

MANIPURI PARTS OF SPEECH ANNOTATED CORPUS



মনিপুরী
ৰাহৈ কয়াং শক্তাক ৰাহৈকোন

MANIPURI 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:

**Amom Nandaraj Meetei
Yumnam Premila Chanu
Dr. Narayan Choudhary
Rajेश N.**

**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

Manipuri Parts of Speech Annotated Corpus

Authors: Amom Nandaraj Meetei, Yumnam Premila Chanu, Dr. Narayan Choudhary, Rajesha N.

e-ISBN: 978-81-69175-43-2

CIIL Publication No.: 1697

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: Amom Nandaraj Meetei

Table of 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
1.6	Reference	Error! Bookmark not defined.
2	Manipuri PoS Annotated Corpus	5
2.1	Grammatical peculiarities of Manipuri	6
2.2	LDC-IL Customized BIS Tagset	6
2.3	Manipuri PoS Tagging	8
2.4	Summary of LDC-IL Manipuri PoS Data	15

Figures

No table of figures entries found.

Tables

Table 1-1: BIS Tagset	4
Table 2-1: Manipuri Tagset	7

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 level 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 Conventio**	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 has been able to achieve both standardization 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 have their own grammatical or structural peculiarities; so those characteristics need to be considered and it cannot always 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 MANIPURI POS ANNOTATED CORPUS

Amom Nandaraj Meetei

2.1 GRAMMATICAL PECULARITIES OF MANIPURI

Typologically, Manipuri is a morphologically rich language exhibiting highly productive word-formation processes. Lexical items in Manipuri often undergo inflection, derivation, and compounding, which in turn results in a vast and potentially open-ended vocabulary. Owing to these productive morphological operations, a single root or lemma can give rise to wide range of word forms, each encoding distinct grammatical or semantic functions.

From a Part-of-Speech (PoS) perspective, this morphological richness provide considerable analytical challenges. Multiple grammatical categories- such as case, number, gender, definiteness, tense, aspect, mood, person, and voice- may be simultaneously articulated within a single morphological word. Consequently, PoS tagging in Manipuri requires both the identification of the syntactic category of a token and the interpretation of the complex morphological information embedded within it.

Furthermore, corpus evidence shows that certain independent lexical items, particularly demonstratives, may undergo grammaticalization, eventually emerging as bound suffixal forms through phonological reduction and morphologization.

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 the LDC-IL–annotated Manipuri corpus, certain categories- such as relative pronouns, relative demonstratives, verbal auxiliaries, infinitives, participles, and classifiers—are not employed in Manipuri PoS tagging, as these categories are either rare or absent in the grammatical structure of the language (Linguistic Standards for Indian Language (IL) Corpora Annotation: PoS Categories and Labels, June 12, 2010).

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	ঝুম
1.2		Proper		NNP	N__NNP	ইম্ফাল
1.3		Nloc		NST	N__NST	মমাং

2	Pronoun			PR	PR	
2.1		Personal		PRP	PR__PRP	ঐ
2.2		Reflexive		PRF	PR__PRF	ইশা
2.3		Reciprocal		PRC	PR__PRC	মশেল
2.4		Wh-word		PRQ	PR__PRQ	কনা
2.5		Indefinite		PRI	PR__PRI	কনাগুমা
3	Demonstrative			DM	DM	
3.1		Deictic		DMD	DM__DMD	অসি
		Wh-word		DMQ	DM__DMQ	করমা
3.2		Indefinite		DMI	DM__DMI	কনাগুমা
4	Verb			V	V	
4.1		Main		VM	V__VM	
4.1.1			Finite	VF	V__VM__VF	তৌই
4.1.2			Non-finite	VNF	V__VM__VNF	তৌরগা
4.1.3			Gerund	VNG	V__VM__VNG	তৌবা
5	Adjective			JJ		অফবা
6	Adverb			RB		Only manner adverbs
7	Postposition			PSP		
8	Conjunction			CC	CC	
8.1		Co-ordinator		CCD	CC__CCD	অমসুং
8.2		Subordinator		CCS	CC__CCS	
8.2.1			Quotative	UT	CC__CCS__UT	হায়না
9	Particles			RP	RP	
9.1		Default		RPD	RP__RPD	খক
9.2		Interjection		INJ	RP__INJ	হৈমা
9.3		Intensifier		INTF	RP__INTF	য়ান্না
10	Quantifiers			QT	QT	
10.1		General		QTF	QT__QTF	খর
10.2		Cardinals		QTC	QT__QTC	অমা
10.3		Ordinals		QTO	QT__QTO	অনিশুবা
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		Echo words		ECH	RD__ECH	

Table 2-1: Manipur 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. The Part-of-Speech (PoS) tags for the Manipuri language were subsequently finalized by Amom Nandaraj Meetei from the Language Data Consortium for Indian Languages, Central Institute of Indian Languages, Mysore.

