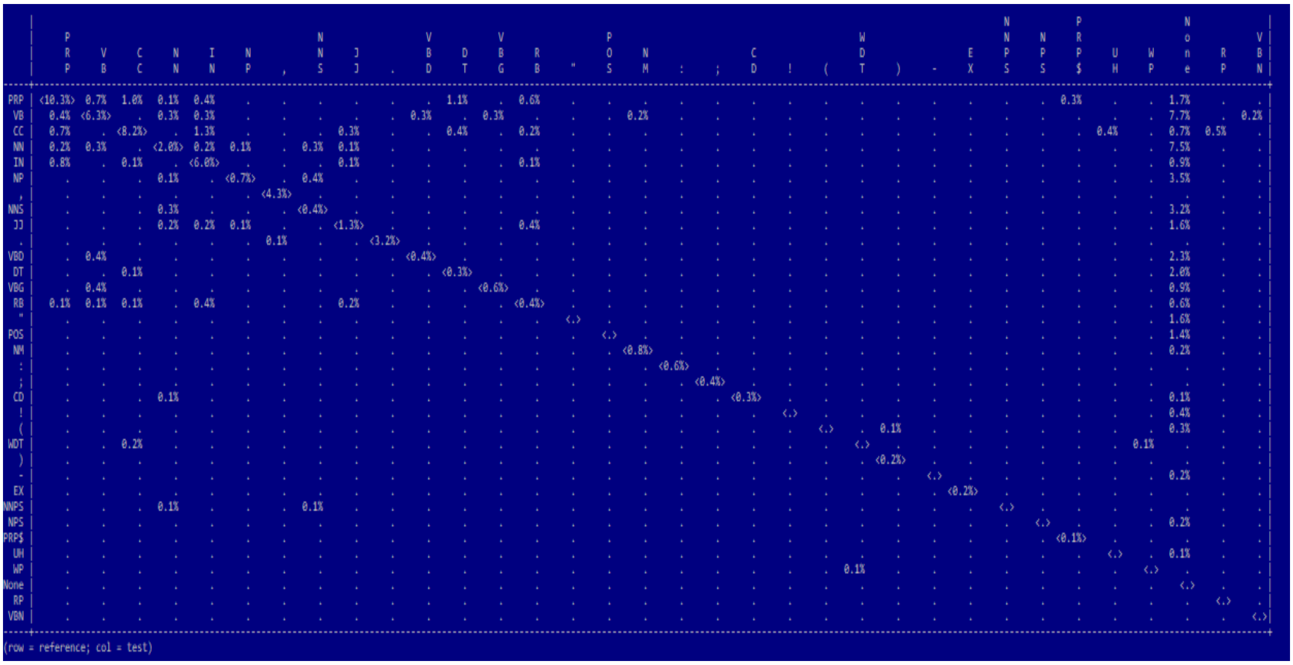
```
     |                                                                                                          N                             N                             P                                                            |
     |                                                                                                          N     N           N     o     P     P     P     R           R                             V     V     V     W       |
     |                                                        C     C     D     E     I     J     N     N     P     N     N     P     n     D     O     R     P     R     B     R     U     V     B     B     B     D     W |
     |    !     "     (     )     ,     -     .     :     ;     C     D     T     X     N     J     M     N     S     S     P     S     e     T     S     P     $     B     R     P     H     B     D     G     N     T     P |
-----+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
   ! |   <.>    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  0.4%     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
   " |    .    <.>    .     .     .     .     .     .     .     .     .     .     .     .     .     .  1.4%     .     .     .     .     .     .     .     .     .     .     .     .     .  0.2%     .     .     .     .     . |
   ( |    .     . <0.1%>    .     .     .     .     .     .     .     .     .     .     .     .     .  0.3%     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
   ) |    .     .  0.2%    <.>    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
   , |    .     .     .     . <4.3%>    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
   - |    .     .     .     .     .    <.>    .     .     .     .     .     .     .     .     .     .  0.2%     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
   . |    .     .     .     .  0.1%     . <3.2%>    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
   : |    .     .     .     .     .     .     . <0.6%>    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
   ; |    .     .     .     .     .     .     .     . <0.4%>    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
  CC |    .     .     .     .     .     .     .     .     . <7.4%>    .  0.4%     .  1.4%  0.3%     .  0.6%     .     .     .     .     .     .  0.4%  0.4%     .  0.2%     .  0.5%  0.5%  0.6%     .     .     .     .     . |
  CD |    .     .     .     .     .     .     .     .     .     . <0.3%>    .     .     .     .     .  0.2%     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
  DT |    .     .     .     .     .     .     .     .     .  0.1%     . <0.2%>    .     .     .     .  2.1%     .     .     .     .     .     .     .  0.1%     .     .     .     .     .     .     .     .     .     .     . |
  EX |    .     .     .     .     .     .     .     .     .     .     .     . <0.2%>    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
  IN |    .     .     .     .     .     .     .     .     .  0.1%     .  0.1%     . <5.7%>  0.1%     .  0.9%     .     .     .     .     .     .  0.4%  0.6%     .  0.1%     .     .     .  0.1%     .     .     .     .     . |
  JJ |    .     .     .     .     .     .     .     .     .     .     .     .     .  0.2% <1.2%>    .  1.7%     .     .  0.1%     .     .     .     .     .     .  0.4%  0.1%     .  0.1%  0.1%     .     .     .     .     . |
  NM |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     . <0.8%>  0.2%     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
  NN |    .     .     .     .     .     .     .     .     .     .     .     .     .  0.2%  0.1%     . <9.5%>    .  0.3%  0.2%     .     .     .     .  0.2%     .     .     .     .     .  0.3%     .     .     .     .     . |
NNPS |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  0.1%    <.>  0.1%     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
 NNS |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  3.5%     . <0.4%>    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
  NP |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  3.4%     .  0.4% <0.6%>    .  0.1%     .     .     .     .     .     .     .     .  0.2%     .     .     .     .     . |
 NPS |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  0.2%     .     .     .    <.>    .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
