POS Tagging in Low-Resource Maithili Language: Specific Challenges and Nuances

Shivani Priya, Shruti Jha, Urmila Jha, Deepali Tiwari, Jyoti, Girish Nath Jha

School of Sanskrit and Indic Studies, Jawaharlal Nehru University,
New Delhi-110067
{priyashivani683, 17shrutijha, urmilajha006, deepali0128, girishjha}@gmail.com,
jyotiraj@mail.jnu.ac.in

Abstract

Part-of-Speech (POS) tagging is a key step in Natural Language Processing (NLP), laying the groundwork for more advanced syntactic and semantic tasks. Despite Maithili's status as an Indo-Aryan language with a rich literary tradition and official recognition in India, computational resources for it are still very limited. In this paper, the creation of an annotated corpus of 25,000 sentences drawn from the fields of health, tourism, and administration is described with the hierarchical tagset currently used for Maithili. This paper also indicates that standard tagsets, typically adapted from English or Hindi, fail to capture the linguistic nuances of Maithili. This underscores the need for a dedicated tagging framework that considers characteristics like vocative particles, verbal nuances, honorific complexities.

Keywords: Parts of Speech, Natural Language Processing, Maithili, annotation

1.Introduction

Tagging became a large part of Natural Language Processing (NLP) that attributes grammatical labels or noun, verb, adjective etc. to words within a sentence. The process is central to most NLP tasks, such as parsing, machine translation, and text-to-speech synthesis since it gives morpho-syntactic information (Hardie, 2003) (Priyadarshi et al., 2023) regarding words and their distribution in a sentence. Surface form of a text is converted into morpho-syntactic labelled form by POS annotation. It is used as an interface between raw corpus data and a more abstract language processing with the conversion of unstructured text into structure-analyzed data at the morpho-syntactic level. However, in the instance of Maithili, the Indo-Aryan language predominantly spoken on a vast, expansive and heavily populated territory, in the state of Bihar in India and in Nepal, this medial role is brought to the fore. Despite having a rich literary tradition and being given institutional acknowledgment in the Schedule VIII of the Indian Constitution, Maithili is still considered relatively poor in the area of computational resources. It is a morphologically inflectional language with gender, number, case, honorific inflexions, and complex verb forms of agreement (Kumar and Chaudhary, 2025).

In this type of a wide-ranging system, POS annotation facilitates the syntactic analysis, Part of Speech (POS) or morpho-syntactic semantic processing and creation of sophisticated NLP tools. Therefore, in Maithili, it is a preliminary move towards the development of annotated corpora and the creation of high-level language technologies. In Maithili, annotation of POS is usually performed with the help of the standardized tagsets like the Bureau of Indian Standards (BIS) tagset to provide consistency and interoperability between the languages of India (Gopal, 2011). Labelled language resources for Maithili will support many kinds of NLP applications, such as machine translation, information retrieval, morphological analysis, and parsing.

2. Maithili Linguistic features

2.1 Rich morphology

Maithili has a highly inflection morphology. It displays conjugation of verbs in terms of tense (past, present, future), person (first, second, third), mood (indicative and imperative) and aspect (perfective and imperfective). Another indication of Maithili is that non-verbal agreement is responsive to the honorific status of the subject (Jha et. al. 2018).

Example: अहाँ जाइ छी।
(You (HON) are going.)

The auxiliary 'छी' marks honorific agreement.

Proceedings of WILDRE-8, pages 67–74
12 May 2026.

2.2 Relatively free word order

Maithili has a relatively free word order language, which is frequently determined by constituent structure and discourse prominence as opposed to strict syntactic patterns. It follows that, positional factors (commonly applied to fixed word order language POS tagging) are not absolutely reliable. Therefore, categorization should be done correctly by relying on functional and morphological diagnostics as opposed to a linear position.

Example: **a) SOV-**

हम घर जाइत छी।

“I am going home.”

b) SVO-

हम जाइत छी घर।

Acceptable in spoken/discourse context

c) OSV-

घर हम जाइत छी।

Home focused, topicalization

2.3 Particles

Particles such as ‘तऽ’, ‘सेहो’, ‘तइयो’, and ‘बादो’ inclusive in the Maithili language. These aspects add pragmatic senses like emphasis, contrast or continuation and not lexical senses.