2.3 MANIPURI POS TAGGING

Manipuri PoS tagging is carried out using the BIS tagset, with necessary adaptations from the LDC-IL guidelines to accommodate the linguistic structure of Manipuri. Although Manipuri is an agglutinative language, preprocessing steps such as word segmentation were not performed prior to tagging in order to conform to the tagset framework. This approach may to a certain extent account for the presence of certain morphological elements that surface through processes of grammaticalization. Nevertheless, to enhance tagging accuracy, context-based tag assignment is employed, thereby ensuring greater consistency across the corpus.

2.3.1 Guidelines

- Manipuri PoS Tagging is done based on BIS tagset with customization to capture specific grammatical and morphological features.
- Although Manipuri is an agglutinative language, no preprocessing was performed to account for morphological elements arising from grammaticalization.
- Complex words were left unsegmented to preserve incorporation phenomena and ensure accurate PoS categorization.
- Compound words were examined carefully to determine appropriate tagging at the word level.
- Spelling mistakes were rectified.
- Extra spaces were removed where tokens had been incorrectly split by spaces.
- Missing spaces were inserted where necessary.
- Validation has been 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

By tokenization we refer to the process of segmenting a text into individual tokens or lexical units. This step is crucial because the accuracy of PoS-tagged data largely depends on the correctness of the input tokens. PoS tagging involves assigning morpho-syntactic information to lexical items in a text. It does not encode semantic, discourse, or pragmatic information.

In an agglutinative language such as Manipuri, the process of tokenization presents significant challenges. The copular verb *নি* /ni/ is often realized morphologically as a suffix and can attach to a wide range of PoS categories. For instance, the word *-জুমনি* /jumni/ consists of two grammatical elements: *-জুম* /jum/ ‘house’ (noun) and the copula *নি* /ni/ (verb).

Furthermore, several Spatio-Temporal Nouns undergo a grammaticalization process in which an originally independent lexical item gradually develops into a bound morpheme through the pathway word → clitic → affix. At the initial stage, the element functions as an independent lexical item referring to a concrete object, ultimately resulting in a process of lexicalization or concretion. This can be illustrated by the following example:

জুমগী মথক /jumgi mətʰək/ “above the house” → *জুম মথক* /jum mətʰək/ “above the house” → *জুমথক* /jumtʰək/ “roof (of the house)”.

Similarly, deictic elements indicating proximal or distal reference may undergo the same process of morphologization. Continuing with the previous example:

জুমথক অদু /jumtʰək ədu / “that roof” → *জুমথক্তু* /jumtʰəktu/ “the roof”.

Finally, the copular suffix can attach to the entire form:

জুমথক্তুনি /jumtʰəktuni/ “it is the roof”.

From a morpho-syntactic perspective, this single word contains multiple grammatical categories: Common Noun (N_NN) + Spatio-Temporal Noun (N_NST) + Deictic (DM_DMD) + Finite Verb (V_VM_VF). Segmenting such forms into their corresponding lexical components would require the insertion of intermediate phonological elements, which results in an artificial representation that does not accurately reflect the natural morpho-phonological structure of the language. Such forms encode multiple grammatical categories within a single token, making segmentation into separate lexical units computationally problematic.

2.3.3 Selecting a Tag

A lexical item has two fundamental aspects: form and function. The functional behavior of a word offers crucial cues for subsequent stages of linguistic processing. Therefore, the present approach attempts to maintain a systematic balance between the formal structure and the functional role of a word.

The grammatical category of a word is decided by examining its morphological, syntactic, and semantic properties. Morphological features often provide primary evidence for category assignment. However, lexical items may display category flexibility in context, whereby their functional role diverges from their canonical morphological class. Such variations are usually identified through syntactic distributional cues. For instance, a noun may assume an adjectival function when it occurs immediately before another noun. Semantic interpretation can also contribute to determining the appropriate lexical category.

A relevant example in Manipuri is the form অদুনা /ədunə/. Morphologically, it consists of the distal demonstrative অদুনা /ədu/ combined with the nominative case marker না /nə/, which may convey an agentive interpretation. Nevertheless, in certain discourse contexts, the same form undergoes functional reanalysis and operates as a discourse-linking conjunction. Thus, depending on the syntactic and discourse environment, অদুনা /ədunə/ may be annotated either as a demonstrative (DM_DMD) or as a subordinating conjunction (CC_CCS). The following examples illustrate these two functional realizations.