None |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .    <.>    .     .     .     .     .     .     .     .     .     .     .     .     .     . |
 PDT |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .    <.>    .     .     .     .     .     .     .     .     .     .     .     .     . |
 POS |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  1.3%     .     .     .     .     .     .    <.>    .     .     .     .     .     .  0.1%     .     .     .     .     . |
 PRP |    .     .     .     .     .     .     .     .     .  0.8%     .  1.0%     .  0.6%     .     .  1.5%     .     .  0.1%     .     .  0.1%  0.5% <9.6%>  0.3%  0.5%     .  0.4%  0.2%  0.4%     .     .     .     .     . |
PRP$ |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . <0.1%>    .     .     .     .     .     .     .     .     .     . |
  RB |    .     .     .     .     .     .     .     .     .  0.1%     .     .     .  0.2%  0.2%     .  0.6%     .     .     .     .     .     .     .  0.1%     . <0.3%>  0.1%     .     .  0.3%     .     .     .     .     . |
 RBR |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .    <.>    .     .     .     .     .     .     .     . |
  RP |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .    <.>    .     .     .     .     .     .     . |
  UH |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  0.1%     .     .     .     .     .     .     .     .     .     .     .     .    <.>    .     .     .     .     .     . |
  VB |    .     .     .     .     .     .     .     .     .     .     .     .     .  0.4%  0.3%  0.2%  7.4%     .     .  0.1%     .     .     .  0.2%  0.2%     .     .     .     .     . <6.5%>  0.2%  0.2%  0.3%     .     . |
 VBD |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  2.3%     .     .     .     .     .     .     .     .     .     .     .     .     .  0.4% <0.4%>    .     .     .     . |
 VBG |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  0.9%     .     .     .     .     .     .     .     .     .     .     .     .     .  0.4%     . <0.4%>  0.2%     .     . |
 VBN |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .    <.>    .     . |
 WDT |    .     .     .     .     .     .     .     .     .  0.2%     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .    <.>  0.1% |
  WP |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  0.1%    <.>|
-----+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
(row = reference; col = test)
```

*2) Confusion Matrix for Hidden Markov Model (HMM) tagger*



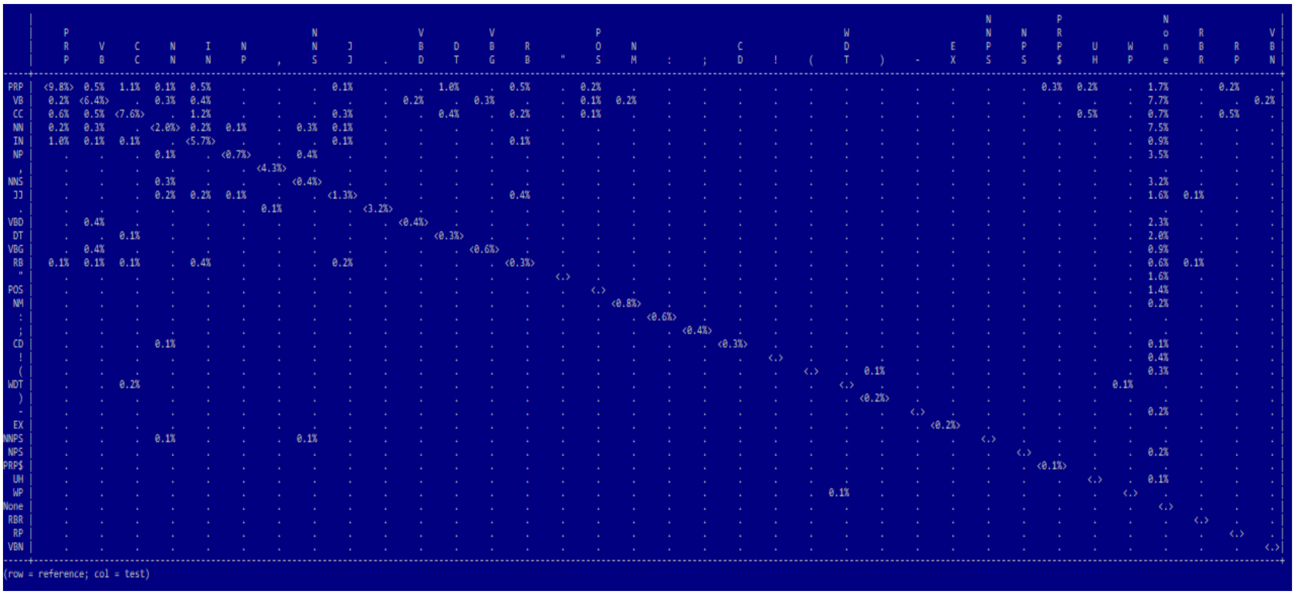
```
     |                                                                                                                               N                                           P                                                   |
     |                                                                                                                               N      N             N      P      P      R                           V      V      W        |
     |                                                                  C      C      D      E      I      J      N      N      P      N      N      P      O      R      P      R      U      V      B      B      D      W |
     |      !      "      (      )      ,      -      .      :      ;      C      D      T      X      N      J      M      N      S      S      P      S      S      P      $      B      H      B      D      G      T      P |
-----+-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
   ! |    <.>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   0.4%     .      .      .      .      .      .      .      .      .      .      . |
   " |     .     <.>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   1.6%     .      .      .      .      .      .      .      .      .      .      . |
   ( |     .      .     <.>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   0.4%     .      .      .      .      .      .      .      .      .      .      . |
   ) |     .      .      .     <.>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   0.2%     .      .      .      .      .      .      .      .      .      .      . |
   , |     .      .      .      .     <.>     .      .      .      .      .      .      .      .      .      .      .      .      .      .   4.3%     .      .      .      .      .      .      .      .      .      .      . |
