ISSN 2007-9737

Identification of POS Tags for the Khasi Language based on Brill's Transformation Rule-Based Tagger

Sunita Warjri[1], Partha Pakray[2], Saralin A. Lyngdoh[3], Arnab Kumar Maji[1]

[1] North-Eastern Hill University,
Department of Information Technology,
India

[2] National Institute of Technology, Assam,
Department of Computer science and Engineering,
India

[3] North-Eastern Hill University,
Department of Linguistics,
India

{sunitawarjri, parthapakray, saralyngdoh, arnab.maji}@gmail.com

Abstract. Khasi is a Mon-Khmer language that belongs to the Austro-Asiatic language family. Khasi language is spoken by the indigenous people of the state Meghalaya in the North-Eastern part of India. The main purpose of this paper is to develop Part-of-Speech (PoS) tagger for the Khasi language using a Rule-based approach. To work on POS tagging, one needs a grammatically tagged corpus. However, the Khasi language does not have a standard corpus for PoS tagging. Therefore, another aim or purpose of this paper is to develop a Khasi lexicon or POS corpus and using the Rule-Based Brill's Transformation to automatically tag the given Khasi text. While anticipating the challenges in building such a corpus, this paper has brought out an analysis based on the Khasi corpus of around 1,03,998 words in its initial phase. We also show in this paper how the Khasi corpus is created. By using Brill's Transformation rule-based learning on the created corpus in this preliminary study, accuracies of 97.73% and 95.52% were obtained on validating data and testing data respectively. This work is the first attempt to investigate POS tagging using the rule-based model with the designed Khasi POS corpus.

Keywords. Natural language processing (NLP), computational linguistic, part-of-speech (PoS), PoS tagging, Khasi language, Khasi corpus, lexical morphology, transformation rule-based tagging.

1 Introduction

Natural Language Processing (NLP) is a zone of Machine Learning, where Natural Languages are made to interact with computer systems. This process of interaction involves linguistic analysis that combines with the computational power of Artificial Intelligence (AI) and Computer Science (CS), i.e. NLP involves the power of natural language theory and its applied component. To make the computer automatically understand the natural language either in the form of text or speech, many fields of NLP is involved. One such field is Part-of-Speech (PoS) Tagging or PoST.

PoS tagging means assigning grammatical class information to the words of a given sentence, according to its context. A system that results in identifying tags for the given input text is called PoS Tagger. POS tagging is also known as grammatical tagging or word category disambiguation. Part-of-Speech (PoS) tagging system is a very strong backbone of NLP as it can be used for other fields of NLP such as Named Entity Recognition (NER), Information

Extraction, Translation of language, and other NLP applications.

To proceed towards PoS tagging, one needs to study in detail the nature of the preferred language and to understand its structure and its grammatical flow. For PoS tagging, tagsets are needed. Tagset is a set that consists of tags or labels. These tags or labels describe the classes of grammatical part-of-speech, which can be used for annotating words of a particular language. For example, tags are basically labels such as NN, ADJ, which represent Nouns and Adjectives respectively.

One of the PoS taggers was built by E. Brill [6]. This PoS tagger was developed by using the rule-based algorithm. The rule-based tagging approach is still commonly used today for tagging for languages. The rule-based tagging is also called a Transformation based tagging method. This transformation-based tagging automatically tags the given input words and produces the word along with its belonging tag as an output. Contextual rules and regular expression rules are the main components for developing the transformation rule-based algorithm. This algorithm is the oldest method towards checking on the context rules.

Other tools that perform POS tagging include Stanford Log-linear Part-of-Speech Tagger [25], Tree Tagger [21], Hidden Markov Model (HMM) based Tagger [4], and etc. Though, there are other tools for the PoS tagging method, Brill's transformation rule-based method has been opted and used in this paper.

As each language differs in terms of the utterance, spelling system of a language, writing, and also the flow and formation of the sentences. This makes the grammatical rules to be different in different languages. Therefore, in this paper, we perform research and analyze the Khasi language using this transformation rule-based approach. In Brill's transformation rule-based method, one does not have to design the complex detailed rules for a particular language. The transformation ruled based learning has the property of complex crossing with the manually generated rules and the machine-learned rules from the corpus.

Transformation rule-based tagger is said to be more accurate than Hidden Markov Model (HMM) based Tagger [4, 13] in terms of producing and generating high accuracy and also to tag the given words correctly. However, some related research works show HMM to be more accurate than the rule-based method [11, 12, 18]. This is indicative of the fact that shows that different languages exhibit different flow on different systems.

Research works or literature on Khasi from Computational Linguistics (CL) perspectives are inadequate. PoS tagging being the most important initial task towards other fields of NLP, we believe that this research is going to contribute to the realm of NLP study of Khasi language. It will help this indigenous language of India to develop and get recognition in the world as well as in the country.

The paper is organized as follows: Section 2 describes the existing related works on POS Tagging; Section 3 describes methodology used in Khasi Part-of-speech Tagging (KPOST); Section 4 describes some of challenges for Khasi corpus building; Section 5 shows the experimental results ; Whereas Section 7 consists of Conclusions and some future perspectives of the work.

2 Existing Literature on Rule-Based Tagger

The rule-based approach has been used extensively in many languages. In this section, we will have a brief discussion on some of the NLP related works, that have used the rule-based methodology.

In paper [15], the rule-based and statistical approach has been introduced in the study of the Arabic language. Pre-processing of the lexicon has been done before tagging automatically, as the Arabic language is morphologically rich. In this paper, the corpus has been created manually using the tag set comprising of 131 tags. Data were taken from newspaper and published paper for building the corpus: from "AlJazirah" newspaper 59,040 words, from "Al-Ahram" newspaper 3,104 words, from "Al-Bayan" newspaper 5,811 words, and from "Al-Mishkat" published paper in social science 17,204 words were taken. After disambiguation of the ambiguous words, the accuracy of around 90% is achieved from this statistical tagger.

ISSN 2007-9737

In paper [1], tagging on Modern Arabic text has been carried out using the Transformation-based learning technique. The Arabic Tree Bank (ATB) [10] has been used as a corpus in this paper. The corpus consists of 770k words. The annotated corpus includes the syntactic trees and morphological analysis also. Experiments are conducted twice in this research work. The training accuracy of 98.50% is achieved from experiment 1 and 97.90% from experiment 2. Evaluation of testing data has also been carried out and the accuracy 96.90% and 96.15% are obtained for experiment 1 and experiment 2 respectively.

In paper [9], a discussion is made on the rule-based technique for Part-of-Speech (PoS) tagging and Named Entity Recognition (NER), for the Arabic language. For PoS tagging, the lexicon phase and morphological phase have been used. The POS tagger had been tested on 793 words, out of which 679 words are correctly tagged. For NER tagging, 480 words are tagged correctly from 490 different words.

