

A Silver-Standard POS Corpus and Baseline Models for Tagin, a Low-Resource Language

Tungon Dugi*, Koj Sambyo

National Institute of Technology Arunachal Pradesh,
Department of Computer Science and Engineering,
India

tungondugi@gmail.com

Abstract. This study presents the design of a machine-learning-based part-of-speech (POS) tagger and a silver-standard POS tagged corpus for the low-resource Tagin language. Due to the lack of digitally available text data, applying modern NLP techniques to the Tagin language remains extremely challenging. To address this, we constructed a POS tagged corpus comprising 5,000 sentences collected from legacy Tagin dictionaries and grammar books, with 23136 tokens and 4139 unique word forms. We employed 21 UD v2-compatible POS tags with Tagin-specific morphosyntactic adaptations. For benchmarking, we evaluated four classical sequence-labeling models (HMM, CRF, Average Perceptron, and Maximum Entropy). Across all evaluation metrics, the CRF model achieved the best performance over all the metrics, with a word level accuracy of 81.80%, and OOV accuracy of 56.76%. The worst-performing model was the HMM, with a word accuracy of 77.18% and a notably low OOV accuracy of 39.62%. We also discuss recently proposed POS taggers for low-resource languages and other relevant NLP studies for comparative purposes. To promote reproducibility and future research, we have released the dataset and model details in a publicly accessible repository.

Keywords. Part-of-speech tagging, low-resource language, Tagin language, conditional random fields, sequence labeling.

1 Introduction

One of the most important jobs in the field of Natural Language Processing (NLP) is Part-of-Speech (POS) tagging. In order to process natural human language computationally, it should be annotated according to the language's grammatical construct.

A text corpus tagged with grammatical information is particularly useful in NLP research. It serves as the fundamental knowledge base for other larger upstream NLP tasks like Named Entity Recognition (NER), Sentiment Analysis, Neural Machine Translation (NMT), and text-to-text generation. So having an annotated text corpus for a particular language is invaluable.

POS tagging entails giving each word in a phrase or a sentence a grammatical class, such as adverb, pronoun, verb, or noun. It is important for a model to understand the underlying grammatical norms of the language so that it can be used to extract information as in information retrieval, speech recognition, text-to-speech, semantics processing, word sense disambiguation, and NMT. Creating a POS tagged corpus is difficult as one particular word can have multiple meanings in context of the sentence.

For example: In the sentence “The record₁ will record₂ your voice.”, “record₁” is a Noun and “record₂” is a Verb in simple future tense. Similarly, in the Tagin sentence: “Svlu nyido₁ homa; svlu donyi nyido₂”(It's not raining today; it's sunny.), the word “nyido₁” is a Noun and the “nyido₂” is a Verb in present tense. So there may be more than one POS tag for a word, this phenomenon is called homograph ambiguity and is a common trait of a language. POS tagging solves this by properly classifying the word in a sentence with a particular POS tag.

While numerous POS-annotated corpora and robust taggers exist for high-resource languages such as English, Spanish, and French, such resources are virtually non-existent

for low-resource languages like Tagin. This gap underscores the digital imperative for enhanced research and development in this subject, beginning with the essential foundation of morphosyntactic annotation.

In recent years, several apex tribal organizations in Arunachal Pradesh have developed primary-level textbooks¹ in their respective mother tongues to facilitate indigenous language education and preservation. This initiative is good for the future of NLP research in these domains, as they are indirectly generating digital resources for their mother tongue. While these printed materials are a vital step for cultural survival, they remain largely inaccessible to modern linguistic technologies.

Therefore, building these foundational digital corpora and baseline evaluations now will provide the necessary infrastructure to process and analyze the growing volume of indigenous digital data effectively.

1.1 Tagin Language

The Tagin language is spoken by the Tagin tribe of Arunachal Pradesh, India and South-Tibet, China. It belongs to the Tani branch of the Sino-Tibetan language family, which includes languages like Adi, Nyishi, Apatani, Mising, and Galo spoken in the same region. According to the 2011 Census of India, the total population of Tagin tribal people is 62,931. It is listed as Endangered/Unsafe on the list of UNESCO Atlas of the World's Languages in Danger. One of the difficulties in digitization of Tagin language is the need for a script system. It uses modified Roman Script, which covers pronunciation of Tagin vocabularies.

It follows subject-object-verb (SOV) word order pattern mostly and distinguishes between long and short sounds, an interesting characteristic which utilizes varying letter lengths, allowing certain words to convey different meanings when spoken with long or short tones. Central vowel sounds, such as [i] and [ə] are also found in spoken Tagin language.

¹<https://scertarunachal.nic.in/UserPortal/BookDetails.aspx>

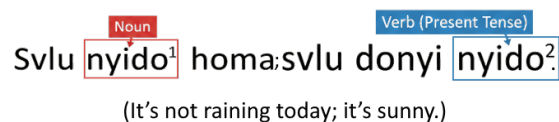


Fig. 1. Context-dependent meaning variation of a polysemous word, illustrating lexical ambiguity in the Tagin language

1.1.1 Complexities in POS-tagging for Tagin

Tagin POS tagging is very challenging due to the absolute scarcity of digital text data.

The language lacks orthographic standardization, lexical items exhibit functional shift, where a single word form displays dual-category behavior, which, compounded by significant lexical variation, creates substantial challenges for automated disambiguation. But the lack of digital resources remains the primary challenge for the development of Tagin POS taggers. Here is a quick overview of these problems:

— **Lexico-Grammatical Ambiguity:** This challenge arises due to the dual behavior of some Tagin words. A word may have different grammatical tags according to its position in a sentence. In Figure 1 the word 'nyido' acts as a Noun and a Verb exhibiting varying meanings 'rain' and 'to ascend' respectively. Since the written Tagin is an informal language, it lacks standard spelling. Writing a word in multiple ways increases the number of vocabularies in a corpus, however, the token's frequency declines, creating sparsity issues.

The accuracy of the model is directly reduced by this sparsity, because many variations of the same term must have the same POS tag assigned by the system. Like the common Tagin verb, 'kabmabv' (*don't cry*) (Figure 2) can be written as 'kabmab', 'kapmab', 'kabyo', or 'kapyo'. Hence, this behavior needs to be checked as sequence generation models rely heavily on exact token matching and may underperform if there are lots of other versions of the same word.

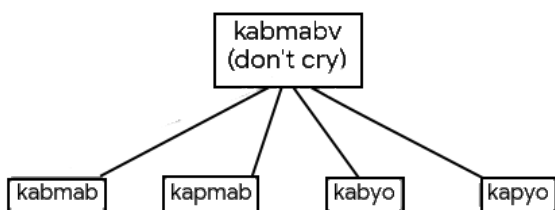


Fig. 2. The figure shows the central term and its orthographic or dialectal variations of the Tagin verb *kabmabv* (meaning "don't cry"): *kabmab*, *kapmab*, *kabyo*, and *kapyo*

- **Ambiguity in Accent:** Most common challenge in Tagin orthography is the repetitive nasal-vowel alternation, where a long vowel sound is used interchangeably with a nasal vowel sound. This is evident in accentual variants such as 'bvvnām' vs. 'bvngnām' (to carry), 'mwvvnām' vs. 'mwngnām' (to introspect), and 'chimma' vs. 'chingma' (don't know). Furthermore, variations in terminal suffixes, such as 'mwvvnw' vs. 'mwngnw' (desire of doing something), demonstrates that dialectal phonology directly impacts word-final orthography. For a POS tagger, these variants trigger high Out-of-Vocabulary (OOV) rates, as the model fails to map these distinct surface forms to a single lemma or grammatical category.
- **Resource Scarcity:** Since Tagin has been classified as a vulnerable language by UNESCO, it is extremely under-resourced. There are no language tools, no digital training dataset for the POS tagging, hence developing one is crucial and is the need of the hour.