   - |     .      .      .      .      .     <.>     .      .      .      .      .      .      .      .      .      .      .      .      .   0.2%     .      .      .      .      .      .      .      .      .      .      . |
   . |     .      .      .      .      .      .     <.>     .      .      .      .      .      .      .      .      .      .      .      .   3.3%     .      .      .      .      .      .      .      .      .      .      . |
   : |     .      .      .      .      .      .      .     <.>     .      .      .      .      .      .      .      .      .      .      .   0.6%     .      .      .      .      .      .      .      .      .      .      . |
   ; |     .      .      .      .      .      .      .      .     <.>     .      .      .      .      .      .      .      .      .      .   0.4%     .      .      .      .      .      .      .      .      .      .      . |
  CC |     .      .      .      .      .      .      .      .      .  <0.5%>     .      .      .   0.1%     .      .      .      .      .  11.9%     .      .   0.1%     .      .      .      .      .      .      .      . |
  CD |     .      .      .      .      .      .      .      .      .      .     <.>     .      .      .      .      .      .      .      .   0.4%     .      .      .      .      .      .      .      .      .      .      . |
  DT |     .      .      .      .      .      .      .      .      .      .      .     <.>     .      .      .      .      .      .      .   2.4%     .      .      .      .      .      .      .      .      .      .      . |
  EX |     .      .      .      .      .      .      .      .      .      .      .      .     <.>     .      .      .      .      .      .   0.2%     .      .      .      .      .      .      .      .      .      .      . |
  IN |     .      .      .      .      .      .      .      .      .      .      .      .      .  <0.5%>     .      .      .      .      .   7.4%     .      .      .      .      .      .      .      .      .      .      . |
  JJ |     .      .      .      .      .      .      .      .      .      .      .      .      .      .  <0.1%>     .      .      .      .   3.7%     .      .      .      .      .      .      .      .      .      .      . |
  NM |     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .      .      .   1.0%     .      .      .      .      .      .      .      .      .      .      . |
  NN |     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .  <0.3%>     .      .  10.4%     .      .      .      .      .      .      .      .      .      .      . |
NNPS |     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .   0.2%     .      .      .      .      .      .      .      .      .      .      . |
 NNS |     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>  3.9%     .      .      .      .      .      .      .      .      .      .      . |
  NP |     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .  <4.7%>     .      .      .      .      .      .      .      .      .      .      . |
 NPS |     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   0.2%    <.>     .      .      .      .      .      .      .      .      .      . |
 POS |     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   1.4%     .     <.>     .      .      .      .      .      .      .      .      . |
 PRP |     .      .      .      .      .      .      .      .      .      .      .   0.1%     .   0.3%     .      .      .      .      .  14.6%     .      .  <0.7%>  0.1%   0.3%     .      .      .      .      .      . |
PRP$ |     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   0.1%     .      .      .     <.>     .      .      .      .      .      .      . |
  RB |     .      .      .      .      .      .      .      .      .      .      .      .      .   0.1%     .      .      .      .      .   1.6%     .      .      .      .  <0.1%>     .      .      .      .      .      . |
  UH |     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   0.1%     .      .      .      .      .     <.>     .      .      .      .      . |
  VB |     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .  15.5%     .      .      .      .      .      .  <0.2%>     .      .      .      . |
