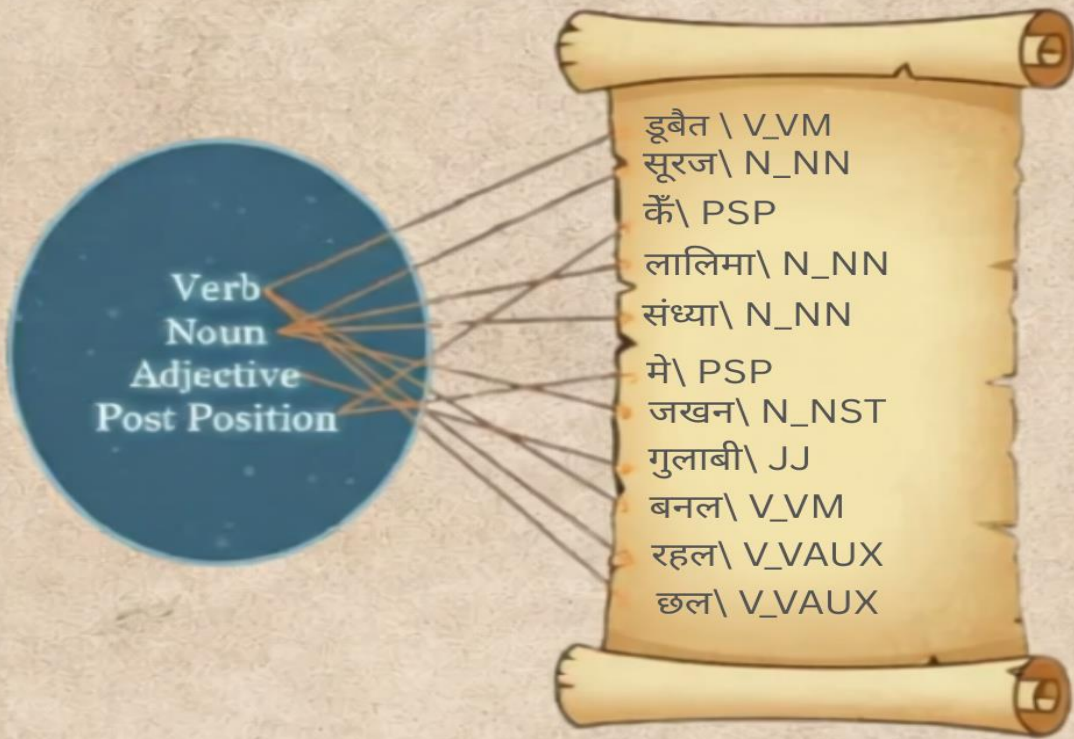


मैथिली शब्द-भेद एनोटेटेड कॉर्पस

मैथिली शब्द-भेद एनोटेटेड कॉर्पस



**MAITHILI
PARTS OF SPEECH
ANNOTATED CORPUS**

MAITHILI PARTS OF SPEECH ANNOTATED CORPUS



34

Annotated, quality language data (both-text & speech) and tools in Indian Languages to Individuals, Institutions and Industry for Research & Development - Created in-house, through outsourcing and acquisition.

Authors:

Dinesh Mishra

Dr. Narayan Choudhary

Rajेश N.

Manasa G.

**Linguistic Data Consortium for Indian Languages
Central Institute of Indian Language
Mysuru, India-570006**

भारतीय भाषा संस्थान

उच्चतर शिक्षा विभाग, शिक्षा मंत्रालय, भारत सरकार, मानसगंगोत्री, हुन्सूर रोड, मैसूरु-570006, कर्नाटक

CENTRAL INSTITUTE OF INDIAN LANGUAGES

Department of Higher Education, Ministry of Education, Government of India

Manasagangotri, Hunsur Road, Mysuru-570006, Karnataka

www.ciil.gov.in :: e-✉ ciil.publication@gmail.com :: ☎ +91-821-234-5182

Maithili Parts of Speech Annotated Corpus

Authors: Dinesh Mishra, Dr. Narayan Choudhary, Rajesha N., Manasa G.