In paper [16], authors present POS tagging about the Sindhi language, where the Rule-based approach has been applied. In this work, rules are framed to disambiguate the words and a lexicon is developed for the Sindhi language. Tags set consisted of 67 different tags, which are used for designing supervised corpora. The training corpus, which is used in this paper, has 26366 words and a testing corpus of 6783 words are used. An accuracy of 96.28% is achieved for the Sindhi language. A Sindhi linguist has verified the data analyzed in this paper.

Several Indian languages are also explored using this methodology. Some of them are described in this section. In paper [23], POS tagging for Manipuri language has been discussed using the Rule-based method. The author has segmented the affixes from the roots by using the stripping technique. Fewer corpus resources are used in this work. Using some POS rules, the system is able to achieve an accuracy of 50% for 100 words with 5 rules, 77% for 500 words with 15 rules and 85% for 1000 words with 25 rules.

In paper [10], POS tagging using a rule-based approach with layered tagging has been introduced for Bangla language. Morphological analysis is performed for Nouns and Verbs. Based upon the postfix of words, morphological analysis is generated. With regard to morphological features, the words were classified to post positions, numbers, and classifiers. The verbs are classified according to the morpho-syntactic formation of the root and then classified to honorific and persons. A four-level layered architecture of the work has been discussed in this work. Assignment of tags to words at level 1, rules to disambiguate at level 2, multiple word categories of the verb at level 3 and in level 4, chunking of words are performed.

In paper [12], Rule-Based Part-of-Speech Tagger has been used to analyze the Hindi language. Corpus for this Tagger consists of 26,149 tagging words with 30 tags. The data are collected for the corpus consists of short stories, news, and essays. Using a rule-based tagger, the system achieved 87.55% as overall accuracy on the testing data.

Also, Recall, Precision, and F-Measure parameters are used for the result computation. The Recall score, that has been achieved in this work, is 92.84% for news, 87.32% for essays and 88.99% for short stories. The precision score is 89.94% for news, 81.36% for essay 85.11% for short stories. An F- score of 91.37% for news, 84.23% for essay and 87.06% for short stories, is achieved in this work.

In paper [14] also, PoS tagging for Bangla language using the rule-based approach has been explored. This tagging system is based on words suffix rules and stemming technique. Corpus having 45,000 words along with their corresponding tags is used in this system. Using some ruleset with a verb dataset, the accuracy of 93.7% is achieved, which is significantly improved accuracy than the other existing POS tagging system for Bangla.

In paper [22], POS tagging and morphological analysis for the Tamil language has been discussed. The alignment and projection techniques have been used to project the POS tags. Lemmatization (i.e. to analyze vocabulary and morphological words correctly) and induction methods are also employed for getting the root words from English to Tamil. During the testing phase, 85.56% accuracy is achieved for Bible

corpus and accuracy of 83% for CIIL corpus. An improvement of the system is also proposed, which obtained the accuracy of 92.48% for Bible test corpus.

The achieved accuracy is having an improvement of 7% in comparison to the previous accuracy of 85.56%.

In the paper [19], Part-of-Speech tagging for the Indonesian language is presented using rule-based approach. In this work, the Indonesian large dictionary or KBBI is utilized with some morphological rules for POS tagging. Using PAN Localization corpus in 4 parts for Indonesian an average accuracy rate of 87.4% is achieved.

The paper [20], describes part-of-speech tagging for the Indonesian language using a rule-based approach. The manually tagged corpus is used in this work which consists of roughly 250.000 tokens. Using the corpus the system yields an accuracy of 79%.

Some PoS tagging work on Khasi language has also been done. In paper [24], Khasi POS tagging based on HMM tagger had been discussed. The corpus consists of 86,087 tokens with 5,313-word types. NLTK tool tagger has also been applied to the same corpus in [24]. Accuracy of 86.76%, 88.23%, 88.64%, 89.7%, 95.68% is obtained for Baseline Tagger, NLTK Bigram Tagger, NLTK Trigram tagger, NLTK Tagger, and HMM POS Tagger respectively.

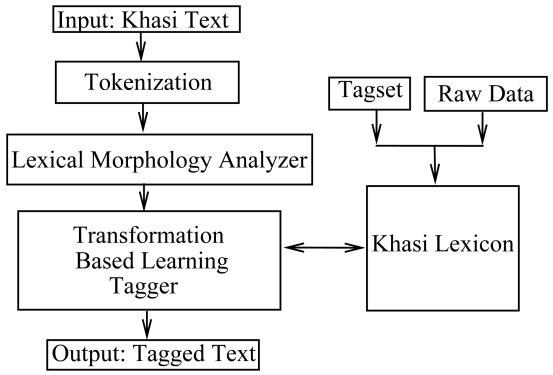

Fig. 1. Architecture for Khasi Part-of-Speech tagging (KPOST)

In paper [28], the Khasi POS tagger has been developed based on the HMM method. The lexicon consists of around 7,500 words for training and 312 words for testing data. Using the manually tagged Khasi lexicon on the HMM-based POS tagger, an accuracy of 76.70% was obtained. Also POS tagging using CRF for Khasi had been discussed in paper [29, 30].

Concerning Rule-based, it is also used in different fields of NLP such as Machine translation. In paper [2], noun phrase translation using a rule-based approach has been discussed. This automatic translation was done from the Punjabi language to the English language.

For this purpose, many steps are taken place such as: 1. Pre-processing, 2. Tagging, 3. Ambiguity Resolution, 4. Translation, and, 5. Synthesis. Around 2000 phrases are used for training the system and 500 sentences are used for testing. The overall translation accuracy, which is achieved by the system is around 85%.

3 Methodology for Khasi Part-of-Speech Tagging (KPOST)

In this paper, the POS tagger for Khasi language, based on Brill's Transformation and Khasi morphological rules, have been employed. In the subsections below, there is a brief discussion on the methods that have been carried out in this work. Architecture for KPOST is shown in Figure 1.

3.1 Tokenizer

This is the first step for POS tagging. It is used for separating the required words, symbols, and punctuation of the given text by assigning space between words. In our designed corpus, we have split some of the words for more clarifications. Therefore, for the given input text, split or replacement of words, is also done accordingly as per requirement based on our designed corpus.

ISSN 2007-9737

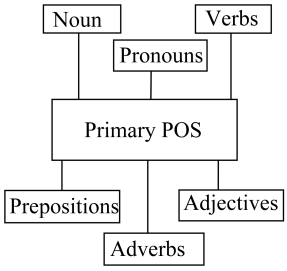

Fig. 2. Classes of Primary Parts-of-Speech

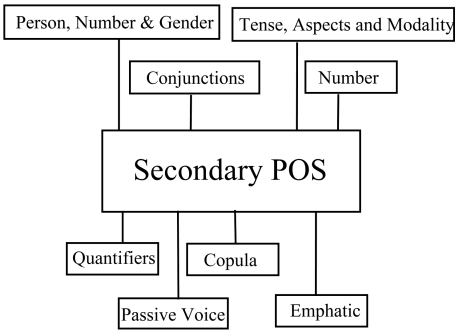

Fig. 3. Classes of Secondary Parts-of-Speech

3.2 Khasi Lexicon

Lexicon is a large collection of vocabulary data, along with its corresponding information, which is used for analyzing computational linguistic. In this work, for building the Khasi lexicon, we have used tag sets comprising of 53 tags. Out of 54 tags discussed in [27], we have used only 53 tags. The tag-sets that we have used in our proposed work can be found in Table 1. Simple graphical representations of the proposed grammatical classes used in the designed tag sets are shown below in Figures 2 and 3.