 VBD |     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   3.2%     .      .      .      .      .      .      .     <.>     .      .      . |
 VBG |     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   1.8%     .      .      .      .      .      .   0.1%     .     <.>     .      . |
 WDT |     .      .      .      .      .      .      .      .      .   0.1%     .      .      .      .      .      .      .      .      .   0.2%     .      .      .      .      .      .      .      .      .     <.>     . |
  WP |     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   0.1%     .      .      .      .      .      .      .      .      .      .     <.>|
-----+-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
(row = reference; col = test)
```

*3) Confusion Matrix for Unigram tagger*



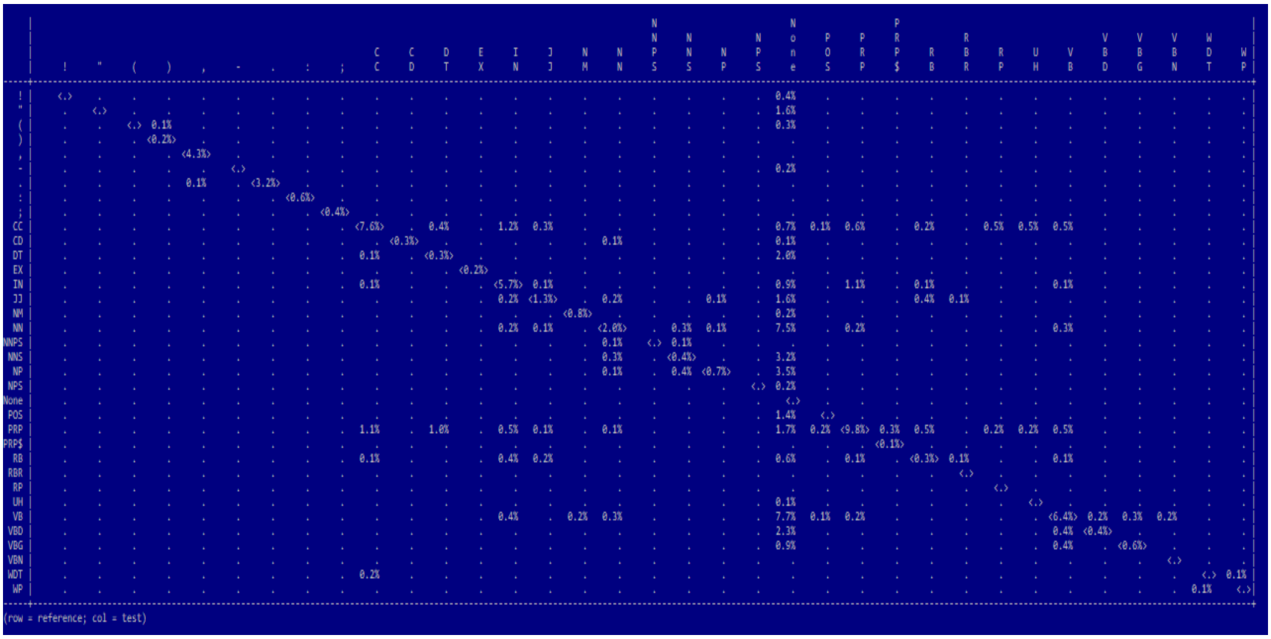
```
     |                                                                                                                                                                                            N             P                    N               |
     |      P                                                N                    V             V                    P                                                W                           N      N      R                    o             V |
     |      R      V      C      N      I      N             N      J             B      D      B      R             O      N                    C                    D                    E      P      P      P      U      W      n      R      B |
     |      P      B      C      N      N      P      ,      S      J      .      D      T      G      B      "      S      M      :      ;      D      !      (      T      )      -      X      S      S      $      H      P      e      P      N |
-----+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
 PRP | <10.3%>  0.7%   1.0%   0.1%   0.4%      .      .      .      .      .      .   1.1%      .   0.6%      .      .      .      .      .      .      .      .      .      .      .      .      .      .   0.3%      .      .   1.7%      .      . |
  VB |   0.4%  <6.3%>     .   0.3%   0.3%      .      .      .      .      .   0.3%      .   0.3%      .      .      .   0.2%      .      .      .      .      .      .      .      .      .      .      .      .      .      .   7.7%      .   0.2% |
  CC |   0.7%      .  <8.2%>     .   1.3%      .      .      .   0.3%      .      .   0.4%      .   0.2%      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   0.4%      .   0.7%   0.5%      . |
  NN |   0.2%   0.3%      .  <2.0%>  0.2%   0.1%      .   0.3%   0.1%      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   7.5%      .      . |
  IN |   0.8%      .   0.1%      .  <6.0%>     .      .      .   0.1%      .      .      .      .   0.1%      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   0.9%      .      . |
  NP |      .      .      .   0.1%      .  <0.7%>     .   0.4%      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   3.5%      .      . |
   , |      .      .      .      .      .      .  <4.3%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      . |
 NNS |      .      .      .   0.3%      .      .      .  <0.4%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   3.2%      .      . |
  JJ |      .      .      .   0.2%   0.2%   0.1%      .      .  <1.3%>     .      .      .      .   0.4%      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   1.6%      .      . |
   . |      .      .      .      .      .      .   0.1%      .      .  <3.2%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      . |
 VBD |      .   0.4%      .      .      .      .      .      .      .      .  <0.4%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   2.3%      .      . |
  DT |      .      .   0.1%      .      .      .      .      .      .      .      .  <0.3%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   2.0%      .      . |
 VBG |      .   0.4%      .      .      .      .      .      .      .      .      .      .  <0.6%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   0.9%      .      . |
  RB |   0.1%   0.1%   0.1%      .   0.4%      .      .      .   0.2%      .      .      .      .  <0.4%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   0.6%      .      . |
   " |      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   1.6%      .      . |