2.4 Demographic variations

Maithili has an enormous demographic variation based on the area, age, gender/socio-educational dimension and this has a direct impact on the part-of-speech annotation. The regional dialect is also observed in the realisation of the auxiliaries and particles such that the same progressive or aspectual construction can be formed with the alternative auxiliaries in the different dialect such that it leads to the existence in the possible differences in the marking of the auxiliary verbs, and the subsequent confusion of the main and auxiliaries.

Example:

ओ कहैत अछि।

ओ कहै है।

“He says.”

काज कऽ लियऽ।

काज कए लियऽ।

“Do the work.”

These examples demonstrate that demographic variation in Maithili produces multiple surface forms for the same syntactic category, requiring POS annotation guidelines that rely on functional and contextual interpretation rather than purely formal criteria.

3. Annotated data-set

In case of annotation of Maithili POS, the data set was tagged in such a way that it would include domain variety and acceptable linguistic coverage. Texts were collected from various domains, including health, tourism, and administration. This multi-domain strategy was followed so as to represent the difference in vocabulary, morphology and syntactic structures among the various registers of usage. The total number of data tagged was approximately 25k sentences, as follows-

S. No.	Domain	Data
1.	Administration	7000
2.	Law	5267
3.	Education	6310
4.	Tourism	838
5.	Health	1008
6.	Agriculture	763
7.	Technology	567

Table 1: Domains and data

4. Tagsets used for tagging

The ILCI Annotation Tool (ILCIANN) (Kumar et al., 2012) is a server-based web application developed under the Indian Languages Corpora Initiative to facilitate large-scale word-level annotation for Indian languages. It is designed to be useful in producing annotated corpora to support NLP, particularly in less-resource languages, through an annotated centralized, uniform environment. It allows annotation of POS through manual means in standardized groups of tags (Bureau of Indian Standards (BIS) tagset and a little automatic tagging in closed grammatical categories) to minimize the workload of the annotators. The tool ensures that annotations will be stored on a central server in sentence-by-sentence form, reducing the data loss and gaps.

Sl. No.	Category		Label	Annotation Convention	Examples
	Top Level	**Sub-type**			
1.	**Noun**		N	N	
1.1		Common	NN	N_NN	किताब, गाछ
1.2		Proper	NNP	N_NNP	राम, मधुबनी
1.3		Nloc	NST	N_NST	नीचाँ, आगू
2	**Pronoun**		PR	PR	
2.1		Personal	PRP	PR_PRP	आहाँ, हम
2.2		Reflexive	PRF	PR_PRF	अपना, अपना-आप
2.3		Relative	PRL	PR_PRL	जकर, एकर
2.4		Reciprocal	PRC	PR_PRC	आपस, परस्पर
2.5		Wh-word	PRQ	PR_PRQ	केकर, कतय
2.6		Indefinite	PRI	PR_PRI	कोनो,
3	**Demonstrative**		DM	DM	
3.1		Deictic	DMD	DM_DMD	ओ, एहि
3.2		Relative	DMR	DM_DMR	जकर, जे, जेना, जकाँ
3.3		Wh-word	DMQ	DM_DMQ	के, कखन
3.4		Indefinite	DMI	DM_DMI	कोनो, कोई
4	**Verb**		V	V	
4.1		Main	VM	V_VM	चलब, देखब
4.2		Auxiliary	VAUX	V_VAUX	अछि, थिक
5	**Adjective**		JJ	JJ	सुन्दर, नीक

6	**Adverbs**		RB	RB	अचानक, धीरे
7	**Postpositions**		PSP	PSP	केर, लेल
8	**Conjunctions**		CC	CC	
8.1		Co-ordinator	CCD	CC_CCD	आओर, मुदा
8.2		Subordinate	CCS	CC_CCS	जँ-तँ, जखन-तखन
9	**Particles**		RP	RP	
9.1		Classifier	CL	RP_CL	बला, टा
9.2		Default	RPD	RP_RPD	तऽ, सेहो
9.3		Interjection	INJ	RP_INJ	अरे, हे
9.4		Intensifier	INTF	RP_INTF	एतेक, बड
9.5		Negation	NEG	RP_NEG	नहि, मत
10	**Quantifiers**		QT	QT	
10.1		General	QTF	QT_QTF	खूब, कनि
10.2		Cardinals	QTC	QT_QTC	एक, दू
10.3		Ordinals	QTO	QT_QTO	पहिल, दोसर
11	**Residuals**		RD	RD	
11.1		Foreign word	FW	RD_FW	Website, Link
11.2		Symbol	SYM	RD_SYM	$, &
11.3		Punctuation	PUNC	RD_PUNC	।,?
11.4		Unknown	UNK	RD_UNK	
11.5		Echo words	ECH	RD_ECH	अलग-थलग, हुइल-माइल