e-ISBN: 978-81-69175-29-6

CIIL Publication No.: 1699

First published: AD 2026 March :: Phalguna 1947 Śaka

© Central Institute of Indian Languages, Mysuru 2026

Publisher: Prof. Shailendra Mohan, Director, CIIL

Printing & Publication Unit

Aleendra Brahma, Officer-in-Charge, Printing Press & Publication Unit

B Manju Bhargavi, Unit-in-Charge, Printing Press & Publication Unit

M.N. Chandrashekar, Compositor, Printing Press & Publication Unit

Cover design: Dinesh Mishra

Contents

1.	PoS Annotated Corpus Based on BIS Tagset.....	1
1.1	Part-of-Speech.....	1
1.2	Part-of-Speech Annotation.....	2
1.3	BIS Tagset.....	2
1.4	PoS Annotation in LDC-IL Data.....	4
1.5	Summary of LDC-IL PoS Data.....	5
2	Maithili PoS Annotated Corpus.....	6
2.1	Grammatical Peculiarities of Maithili.....	6
2.2	LDC-IL Customized BIS Tagset.....	6
2.3	Maithili PoS Tagging.....	8
2.4	Summary of LDC-IL Maithili PoS Data.....	14

Tables

Table 1-1: BIS Tagset	4
Table 1-2: LDC-IL PoS Status.....	5
Table 2-1: Maithili Tagset	7
Table 2-2: Maithili PoS Status.....	14
Table 2-3: Distinct Tag Count of Maithili	14

1. POS ANNOTATED CORPUS BASED ON BIS TAGSET

Rejitha K .S., Narayan Kumar Choudhary

The Linguistic Data Consortium for Indian Languages (LDC-IL) has developed Parts-of-Speech annotated corpus for Scheduled Indian languages. The corpus is annotated with Part-of-Speech (PoS) tags based on the Bureau of Indian Standards (BIS) PoS Tagset. This data is a significant resource for natural language processing and linguistic research in Indian Languages.

Corpus annotation refers to the process of providing additional linguistic information to a corpus. It offers greater insight into the structure and content of the texts. One of the common annotation is parts-of-speech (PoS) tagging, which involves marking words with labels corresponding to their syntactic categories. This process can disambiguate words that are spelled identically but serve different grammatical roles. For example, the word "tear" can be a noun (a drop of fluid from the eye) or a verb (to rip or break apart) depending on the usage. In the sentence "She wiped away a tear from her eye." tear act as noun and in "He will tear the paper into pieces." tear act as verb. Annotation enriches the corpus with labels by adding meaningful grammatical categories to the data to help a deeper understanding and analysis of the text. By adding these labels, the corpus becomes more interpretable so it is easier to use in computational analyses, train machine learning models etc.

1.1 PART-OF-SPEECH

There are eight universally recognized primary parts of speech such as noun, pronoun, verb, adjective, adverb, preposition, conjunction, and interjection. These fundamental lexical categories define the grammatical function of words within a sentence. These categories facilitate the construction of syntactically accurate communication by providing both structural and semantic clarity.

The eight principal parts of speech and their respective functions are as follows:

1. Noun: Denotes a person, place, object, or abstract concept (e.g. Ramu, window, India, anger).
2. Pronoun: Serves as a substitute for a noun to prevent redundancy (e.g. he, it, I).
3. Verb: Conveys an action or a state of being (e.g. playing, is, run).
4. Adjective: Modifies or qualifies a noun or pronoun, providing additional descriptive detail regarding its attributes. (e. g. beautiful, happy)
5. Adverb: Modifies a verb, adjective, or another adverb, offering information about manner, time, place, or degree. (e. g. beautifully, early, everywhere)
6. Preposition: Establishes a relationship between a noun or pronoun and another element within the sentence, frequently indicating spatial or temporal context. (e.g. in, on, at, for)
7. Conjunction: Functions to connect words, phrases, or clauses, facilitating syntactic cohesion (e.g., and, but).
8. Interjection: Conveys strong emotion or exclamation, typically positioned at the beginning of a sentence (e.g., Wow!).

1.2 PART-OF-SPEECH ANNOTATION

PoS annotation is the process of labelling each word in a text with its corresponding part of speech, such as: Noun (N), Verb (V), Adjective (ADJ), Adverb (ADV), Pronoun (PRON), Preposition (PREP), Conjunction (CONJ) etc.