Figures 2 and 3 represent the visual art of classes of Part-of-Speech (POS), used for categorizing Khasi text in this work. Figure 2 represents the Class of Primary Part-of-Speech and Figure 3 represents the Class of Secondary Parts of Speech. More details regarding the grammatical classes used for POS tagging, designed tag-sets of Khasi language can be found in [27].

The Khasi corpus has been built manually, by tagging the raw text or words to its corresponding tags by using the designed tagsets. The manually tagged data represent the POS information of each Khasi word in the corpus. The raw data for corpus have been collected form Khasi newspapers [17], which are available online. The raw data that are collected mainly comprises of the political news and article news. Prepossessing of the corpus context, such as splitting and orthography correction are done manually.

However, the information maintained in the Khasi corpus or lexicon which we have discussed in paper [28] differs from the corpus data used in this paper. The corpus size is more in this present work than the earlier [28] as well as changes are made is few tagged words.

In this work, we have to build a Khasi corpus manually, which consists of around 1,03,998 words, with 6645 numbers of distinct words. For example, some Tagged Khasi sentences from the corpus are shown in Table 1.

These manually tagged Khasi words in the lexicon corpus are created under the observation and verification of a linguistic expert from the Department of Linguistics, North-Eastern Hill University, Shillong, India.

As this research work is an initial work towards computational Linguistic for Khasi, as a remark, We are releasing some of the designed Khasi POS corpus data. The corpus is available online at [26].

3.3 Lexical Analyzer

The Lexical analyzer is used to analyze the given input words, according to the predefined lexical rules. Lexical rules are used as foundation for analyzing formation of words. Lexical rules are the basic special compositions or building blocks of a word vocabulary. Lexical rules are recorded to generate a productive resource in computational linguistic. There are three types of Lexical rules [3, 5]. They are 1. Inflectional (lexeme to word), 2. Derivational (lexeme to lexeme), and 3. Post-Inflectional (word to word). To analyze morphological processes of Khasi words, Derivational method is widely used in this work. Some prefixes of the words are

Table 1. Khasi Lexicon: Manually tagged dataset for training

Sentences	Khasi Lexicon
1	Shi/QNT sngi/CMN hadien/ADP ba/COM la/VST dep/ITV ka/3PSF jingïathep/ABN vote/FR ba/COM n/VFT jied/TRV ïa/IN ki/3PPG MDC/FR jong/POP ka/3PSF KHADC/FR bad/COC JHADC/FR ,/SYM ka/3PSF lyer/CMN ka/3PSF sdang/TRV ba/COM n/VFT beh/TRV ba/COM ka/3PSF jingïakhun/ABN ka/3PPG n/VFT long/CO hi/EM hapdeng/ADP ka/3PSF Congress/FR ,/SYM ka/3PSF UDP/FR bad/COC ka/3PSF NPP/FR ha/IN KHADC/FR bad/COC ha/IN JHADC/FR ka/3PSF jingïakhun/ABN ka/3PPG n/VFT long/CO hapdeng/ADP ka/3PSF UDP/FR bad/COC ka/3PSF NPP/FR ./.
2	Ki/3PPG ba/COM bun/QNT ki/3PPG riew/CMN sngap/ITV lyer/CMN ki/3PPG la/VST sdang/TRV ba/COM n/VFT ïathuh/TRV lypa/AD ruh/COC ne/CRC ba/COM n/VFT ‘/SYM predict/FR ’/SYM ba/COM na/IN kata/DMP ka/3PSF Constituency/FR yn/VFT jop/ITV uta/DMP ,/SYM yn/VFT jop/ITV kata/DMP ./.
3	Ki/3PPG n/VFT don/CO napdeng/ADP kine/DMP ki/3PPG ba/COM lah/VSP ba/COM n/VFT dei/CO bad/COC ki/3PPG n/VFT don/CO ruh/COC ki/3PPG ba/COM n/VFT lait/ITV ./.

also analyzed, by using lexical rules to identify the POS categories of the given words. The morpho-syntactic processes, that are engaged for the analysis of lexical resource, are nominalization, agentivization, and causation. In Khasi, a series of prefixes indicate the existence of double morphological rules. For instance, nominalization, followed by causativization and by root word, as in *“Nongpynsum"* which means “one who causes/give a bath to somebody”.

NONG+PYN+SUM $\rightarrow$ NONG is Agentive, PYN is Causation and SUM is Verb.

Some words are followed by single morpho-syntactic rules such as nominalization or causativization, words like *“Nongsum"* and *“Pynsum"*. Word *“Nongsum"* mean “one who bathes" and word *“Pynsum"* means “to cause/give a bath to somebody".

NONG+SUM $\rightarrow$ NONG is Agentive, and SUM is Verb.

PYN+SUM $\rightarrow$ PYN is Causation, and SUM is Verb.

Table 2, shows some examples of Khasi prefixed words used in generating morphological lexical rules.

3.4 Brill’s Transformation-based Learning (TBL)

Transformation-based learning (TBL) is an algorithm that uses the rule-based technique to categorize the part-of-speech classes and tag the given words automatically. The transformation-based Rule for POS tagging was first introduced by E. Brill [7, 8]. Using the Lexicon or Gold corpus, it will learn automatically the actual information that is linguistically hidden in sentences of the languages. This complex property maintains the flow of the grammatical class orders of sentences. This learned information helps the system to tag the words appropriately, according to their corresponding meanings.

In the present paper, analyzes accounted are based on the Transmission Rule-Based method of identification of POS tagging introduced by Brill. This model, at the initial stage, uses the un-annotated Corpus (or Raw data), which consists of input sentences that are not tagged or labeled. These sentences are then passed to the Initial State Annotator. Here, each word of the input text is annotated i.e. tags are assigned. To generate the possible tags for each word, this method makes use of a dictionary or lexicon (train data). The Temporary Corpus is the output that comes out from the Initial State Annotator. The Gold Corpus consists of the texts which are

Table 2. Some of Khasi Lexical morphology rules

Sl. No.	**Lexical morphology**	**Khasi Words**	**Meaning**	**POS Tags**
1	Nomilization	Jingpule Jingbatai	Studies Instructions	Jing->ABN
		Jinglum jingkhuid	Collections Cleanliness	jing->ABN
2	Causation	pynkulmar pynshai	To make trouble To make clear	Pyn->CAV
		pyntreikam pynkhuid	To make it work To make it clean	pyn->CAV
3	Agentive	Nongjop Nongthoh	Winner Writer	Nong->CMN
		Nongaïlam Nongrem	Leader Loser	nong->CMN

already tagged manually. This Khasi Lexicon is used to compare with the Temporary Corpus using some rules.