In response to these challenges, this study presents the first silver-standard, manually annotated POS-tagged corpus for the Tagin language. By meticulously labeling the morphosyntactic features of a diverse text collection, we provide a foundational resource that accounts for the language's unique accentual variations and lexical ambiguities by bypassing automated tagging to avoid error propagation.

Furthermore, we established a performance benchmark through a baseline evaluation using

standard machine learning models. This work not only bridges the current digital division for Tagin but also provides a repeatable methodology for developing computational resources for other under-resourced Tani languages.

1.2 Contributions

The main contribution of this study includes:

- We present a meticulously curated, fully manually annotated silver-standard POS tagged corpus for the Tagin language.
- We establish a performance benchmark for Tagin NLP by evaluating the POS tagged corpus against traditional machine learning architectures.
- By releasing these resources, we provide the computational infrastructure necessary for downstream applications, thereby supporting the preservation of Tagin in the digital age.

The subsequent sections of this work are structured as follows: Section 2 examines related work in this subject; Section 3 covers the dataset and targetset utilized in this study; Section 4 and Section 5 outlines the models employed in this study; Section 4 and Section 5 details the models employed in this study; and Section 6 shows the results and discussion of the study and finally Section 7 concludes with directions for future research.

2 Literature Review

The scenario of computation linguistics is currently navigating a pivotal transition where the limitations of massive artificial neural network architectures are being re-studied in the context of under-represented and low-resource languages [31].

While deep learning architectures have achieved state-of-the-art performance for high-resource languages such as English and Dutch, often reaching up to 97% accuracy, these models frequently depend on vast annotated corpora that are not available for all the 7,000 world's languages

[8]. POS tagging, a foundational natural language processing (NLP) task that involves assigning corresponding grammatical category, remains a critical prerequisite for higher-level applications like machine translation, information retrieval, and named entity recognition [8].

Traditional machine learning algorithms like, Hidden Markov Model (HMM), Conditional Random Fields (CRF), Structured Average Perceptrons, and Maximum Entropy (MaxEnt), have bubbled-up as superior alternatives in low-resource scenarios [31].

Due to the generative nature and reliance on the Markov assumption, the HMM is persistently in use as a probabilistic approaches to POS tagging [18]. HMM assumes that the observed tokens in a phrase/sentence are projected by a sequence of hidden states, which encodes the parts of speech [17]. In this method, the model maximizes the joint probability by predicting the most probable tag sequence for a given order of words.

This probability is then decomposed into transition probabilities, and emission probabilities, which captures the likelihood of a tag following another, and the likelihood of a specific word being associated with a specific tag respectively [3].

The primary impediment to HMM performance in low-resource settings is the sparse-data problem, where unseen tag transitions or word-tag emissions result in zero probabilities that can invalidate the entire sequence calculation [17]. Researchers have developed several smoothing techniques to mitigate this.

Laplace smoothing, or add-one smoothing, is a primitive yet frequently used method that adds a small constant to all frequency counts [17]. More sophisticated approaches include linear interpolation, which estimates trigram probabilities by taking a weighted average of unigram, bigram, and trigram counts. Kneser-Ney smoothing is another advanced technique often investigated for languages like Yoruba to handle the absence of specific n-grams in limited corpora [3].

For out-of-vocabulary (OOV) words, which are particularly prevalent in morphologically rich languages like Uzbek, Assamese, and Punjabi, HMMs often employ suffix and prefix analysis [17]. Because the grammatical category of a word in

an agglutinative language is frequently encoded in its endings, models can guess the POS tag based on trailing character strings [5]. The TnT tagger, for instance, uses a suffix analysis of up to 10 characters to generate probability distributions for unknown words, which significantly improves disambiguation for inflected word forms [17]. In Punjabi, a bigram HMM reached 90.11% accuracy on a corpus of approximately 20,000 words by utilizing these statistical heuristics [5].

The HMM has shown promising results across diverse linguistic families with low training data. The application of HMM with a bigram approach, in Arabic language has achieved 95.8% accuracy with the incorporation of several lexical models for prefix and suffix probability [5]. For the Awngi language with 188,760 words [14], 93.54% unigram accuracy and 94.77% bigram accuracy were reported.

[3] reported that HMM yielded highly promising results for the low-resource language Yoruba, but it was slightly outperformed by Conditional Random Fields, with a margin of 3.05% due to its ability to recognize more OOV words. This highlights the limitations of HMM, Which limits the capacity to capture long-distance dependencies compared to discriminative models [4].

For languages with complex morphological features and contextual cues, the Conditional Random Fields (CRF) has emerged as a dominant alternative of HMM [32]. CRF, unlike HMM, is a discriminative, undirected, graphical probabilistic model, that directly models the conditional probability of a tag sequence on the entire observation sequence [17], and overcomes global normalization problem, which is a built-in properties of directed models like Maximum Entropy Markov Models (MEMM), where states with few transitions from a point are disproportionately favored [28].

[32] recorded a test-set accuracy of 92.12% using a feature-based CRF in the study of the Khasi language, with a token-level feature set, such as lowercase conversion, digit indicators, as well as contextual windowing features, that examined the properties of the past and the future tokens.

And due to the utilization of linguistic clues like the '-iy' suffix as a definitive characteristic

of adjectives, the CRF achieved 88%, with a 6% point difference in comparison to the HMM implementation for the Uzbek language [17].

Also in a study on Malay language [1], the CRFs outperformed HMM by capturing more sequential dependencies and improved performance on OOV words.

The SoMeWeTa tagger [14], which is based on a structured average perceptron evaluated on the Magahi and Bhojpuri languages has achieved accuracies of 90.70% and 94.08% respectively. The advantage of the structured perceptron in low-resource settings is that, it supports high-dimensional feature spaces with considerably lower training times than CRFs [24, 28], which makes it an ideal framework for working with limited computing resources or needing to perform rapid model design.

The Maximum Entropy (MaxEnt) framework with multinomial logistic regression in terms of sequence labelling, follows the principle that the best model is as uniform as possible while satisfying all known parameters from the training data.

In a language like Odia [12], with rich morphology and limited annotated texts corpus, the researchers proposed a POS tagger based on a hybrid method, with the combination of machine learning and linguistic rules. In order to identify the most probable tag sequence efficiently during testing, they employed a beam search algorithm, with a beam size of $k = 2$. With the integration of guided linguistic rules with standalone MaxEnt models allow the tagger to resolve confusion that might be missed by the purely statistical methods in data insufficient scenarios. In comparative studies, this modified MaxEnt model demonstrated dominance over basic standalone HMM and rule-based systems for Odia text.