Each word is tagged based on its syntactic role in a sentence. For example:
She/PRON eats/V lunch/N at/PREP noon/N ./SYM

PoS annotation is foundational in Natural Language Processing and used in grammar checking, machine translation, information retrieval, corpus linguistics, and language documentation.

1.3 BIS TAGSET

The BIS tagset is developed by the Bureau of Indian Standards. It is a standardized PoS tagset for Indian languages. The document is [Doc No: LITD 20 \(16341\) WC](#) provides a uniform tag structure to annotate Indian language corpora.

BIS tagset is a hierarchical tagset which has 3 levels of tags. The first level represents the grammatical category, encompassing 11 primary tags. The second level is classified into 32 distinct subcategories, while the third level outlines 10 specialized categories. The different levels of tags minimize the redundancy by incorporating both lexical and syntactic information. Many PoS tags in the BIS are assigned specifically to individual lexical items.

To accommodate the diverse range of Indian languages within a single tagset, the BIS developed a superset of PoS tags specifically designed for Indian languages. All categories or subcategories within the BIS PoS tagset may not be applicable to every Indian language. Due to the linguistic diversity and structural variances among Indian languages, some categories may be relevant to certain languages while others might not be, depending on their distinctive grammatical features and grammatical constructions.

Sl. No	Category			Label	Annotation Convention **	Remarks
	Top level	Subtype (level 1)	Subtype (level 2)			
1	Noun			N	N	
1.1		Common		NN	N NN	
1.2		Proper		NNP	N NNP	
1.3		Verbal		NNV	N NNV	The verbal noun type is only for languages such as Tamil and Malayalam)
1.4		Nloc		NST	N NST	
2	Pronoun			PR	PR	

2.1		Personal		PRP	PR PRP	
2.2		Reflexive		PRF	PR PRF	
2.3		Relative		PRL	PR PRL	
2.4		Reciprocal		PRC	PR PRC	
2.5		Wh-word		PRQ	PR PRQ	
2.6		INDEFINITE		PRI	PR PRI	
3	Demonstrative			DM	DM	
3.1		Deictic		DMD	DM DMD	
3.2		Relative		DMR	DM DMR	
3.3		Wh-word		DMQ	DM DMQ	
3.4		Indefinite		DMI	DM DMI	
4	Verb			V	V	
4.1		Main		VM	V VM	
4.1.1			Finite	VF	V VM VF	
4.1.2			Non-finite	VNF	V VM VNF	
4.1.3			Infinitive	VINF	V VM VINF	
4.1.4			Gerund	VNG	V VM VNG	
4.2		Verbal		VN	V VN	<i>paTittam, naTattam, naTanam</i>
4.2		Auxiliary		VAUX	V VAUX	
4.2.1			Finite	VAUX	V VAUX VF	
4.2.2			Non-finite	VNF	V VAUX VN F	
4.2.3			Infinitive	VINF	V VAUX VI NF	
4.2.4			Gerund	VNG	V VAUX VN G	
4.2.5			Participle noun	VNP	V_VAUX_VNP	
5	Adjective			JJ		
6	Adverb			RB		Only manner adverbs
7	Postposition			PSP		
8	Conjunction			CC	CC	
8.1		Co-ordinator		CCD	CC CCD	
8.2		Subordinator		CCS	CC CCS	
8.2.1			Quotative	UT	CC CCS UT	
9	Particles			RP	RP	

9.1		Default		RPD	RP RPD	
9.2		Classifier		CL	RP CL	
9.3		Interjection		INJ	RP INJ	
9.4		Intensifier		INTF	RP INTF	
9.5		Negation		NEG	RP NEG	
10	Quantifiers			QT	QT	
10.1		General		QTF	QT QTF	
10.2		Cardinals		QTC	QT QTC	
10.3		Ordinals		QTO	QT QTO	
11	Residuals			RD	RD	
11.1		Foreign word		RDF	RD RDF	A word written in script other than the script of the original text
11.2		Symbol		SYM	RD SYM	For symbols such as \$, & etc
11.3		Punctuation		PUNC	RD PUNC	Only for punctuations
11.4		Unknown		UNK	RD UNK	
11.5		Echowords		ECH	RD ECH	

Table 1-1: BIS Tagset

1.4 PoS ANNOTATION IN LDC-IL DATA

LDCIL works on building annotated linguistic corpora for Indian languages. These corpora are automatically tagged and then verified by linguistic experts to ensure accuracy and consistency.