 POS |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   1.4%      .      . |
  NM |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .  <0.8%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .   0.2%      .      . |
   : |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .  <0.6%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      . |
   ; |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .  <0.4%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      . |
  CD |      .      .      .   0.1%      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .  <0.3%>     .      .      .      .      .      .      .      .      .      .      .   0.1%      .      . |
   ! |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .      .      .      .      .      .      .      .      .      .   0.4%      .      . |
   ( |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .   0.1%      .      .      .      .      .      .      .   0.3%      .      . |
 WDT |      .      .   0.2%      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .      .      .      .      .      .      .   0.1%      .      .      . |
   ) |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .  <0.2%>     .      .      .      .      .      .      .      .      .      . |
   - |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .      .      .      .      .      .   0.2%      .      . |
  EX |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .  <0.2%>     .      .      .      .      .      .      .      . |
NNPS |      .      .      .   0.1%      .      .      .   0.1%      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .      .      .      .      .      .      . |
 NPS |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .      .      .   0.2%      .      . |
PRP$ |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .  <0.1%>     .      .      .      .      . |
  UH |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .   0.1%      .      . |
  WP |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   0.1%      .      .      .      .      .      .      .     <.>     .      .      . |
None |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .      . |
  RP |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     . |
 VBN |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>|
-----+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
(row = reference; col = test)
```

*4) Confusion Matrix for Bigram tagger*



```
     |                                                                                                                                                                                            N             P                    N                      |
     |      P                                                N                    V             V                    P                                                W                           N      N      R                    o      R             V |
     |      R      V      C      N      I      N             N      J             B      D      B      R             O      N                    C                    D                    E      P      P      P      U      W      n      B      R      B |
     |      P      B      C      N      N      P      ,      S      J      .      D      T      G      B      "      S      M      :      ;      D      !      (      T      )      -      X      S      S      $      H      P      e      R      P      N |
-----+---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
 PRP |  <9.8%>  0.5%   1.1%   0.1%   0.5%      .      .      .   0.1%      .      .   1.0%      .   0.5%      .   0.2%      .      .      .      .      .      .      .      .      .      .      .      .   0.3%   0.2%      .   1.7%      .   0.2%      . |
  VB |   0.2%  <6.4%>     .   0.3%   0.4%      .      .      .      .      .   0.2%      .   0.3%      .      .   0.1%   0.2%      .      .      .      .      .      .      .      .      .      .      .      .      .      .   7.7%      .      .   0.2% |
  CC |   0.6%   0.5%  <7.6%>     .   1.2%      .      .      .   0.3%      .      .   0.4%      .   0.2%      .   0.1%      .      .      .      .      .      .      .      .      .      .      .      .      .   0.5%      .   0.7%      .   0.5%      . |