The tagger used two types of transformation processes which consist of rules: the Lexical and the Contextual Learner. In the Lexical learner module, the most frequent tag is used to assign for the known words, for the unknown words this module uses some system-generated rules to tag. Basically, in this architecture, the words are tagged by following or looking at the Khasi Lexicon or corpus. For particular words, that are not present in the corpus and the generated rules are unable to tag it, then that word is tagged as UNK (unknown).

In the Contextual learner module, Khasi Lexicon and some rules are used, to achieve high accuracy in reading and representation of the tagger. These rules are based on the flow of the context. As a word may belong to more than one tag, these rules act as a very important process, as they are used to identify the correct tags. These rules are automatically created by the Transformation-Based Error-Driven Learning (TEL). The rules are used to change the incorrect tags by applying the rules repeatedly.

Thus the Brill's transformation-based method uses some remarkable predefined set of rules as well as automatically generated rules that are learned by the system while training the data. Simple steps are given below which is followed in building the Khasi POS tagger:

— **Step 1.** Load the Corpus file.

— **Step 2.** Converting data from "word/tag" pair to a list containing tuple. i.e,[(word1,tag),(word2,tag)].

— **Step 3.** Set some predefined rules or lexical rules.

— **Step 4.** Feed the corpus data to BrillTagger-Trainer() for training data.

— **Step 5.** Evaluating and print the accuracy.

— **Step 6.** Printing the confusion matrix of standard tags with the test tags.

— **Step 7.** Calculate the Precision, Recall, F1-score.

In this experiment our **Input** to the system is the Khasi POS corpus (training data, validating data, test data) and raw data(Un-tagged words). The **Output** achieved from the system are Accuracy and Tagged words. Some of the predefine rules are as discussed in Sub-section 3.3 and also shown in Table 2. The achieved results whit the data are shown and discussed in Section 5.

4 Challenges on Corpus Building

In this section, we discuss some of the challenges, we met while building the corpus. These problems are (i) problems of spellings and orthography, and

(ii) problems of Ambiguity. As we have collected the raw data from online Khasi newspapers, we have found out that words which have the same meanings are spelled differently in different newspapers and in different articles. This is due to the reason that standardization of spellings and orthography of Khasi is still to be established. Khasi orthography has 23 alphabets and 6 vowels, the vowels are: “a e i ï o u". The Khasi alphabets are shown in Table 3.

Throughout our corpus building process, we found the same words are spelled differently by different writers. Such as, in some words, an alphabet is placed with *“ i "* instead of *“ ï "* and similarly, for the alphabets *“ n "* or *“ ñ "*. In the Khasi newspapers context, orthography problems are found to a large extent. Some such words are shown in Table 4.

Another technical problem that we have encountered is the presence of ambiguous words based on context. Some of such ambiguous words in Khasi are shown in Table 5, with their corresponding tags which are tag using the designed tagsets For instance, some words that are spelled alike but pronounced differently and have distinct meanings. The word like *"hadien"* which means *"after"* can be the Preposition or *"hadien"* which also mean *"behind"* can be a grammatical class for Adverb of Place.

Table 3. Khasi alphabet

a	b	k	d	e	g	ng	h	i	ï	j	l
m	n	ñ	o	p	r	s	t	u	w	y	

Table 4. Orthographic words in Khasi Language

Words	Orthography
ïing	ïing **or** ïing
ïathuh	ïathuh **or** iathuh
pynïoh	pynïoh **or** pynioh
Hynñiew	Hynñiew **or** Hynniew
Hynñiewtrep	Hynñiewtrep **or** Hynniewtrep
Rilum	Rilum **or** Ri lum
Kin	Kin **or** Ki yn

Table 5. Ambiguous words in Khasi Language

Khasi word	Meaning	Corresponding POS tags
hadien	after	IN
hadien	behind	ADP
shitom	Sick	CMN
shitom	tough/hard/difficult	ADJ
rong	color	CMN
rong	carry/blown away	TRV
mar	as soon as	AD
mar	material	MTN
mar	distribute	AD

Table 6. Most common words count

Sl.No.	Khasi words	Count in Khasi Corpus
1	ka	12461
2	ki	6888
3	ba	5286
4	u	4397
5	ïa	4124
6	la	2965
7	ban	2771
8	ha	2231
9	bad	1638
10	jong	1457
11	ngi	1006
12	na	793
13	dei	667
14	ruh	612
15	kane	598

5 Experimental Results

In the subsequent subsections, a brief discussion is presented based on the experimental work conducted on the Khasi lexicon or corpus along with the achieved results and its analysis.

5.1 Results

In this work, the method of Transformation-based learning (TBL) has been used to identify POS classes for the Khasi language. In comparison to the Khasi language, this is the first attempt to investigate POS tagging using the rule-based model with the designed Khasi POS corpus. The designed Khasi POS corpus consists of 4,580 sentences.

Table 6, presents the 15 most common words count in the corpus. We also present the

Table 7. Distribution of PoS Tags in the Designed Khasi corpus

Sl.No.	Tags	Desctription	Count in Khasi Corpus
1	QNT	Quantifiers	2056
2	CMN	Common nouns	9101
3	ADP	Adverb of Place	1048
4	COM	Complementizer	8381
5	VST	Verb, past tense	2945
6	ITV	Intransitive verb	1672
7	3PSF	3rd Person singular Feminine	12461
8	ABN	Abstract nouns	3166
9	FR	Foreign words	3516
10	VFT	Verb, future tense	4143
11	TRV	Transitive verb	2867
12	IN	Preposition	5134
13	3PPG	3rd Person plural common	6888
14	POP	Possessive Pronoun	1881
15	COC	Coordinating conjunction	1418
16	SYM	Symbols	4636
17	CO	Copula	5286
18	EM	Emphatic	712
19	AD	Adverb	4322
20	CRC	Correlative conjunction	304
21	DMP	Demonstrative Pronouns	1464
22	VSP	Verb, past perfective participle	203
23	ADD	Adverb of degree	591
24	NEG	Negation	2007
25	ON	Ordinal number	74
26	SUC	Subordinating conjunction	919
27	SPA	Superlative Adjective marker	752
28	CN	Cardinal Number	1127
29	RFP	Reflexive Pronouns	103
30	CAV	Causative Verb	1116
31	ADJ	Adjective	1575
32	ADF	Adverb of frequency	71
33	3PSM	3^{rd} Person singular Masculine gender	4397
34	PPN	Proper nouns	2713
35	MOD	Modalities	421
36	CLF	Classifier	369
37	ADT	Adverb of Time	921
38	DTV	Ditransitive verb	583
39	RLP	Relative Pronouns	93
40	CLN	Collective nouns	377
41	1PSG	1^{st} Person singular common gender	229
42	VPP	Verb, present progressive participle	539
43	3PSG	3^{rd} Person singular common Gender	199
44	PAV	Passive Voice	178
45	1PPG	1^{st} Person plural common gender	1009
46	CMA	Comparative Adjective marker	199
47	INP	Interrogative Pronouns	275
48	2PG	2^{nd} Person singular/plural common gender	145
49	MTN	Material nouns	114
50	DTV	Ditransitive verb	45
51	ADM	Adverb of Manner	7
52	2PF	2^{nd} Person singular/plural Feminine	7
53	2PM	2^{nd} Person singular/plural Masculine gender	5

distribution of tokens concerning POS-tags in the designed Khasi corpus along with its specifications as shown in Table 7. For more details regarding the tagsets, it can be found in [27].