LDC-IL has developed the PoS tagged datasets in multiple scheduled Indian languages, including Assamese, Bengali, Bodo, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Nepali, Odia, Punjabi, Tamil, Telugu, and Urdu. This wide coverage highlights LDC-IL's efforts toward building a rich and standardized linguistic resource for diverse Indian languages.

In developing these corpora, LDC-IL adopted the Bureau of Indian Standards tagset as the underlying framework. Since Indian languages are structurally diverse and do not always share identical grammatical categories, the BIS tagset was customized to accommodate the unique linguistic features of each language. Certain morphological or syntactic categories may be present in one language but absent in another. For example, while some languages distinguish between honorific and non-honorific verb forms, others may not.

LDC-IL developed the PoS data in multiple languages based on the guidelines prepared by the Language Technologies Research Centre, IIIT Hyderabad. The document, [AnnCorra: Annotating Corpora – Guidelines for POS and Chunk Annotation for Indian Languages](#), was created by Akshar Bharati, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal, and released on 15-12-2006. These annotation standards define a PoS tagging scheme for Indian languages consisting of 26 part-of-speech tags.

LDC-IL has been able to achieve both standardizations across languages and language-specific accuracy. This data can be comparable among the languages because it shares the same framework and are tagged with similar grammatical structures. This ensures that the corpora are linguistically meaningful, making them valuable resources for computational linguistics, natural language processing, and linguistic research.

LDC-IL developed PoS data on 17 Indian scheduled languages and they are from different language families. Each language family or individual languages keep their grammatical or structural peculiarities so those characteristics need to be considered and it always cannot be kept under certain unique parameters. Dravidian languages such as Kannada, Malayalam, Tamil, and Telugu use the same grammatical categories because structurally they are similar. Indo-Aryan and Tibeto-Burman families use their own customized tagsets according to their convenience.

1.5 SUMMARY OF LDC-IL PoS DATA

LDC-IL developed PoS Data in 17 Scheduled Indian languages. The details about each language peculiar grammatical characteristics and data size are mentioned in the following chapters. LDC-IL PoS annotated data based on BIS framework has 12668303 tags. The summary of LDC-IL parts of speech data is given below.

Dataset Summary	Aggregate Count
Total PoS Tags	12668303
Total Sentences	928592
Total Word Count	10795498
Total File Count	13829

Table 1-2: LDC-IL PoS Status

2 MAITHILI POS ANNOTATED CORPUS

Dinesh Mishra

2.1 GRAMMATICAL PECULIARITIES OF MAITHILI

Maithili is a morphologically rich language characterized by extensive word formation processes. Words in Maithili commonly undergo inflection, derivation, and compounding, resulting in an extremely large and virtually open-ended vocabulary. Because of these productive morphological operations, a single root or lemma can generate a wide range of wordforms, each encoding distinct grammatical functions.

From a Part-of-Speech (POS) perspective, this morphological complexity poses significant challenges. Multiple grammatical categories—such as case, number, gender, tense, aspect, mood, person, and voice—may be simultaneously encoded within a single word. As a result, PoS tagging in Maithili requires not only identifying the syntactic class of a token but also understanding the layered morphological information embedded within it.

A corpus effectively captures the linguistic nuances and unique characteristics of a language when it systematically represents the morphological, syntactic, and semantic variables.

2.2 LDC-IL CUSTOMIZED BIS TAGSET

Based on the BIS tagset and LDC-IL–annotated Maithili texts, certain categories such as indefinite pronouns and classifiers are not applied in Maithili tagging, as these features are either rare or not present in Maithili grammatical structure. Linguistic Standards for Indian Language (IL) Corpora Annotation: PoS Categories and Labels (June 12, 2010).