  NN |   0.2%   0.3%      .  <2.0%>  0.2%   0.1%      .   0.3%   0.1%      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   7.5%      .      .      . |
  IN |   1.0%   0.1%   0.1%      .  <5.7%>     .      .      .   0.1%      .      .      .      .   0.1%      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   0.9%      .      .      . |
  NP |      .      .      .   0.1%      .  <0.7%>     .   0.4%      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   3.5%      .      .      . |
   , |      .      .      .      .      .      .  <4.3%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      . |
 NNS |      .      .      .   0.3%      .      .      .  <0.4%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   3.2%      .      .      . |
  JJ |      .      .      .   0.2%   0.2%   0.1%      .      .  <1.3%>     .      .      .      .   0.4%      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   1.6%   0.1%      .      . |
   . |      .      .      .      .      .      .   0.1%      .      .  <3.2%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      . |
 VBD |      .   0.4%      .      .      .      .      .      .      .      .  <0.4%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   2.3%      .      .      . |
  DT |      .      .   0.1%      .      .      .      .      .      .      .      .  <0.3%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   2.0%      .      .      . |
 VBG |      .   0.4%      .      .      .      .      .      .      .      .      .      .  <0.6%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   0.9%      .      .      . |
  RB |   0.1%   0.1%   0.1%      .   0.4%      .      .      .   0.2%      .      .      .      .  <0.3%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   0.6%   0.1%      .      . |
   " |      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   1.6%      .      .      . |
 POS |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   1.4%      .      .      . |
  NM |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .  <0.8%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .   0.2%      .      .      . |
   : |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .  <0.6%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      . |
   ; |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .  <0.4%>     .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      . |
  CD |      .      .      .   0.1%      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .  <0.3%>     .      .      .      .      .      .      .      .      .      .      .   0.1%      .      .      . |
   ! |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .      .      .      .      .      .      .      .      .      .   0.4%      .      .      . |
   ( |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .   0.1%      .      .      .      .      .      .      .   0.3%      .      .      . |
 WDT |      .      .   0.2%      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .      .      .      .      .      .      .   0.1%      .      .      .      . |
   ) |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .  <0.2%>     .      .      .      .      .      .      .      .      .      .      . |
   - |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .      .      .      .      .      .   0.2%      .      .      . |
  EX |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .  <0.2%>     .      .      .      .      .      .      .      .      . |
NNPS |      .      .      .   0.1%      .      .      .   0.1%      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .      .      .      .      .      .      .      . |
 NPS |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .      .      .   0.2%      .      .      . |
PRP$ |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .  <0.1%>     .      .      .      .      .      . |
  UH |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .   0.1%      .      .      . |
  WP |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .   0.1%      .      .      .      .      .      .      .     <.>     .      .      .      . |
None |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .      .      . |
 RBR |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     .      . |
  RP |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>     . |
 VBN |      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      .     <.>|