From the designed Khasi corpus 84,972 manually tagged words are used as training data for the system, out of which 6,645 are the distinct Khasi words. For validating data 14,686 tagged words are used and 4340 words are used as testing data.

Using the validated data alongside the training data on the system an accuracy of 97.73% is yielded as performance. Due to the non-availability of lexicon or corpus, therefore we have created

Table 8. Validation result of our proposed Khasi POS tagger

Sl. No.	Khasi Lexicon Size	Rules Generated	Validation Accuracy
1	Training data (20,280 tokens) Validating data (5,323 tokens)	15005	86.79%
2	Training data (30,580 tokens) Validating data (5,323 tokens)	20102	88.84%
3	Training data (40,920 tokens) Validating data (5,323 tokens)	26017	91.83%
4	Training data (84,972 tokens) Validating data (14,686 tokens)	53,577	97.73%

Table 9. Validating result achieved using state-of-art and proposed TBL system for the Khasi Lexicon

Sl. No.	Khasi Corpus	Technique	Accuracy
1.	Training data (84,972 tokens) Validating data (14,686 tokens)	NLTK Bi-gram	88.88%
2.		NLTK Tri-gram	88.15%
3.		combining (Bi-gram+Trigram)	92.55%
4.		TBL **(Proposed work)**	97.73%

a Khasi lexicon or corpus. It is expected that the accuracy will increase further if more data are added to the corpus.

Table 8 represents the different validation results of the Transformation rule-based learning method for Khasi POS tagging using the designed Khasi lexicon. The result shows that as the data in the lexicon increase and fed to the model, the validation accuracy and generated rules are also increased.

Table 10. Some sample rules generated by the system

TBL Generated rules
11 11 0 0 | COM->None if Word:ba@[0] & Word:@[1] & Word:n@[2]
5 5 0 0 | COM->3PSF if Word:@[0] & Word:ka@[-1]
5 5 0 0 | 3PSF->None if Word:ka@[0] & Word:@[1]
4 5 1 0 | ITV->TRV if Word:bynrap@[0] & Word:shuh@[1] & Word:shuh@[2]
4 4 0 0 | TRV->ITV if Word:ong@[0] & Word:.@[1] & Word:“@[2]
4 4 0 0 | TRV->ITV if Word:nonghikai@[1,2,3]
4 4 0 0 | .->SYM if Word:.@[-3,-2,-1]
3 3 0 0 | 3PPG->None if Word:ki@[0] & Word:@[1] & Word:ba@[2]
3 3 0 0 | CMN->TRV if Word:pule@[0] & Word:-@[1] & Word:puthi@[2]

We have also compared the achieved results of the proposed TBL system with some state-of-the-art techniques shown in Table 9. From Table 9, we can observe that the TBL system has outperformed the state-of-the-art method.

Table 10 represents some sample sets of rules generated by the TBL system for Khasi which is generated from the trained Khasi lexicon. A comparison of our achieved result from the designed Khasi POS corpus using rule-based, with other related work on POS tagging that uses a rule-based approach is presented in Table 11.

Table 11 describes the corpus size used during the experiment, the different languages, and the accuracy achieved in percentage. From the table, we can observe that the proposed Khasi POS tagging which is experimented with the manually designed Khasi corpus give accurate result in comparison to other languages. We strongly suspect that more accuracy may be obtained, if more data are added to the training corpus.

The precision, recall, F-score are calculated from the confusion matrix so that the false positives, true positives and false negatives values can be accessed. The confusion matrix for training data with 40,920 tokens and validating data with 5,623 tokens is shown in Table 12.

Table 11. Comparison with some existing work on Khasi language for rule-based tagging

Sl. No.	Corpus	Generated Rules	Accuracy	Language References
1	Corpus data of around 85,159 tokens	-	90%	Arabic [15]
2	corpus consists of around 770k words	experiment 1 - 255 and experiment 2 - 1500	Training experiment 1 - 98.50% Training experiment 2 - 97.90% Testing experiment 1 - 96.9% Testing experiment 2 - 96.15%	Arabic [1]
3	Training data (26366 words) Testing data (6783 words)	-	96.28%	Sindhi [16]
4	corpus consists of 1000 words	For 100 words - 7 For 500 words - 15 and For 1000 words - 25	For 100 words - 50% For 500 words - 77% and For 1000 words - 85%	Manipuri [23]
5	Corpus of 26,149 words	-	87.55%	Hindi [12]
6	Corpus of 45,000 words	-	93.7%	Bangla [14]
7	KBBI	-	87.4%	Indonesian [19]
8	250,000 tokens	-	79%	Indonesian [20]
9	training data (84,972 tokens) validating data (14,686 tokens) testing data (4,340 tokens)	53528	validating $\rightarrow 97.73\%$ testing $\rightarrow 95.52\%$	Khasi (Our proposed KPOST)

In Table 12, Rows represent the actual values and column represent the predicted values. The true positive value can be identified as the intersection of row and column. Such as 1PSG in Row and 1PSG in the column the true positive value is 8.

From the table, the wrongly tagged result which is counted as false positive and false negative can be found. Such as, from the actual values, we can see there are 3 wrongly tagged outputs for 3PSF as 3PSM, which is counted as false positive for 3PSM. Similarly, there are 3 false negatives for 3PSF.

Again to have a close look at results we present the confusion matrix for training data with 84,972 tokens and validating data with 14,686 tokens in Table 13.