Sl. No	Category		Label	Annotation Convention**	Examples
	Top level	Subtype (level 1)			
1	Noun		N	N	
1.1		Common	NN	N_NN	टिकुली
1.2		Proper	NNP	N_NNP	महेश
1.3		Nloc	NST	N_NST	आगूमे
2	Pronoun		PR	PR	
2.1		Personal	PRP	PR_PRP	हमर
2.2		Reflexive	PRF	PR_PRF	अपन
2.3		Relative	PRL	PR_PRL	जिनकर
2.4		Reciprocal	PRC	PR_PRC	एक-दोसरकेँ

2.5		Wh-word	PRQ	PR_PRQ	कहाँ
3	Demonstrative		DM	DM	
3.1		Deictic	DMD	DM_DMD	ई
3.2		Relative	DMR	DM_DMR	जनिका
3.3		Wh-word	DMQ	DM_DMQ	किनका
4	Verb		V	V	
4.1		Main	VM	V_VM	खाइत, जाइत
4.3		Auxiliary	VAUX	V_VAUX	अछि, छल
5	Adjective		JJ		पुरान, नीक
6	Adverb		RB		फेर, पूर्णतः
7	Postposition		PSP		सँ, केँ
8	Conjunction		CC	CC	
8.1		Co-ordinator	CCD	CC_CCD	मुदा, आ
8.2		Subordinator	CCS	CC_CCS	तँ, जँ
9	Particles		RP	RP	
9.1		Default	RPD	RP_RPD	हँ, सन, धरि,
9.3		Interjection	INJ	RP_INJ	अरे, ओह, उफ
9.4		Intensifier	INTF	RP_INTF	बहुत, ज्यादा
9.5		Negation	NEG	RP_NEG	न, नहप, मत
10	Quantifiers		QT	QT	
10.1		General	QTF	QT_QTF	कनेक, बहुत,
10.2		Cardinals	QTC	QT_QTC	एक, एकटा, दुई, बीसगोट
10.3		Ordinals	QTO	QT_QTO	पहिल, दोसर
11	Residuals		RD	RD	
11.1		Foreign word	RDF	RD_RDF	A word written in script other than the script of the original text
11.2		Symbol	SYM	RD_SYM	\$, &
11.3		Punctuation	PUNC	RD_PUNC	, .
11.4		Unknown	UNK	RD_UNK	
11.5		Echowords	ECH	RD_ECH	जलखे- (तलखे), मोंट-(सोंट)

Table 2-1: Maithili Tagset

The annotation is to be done using the lowest level tag of the type hierarchy. Once the lower level tag is selected, the higher level tags should be stored automatically.

The list was prepared in the PoS standards committee meeting held on June 12, 2010 at IIIT, Hyderabad.

2.3 MAITHILI POS TAGGING

Maithili PoS tagging is performed using the BIS tagset, with necessary adaptations from LDC-IL guidelines to suit Maithili's linguistic structure. To improve tagging accuracy context-based tag assignment is carried out to ensure the consistency across the corpus.

2.3.1 Annotation Guidelines

- Maithili PoS Tagging is done based on BIS tagset with customazation to capture specific grammatical and morphological features.
- Compound words were examined carefully to determine appropriate tagging at the word level.
- Spelling mistakes rectified.
- Extra spaces were removed where tokens had been incorrectly split by spaces.
- Missing spaces were inserted where necessary.
- Validation has done by using ambiguous word list.
- Contextual validation was performed to assign the correct tag
- Manual verification was conducted to maintain data quality
- Disagreements were resolved through discussion

2.3.2 Tokenization

Tokenization is the process of breaking down a sentence into its constituents. It is important because accuracy of the PoS tagged data depends on the input words. PoS tagging incorporates the morpho-syntactic information to the lexical item. It does not insert any semantic, discourse or pragmatic information. On the word-level, tokenization refers to breaking down of a sentence into the constituent words. For example, हम भोजन करैत छी will be tokenized into 4 tokens- हम | भोजन | करैत | छी

In inflectional languages such as Maithili, tokenization helps in breaking down the morphological and syntactic status of the words. Furthermore, in compound words, tokenisation helps understand the layered meaning of a word.