Table 2: Tagset for Maithili language

Example:

- सरकारी\JJ स्कूल\N_NN मे\PSP प्रवेश\N_NN लेल\PSP ‘प्रवेश\N_NN सप्ताह’\N_NNP शुरू\N_NN भऽ\V_VM गेल\V_VAUX अछि\V_VAUX ।\RD_PUNC
  “Admission week has started for admission in government schools (Priyadarshi Ankur, 2023).”

- चाउरक\N_NN क्वालिटी\N_NN नीक\JJ भेला\V_VM पर\PSP बजार\N_NN मे\PSP एकर\DM_DMD दाम\N_NN सेहो\RP_RPD नीक\JJ भेटत\V_VM\।\RD_PUNC
  “If the quality of rice is good, it will fetch a good price in the market.”

5. Nuances found in the process

There were also some linguistic nuances that caused discrepancies in the process of POS annotation. The Penn Treebank tagset of English was used as a source of influence by annotators, who sometimes think of intensifiers (बड, खूब, बेसी) as general adverbs, overlooking their distinct functional role in Maithili.

Example: एहि\DM_DMD कीटनाशक\N_NN कें\PSP खेती\N_NN मे\PSP बड\RP_INTF उपयोग\N_NN अछि\V_VAUX ।\RD_PUNC

“This pesticide has many uses in farming.”

Here, the word ‘बड’ has been tagged as an intensifier.

ओ\DM_DMD बड\RB सुन्दर\JJ बचिया\N_NN छैक\V_VAUX ।\RD_PUNC

“She is a very beautiful girl.”

Here, even though the word ’बड’ acts as an intensifier but the tagger has tagged it as adverb because of the influence of Penn Treebank from English.

Additionally, some words were tagged based on their lexical identity rather than the sentence context, whereas in the sentence, the word functioned as a proper noun.

Example: भारतीय\JJ जनता\N_NN पार्टी\N_NN was tagged as an adjective while the word in the context was functioning as a proper noun.

Another instance is in conjunct verbs like शुरु\V_VM कएल\V_VAUX गेल\ V_VAUX रहय\V_VAUX ।\RD_PUNC (was started), the noun ‘शुरु’ (start) is frequently mislabelled as a verb instead of noun. This occurs because English linguistic interference treats "started" as a single verbal unit.

Moreover, as a result of the uncertainty between Main Verbs and Light Verbs in Compound Verb Constructions (CVCs), a tagger can tag auxiliary verbs as a main verb, although the auxiliary verb loses its literal sense to give aspectual state in a sentence.

Example:

मारि\V_VM देब\V_VM (will kill). In this case, the word ‘मारि’ is the main verb, and ‘देब’ is the light verb showing the completion of the action. If ‘देब’ is tagged as a Main Verb, the system may interpret the sentence as having two distinct actions ("killing" and "giving") rather than one unified event.

6. Challenges

6.1 Low-resource language

Maithili is generally a low resource language in the field of computational linguistics, where there are not many annotated corpus, standardized tagsets, and pre-trained language models. The issue is also aggravated by data sparsity which can impact especially when it comes to processing rare morphological forms or dialectal variation as well as discourse particles. locative forms like घरमे, गाममे, or dialectal variants such as घरम’ काज कऽ, काज कए may not appear frequently enough for the system to robustly learn the postpositional pattern ‘-मे/-म’ ‘कऽ/कए’ as demographic variation.

6.2 Cliticization and particles

Not all words are morphologically simple but represent complex structures that are formed by a component (deictic, pronoun or adverb)

and enclitic particles such as 'नो' 'यो' 'हु' 'बे' in 'तखनो', 'तइयो', 'हमहु', 'हेबे करतै', creating a structural ambiguity. The tagger has to determine whether to assign a single unified tag (e.g., RB or PRP or PSP) or to recognize and segment the internal components (e.g., PRP + RP).