One of the most significant insights from recent literature is that the optimal choice of machine learning model is often dictated by the morphological type of the language. Languages are typically categorized along a scale from isolating to agglutinating and fusional.

In agglutinative languages like Uzbek, Turkish, or the Nguni languages (isiZulu, isiXhosa), a single word can consist of many morphemes

representing complex grammatical relations [21]. This led to an extremely large vocabulary size and high rate of unique word forms, causing standard word-based HMMs to fail due to high OOV rates. To combat this, researches have moved toward 'morphological disambiguation' (MD), which seeks to predict not just the POS tag but also the root and the full sequence or morphemes.

The development of high-quality, 'gold-standard' annotated corpora is the essential catalyst for moving low-resource languages into the modern NLP ecosystem. Recent efforts have focused on languages with minimal digital footprints, utilizing innovative methodologies to overcome the high cost and time requirements of manual annotation.

Researchers [19] created a parallel POS evaluation dataset for three very low-resource Indian languages (Angika, Magahi, and Bhojpuri) alongside Hindi. This effort produced the first UD compliant datasets for Angika and Magahi, comprising 708 parallel sentences per language.

The creation of a parallel corpora allows for direct examination of cross-lingual transfer gaps and the effectiveness of transferring knowledge from a related high-resource language like Hindi.

For the Indo-Aryan language Dogri [30], a significant gap was filled with the creation of a manually annotated corpus containing approximately 400,000 tokens across 25,000 sentences.

The project utilized a domain-specific subset of the Linguistic Data Consortium for Indian Languages (LDC-IL) and a specialized web platform called "DogriTag" for semi-automated suggestions and audit tracking. The quality of the corpus was validated through inter-annotator agreement analysis, yielding a high Cohen's Kappa score of 0.89.

A gold-standard dataset for Javanese [2] was developed, consisting of 1,000 sentences (14,323 words) formatted for four key NLP tasks, including POS tagging and dependency parsing. This follows the UD annotation scheme in the CoNLL-U format and was integrated into the Universal Dependencies v2.12 release.

The data collection involved annotation by the native speakers and texts collected from grammar books, online and news sources.

In the study on Lao language [25], a test set of 1,000 gold-standard sentences was released as part of a larger 10,000 sentence corpus, which addressed the lack of digital text data. They applied a semi-automatic annotation approach, where GPT-4o was used for initial pre-annotation, followed by thorough human verification by native speakers.

Character n-grams as input features were used in a shared task (NLAPOST2021) for the Nguni languages [11], which demonstrated significantly better performance than the full words as input features. By adding the vector representation of subword units, the model was able to infer the grammatical features of words they had never seen before by recognizing the common prefixes and suffixes. The best system for IsiXhosa language reached 95% accuracy using a combination of BiLSTM and CRF with character 2-grams.

Developing digital tools for low-resource and indigenous languages is not only a technical exercise, but is a vital step toward digital inclusion and language revitalization. For the development of speech recognition and machine translation systems, the construction of a POS tagger is seen as a foundational resource for the languages like Nasa Yuwe [22], where no tagged corpus previously existed.

The renewed growth of HMMs, CRFs, and Structured Average Perceptrons in the low-resource context is driven by their inherent robustness, their ability to incorporate linguistic domain knowledge through feature engineering, and their superior performance when training data is minimal.

As the NLP community increasingly recognizes that deep learning is not a 'one-size-fits-all' solution, these foundational machine learning algorithms will remain indispensable tools in effort to document and digitally empower the world's diverse linguistic heritage.

3 Dataset

The main challenge in building downstream NLP tools for low-resource languages like Tagin is the lack of digital data. Tagin is a low-resource language without any digital data or text corpus

available in any form. There are some old grammar books, dictionaries, etc., but a dedicated corpus for the NLP research is near to zero.

So, in order to build NLP tools, specifically a POS tagger for Tagin language, we have curated a POS tagged corpus manually and by using old books and community magazines. The following sections highlight the process of data acquisition for POS tagged Tagin corpus.

3.1 Tagin Text Corpus

We initially mined texts from old Tagin grammars, Tagin-English dictionaries and Tagin primary-level textbooks using optical character recognition tools (OCR) by manually scanning each page and converting them to PDFs.

This helped in achieving the initial digital Tagin lexicon and its grammatical category, and with the help of these lexicons we have started manually translating sentences from Tatoeba² English monolingual corpus to Tagin. From this we have generated 15K Tagin-English parallel sentences.

These unprocessed sentences were cleaned and pre-processed in many stages. Initially, we removed any unwanted characters which can be generated while performing OCR, and we used certain sentence delimiters like '.' and '?', etc., so that each sentence can easily be identified while constructing a POS tagged corpus.

Due to orthographic variability in the source vocabularies, we applied a manual spelling normalization step to ensure consistent lexical representations for model training. Since Tagin is a non-written language without its native script, different sources have different spellings for the same word, so spelling normalization was extremely important.

We followed the modified Roman script (MRS) released by the apex community organization of the Tagin community, which is shown in figure 3.

²<https://tatoeba.org/en/downloads>

Table 1. CRF POS tagging performance and methodology details for various low-resource languages

Language	LCRF Accuracy	Feature Engineering Detail
Bodo	75.2%	Linear-chain structure [27]
Odia	~82%	Universal POS tag set [28]
Uzbek	88.0%	L-BFGS optimized, suffix-aware [17]
Yoruba	84.66%	Linear-chain, CRF++ tool [3]
Khasi	92.12%	Lexical and orthographic features [32]

Tagin	A	E	I	O	U	V	W
	a	e	i	o	u	v	w
English	A	E	I	O	U	-	-
	a	e	i	o	u	-	-
Devanagari	अ	ए	इ	ओ	उ	-	-
Phonetic	[a]	[e]	[i]	[o]	[u]	[ə]	[ɪ]

Tagin	B	C	D	F	G	H	J	K	L	M	N	P	Q	R	S	T	X	Y	Z
	b	c	d	f	g	h	j	k	l	m	n	p	q	r	s	t	x	y	z
English	B	C	D	F	G	H	J	K	L	M	N	P	Q	R	S	T	X	Y	Z
	b	c	d	f	g	h	j	k	l	m	n	p	q	r	s	t	x	y	z
Devanagari	ब	-	द	-	ग	ह	ज	क	ल	म	न	प	क	र	स	त	-	य	-
Phonetic	[b]	-	[d]	-	[g]	[h]	[dʒ]	[k]	[l]	[m]	[n]	[p]	-	[r]	[s/e]	[t]	-	[j]	-

Tagin	Ch	Ng	Ny
	ch	ng	ny
English	Ch	Ng	Ny
	ch	ng	ny
Devanagari	च	ङ	ञ
Phonetic	[tʃ]	[ŋ]	[ɲ]

Fig. 3. Tagin Modified Roman Script (MRS)

3.2 Creation of Tagin Digital Lexicon

In order to know the basic grammatical category of each lexicon of Tagin language (so that a benchmark POS-labeled dataset can be constructed), we need to build an end-to-end digital lexicon.