Some compound words often pose a challenge to the tokenization process and need to be pre-processed. For example, the expressions such as मने-मन 'mane-man' (Noun+ Noun), हुकूमत-ए-बर्तानिया 'hukumat-e-bartaaniya' (Noun+Postposition+Proper Noun) have multiple grammatical categories but they function as a single token, tagged as a noun.

It is interesting to note that though compound words in Maithili are written using a hyphen, non-hyphenated instances can also be found. For eg: the following words have two instances of the same category (Noun). Though the grammatical combination is noun+noun, the words function as single tokens- समयसीमा ‘samay seema’, पत्थरदिल ‘patthardil’

2.3.3 Selecting a Tag

A word has two important elements; one is its form and other is its function. The function of a word offers significant clues for the subsequent stages of processing. So here adopt a balance between the form and the function of a word in a systematic and consistent way.

Category of a word is decided by utilizing the morphological, syntactic and semantic functions. A word’s category can be easily identified by morphological properties, but functionally it can act as another category due to context. This variation is extracted by syntactic clues. For example some pronouns function as demonstratives if they occur before a noun.

ओ घरक बहार ठढ़ अछि। (He is standing outside the house.)

ओ किताब टेबुल पर राखि दियो। (Keep that book on the table.)

In the first sentence, the token ‘ओ’ serves as a personal pronoun. However, in the second instance, it works as a demonstrative because it modifies a common noun (किताब). Similarly, indefinite pronouns such as कोनो and केओ also function as demonstratives when preceding a noun.

In Maithili, there exists a plethora of words which differ in their function depending upon the neighboring linguistic environment such as ‘डाल’, ‘होयब’, ‘नाबालिग’ etc.

2.3.4 Observation on PoS Tagging of LDC-IL Maithili Text Corpora

PoS tagging gives the lexical category and it gives some surrounding information. That means it taken care the form and function of the words. If it is possible to give the formal and functional features then it will be very useful for higher level of processing. The PoS data can be classified into rightly tagged words, ambiguously tagged words and untagged words.

Most of the noun, pronoun, adjectives, finite verb, non- finite verb, adverb etc. are rightly tagged. Ambiguity is a common feature of natural language and which arises at the lexical level. The efficiency of a PoS tagger can be evaluated based on how well it handles lexical ambiguity. Many lexical items give more than one sense and it can appear in different categories. If the contextual meaning and idiomatic expressions are taken into account the order of ambiguity will increases. Words remain untagged or are assigned an unknown tag due to a lack of information about the word or an inadequate definition of the tagset. This can occur with foreign words written in a foreign script, foreign words in transliterated form that are unfamiliar to the language, dialectal variations, and similar cases.

2.3.4.1 Inconsistency

In Maithili PoS tagged corpus, inconsistencies often arise due to the morpho-syntactic flexibility present in the language where though expressions share semantic similarity, they exhibit structural differences. As discussed earlier, compound forms have varying instantiations. Notably, forms such as तरकारीवाला and तरकारी वाला or दूधवाला and दूध वाला convey similar meanings but they differ in tokenization and subsequent tagging- The following examples shows the same elements can appear separately or as a single combined token with their respective PoS tags.

तरकारीवाला\N_NN आ तरकारी\N_NN वाला\RP_RPD या दूधवाला\N_NN आ दूध\N_NN वाला\RP_RPD

The phenomenon of zero derivation in Maithili also often creates ambiguity in PoS tag assignment. However, examining the context solves this problem. Let us look at the word ‘बस’ to understand this further.

ओ बसमे\N_NN बैस गेलीह।
 बसमे\RP_RPD, हम तुरन्त गेलहु आ तुरन्ते एलहुँ।
 ओ जर्मनीमे बैस\V_VM गेल छथि।

This example demonstrates that the same token functions as a noun, particle or a verb depending upon the context. More examples are presented in the ambiguity subsection. Though this results in inconsistent tags for the same token, these variations highlight that inconsistencies in Maithili POS-tagged corpora stem not only from annotation practices but also the rich linguistic properties such as variable tokenization and category fluidity.