From the confusion matrix, one can calculate the recall, F-score, and precision of the validating data after training the model. The formula for calculating the recall, F-score, and precision can be express as below:

$$precision = \frac{TP}{TP + FP}, \tag{1}$$

$$recall = \frac{TP}{TP + FN}, \tag{2}$$

$$fscore = \frac{2 * (precision * recall)}{precision + recall}, \tag{3}$$

Table 12. Confusion matrix table of validating data 5,623 tokens when trained with 40,920 tokens

Tags	.	1PPG	1PSG	3PPG	3PSF	3PSG	3PSM	ABN	AD	ADD	ADJ	ADP	ADT	CAV	CLF	CLN	CMA	CMN	CN	CO	COC	COM	CRC	DMP	DTV	EM	FR	IN	ITV	MOD	MTN	NEG	ON	PAV	POP	PPN	QNT	RFP	RLP	SPA	SUC	SYM	TRV	VFT	VPP	VSP	VST	XX
.	324																																																	
1PPG		8																																																
1PSG			8																																															
3PPG				247																																														
3PSF					650		3											1																																	
3PSG						6	1																																										
3PSM					31		340											1																																	
ABN								145										2																																	
AD									1305	6								5																																	
ADD									5	36								2					3																											
ADJ								2	12	2	76	2						21																																	
ADP									1			13						3											1																					
ADT									1		1		48					3											1																					
CAV									1					75																																			
CLF															9	3																																	
CLN																22					3																												
CMA																	2																																
CMN							1		1		7	1						602		2				4		1		1																				
CN																			56										1																				
CO																				97																													
COC																					160							3																				
COM																						339																										
CRC																							2																									
DMP																		1						71																								
DTV																		1							15			1																				
EM																										39																							
FR																		78	1								2391																					
IN									9			10						1							1			434																				
ITV						1												6										48																				
MOD																													8																			
MTN																														0																		
NEG																															33																	
ON																																0																
PAV																																	4															
POP									1																									95														
PPN																		65		3				7											179													
QNT										1								2	2																	63												
RFP																																				6											
RLP																																			1	0										
SPA																																					20									
SUC									1									1		6																	25		1	1					
SYM																		1																				196								
TRV									8		2			1				28						2				26										1	2242							
VFT																																							144							
VPP									8																														1	9						
VSP																																								0						
VST																																	2								173				
XX																				1																							1				0	

where *TP is true positives*, *FP is false positives* and *FN is false negatives*.

We present Table 14, which shows the recall, f-score, and precision of each tag of validating data produced by the system when used the training data of 84,972 tokens and validating data of 14,686 tokens. Where we can observe the performance of the system as the data in the corpus is increased.

To have a good view of the system for the designed Khasi corpus we present the graphical view Figure 4 and 5. Figure 4, is the graphical representation of the F1-score of each tag using 40,920 tokens as training data in the proposed

Table 13. Confusion matrix table for validating data 14,686 tokens with 84,972 tokens as training data

Tags	.	1PPG	1PSG	2PG	3PPG	3PSF	3PSG	3PSM	ABN	AD	ADD	ADF	ADJ	ADP	ADT	CAV	CLF	CLN	CMA	CMN	CN	CO	COC	COM	CRC	DMP	DTV	EM	FR	IN	INP	ITV	MOD	NEG	ON	PAV	POP	PPN	QNT	RFP	RLP	SPA	SUC	SYM	TRV	VFT	VPP	VSP	VST	XX
.	623																																																	
1PPG	44																																																	
1PSG		12																																																
2PG			7																																															
3PPG				7306																																														
3PSF				22	1703	9												1										1																							
3PSG						7	1																																											
3PSM					31		902																																											
ABN								327										2																																
AD									3854	3								2																																
ADD									3	122								1						1																										
ADF										1																																							
ADJ									8	2291							5										1	1																						
ADP									5		56																	8																					
ADT									2	1		113																2																					
CAV													182																																				
CLF														32	2																																		
CLN															55			3																															
CMA																15																																	
CMN							1			1		6	3			1		1434		1						4	1																						
CN																		201																															
CO																			2791																														
COC									1											389																														
COM																			1	3	910																												
CRC																						23																											
DMP																					1		198																										
DTV																								29		1																						
EM																									87																							
FR																		14								1021																						
IN									1			2									2			2		1163																						
INP																							1		11																						
ITV						1						1															231																					
MOD																												26																				
NEG																											1		97																			
ON																														5																	
PAV																															14																
POP									2																							233															10	
PPN																		2								4							549														
QNT									1		1							1																146	1												
RFP																																		12													
RLP																																			2												
SPA																																				46											
SUC									1											10																	93										
SYM						1																				1												640								
TRV										4		1						4									6												1	6921							
VFT																				2									1										1	419	1					
VPP																																									1		30				
VSP																												1																3			
VST																													1		1													1	429	
XX																																															5

tagging system for Khasi. Figure 5, is the graphical representation of the F1-score of each tag using 84,972 tokens as training data in the proposed tagging system for Khasi.

From Figure 4 and 5 we can easily observe the tagging system it improve as we increased the numbers of data in the Khasi corpus.

5.2 Performance Analysis

To check the performance of the system, testing data is used. The testing data are the raw data which are un-tagged. These raw data are given to the system to tag them automatically by learning the rules and tags from the given Khasi lexicon

ISSN 2007-9737

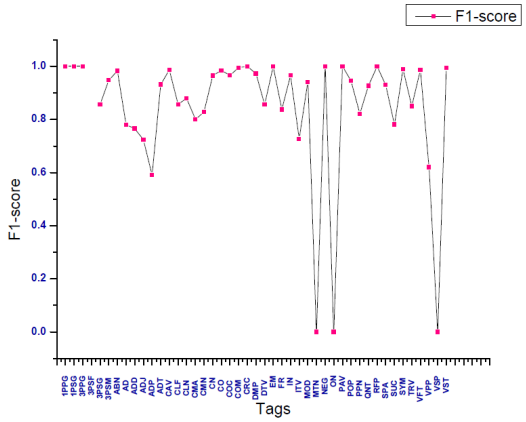

Fig. 4. Graphical representation of f1-score on individual POS tags using 40,920 tokens in the proposed system

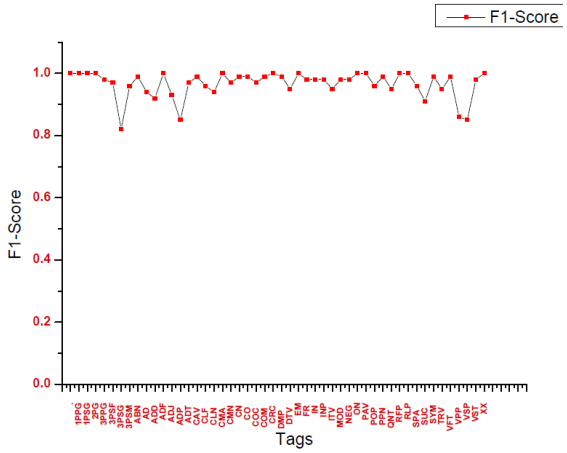

Fig. 5. Graphical representation of f1-score on individual POS tags using 84,972 tokens tokens in the proposed system

that act as gold corpus. As an output for the input raw Khasi data, the system produces the tags as an output for each word. Table 15 shows the sample resulting tags for particular words, which are given as input Khasi text. Table 15 shows the input raw Khasi text in the first row, and the second row represents its corresponding tags with its particular words. The output results are shown and verified by the Linguistic expert and found to be satisfactory.