A lexicon can be defined as a systematically structured list of words with grammatical information. Building one is rewarding yet time consuming, so we leveraged an old Tagin-English dictionary by employing OCR technology to scan the entire Tagin-English dictionary.

The lexicon was rather modest, comprising of 3000 terms; yet, it provided us with a basis. The lexicon is in the CSV format, where the

basic grammatical categories of each word are represented in a morphological features column. Figure 4 shows a snapshot of the lexicon, where each word in the lexicon is annotated with basic morphological categories, such as, case, aspect, tense, person etc.

This served as the lookup table for building a silver-standard POS-tagged corpus for Tagin by lexicon-based annotation of words.

3.3 Part-of-Speech Tagset

POS tagging is the process of assigning grammatical information (like Noun, Pronoun, Verb, Adjective, etc.) to a word in a language,

which in turn gives more functional linguistic knowledge about a particular word in a sentence.

Based on the context of a sentence, a word can have more than one POS tag, like the word 'record' in the sentence 'the record will record your voice', and the word 'nyido' in Tagin sentence, 'Svlu nyido homa, svlu donyi nyido'. POS tagging is one of the main foundations in the field of NLP research, as it is used extensively in many tasks like machine translation, named entity recognition, tokenization, and so on.

A POS tagger is a computational system which uses various mathematical functions and computational models to automatically assign POS tags to each word in a sentence, few examples includes Bodo-POS-tagger [27], Khasi-POS-tagger [23, 33], Pashto-POS-tagger [15], Roman Urdu-POS-tagger [13], and Urdu-POS-tagger [16]. Building a POS-tagger for a language is a complex and laborious task especially if it involves only manual annotation strategy and also if the language is extremely low-resourced. One has to build a corpus from zero, then clean the corpus, select POS tagset and select a benchmark algorithm.

For this experiment, we have employed 21 distinct POS tags, designed to be largely compatible with the Universal Dependencies [26] (UD) v2 framework, while incorporating targeted adaptations to account for the morphosyntactic characteristics of the Tagin language. Core UD categories - NOUN, VERB, ADJ, ADV, PRON, PROPN, DET, NUM, ADP, AUX, CCONJ, INTJ, and COP; are directly adopted, facilitating cross-linguistic comparability and reuse for UD-based tools.

In addition, the tagset introduces language-specific distinctions not explicitly encoded in UD v2, such as CLF for classifiers, NEG for explicit negation markers, and PART-INT for interrogative particles, all of which play a salient grammatical role in Tagin language. The addition of PHRASE tag in the tag set enables the annotation of multiword expressions that function as a single word unit, addressing challenges led by orthographic variation. Furthermore, tags such as PREP and CONJ are retained to maintain continuity with legacy annotations and annotator

Table 2. Comprehensive List of Hybrid POS Tags for the Tagin Language

Tag	Categories
PREP	Preposition
PART	Particle
NUM	Numeral
ADJ	Adjective
NOUN	Noun
CCONJ	Coordinating Conjunction
PUNCT	Punctuation
INTJ	Interjection
PROPN	Proper Noun
PART-INT	Particle / Interjection
PRON	Pronoun
PHRASE	Phrase-level construct
CONJ	Conjunction (general)
DET	Determiner
CLF	Classifier
COP	Copula Verb
ADP	Adposition (preposition/postposition)
VERB	Verb
ADV	Adverb
NEG	Negation
AUX	Auxiliary Verb

intuition, despite their partial function overlap with ADP, and CCONJ in UD v2.

Overall, the resulting tagset represents a hybrid annotation scheme that balances UD v2 interoperability with descriptive adequacy for a low-resource, typologically underrepresented language, thereby supporting both linguistic analysis and downstream NLP applications. The complete tagset used for this experiment is shown in table 2.

4 Models

The models we employed in our investigation are covered in this section. We studied the performance of four models for POS tagging of Tagin texts, including Structured Averaged Perceptron (AvgPercep), Hidden Markov Model (HMM) with Viterbi algorithm, Maximum Entropy (MaxEnt) model, and finally the Conditional Random Field (CRF) model. All chosen

```

tagin_pos_corpus.csv > data
1 Word,POS,MorphFeatures
2 a,ADP,Case=Acc
3 aaba,VERB,Aspect=Simple|Tense=Past
4 aadww,VERB,Aspect=Simple|Tense=Past
5 aadww,VERB,Aspect=Simple|Person=3|Tense=Present
6 aadww ku,VERB,Person=3|Tense=Present
7 aadw ho,NOUN,
8 aadu,VERB,Aspect=Simple|Tense=Present
9 la,VERB,Aspect=Simple|Tense=Present
10 aadum,VERB,

```

Fig. 4. Tagin Digital Lexicon

Table 3. Tagin Silver-Standard Corpus Statistics

Statistic	Value
Total Sentences	5,000
Total Tokens	23,136
Vocabulary Size	4,139
Type-Token Ratio	0.1789
Average Sentence Length	4.63 words

algorithms are computationally efficient, this makes them ideal for digital imperative, as they can be trained on smaller, manually curated 'Silver-Standard' datasets.

4.1 Hidden Markov Model (HMM)

An HMM [6, 29, 9] is a probabilistic sequence model in which hidden states are inferred from observations. It is one of the widely used statistical models for POS tagging. We implemented a second-order HMM which is defined by three primary probability distributions:

1. Initial State Probabilities (π): The likelihood of a tag occurring at the beginning of a sequence.
2. Transition Probabilities (A): The probability of moving from one POS tag to another ($P(t_i|t_{i-1})$).
3. Emission Probabilities (B): The probability of a specific word (surface form) being generated by a tag ($P(w_i|t_i)$).

During the evaluation phase, the Viterbi algorithm, a dynamic programming approach, was applied to find the optimal tag sequence $T = (t_1, t_2, \dots, t_n)$ for a given input sentence

Table 4. Top 10 most frequent POS tags in the corpus

Tag	Frequency	Percentage (%)
VERB	3,635	15.71
NOUN	3,589	15.51
PRON	3,271	14.14
ADV	1,876	8.11
AUX	1,833	7.92
ADJ	1,615	6.98
PART	1,489	6.44
DET	1,426	6.16
COP	1,241	5.36
PROPN	1,044	4.51

$W = (w_1, w_2, \dots, w_n)$. The algorithm maximizes the joint probability:

$$P(T, W) = P(t_1) \prod_{i=2}^n P(t_i|t_{i-1}) \prod_{i=1}^n P(w_i|t_i). \quad (1)$$

This choice of algorithm is essential for Tagin, as it lets the model to utilize the syntactic-flow of the Tagin language to resolve lexical ambiguities like the Noun/Verb distinction in words as 'nyido'.

4.2 Structured Averaged Perceptron (AvgPercep)

We used a Structured Averaged Perceptron (AvgPercep) [10] as our second baseline model. It is a discriminative linear classifier model which is adapted for the sequence labelling tasks. It directly models the mapping from the feature space to the tag sequence, unlike the HMMs that model the joint probability $P(T, W)$.