6.3 Honour displayed in verb forms

The rich honorific system that Maithili has also makes tagging more difficult. Pronouns such as 'तूँ', 'ई', and 'अहाँ' trigger distinct verb forms, increasing morphological variation and data sparsity. In Maithili traditions, daughter-in-laws often address their relatives-in-law in the third person, using corresponding verb forms to express respect. Eg- 'ई बैसौथ' (You "hon." sit.), 'ई लऊथ' (You "hon." take). Here, forms that function as third-person pronouns pragmatically serve as an honorific second-person address.

Another instance of this is 'अपनेक बैसियौ' (you "hon." sit), 'अपने' is reflexive but here functions as a personal pronoun that marks honour.

Treating honorificity solely as a semantic feature fails to capture its grammatical impact, leading to loss of information in POS annotation. This creates a significant ambiguity for POS tagging systems.

6.4 Intensifier and plurality

A challenge in Maithili POS tagging arises from the language's strategy of marking plurality. In the language, plurality is marked by 'सब'(all), as in 'बच्चा सब' (children), but 'सब' also acts as an intensifier when attached to 'सँ' (सबसँ) (superlative degree). The same problem is faced with the word 'बेसी' (very/many). The word 'बेसी' functions as an intensifier and a quantifier, like 'बेसी' चोट' (bad injury), 'बेसी' acts as an intensifier that shows the intensity of the injury, while in 'बेसी' कऽ आटा'(more flour) 'बेसी' here acts as a quantifier denoting the quantity of 'आटा' (flour).

6.5 Morphological case marker ambiguity

The word 'के' (of) is semantically versatile. At some places it acts as a genitive marker, like in 'सोमनाथकेँ मन्दिर' (Temple of Somanath), as an accusative marker, like in 'श्यामकेँ बजाऊ' (call Shyam), and as a wh-word, like in 'के आयल?' (who came). A tagger might default to a frequentist approach (e.g., always tagging 'के' as a Postposition), thereby missing the subtle genitive and other nuances.

6.6 Functional Overlap

Another major challenge in Maithili POS tagging is the functional overlap that bridges the gap between closed-class postpositions and open-class verbal roots. A prime example is the lexical item लेल', which exhibits functional polysemy across different syntactic environments. In phrases such as 'रामक लेल' (for Ram), 'लेल' functions as a postposition (PSP) denoting purpose or beneficence. Conversely, in the construction 'लेल गेल' (was taken/received), it functions as the main verb (VM) derived from the root "to take."

In the same way, the word 'क' also acts as a postposition denoting the accusative or genitive aspect of the word, like in the phrase 'रामक किताब' (book of Ram), whereas in the phrase 'काज कऽ लेलहुँ', 'क' is acting as a main verb form meaning "did the work".

6.7 Absence of Reduplication tag

A critical limitation in existing tagging frameworks is the absence of dedicated tags for reduplication, forcing a conflict between a word's literal lexical category and its contextual functional role. Maithili frequently employs morphological reduplication to encode distributive, frequentative, or aspectual nuances that are absent in the individual roots.

Example:

'घरे-घरे' (in every house) 'चलैत-चलैत' (while walking). Here the phrase 'घरे-घरे' seems like noun but it is reduplicated, so tagger gets confused while tagging.

7. Suggested future work

7.1 Addition of honorific tag

Maithili Honorificity has a direct influence on the agreement morphology and syntactic structure. Verbs are different in terms of amounts of respect and form paradigmatic contrasts that are grammatically inevitable in situation. The fact that all such forms are treated under one POS label distorts morpho-syntactic information. Suggested future work could explore incorporating an HON (honorific feature) within the POS framework or introducing feature-level annotation (e.g., PRON[+HON], VERB[+HON]).

7.2 Tagging of address words

Reclassification of address words like **'यौ', 'हौ', 'रौ', 'रे'** (hey!). These objects fall under the general group of interjections. But interjections usually have spontaneous emotional or expressive meaning (e.g. surprise, pain, exclamation). Conversely, these Maithili forms are mainly used as an address or vocative particle to attract attention, indicate interpersonal stance or as a part of a direct address construction. It would be possible to introduce special tags like VOC.PART (vocative particle) or ADDR (address marker) to have a more functional differentiation.