The testing accuracy is shown in Table 16, using the proposed Transformation based learning (TBL) approach on the developed Khasi lexicon. We have checked the accuracy of our work in a phase-wise manner, where the Input data and output data are shown in Table 15. With the training data of 20,280 tokens and testing data of 970 tokens, the system yields an accuracy of 70.39%.

Whereas, with 30,580 tokens used for training and 1409 tokens used for testing, the accuracy of 79.60% is achieved, again for training data of 40,920 tokens and testing data of 3,122 tokens 90.97% accuracy is achieved. The accuracy increases gradually with the corpus size as shown in Table 16.

It is observed that 95.52% accuracy is achieved, in case of training data that consists of 84,972 tokens and testing data that consists of 6,643 tokens.

6 Discussion

This paper presents a brief discussion on the design Khasi POS corpus. Using the manually designed POS corpus several experiments are executed by dividing the corpus into small sections. The experiment shows an improvement in the performance regarding the identification of tags and it also increases in accuracy results, when the corpus size is increased.

This work is the first instance that uses the Transformation-based learning (TBL) approach for PoS tagging in the Khasi language. This work is indeed a resource work building for Khasi language towards the NLP perspective.

7 Conclusion and Future Works

In this paper, the transformation based method for POS tagging on the designed POS corpus for the Khasi language is presented. The designed Khasi corpus with some challenges and other related POS tagging issues of computing Khasi are also addressed in this paper. This paper presents the first application of Brill Tagger for POS in Khasi.

The accuracy results of validating data and testing data are shown in the tables and the different scores for validating data are also presented. We have also compared the accuracy

Table 14. The tags precision, recall and f-score for training data of 84,972 tokens and validating data of 14,686 tokens

Sl.no.	Tags	Precision	Recall	F1-score
1	.	1.0	1.0	1.0
2	1PPG	1.0	1.0	1.0
3	1PSG	1.0	1.0	1.0
4	2PG	1.0	1.0	1.0
5	3PPG	0.97	0.99	0.98
6	3PSF	0.97	0.98	0.97
7	3PSG	0.77	0.87	0.82
8	3PSM	0.98	0.95	0.96
9	ABN	1.0	0.99	0.99
10	AD	0.94	0.94	0.94
11	ADD	0.96	0.89	0.92
12	ADF	1.0	1.0	1.0
13	ADJ	0.93	0.93	0.93
14	ADP	0.90	0.81	0.85
15	ADT	1.0	0.94	0.97
16	CAV	0.99	1.0	0.99
17	CLF	1.0	0.94	0.96
18	CLN	0.94	0.945	0.94
19	CMA	1.0	1.0	1.0
20	CMN	0.97	0.97	0.97
21	CN	0.99	1.0	0.99
22	CO	0.99	0.99	0.99
23	COC	0.96	0.99	0.97
24	COM	0.98	0.99	0.99
25	CRC	1.0	1.0	1.0
26	DMP	0.99	0.99	0.99
27	DTV	0.93	0.96	0.95
28	EM	1.0	1.0	1.0
29	FR	0.99	0.98	0.98
30	IN	0.98	0.99	0.98
31	INP	1.0	0.91	0.95
32	ITV	0.96	0.84	0.90
33	MOD	0.96	1.0	0.98
34	NEG	0.97	0.98	0.989
35	ON	1.0	1.0	1.0
36	PAV	1.0	1.0	1.0
37	POP	0.97	0.95	0.96
38	PPN	0.99	0.98	0.99
39	QNT	0.93	0.97	0.95
40	RFP	1.0	1.0	1.0
41	RLP	1.0	1.0	1.0
42	SPA	0.93	1.0	0.96
43	SUC	0.93	0.89	0.91
44	SYM	0.99	0.99	0.99
45	TRV	0.93	0.97	0.95
46	VFT	0.99	0.98	0.99
47	VPP	0.78	0.96	0.86
48	VSP	1.0	0.75	0.85
49	VST	0.975	0.99	0.98
50	XX	1.0	1.0	1.0

Table 15. Input Text and the output Khasi text with the tags

Input: Khasi Text	'Ka jylla nalor ka long kaba la die kynrei ne pathar bha ïa ka kyïad, bluit wan ban mih paw pat sa ka jingïoh ban die pathar ki nongkhaïi sa ïa u drok, uba lehse ba long uwei na ki jait jingdih, uba ki khun samla kin hap ban long kiba peitngor ym tang ba ki ngat ïalade, hynrei ba kin shim khia naka bynta ban peitngor pat ïala ki para samla, ba kin nym shah shong kulai sa ha une uwei pat u jait jingdih pynbuaid uba ka pyrthei ruh ka dum buit ban tem ïaka'
Output	Ka/3PSF jylla/CMN nalor/ADD ka/3PSF long/CO ka/3PSF ba/COM la/VST die/TRV kynrei/AD ne/CRC pathar/ADJ bha/AD ïa/IN ka/3PSF kyïad/CMN (/SYM bluit/AD wan/ITV ba/COM n/VFT mih/ITV paw/AD pat/AD sa/VFT ka/3PSF jingïoh/ABN ba/COM n/VFT die/TRV pathar/AD ki/3PPG nongkhaïi/CMN sa/VFT ïa/IN u/3PSM drok/FR ,/SYM u/3PSM ba/COM lehse/AD ba/COM long/CO u/3PSM wei/CN na/IN ki/3PPG jait/ADJ jingdih/ABN ,/SYM u/3PSM ba/COM ki/3PPG khun/CMN samla/CMN ki/3PPG n/VFT hap/TRV ba/COM n/VFT long/CO ki/3PPG ba/COM peitngor/TRV ym/NEG tang/ADD ba/COM ki/3PPG ngat/TRV ïa/IN lade/RFP (/SYM hynrei/SUC ba/COM ki/3PPG n/VFT shim/TRV khia/AD na/IN ka/3PSF bynta/CMN ba/COM n/VFT peitngor/TRV pat/AD ïa/IN la/POP ki/3PPG para/CMN samla/CMN ,/SYM ba/COM ki/3PPG n/VFT nym/NEG shah/PAV shong/TRV kulai/CMN sa/VFT ha/IN une/DMP u/3PSM wei/CN pat/AD u/3PSM jait/ADJ jingdih/ABN pynbuaid/CAV ba/COM ka/3PSF pyrthei/CMN ruh/COC ka/3PSF dum/ADJ buit/CMN ba/COM n/VFT tem/TRV ïa/IN ka/3PSF

results of our work and other research work concerning rule-based pos tagging. From the comparison table, the proposed approach for Khasi POS tagging shown in this paper can be claimed that it generates more accurate POS tag sets, and produce higher accuracy result than the existing works.

The performance analysis of the output produced by the tagger are also discussed. Therefore, in the future, more tokens on the Khasi lexicon or corpus for POS tagging will be introduced. A thorough investigation of the corpus to account for the problems of ambiguities, and orthography encountered in this research will be addressed extensively.

This research paper will eventually contribute to the development of standard Khasi gold corpus.