In Manipuri
মপাদু\N_NN থবক\N_NN চংখে\V_VM_VF ;\RD_PUNC অদুনা\CC_CCS মচাশিংদু\N_NN লাংবা\V_VM_VGN হৌরক্লে\V_BM_VF \RD_PUNC
নুপা\N_NN অদুনা\DM_DMD নোক্লে\V_VM_VF \RD_PUNC
In IPA
məpadu\N_NN tʰəbək\N_NN tʃəkʰəre\V_VM_VF; ədunə\CC_CCS mətʃaʃiŋdu\N_NN laŋbə\V_VM_VGN həurəkʰle\V_VM_VF.
nupa\N_NN ədunə\DM_DMD nokʰle\V_VM_VF.\RD_PUNC

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

Part-of-Speech (PoS) tagging assigns lexical categories to words while also capturing certain contextual information. In doing so, it accounts for both the formal properties and functional roles of lexical items. When formal and functional features are explicitly encoded, the annotated data become significantly more useful for higher-level linguistic processing, such as syntactic parsing and semantic analysis. PoS-tagged data can generally be divided into three categories: correctly tagged words, ambiguously tagged words, and untagged words.

Most lexical items, including nouns, pronouns, adjectives, finite verbs, non-finite verbs, and adverbs, are typically assigned clear and unambiguous tags. However, lexical ambiguity is an inherent characteristic of natural language and commonly arises at the word level. The effectiveness of a PoS tagger is therefore often evaluated based on its ability to handle such

ambiguity. Many lexical items have multiple meanings and may appear in different grammatical categories depending on the context. When contextual interpretation and idiomatic expressions are taken into account, the degree of ambiguity may further increase.

In some cases, words remain untagged or are assigned an unknown tag due to insufficient lexical information or limitations in the tagset design. This situation may arise with foreign words written in non-native scripts, transliterated loanwords unfamiliar to the language, dialectal variants, and other similar forms.

2.3.4.1 Inconsistency

The same lexical item may appear in different forms within the corpus. For instance, in Manipuri, certain words may occur as compound forms in some contexts, while in other contexts the same elements may appear as separate lexical items. Although these variants often convey the same semantic meaning, they may exhibit different morpho-syntactic properties.

Language names, which are typically categorized as proper nouns, may also function adjectivally depending on their syntactic position. For example, in the phrase মনিপুরী লৈশাবীশিং /mənɪpuri ləɪsəbiʃiŋ/ ‘Manipuri maidens’, the word মনিপুরী /mənɪpuri/ functions as an adjective modifying the head noun লৈশাবীশিং /ləɪsəbiʃiŋ/ ‘maidens’. In contrast, in the expression মনিপুরী তম্মু /mənɪpuri təmму/ ‘learn Manipuri’, the same lexical item মনিপুরী /mənɪpuri/ functions as a proper noun referring to the language.

This example demonstrates that the grammatical category of a word is often determined by its syntactic distribution and linguistic environment rather than by its lexical form alone.

2.3.4.2 Ambiguity

Natural language is inherently characterized by ambiguity. Ambiguity arises when a single lexical item or sequence of words is associated with more than one interpretation within a linguistic system (Chierchia and McConnell-Ginet 1990:32). In linguistic analysis, lexical ambiguity is usually classified into two major types: syntactic lexical ambiguity and semantic lexical ambiguity. Syntactic lexical ambiguity refers to ambiguity in grammatical category, where a lexical item may belong to more than one part of speech, such as functioning both as a noun and a verb. Semantic lexical ambiguity, on the other hand, occurs when a single lexical item is associated with multiple meanings within the semantic system of a language.

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

JJ & N_NN

In Manipuri

অইবনা\N_NN লৈখিদবদা\V_VM_VGN মীয়ান্না\N_NN অৱাবা\JJ মপাউ\N_NN ফোংদোক্লরক্লি\V_VM_VF I\RD_PUNC
ইচাওসিনা\N_NN মরম\N_NN ওইদুনা\V_VM_VNF মীয়ান্না\N_NN অৱাবা\N_NN কয়া\QT_QTF মায়োক্লরি\V_VM_VF I\RD_PUNC
In IPA
əɪbənə\N_NN ləɪkɪdəbədə\V_VM_VGN mijamnə\N_NN əwabəJJ məpau\N_NN p ^h oŋdoknəɾəklɪ\V_VM_VF .I\RD_PUNC
ɪʃausinə\N_NN məɾəm\N_NN oɪdunə\V_VM_VNF mijamnə\N_NN əwabə\N_NN kəjə\QT_QTF məjoknəɾɪ\V_VM_VF .I\RD_PUNC

DM_DMD & PR_PRP

In Manipuri
লাইরিক\N_NN অসিনা\DM_DMD হেন্না\RB থাই\V_VM_VF I\RD_PUNC
পাউ\N_NN অদু\DM_DMD অসিনা\PR_PRP তস্থনি\V_VM_VNG I\RD_PUNC
In IPA
laɪrɪk\N_NN əsɪnə\DM_DMD hennə\RB t ^h aɪ\V_VM_VF .I\RD_PUNC
pau\N_NN ədu\DM_DMD əsɪnə\PR_PRP təstʰəni\V_VM_VNG .I\RD_PUNC