The Perceptron has the ability to incorporate overlapping, non-independent features. In order to prevent the model from overfitting, and to ensure stability in a low-resource setting, we implemented an averaging weight update rule, where, w^i represents the weight vector at i^{th} iteration, and the final weights w_{avg} are calculated as:

$$w_{avg} = \frac{1}{N} \sum_{i=1}^N w^{(i)}, \quad (2)$$

where N is the total number of training steps.

This averaging process smoothens the decision boundary and has been shown to achieve performance competitive with Conditional Random Fields (CRFs) while maintaining significantly lower computational overhead.

4.3 Maximum Entropy (MaxEnt) Model

The third baseline is a Maximum Entropy (MaxEnt) [7] model, implemented via multinomial logistic regression with a log-linear framework. MaxEnt models are particularly suited for POS tagging in low-resource contexts because they do not assume feature independence, allowing for the integration of diverse linguistic cues into a single probabilistic model.

The MaxEnt model estimates the conditional probability of a tag t given a word and its context c using the following exponential form:

$$P(t|c) = \frac{1}{Z(c)} \exp \left(\sum_i \lambda_i f_i(c, t) \right), \quad (3)$$

where $f_i(c, t)$ are binary feature functions, λ_i are the learned weights (parameters), and $Z(c)$ is the normalization constant.

The model is trained to find the parameter set λ that maximizes the entropy of the distribution, subject to the constraints that the expected values of the features in the model match their observed frequencies in the Tagin silver-standard corpus.

4.4 Conditional Random Fields (CRF)

The primary baseline for our sequence labeling task is the Conditional Random Field (CRF) [20] model.

CRFs are discriminative probabilistic graphical models that are particularly effective for POS tagging because they model the entire sequence of tags T globally, conditioned on the input sequence W .

Unlike HMMs, CRFs do not assume feature independence, allowing for the integration of rich, overlapping linguistic cues. The CRF model

calculates the conditional probability of a tag sequence $\mathbf{v}$ given an input sequence $\mathbf{u}$ as:

$$P(\mathbf{v}|\mathbf{u}) = \frac{1}{Z(\mathbf{u})} \exp \left(\sum_{j=1}^k \sum_{i=1}^n \lambda_j f_j(v_{i-1}, v_i, \mathbf{u}, i) \right), \quad (4)$$

where f_j represents a set of feature functions, λ_j are the learned weights, and $Z(\mathbf{u})$ is the normalization factor.

By modeling the transition between tags (v_{i-1}, v_i) and the relationship between tags and words simultaneously, CRF captures the syntactic flow of the Tagin language more effectively than local classifiers.

5 Experimental Setup

This study's main objective is to construct a manually curated silver-standard POS tagged corpus and a POS tagger based on traditional machine learning models for low-resource Tagin language and also to establish a baseline benchmark for future research in this field.

We compared four baseline models (HMM, CRF, AvgPercep and MaxEnt). These models utilize high-dimensional sparse feature vectors. This approach avoids the cold-start problem associated with training embeddings on limited data. By using orthographic and morphological feature functions, we allow the models to capture the semantic and syntactic similarities between accentual variants without the need for large-scale external corpora.

Before training, we split the data (5K POS tagged sentences) into training-testing sets, in the ratio of 80:20 and with a random state of 42 for reproducibility and generability.

In HMM, which is a probabilistic model, we implemented a second-order HMM and in order to address the 'Zero-Frequency Problem', we applied Lidstone Smoothing (also known as additive smoothing) with a parameter of $\gamma = 0.1$. This ensures that the model can handle out-of-vocabulary (OOV) words and unseen tag transitions without resulting in zero-probability errors.

The AvgPercep model is a discriminative linear classifier adapted for sequence labeling tasks, which directly models the mapping from the feature

space to the tag sequence. To address the specific challenges of Tagin, the following features are extracted for each token w_i :

1. Morphological Cues: Suffixes (last 3 characters) and prefixes (first character) to capture nasal vowel alternations like *vvnam* vs. *vngnam*.
2. Contextual Windows: Word-level features for the current (w_i), previous (w_{i-1}), and subsequent (w_{i+1}) tokens.
3. Syntactic Context: The predicted labels of the two past tokens (t_{i-1}, t_{i-2}) in order to stabilize the structural consistency.

In order to ensure stability in a low-resource setting and to prevent overfitting, an averaging weight update rule was implemented.

The model was trained for the 15 iterations with a uniform random seed in each epoch to ensure efficient weight convergence. This iterative tuning allowed the model to learn the dual-behavior patterns of Tagin lexical items by penalizing incorrect tag assignments and boosting the weights of features associated with the original labels.

The MaxEnt model, which is particularly suited for POS tagging in low-resource settings due to the feature dependence, allowing for the integration of diverse linguistic cues into a single probabilistic model, we implemented multinomial logistic regression with a log-linear framework.

To handle the orthographic inconsistencies and the variations in accent of Tagin, we leveraged a high-dimensional feature space transformed via a DictVectorizer into a sparse matrix. It includes features like:

- Internal Morphological Features: Character-level prefixes (length 1-3) and suffixes (length 2-3) to capture Tagin-specific endings like *-nam* and *-vng*.
- Orthographic Identifiers: Binary indicators for title casing, digit presence, and capitalization.
- Contextual Anchors: Lowercase word-forms of the current, preceding, and succeeding tokens to provide local syntactic context.

- Boundary Markers: Explicit BOS (Beginning of Sentence) and EOS (End of Sentence) flags to improve tagging accuracy at sentence boundaries.

We employed the L-BFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) optimization algorithm for parameter estimation. L-BFGS is an iterative quasi-Newton method that is highly efficient for large-scale multinomial problems, ensuring convergence even with the sparse feature matrices generated by the unique vocabulary of the Tagin language.

CRFs are discriminative probabilistic graphical models that are particularly effective for POS tagging because they model the entire sequence of tags globally, conditioned on the given input sequence. Unlike HMMs, CRFs do not assume feature independence, allowing for the integration of rich, overlapping linguistic cues.

To address the variations in Tagin accents and the morphological richness, we implemented a feature extraction pipeline that focuses on word shapes and contextual dependencies:

- Morphological Indicators: We extracted character-level suffixes (length 2 and 3) and prefixes (length 3). These features are vital for capturing Tagin-specific verbal markers (e.g., *-nam*, *-vng*) and nominalizing prefixes, helping the model resolve variations like *bvvnam* vs. *textitbvngnam*.
- Orthographic Identifiers: Binary indicators for title casing, digit presence, and capitalization.
- Local Contextual Windows: The model examines a sliding window of ± 1 words. For each token, we include the lowercase form and capitalization status of the previous and subsequent words. This context is essential for disambiguating words with functional shifts, such as *nyido* (Noun vs. Verb).
- Global Boundary Markers: We use BOS and EOS markers to assist the model in learning sentence-level syntax.