2.3.4.2 Ambiguity

Natural language is ambiguous from its basic nature. Ambiguity arise when a single word or string of words is associated in the language system with more than one meaning (Chierchia and McConnell-Ginet 1990:32). Generally lexical ambiguity is of two types; syntactic lexical ambiguity and semantic lexical ambiguity. Syntactic lexical ambiguity simply refers to ambiguity of category, e.g., noun or verb. Semantic lexical ambiguity means a word has several meanings.

The following are some examples of ambiguity observed in Maithili PoS tagging. This was addressed with careful attention in LDC-IL Maithili PoS tagged text corpus.

JJ, RP_RPD and RB

In Maithili
ई किताब नीक\JJ अछि।
नीक\RP_RPD, ओ कियैक आयल छल?
ओ नीक\RB नाचैत अछि।

In IPA
i: kiʈab ni:k\JJ əcʰi
ni:k\RP_RPD, o kiʃek əjəl cʰəl
o ni:k\RB nacɛʈ əcʰi

N_NN & V_VM

In Maithili
सड़क पर बच्चा\N_NN दौड़ि रहल अछि।
दुनू बच्चा\V_VM आबि गेल।
In IPA
səʈək pər bæcca\N_NN d̪ɔʈi rəʃəl əcʰi
d̪unu: bæcca\V_VM abi gel

N_NN & RP_NEG

In Maithili
मोहनजी वोट\N_NN देबाक लेल गेल छथि।
एहि हालतमे अहाँ बाहर नहि\RP_NEG जाउ।
In IPA
mohənəʃi: wot\N_NN d̪ebak lel gel cʰəʈʰi
ehi hālətəme əɦaː bəɦər nəɦi\RP_NEG ʃau

N_NN & N_NNP

In Maithili
सीमाक उफ\N_NN सुनिकs मोहन भागल चलि आयल।
उफ\RP_INJ ! आई बहुत रास काज करबाक अछि।
In IPA
si:mak upʰ\N_NN sunikəs mohən bʱagəl cəli əjəl
upʰ\RP_INJ ! ai: bəɦuʈ ras kaʃ kərəbak əcʰi

JJ & V_VM

In Maithili
शिक्षिका द्वारा हमरा हरियर\JJ कागज देलनि
ओ हमरा लड़ाईमे हरा\V_VM देलक।
In IPA
ʃikʃika d̪wara həmərə həriʃər\JJ kaɡəʃ d̪ələni
o həmərə ləʈai:me hərə\V_VM d̪ələk

QT_QTO & N_NN

In Maithili
ई नहि, दोसर\QT_QTO साड़ी दिखाउ।
एक लड़की कलम लेलक, दोसर\N_NN किताब लेलक।
In IPA
i: nəɦi, d̪osər\QT_QTO saʈi: d̪ikʰau
ek ləʈəki: kələm lelək, d̪osər\N_NN kiʈab lelək

2.3.4.3 Abbreviation

Abbreviations are shortened forms of words or phrases which are widely used in both written and spoken language. Some examples of abbreviations in Maithili are- के.बी., एल-एल.बी., बी.डी.ओ, बी.ए. etc.

2.3.4.4 Confusion in the definition of Nloc

The Nloc category in Maithili is particularly prone to inconsistency because of the functional flexibility of spatial words. Words encompassed under this category, such as के अंदर, के बाहर, के ऊपर, के नीचा etc, though inherently denote location or time, their syntactic behavior differs across contexts at times.

In Maithili
ओ अंदर\RB गेलाह।
अंदरक\N_NST कोठली।
In IPA
o əndəɾ\RB gelah
əndəɾək\N_NST koṭʰəli:

In the first sentence, अंदर functions as an adverb because it is modifying the verb, whereas in the second sentence, it behaves as a nominal modifier and is tagged as N_NST. नहि अछि

In Maithili
घरक लग\N_NST दुकान अछि।
हमरा लग\PSP कलम नहि अछि।
In IPA
gʱəɾək ləg\N_NST dukan əɕʰi
ɦəmərə ləg\PSP kələm nəɦi əɕʰi

Similarly, in the first example, लग functions as a nominal modifier and is hence tagged as a locative marker. However, in the second instance, the same word takes on a more postpositional function, leading to the variation in the tags. This fluidity of spatial/temporal expressions into nominal, adverbial and postpositional categories makes tag assignment confusing and often leads to inter-annotator disagreement and makes it difficult to enforce categorical boundaries.