-----+---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
(row = reference; col = test)
```

*5) Confusion Matrix for Trigram tagger*



```
     |                                                                                                          N                       N                 P                                                             |
     |                                                                                                          N     N           N     o     P     P     R           R                       V     V     V     W       |
     |                                                          C     C     D     E     I     J     N     N     P     N     N     P     n     O     R     P     R     B     R     U     V     B     B     B     D     W |
     |    !     "     (     )     ,     -     .     :     ;     C     D     T     X     N     J     M     N     S     S     P     S     e     S     P     $     B     R     P     H     B     D     G     N     T     P |
-----+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
   ! |   <.>    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  0.4%     .     .     .     .     .     .     .     .     .     .     .     .     . |
   " |    .    <.>    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  1.6%     .     .     .     .     .     .     .     .     .     .     .     .     . |
   ( |    .     .    <.>  0.1%     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  0.3%     .     .     .     .     .     .     .     .     .     .     .     .     . |
   ) |    .     .     . <0.2%>    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
   , |    .     .     .     . <4.3%>    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
   - |    .     .     .     .     .    <.>    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  0.2%     .     .     .     .     .     .     .     .     .     .     .     .     . |
   . |    .     .     .     .  0.1%     . <3.2%>    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
   : |    .     .     .     .     .     .     . <0.6%>    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
   ; |    .     .     .     .     .     .     .     . <0.4%>    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
  CC |    .     .     .     .     .     .     .     .     . <7.6%>    .  0.4%     .  1.2%  0.3%     .     .     .     .     .     .  0.7%  0.1%  0.6%     .  0.2%     .  0.5%  0.5%  0.5%     .     .     .     .     . |
  CD |    .     .     .     .     .     .     .     .     .     . <0.3%>    .     .     .     .     .  0.1%     .     .     .     .  0.1%     .     .     .     .     .     .     .     .     .     .     .     .     . |
  DT |    .     .     .     .     .     .     .     .     .  0.1%     . <0.3%>    .     .     .     .     .     .     .     .     .  2.0%     .     .     .     .     .     .     .     .     .     .     .     .     . |
  EX |    .     .     .     .     .     .     .     .     .     .     .     . <0.2%>    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
  IN |    .     .     .     .     .     .     .     .     .  0.1%     .     .     . <5.7%> 0.1%     .     .     .     .     .     .  0.9%     .  1.1%     .  0.1%     .     .     .  0.1%     .     .     .     .     . |
  JJ |    .     .     .     .     .     .     .     .     .     .     .     .     .  0.2% <1.3%>    .  0.2%     .     .  0.1%     .  1.6%     .     .     .  0.4%  0.1%     .     .     .     .     .     .     .     . |
  NM |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     . <0.8%>    .     .     .     .     .  0.2%     .     .     .     .     .     .     .     .     .     .     .     .     . |
  NN |    .     .     .     .     .     .     .     .     .     .     .     .     .  0.2%  0.1%     . <2.0%>    .  0.3%  0.1%     .  7.5%     .  0.2%     .     .     .     .     .  0.3%     .     .     .     .     . |
NNPS |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  0.1%    <.>  0.1%     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . |
 NNS |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  0.3%     . <0.4%>    .     .  3.2%     .     .     .     .     .     .     .     .     .     .     .     .     . |
  NP |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  0.1%     .  0.4% <0.7%>    .  3.5%     .     .     .     .     .     .     .     .     .     .     .     .     . |