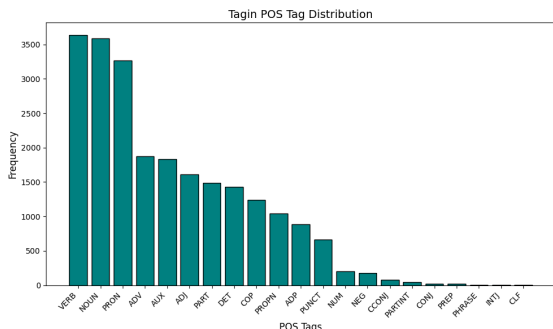


Fig. 5. POS tag distribution in the Tagin Silver-Standard Corpus

The model was trained using the L-BFGS algorithm. To prevent overfitting in our low-resource environment, we applied Elastic Net regularization ($L1$ and $L2$):

- $C1$ ($L1$ Regularization) = 0.1: Encourages sparsity by driving less important feature weights to zero.
- $C2$ ($L2$ Regularization) = 0.1: Prevents individual weights from becoming excessively large.

We allow the CRF to learn transitions that may not have appeared in the training data, improving the model's ability to handle unseen sequences in the Tagin test set.

Despite their differing mathematical architectures, the models used in this study share several fundamental characteristics that make them suitable for sequence labeling in low-resource linguistic environments also, all chosen algorithms are computationally efficient, making them ideal for the digital imperative discussed in this study, as they can be trained on smaller, manually curated silver-standard datasets without requiring massive GPU infrastructure.

6 Results and Discussion

This section provides a rigorous quantitative and qualitative assessment of four classical sequence labeling models (HMM, MaxEnt, AvgPercep, and

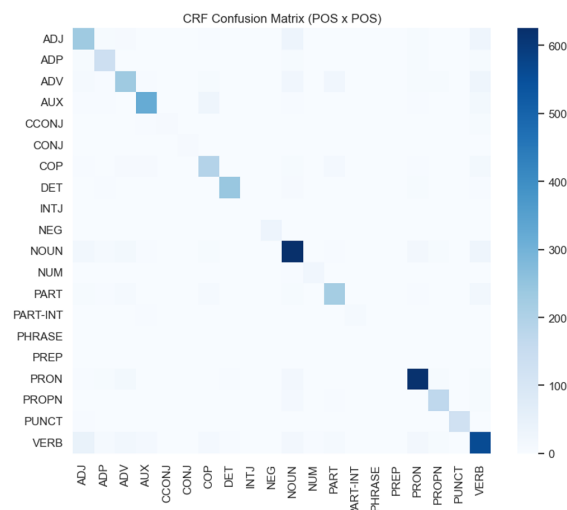


Fig. 6. Confusion Matrix for Tagin CRF POS Tagger

CRF) we have used in this study, evaluated on the 5K Tagin POS annotated sentences.

The evaluation focuses on word-level accuracy, sentence-level accuracy, OOV accuracy, Macro-F1 score, and cross-validation stability to account for class imbalance.

6.1 Quantitative Performance Analysis

Across all evaluation metrics, the CRF model consistently captured long-range contextual dependencies by optimizing the joint probability of the entire label sequence conditioned on the observation sequence. It achieved highest word-level accuracy (81.80%), sentence-level accuracy (53.00%), OOV accuracy (56.76%), and Macro-F1 score (71.41%).

While the AvgPercep and MaxEnt showed a comparable local accuracy, with word accuracies of 80.77% and 80.05%, respectively. They lacked the global normalization capabilities of the CRF model in sentence-level accuracy and Macro-F1.

The worst performing model was the HMM, with a word accuracy of 77.18% and a notably low OOV accuracy of 39.62%.

This performance gap highlights the limitation of generative sequence models in low-resource settings, where OOV tokens are frequent and emission probabilities cannot be reliably estimated.

All models reflected a very low standard deviation in cross-validation (approximately 0.007 – 0.008), exhibiting strong stability. This indicates despite its relatively small dataset size, the results are robust and consistent across different folds.

These results indicate that CRFs are particularly well-suited for Tagin POS tagging, as they effectively model contextual dependencies between adjacent labels while leveraging rich lexical and subword features.

6.2 Qualitative Error Analysis

Several linguistically motivated error patterns are found during manual inspection of misclassified instances that pose difficulty for classical feature-based learners in Tagin.

6.2.1 Nominal-Modifier Ambiguity

In specific syntactic structures (e.g., *subu ador gu jito*), models frequently misclassified the DET (*ador*) as a NOUN. This pattern suggests a high degree of functional overlap where Tagin nouns can occupy modifier slots, leading the models to favor a NOUN-NOUN compound structure over a NOUN-DET sequence.

This is a common problems in low-resource settings that reflect the fluid boundary between nominal and modifier categories.

6.2.2 Adverb vs. Verb/Adjective Ambiguity

The models have shown significant difficulty in differentiating word stems that share functional characteristics across different categories.

A primary example is the word *'nyorsi'*:

- Silver Standard: VERB,
- CRF Prediction: ADJ,
- MaxEnt/Perceptron Prediction: ADV.

Such errors reveals that the feature sets are insufficient to differentiate between actions, qualities, and manners of action in Tagin.

6.2.3 Complex Predicate Structures

The identification of complex predicates, specifically the distinction between main verbs, auxiliary verbs, and adjectival modifiers, underscores the necessity of global sequence constraints, which represents a major bottleneck for discriminative models (MaxEnt/AvgPercep). In the Tagin sentence: *'amin pusuma pasik ro'*: (NOUN ADP VERB AUX):

- MaxEnt predicted NOUN VERB VERB AUX, failing to recognize the adpositional relationship and treating it as a verbal chain.
- CRF and AvgPercep predicted ADJ or PRON for the verbal elements, likely influenced by the preceding nominal context.

6.3 Metric-Specific Evaluations

6.3.1 Out-of-Vocabulary (OOV) Accuracy

The CRF model achieved an OOV accuracy of 56.76%, significantly outperforming the Perceptron (51.24%) and HMM (39.62%).

This achievement is primarily due to the effective use of character-level features (prefixes and suffixes), which allowed the CRF to infer tags better to unseen lexical forms within its global optimization framework.

6.3.2 Macro-F1 versus Word Accuracy

Although word-level accuracies are relatively high, the lower Macro-F1 scores (63.12% – 71.14%) suggests skew class distribution.