7.3 Creation of Maithili-specific Tagset

The existing model of Maithili-POS tagging has borrowed the model of Hindi and English POS tagging. These are insufficient to represent Maithili-specific characteristics, e.g. layered honorificity, vocative particles reduplication tag etc. The design of a linguistically based, Maithili-specific POS tagset, which considers morpho-syntactic and discourse-pragmatic categories, can assist in adding these characteristics to the structure.

8. Conclusion

The Maithili annotated corpus creation illustrates that language-specific frameworks are imperative in Natural Language Processing. Although this experiment was able to tag 25,000 sentences, it was found that conventional tagsets do not always reflect the structural peculiarities of Maithili, including its stratified honorific system, cliticization and functional polysemy.

The study has shown that the reduplication and vocative particle are not specially marked, which results in severe semantic loss, due to the influence of annotator bias. Such issues as the confusion of light verbs and main verbs, and the shift of postpositions when the context is changed prove that the general strategy cannot fit morphologically rich languages.

Finally, in order to enhance the accuracy of the tagging, future efforts should be focused on developing a Maithili-specific tagset. Adding such features as honorific markers and reduplication tags will make sure that the computational models capture the syntactic and pragmatic reality of the language better, which will be a more solid base of advanced Maithili language technologies.

REFERENCES

Bureau of Indian Standards (BIS). (2011). POS Tagset Guidelines for Indian Languages.

Choudhary, N. (2019). *LDC-IL* Maithili Speech Corpus Documentation. *LDC-IL*.

Choudhary, N., & Ramamoorthy, L. (2019). *LDC-IL* Raw Text Corpora: An Overview. In *Language Data Consortium for Indian Languages (LDC-IL)*.

Gopal, Madhav, & Jha, Girish Nath (2011). Tagging Sanskrit Using BIS POS Tagset. In *Information Systems for Indian Languages: International Conference, ICISIL, Communications and Computer and Information Science, Springer,* Volume-139, PP. 191-194.

Hardie, A., Archer, D., McEnery, T., Rayson, P. (2003). Developing an Automated Semantic Analysis System for Early Modern English. *In Proceedings of Corpus Linguistics*, PP. 22-31.

Jha, A. K., Singh, P. P., & Dwivedi, P. (2019, July). The Maithili text-to-speech system. In *IEEE International Conference on Electronics, Computing and Communication Technologies* (*CONECCT 2019*), PP. 1-6. IEEE.

Jha, Girish Nath and Kumari, Sangita (2010). Maithili Bhāṣā. In *Bharatīya Bhāṣā Paricaya,* Volume-2, PP. 453-478.Jha, S.K., Singh, P. P., & Kaul, V. K. (2018). VEA Model. In *Word Formation Process of Maithili MT*.

Kumar, Ritesh, Kaushik, Shiv, Nainwani, Pinkey, Banerjee, Esha, Hadke, Sumedh,

Jha, Girish Nath. (2012). Using the ILCI Annotation Tool for POS Annotation: A Case of Hindi. In *IJCLA Vol.3*, *Number 2, Jul-Dec 2012*, PP. 93-104.

Kumar, S. (2020). Named Entity Recognition in Maithili. Banaras Hindu University, Varanasi.

Kumar, Shantanu, Choudhary, Narayan (2025). Maithili Language Technology: A Survey. In *Language in India,* Volume-25(6), PP. 160-175.

Mundotiya, R. K., Gatla, Praveen, Kanwar, Nikita, Singh, Anil Kumar. (2025). Deep Learning-Based Similar Languages' POS Tagging: Experiments on Bhojpuri, Maithili, and Magahi. In *Lecture Notes in Networks and Systems, Springer*.

Priyadarshi, A., & Saha, S. K. (2023). A study on the performance of recurrent neural network-based models in Maithili part-of-speech tagging. In *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Priyadarshi, A., & Saha, S. K. (2023). A study on the performance of recurrent neural network-based models in Maithili part-of-speech tagging. In *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Singh, Srishti (2015), *Challenges in Automatic POS Tagging of Indian Languages- A Comparative Study of Hindi and Bhojpuri*, Master of Philosophy, Centre for Linguistics, Jawaharlal Nehru University.

Singh, Srishti, Banerjee, E. (2014). Annotating Bhojpuri Corpus using BIS Scheme. In *LREC 2014 workshop WILDRE*.