Table 16. Testing result of Our proposed Khasi POS Tagger

Sl. No.	Khasi Lexicon Size	Rules Generated	Testing Accuracy
1	training data (30,580 tokens) testing data (1409 tokens)	20102	79.60%
2	training data (40,920 tokens) testing data (3122 tokens)	26017	90.97%
3	training data (84972 tokens) testing data (4340 tokens)	53528	95.52%

Acknowledgments

The authors would like to thanks the Government of India, Ministry of Science & Technology, Department of Science & Technology (DST), KIRAN Division, Technology Bhavan, New Delhi, for their supports and financial assistance (Grant: DST/WOS-B/2018/1216/ETD/Sunita(G)) during the study.

References

1. **AlGahtani, S., Black, W., McNaught, J. (2009).** Arabic part-of-speech tagging using transformation-based learning. Proceedings of the Second International Conference on Arabic Language Resources and Tools, 2001, MEDAR, pp. 66–70.
2. **Batra, K. K., Lehal, G. (2010).** Rule based machine translation of noun phrases from punjabi to english. International Journal of Computer Science Issues (IJCSI), Vol. 7, No. 5, pp. 409.
3. **Beard, R. (1987).** Lexical stock expansion. Rules and the Lexicon: Studies in Word Formation. Lublin: Redakcja Wydawnictw KUL, pp. 23–41.
4. **Bhatt, P. M., Ganatra, A. (2009).** Analyzing & enhancing accuracy of part of speech tagger with the usage of mixed approaches for gujarati. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277, Vol. 3878.
5. **Bredenkamp, A., Markantonatou, S., Sadler, L. (1996).** Lexical rules: what are they? Proceedings of the 16th conference on Computational linguistics-Volume 1, Association for Computational Linguistics, pp. 163–168.
6. **Brill, E. (1992).** A simple rule-based part of speech tagger. Proceedings of the third conference on Applied natural language processing, Association for Computational Linguistics, pp. 152–155.
7. **Brill, E. (1994).** Some advances in transformation-based part of speech tagging. arXiv preprint cmp-lg/9406010.
8. **Brill, E. (1995).** Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics, Vol. 21, No. 4, pp. 543–565.
9. **Btoush, M. H., Alarabeyyat, A., Olab, I. (2016).** Rule based approach for arabic part of speech tagging and name entity recognition. Int. J. Adv. Comput. Sci. Appl.(IJACSA), Vol. 7, No. 6, pp. 331–335.
10. **Chakrabarti, D., CDAC, P. (2011).** Layered parts of speech tagging for bangla. Language in India, www. languageinindia. com, Special Volume: Problems of Parsing in Indian Languages.
11. **Ekbal, A., Mondal, S., Bandyopadhyay, S. (2007).** Pos tagging using hmm and rule-based chunking. The Proceedings of SPSAL, Vol. 8, No. 1, pp. 25–28.
12. **Garg, N., Goyal, V., Preet, S. (2012).** Rule based hindi part of speech tagger. Proceedings of COLING 2012: Demonstration Papers, Association for Computational Linguistics, pp. 163–174.
13. **Hasan, F. M. (2006).** Comparison of different POS tagging techniques for some South Asian languages. Ph.D. thesis, BRAC University, Dhaka, Bangladesh.
14. **Hoque, M. N., Seddiqui, M. H. (2015).** Bangla parts-of-speech tagging using bangla stemmer and rule based analyzer. 2015 18th International Conference on Computer and Information Technology (ICCIT), IEEE, pp. 440–444.
15. **Khoja, S. (2001).** Apt: Arabic part-of-speech tagger. Proceedings of the Student Workshop at NAACL, NAACL, pp. 20–25.
16. **Mahar, J. A., Memon, G. Q. (2010).** Rule based part of speech tagging of sindhi language. 2010 International Conference on Signal Acquisition and Processing, IEEE, pp. 101–106.

17. **Mawphor (2017).** Mawphor. https://www.mawphor.com/index.php/. [Online; accessed Nov-2017 to June-2019].
18. **Nisheeth, J., Hemant, D., Iti, M. (2013).** Hmm based pos tagger for hindi. Proceeding of 2013 International Conference on Artificial Intelligence and Soft Computing, Springer, pp. 341–349.
19. **Purnamasari, K., Suwardi, I. (2018).** Rule-based part of speech tagger for indonesian language. IOP Conference Series: Materials Science and Engineering, volume 407, IOP Publishing, pp. 012151.
20. **Rashel, F., Luthfi, A., Dinakaramani, A., Manurung, R. (2014).** Building an indonesian rule-based part-of-speech tagger. 2014 International Conference on Asian Language Processing (IALP), IEEE, pp. 70–73.
21. **Schmid, H. (1999).** Improvements in part-of-speech tagging with an application to german. In Natural language processing using very large corpora. Springer, pp. 13–25.
22. **Selvam, M., Natarajan, A. (2009).** Improvement of rule based morphological analysis and pos tagging in tamil language via projection and induction techniques. International journal of computers, Vol. 3, No. 4, pp. 357–367.
23. **Singha, K. R., Purkayastha, B. S., Singha, K. D. (2012).** Part of speech tagging in manipuri: A rule based approach. International Journal of Computer Applications, Vol. 51, No. 14.
24. **Tham, M. J. (2018).** Challenges and issues in developing an annotated corpus and HMM POS tagger for Khasi. The 15th International Conference on Natural Language Processing, Association for Computational Linguistics, pp. 10–19.
25. **Toutanova, K., Klein, D., Manning, C. D., Singer, Y. (2003).** Feature-rich part-of-speech tagging with a cyclic dependency network. Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology-volume 1, Association for Computational Linguistics, pp. 173–180.
26. **Warjri, S. (2020).** Khasi-corpus. `https://github.com/sunitawarjri/Khasi-Corpus/blob/master/Khasi%20Corpus.txt`.
27. **Warjri, S., Pakray, P., Lyngdoh, S., Kumar Maji, A. (2018).** Khasi language as dominant part-of-speech (POS) ascendant in NLP. International Journal of Computational Intelligence & IoT, Vol. 1, No. 1.
28. **Warjri, S., Pakray, P., Lyngdoh, S., Maji, A. K. (2019).** Identification of POS tag for Khasi language based on Hidden Markov Model POS tagger. Computación y Sistemas, Vol. 23, No. 3.
29. **Warjri, S., Pakray, P., Lyngdoh, S., Maji, A. K. (2021).** Adopting conditional random field (CRF) for Khasi part-of-speech tagging (KPOST). Proceedings of the International Conference on Computing and Communication Systems, Springer, pp. 75–84.
30. **Warjri, S., Pakray, P., Lyngdoh, S. A., Maji, A. K. (2021).** Part-of-speech (POS) tagging using conditional random field (CRF) model for Khasi corpora. International Journal of Speech Technology, pp. 1–12.

Article received on 04/10/2021; accepted on 30/11/2021.
Corresponding author is Partha Pakray.