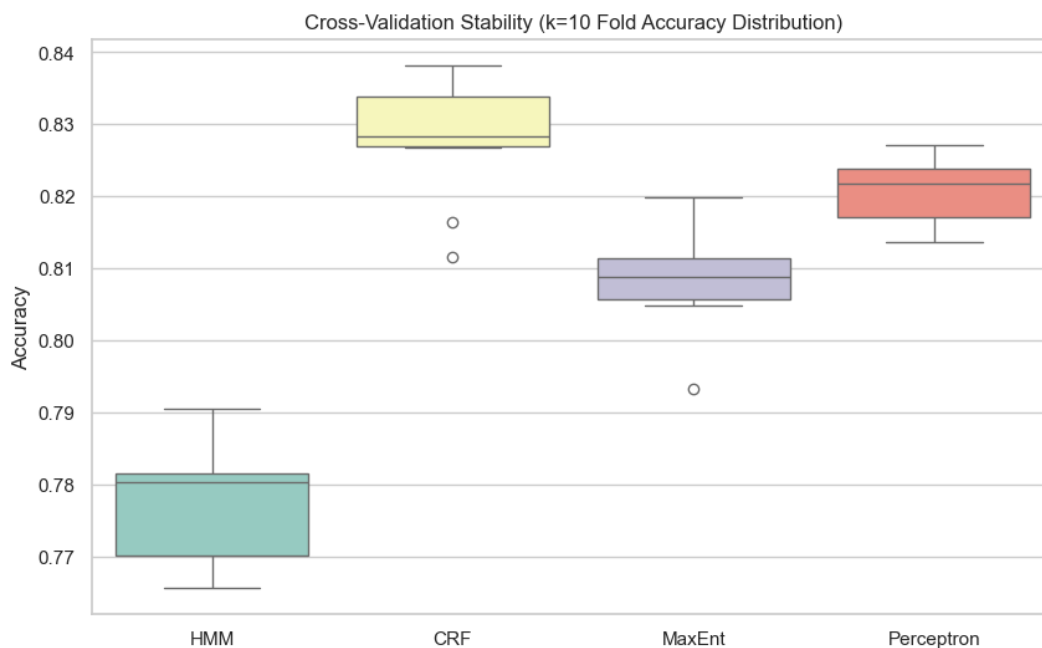
While dominant categories (NOUN, VERB) are handled well, infrequent categories (ADP, PART) suffer from higher error rates, adversely affecting the macro-averaged performance.

6.3.3 Error Propagation

The HMM's low sentence accuracy (40.8%) illustrates the impact of error propagation in generative models. A single incorrect state transition renders the entire sentence incorrect, whereas the CRF's global optimization disrupts these local fluctuations.

Table 5. Performance of different POS tagging models

Model Architecture	Word Acc (%)	Sent. Acc (%)	OOV Acc (%)	Macro-F1 (%)
CRF	81.80	53.00	56.76	71.41
Averaged Perceptron	80.77	48.15	51.24	68.20
MaxEnt	80.05	47.90	50.11	67.54
HMM	77.18	40.80	39.62	63.12

**Fig. 7.** Cross Validation Stability (k=10) for all four POS tagging models

6.3.4 Recommendations for Future Improvements

Although all models showed very similar results across different test splits, several directions for further improvements exist. Like, enhancing the feature engineering could improve OOV handling properly.

Some words in Tagin act like both nouns and modifiers, because of this, the model often gets confused between nouns and determiners, especially in noun groups.

This persistent confusion suggests the need to re-examine annotation consistency in the silver standard Tagin POS corpus. These errors show that Tagin has flexible word usage, which makes POS tagging more difficult.

But given the upward trend suggested by current results, modest data augmentation could yield noticeable improvement, potentially pushing CRF performance beyond 85% word-level accuracy.

Overall, the empirical evidence clearly identifies Conditional Random Fields as the most effective classical architecture for Tagin POS tagging.

It demonstrated the best balance between contextual modeling and morphological generalization, making them particularly suitable for low-resource language settings.

Moreover, these findings establish a strong baseline for future work on neural and multilingual POS tagging for Tagin and related low-resource languages.

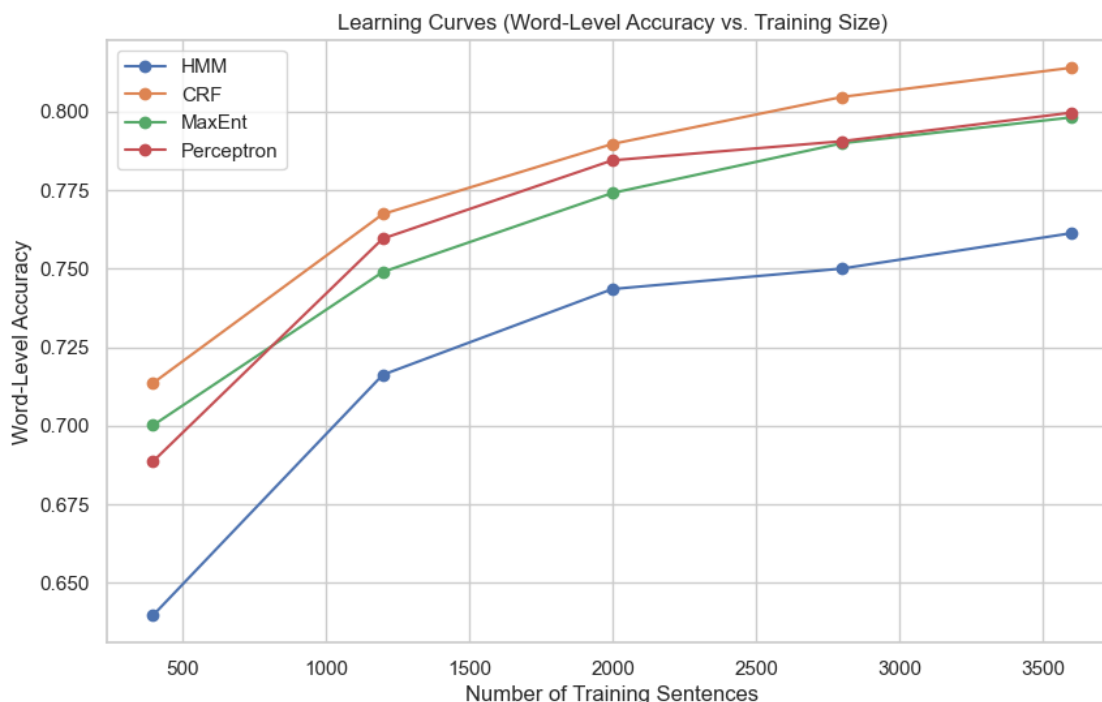


Fig. 8. Word Level Accuracy vs. Training Size

7 Conclusion

This study introduces an initial POS framework for Tagin language is a low-resource language widely spoken by Tagin indigenous people of North-East India and some parts of South Tibet (China).

Despite the constraints of informal written script and a lack of a uniform lexicon, we succeeded in creating an initial silver-standard Tagin POS corpus of 5K sentences and 21 unique POS tags, which is a hybrid of UD v2 scheme and Tagin specific tags. We developed a robust POS tagger using Hidden Markov Models, Conditional Random Fields, Structured Average Perceptron, and Maximum Entropy frameworks.

Our findings demonstrated CRFs outperformed other classical models in all experimental settings. HMM showed significantly lower accuracies, highlighting the limitation of generative sequence models in low-resource settings, where unseen tokens are frequent and emission probabilities cannot be reliably estimated. The CRF model

consistently captured long-range contextual dependencies by optimizing the joint probability of the entire label sequence conditioned on the observation sequence.

This baseline and corpus can guide future Tagin POS research, including the use of deep neural network models, and support the digital preservation and computational inclusion of the Tagin language. Finally, by releasing both the corpus and the annotation guidelines, this work lays the foundation for further Tagin NLP applications.

References

1. Adebayo, B. M., Anbananthen, K. S. M., Muthaiyah, S., Lurudusamy, S. N. (2024). Comparative analysis of deep learning models for part of speech tagging in the Malay language. HighTech and Innovation Journal, Vol. 5, No. 2, pp. 272–281.