 NPS |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .    <.>  0.2%     .     .     .     .     .     .     .     .     .     .     .     .     . |
None |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .    <.>    .     .     .     .     .     .     .     .     .     .     .     .     . |
 POS |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  1.4%    <.>    .     .     .     .     .     .     .     .     .     .     .     . |
 PRP |    .     .     .     .     .     .     .     .     .  1.1%     .  1.0%     .  0.5%  0.1%     .  0.1%     .     .     .     .  1.7%  0.2% <9.8%> 0.3%  0.5%     .  0.2%  0.2%  0.5%     .     .     .     .     . |
PRP$ |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     . <0.1%>    .     .     .     .     .     .     .     .     .     . |
  RB |    .     .     .     .     .     .     .     .     .  0.1%     .     .     .  0.4%  0.2%     .     .     .     .     .     .  0.6%     .  0.1%     . <0.3%> 0.1%     .     .  0.1%     .     .     .     .     . |
 RBR |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .    <.>    .     .     .     .     .     .     .     . |
  RP |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .    <.>    .     .     .     .     .     .     . |
  UH |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  0.1%     .     .     .     .     .     .    <.>    .     .     .     .     .     . |
  VB |    .     .     .     .     .     .     .     .     .     .     .     .     .  0.4%     .  0.2%  0.3%     .     .     .     .  7.7%  0.1%  0.2%     .     .     .     .     . <6.4%> 0.2%  0.3%  0.2%     .     . |
 VBD |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  2.3%     .     .     .     .     .     .     .  0.4% <0.4%>    .     .     .     . |
 VBG |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  0.9%     .     .     .     .     .     .     .  0.4%     . <0.6%>    .     .     . |
 VBN |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .    <.>    .     . |
 WDT |    .     .     .     .     .     .     .     .     .  0.2%     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .    <.>  0.1% |
  WP |    .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .  0.1%    <.>|
-----+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
(row = reference; col = test)
```

## Appendix B
## TBL Generated Rules

### *1) TBL Generated Transformation Rules*

| Rule Number | Rules |
|---|---|
| 1 | IN->PRP if Pos:VB@[-1] & Pos:VB@[1] |
| 2 | VB->VBD if Word:Na@[-1] & Word:aiki@[1] |
| 3 | NN->VB if Pos:VBD@[-1] & Pos:IN@[1] |
| 4 | POS->PRP if Word:za@[-1] |
| 4 | DT->UH if Pos:VBG@[1] |
| 5 | PDT->NN if Pos:EX@[1] |
| 6 | VBN->IN if Pos:,@[1] |
| 7 | DT->PDT if Pos:IN@[2] |
| 8 | PDT->VBZ if Pos:POS@[-1] |
| 9 | POS->PRP if Pos:VBZ@[1] |
| 10 | RP->DT if Pos:?@[2] |
| 11 | VB->VBD if Pos:VBZ@[2] |
| 12 | VB->VBG if Pos:PDT@[2] |
| 13 | CC->IN if Pos:RP@[-3,-2,-1] |
| 14 | NN->VB if Pos:.@[-3,-2,-1] |
| 15 | CC->DT if Pos:PRP@[-1] & Pos:NM@[1] |
| 16 | CC->POS if Pos:VBG@[-1] & Pos:PRP@[1] |
| 17 | NN->NP if Pos:IN@[-1] & Pos:IN@[1] |
| 18 | NN->VB if Pos:VB@[-1] & Pos:DT@[1] |
| 19 | DT->PRP if Pos:VB@[-1] & Pos:VB@[1] |
| 20 | NN->VB if Pos:VB@[-1] & Pos:PRP@[1] |

| | |
|---|---|
| 21 | NN->None if Pos:,@[-1] & Pos:UH@[1] |
| 22 | PRP->IN if Pos:,@[-1] & Pos:JJ@[1] |
| 23 | PRP->IN if Pos:POS@[-1] & Pos:PRP@[1] |
| 24 | VB->IN if Pos:CC@[-1] & Pos:NP@[1] |
| 25 | CC->IN if Pos:NP@[-1] & Pos:IN@[1] |
| 26 | .->None if Word:ce@[-1] |
| 27 | CC->RB if Word:aikin/VB@[-1] |
| 28 | NP->NN if Word:Abdullahi@[-1] |
| 29 | PRP->POS if Word:dawo@[-1] |
| 30 | PRP->POS if Word:zama@[-1] |
| 31 | VBZ->VB if Word:an@[-1] |
| 32 | IN->PRP if Word:yi@[1] |
| 33 | IN->VB if Word:wannan@[1] |
| 34 | PRP->DT if Word:fage@[1] |
| 35 | PRP->DT if Word:kira@[1] |
| 36 | PRP->IN if Word:yaya@[1] |
| 37 | VB->IN if Word:tsirtowa@[1] |
| 38 | VB->NM if Word:aikin/VB@[1] |
| 39 | VB->VBZ if Word:fice@[1] |
| 40 | PRP->VB if Word:za@[-2] |
| 41 | PRP->DT if Word:Sarkin@[2] |
| 42 | PRP->JJ if Word:fi@[2] |
| 43 | NN->VB if Word:dawo@[-3,-2,-1] |
| 44 | NN->VB if Word:Larabci@[1,2,3] |
| 45 | VB->PRP if Word:ce@[-1] & Word:na@[1] |
| 46 | VB->VBD if Word:na@[-1] & Word:aiki@[1] |
| 47 | NN->VB if Pos:VBD@[-1] & Pos:IN@[1] |
| 48 | VBD->VB if Word:karatu@[-3,-2,-1] |