2.3.4.5 Compound Word

Two or more morphemes are combined to form a single unit and the whole token gives a meaning. For example, रात-भर, रंग-बिरंगा, उल्टा-सीधा, अबैत-जाइत, दिल-दिमाग etc. These words are taken as whole and assigned the respective appropriate tag.

2.3.4.6 Dialect Variation

Maithili exhibits significant dialect variations across regions which is characterized by differences in phonology, vocabulary and morphology. Varieties such as Angika, Bajika, etc. though often grouped under the Maithili umbrella, display very unique features setting them apart from Standard Maithili. This dialectical variation greatly influences the composition and annotation of linguistic corpora as speakers often incorporate regional variants into their language. This results in the seeping in of dialectical vocabulary into the Maithili corpora. Some expressions from regional dialects of Maithili found in the corpus are नून, मनई, ढीठ, मांड़-भात, लुकाइ, रोकत etc.

2.3.4.7 Impact of Social Dialect

Dialect variation in Maithili is often shaped by social factors such as age, education, urban/rural setting, cultural exposure etc. Speakers from urban background often code-mix Maithili with English due to which a large amount of English words have entered Maithili corpus. Words such as 'प्लीज', 'कैंसिल', 'हेलो', 'स्लो' etc. have become very common in day-to-day use. Similarly, some cases can be found in our corpus where expressions from other languages such as Kashmiri can also be found in Maithili text corpus. For example, 'कूरी कूरी तबाशीरी' and 'पननि कूरी-लगअयय' are in Kashmiri language but written using Devanagari. Words from other languages such as सुण, रज्जाँ, पैरी etc. can also be observed.

2.3.4.8 Verbal Noun

The word which denoting result of action is tagged as verbal noun category other type of verbal nouns are tagged as noun.

2.3.4.9 Unknown Words

There will be some words in the text which cannot be handled by the tagset. Such words are considered as unknown words and it will pose a problem for the higher level of processing. For example, किमानंदो\RD_UNK, अनालयी\RD_UNK etc.

2.4 SUMMARY OF LDC-IL MAITHILI POS DATA

The status of the LDC-IL Maithili Annotated Parts-of-Speech Corpus and number of distinct tag count are provided below:

Dataset Summary	Aggregate Count
Total PoS Tags	427438
Total Sentences	25431
Total Word Count	371983
Total File Count	96

Table 2-2: Maithili PoS Status

Sl. No.	PoS Category	PoS Tag	Tag Count
1	Common Noun	N_NN	110154
2	Proper Noun	N_NNP	37057
3	Nloc Noun	N_NST	7634
4	Personal Pronoun	PR_PRP	17044
5	Reflexive Pronoun	PR_PRF	2999
6	Relative Pronoun	PR_PRL	2993
7	Reciprocal Pronoun	PR_PRC	106
8	Wh-word Pronoun	PR_PRQ	930
9	Deictic Demonstrative	DM_DMD	6133
10	Wh-word Demonstrative	DM_DMQ	126
11	Relative Demonstrative	DM_DMR	706
12	Indefinite Demonstrative	DM_DMI	984
13	Verb Main	V_VM	53227
14	Verb Auxiliary	V_VAUX	28751
15	Adjective	JJ	35811
16	Adverb	RB	5647
17	Postposition	PSP	15505
18	Co-ordinator	CC_CCD	13519
19	Subordinator	CC_CCS	5141
20	Particles	RP_RPD	5326
21	Interjection	RP_INJ	426
22	Intensifier	RP_INTF	1043
23	Negation	RP_NEG	5463
24	Classifier	RP_CL	574
25	Quantifiers	QT_QTF	5136
26	Cardinals	QT_QTC	6004
27	Ordinals	QT_QTO	1546
28	Foreign word	RD_RDF	94
29	Symbol	RD_SYM	7
30	Punctuation	RD_PUNC	55448
31	Unknown	RD_UNK	869
32	Echowords	RD_ECH	4

Table 2-3: Distinct Tag Count of Maithili