2. **Alfina, I., Yuliawati, A., Tanaya, D., Dinakaramani, A., Zeman, D. (2024).** A gold standard dataset for Javanese tokenization, POS tagging, morphological feature tagging, and dependency parsing. *Forum for Linguistic Studies*, Vol. 6, No. 5, pp. 131–148.
3. **Ayogu, I. I., Adetunmbi, A. O., Ojokoh, B. A., Oluwadare, S. A. (2017).** A comparative study of Hidden Markov Model and Conditional Random Fields on a Yoruba part-of-speech tagging task. *2017 International Conference on Computing Networking and Informatics (ICCNI)*, IEEE, pp. 1–6.
4. **Baishya, D., Baruah, R. (2021).** Highly efficient parts of speech tagging in low resource languages with improved Hidden Markov Model and deep learning. *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 12, No. 10.
5. **Baishya, D., Baruah, R. (2021).** Improving hidden markov model for very low resource languages: An analysis for Assamese parts of speech tagging. *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, IEEE, pp. 142–146.
6. **Baum, L. E., Petrie, T. (1966).** Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, Vol. 37, No. 6, pp. 1554–1563.
7. **Berger, A. L., Pietra, V. J. D., Pietra, S. A. D. (1996).** A maximum entropy approach to natural language processing. *Comput. Linguist.*, Vol. 22, No. 1, pp. 39–71.
8. **Bhattacharjee, P., Raju, R. S., Ahmad, A., Rahman, M. S. (2021).** End-to-end Bangla speech synthesis. *2021 International Conference on Science & Contemporary Technologies (ICSCT)*, IEEE, pp. 1–6.
9. **Church, K. W. (1988).** A stochastic parts program and noun phrase parser for unrestricted text. *Second Conference on Applied Natural Language Processing*, Association for Computational Linguistics, pp. 136–143.
10. **Collins, M. (2002).** Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Association for Computational Linguistics, pp. 1–8.
11. **DHASA Editorial (2022).** NLAPOST2021 1st shared task on part-of-speech tagging for Nguni languages. *Journal of the Digital Humanities Association of Southern Africa*, Vol. 3, No. 01.
12. **du Toit, J. S., Puttkammer, M. J. (2021).** Developing core technologies for resource-scarce Nguni languages. *Information*, Vol. 12, No. 12, pp. 520.
13. **Faheem, A., Ullah, F., Azam, U., et al. (2025).** Part of speech (POS) tagging in Roman Urdu: datasets and models. *Language Resources and Evaluation*, Vol. 59, pp. 4285–4312.
14. **Gaikwad, H., Saini, J. R. (2021).** On state-of-the-art of POS tagger, 'sandhi' splitter, 'alankaar' finder and 'samaas' finder for Indo-Aryan and Dravidian languages. *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 12, No. 4.
15. **Haq, I., Zhang, Y., Qadri, I. A. (2025).** POS tagging of low-resource Pashto language: annotated corpus and BERT-based model. *Language Resources and Evaluation*, Vol. 59, pp. 3243–3265.
16. **Ishaq, M., Asim, M., Ahadullah, others (2025).** Development of novel Urdu language part of speech tagging system for improved Urdu text classification. *Social Network Analysis and Mining*, Vol. 15, No. 1, pp. 58.
17. **Jamoldinova, O., Elov, B. B., Sharipova, M., Miralimova, S., Xusainova, Z. Y., Khudayberganov, N. U. o. (2025).** Statistical POS tagging algorithms (HMM, CRF). *2025 10th International Conference on Computer*

Science and Engineering (UBMK), IEEE, pp. 1750–1755.

18. **Khan, W., Daud, A., Nasir, J. A., Amjad, T. (2016).** A survey on the state-of-the-art machine learning models in the context of NLP. *Kuwait Journal of Science*, Vol. 43, No. 4, pp. 95–113.
19. **Kumar, S., Jyothi, P., Bhattacharyya, P. (2024).** Part-of-speech tagging for extremely low-resource Indian languages. *Findings of the Association for Computational Linguistics: ACL 2024*, Association for Computational Linguistics, pp. 14422–14431.
20. **Lafferty, J. D., McCallum, A., Pereira, F. C. N. (2001).** Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*, Morgan Kaufmann Publishers Inc., pp. 282–289.
21. **Malaviya, C., Wu, S., Cotterell, R. (2019).** A simple joint model for improved contextual neural lemmatization. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, pp. 1517–1528.
22. **Martínez, L. M., Lozada, C. A., Corrales-Muñoz, J. C., Curieux, T. R., Herrera-Viedma, E. E., Peluffo-Ordóñez, D. H. (2018).** Building a Nasa Yuwe language corpus and tagging with a metaheuristic approach. *Computación y Sistemas*, Vol. 22, No. 3, pp. 817–829.
23. **Mitri, A. M., Lyngdoh, E. L., Warjri, S., Saha, G., Lyngdoh, S. A., Maji, A. K. (2025).** Probing a pretrained RoBERTa on Khasi language for POS tagging. *Natural Language Processing*, Vol. 31, No. 2, pp. 230–249.
24. **Nguyen, D. Q., Nguyen, D. Q., Pham, D. D., Pham, S. B. (2016).** A robust transformation-based learning approach using ripple down rules for part-of-speech tagging. *AI Communications*, Vol. 29, No. 3, pp. 409–422.
25. **Nguyen, H., Palongve, T., Nguyen-Quy, C., Le-Trung, K. (2025).** HVULao.NLP: A word-segmented Lao corpus with optional POS tags. *Mendeley Data*, Vol. V2.
26. **Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., Zeman, D. (2020).** Universal dependencies v2: An evergrowing multilingual treebank collection. *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, European Language Resources Association, pp. 4034–4043.
27. **Pathak, D., Narzary, S., Nandi, S., Som, B. (2025).** Part-of-speech tagger for Bodo language using deep learning approach. *Natural Language Processing*, Vol. 31, No. 2, pp. 215–229.
28. **Pattnaik, S., Nayak, A. K. (2022).** A modified Markov-based maximum-entropy model for POS tagging of Odia text. *International Journal of Decision Support System Technology (IJDSST)*, Vol. 14, No. 1, pp. 1–24.
29. **Rabiner, L. (1989).** A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257–286.
30. **Saluja, V., Dongardive, J. (2025).** A POS-tagged corpus for Dogri: Development and annotation using DogriTag. *International Journal of Computer Applications*, Vol. 187, No. 59, pp. 36–43.
31. **Vilares Ferro, D., Darriba Bilbao, V. M., Ribadas Pena, F. J., Graña Gil, J. (2022).** Surfing the modeling of POS taggers in low-resource scenarios. *Mathematics*, Vol. 10, No. 19, pp. 3526.
32. **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*, Vol. 24, No. 4, pp. 853–864.

- 33. Warjri, S., Pakray, P., Lyngdoh, S. A., Maji, A. K. (2022).** Part-of-speech (POS) tagging using deep learning-based approaches on the designed Khasi POS corpus. *ACM Transactions on Asian and Low-Resource*

Language Information Processing, Vol. 21, No. 3, pp. 1–24.

Article received on 21/06/2026; accepted on 17/09/2026.

**Corresponding author is Tungon Dugi.*