N_NN & CC_CCS

In Manipuri
তা\N_NNP চাওবনা\N_NNP চংখিবগী\V_VM_VNG মরম\N_NN ঐদি\PR_PRP খঙদে\V_VM_VF I\RD_PUNC
ইবেমানা\N_NNP চাংয়েংদু\N_NN মায়\N_NN পাকখিদে\V_VM_VF মরমদি\CC_CCS মহাক্লা\PR_PRP লাইরিক\ N_NN চাং\N_NN নায়না\RB পাখিদে\V_VM_VF I\RD_PUNC
In IPA
ta\N_NNP tʃaʊbənə\N_NNP tʃəŋkɪbəgi\V_VM_VNG məɾəm\N_NN əɪdɪ\PR_PRP kəŋde\V_VM_VF .I\RD_PUNC
ɪbəmənə\N_NNP tʃaŋjɛŋdu\N_NN maɪ\N_NN pak ^h ɪde\V_VM_VF məɾəmdɪ\CC_CCS məhaknə\PR_PRP laɪrɪk\N_NN tʃaŋ\N_NN naɪnə\RB pak ^h ɪde\V_VM_VF .I\RD_PUNC

N_NN & N_NNP

In Manipuri

লৈকোনদা\N_NN লৈরাং\N_NN কয়া\QT_QTF শাংলি\V_VM_VF I\RD_PUNC
ওঁরাং\N_NST ইচে\N_NNP লৈরাং\N_NNP ঐখোয়গী\PR_PRP য়ুমদা\N_NN লাক্সমী\V_VM_VF I\RD_PUNC
In IPA
ləikondə\N_NN ləiraŋ\N_NN kəja\QT_QTF satli\V_VM_VF .\RD_PUNC
ŋəraŋ\N_NST iʃe\N_NNP ləiraŋ\N_NNP ək ^h oigɪ\PR_PRP jumdə\N_NN lakləmmi\V_VM_VF .\RD_PUNC

PR_PRQ & DM_DMQ

In Manipuri
মহাক্সা\PR_PRQ করি/PR_PRQ পাম্মী\V_VM_VF ?\RD_PUNC
নহাক\PR_PRQ করি/DM_DMQ নুমিত্তা\N_NN লাকই\V_VM_VF ?\RD_PUNC
In IPA
məhaknə\PR_PRQ kəri\PR_PRQ pammi\V_VM_VF ?\RD_PUNC
nəhak\PR_PRQ kəri\DM_DMQ numittə\N_NN laki\V_VM_VF ?\RD_PUNC

DM_DMI & PR_PRI

In Manipuri
মদু\PR_PRQ কনাগুম্ভা\DM_DMI নুপা\N_NN অমনা\QT_QTC তৌখি\V_VM_VF I\RD_PUNC
কনাগুম্ভা\PR_PRI অমা\QT_QTC অদুদা\RB লৈরে\V_VM_VF I\RD_PUNC
In IPA
mədu\PR_PRQ kənagumbə\DM_DMI nupa\N_NN əmənə\QT_QTC təukhi\V_VM_VF .\RD_PUNC
mədu\PR_PRQ kənagumbə\DM_DMI nupa\N_NN əmənə\QT_QTC təuk ^h i\V_VM_VF .\RD_PUNC
kənagumbə\PR_PRI əmə\QT_QTC ədudə\RB ləire\V_VM_VF .\RD_PUNC

2.3.4.3 Abbreviation

Certain individuals or institutions are commonly referred to by abbreviations of their names. Depending on the linguistic context in which they occur, these forms may be identified as proper nouns or assigned to other grammatical categories. Accordingly, they are tagged based on their contextual function in the sentence..

Manipuri	IPA
জি.সি. তোংব্রা means গিৎচন্দ্র তোংব্রম	/ dʒi.si. tɔŋbra/ means /ɡɪtʃəndrə tɔŋbrəm/
এমসু means ওল মনিপুর স্তুদেস্ত য়ুনিয়ন	/emsu/ means /ol məɪpur student juɲiɲn/

2.3.4.4 Confusion in the definition of Nloc

Words denoting space and time are tagged as N_NST, a subtype of the noun category **Nloc**. In many Indian languages, the ground noun phrase (NP) is followed by axial-part postpositions that express spatial relations such as *above, on, on top of, behind, under, beneath, next to, in front of, between, among, outside, inside, beside, around, and beyond*.

In such constructions, the ground NP is typically marked by different case markers depending on the language. For example, it is marked by the Genitive in Hindi and Manipuri, while in Tamil it may take the Dative marker. Unlike English, which uses prepositions (e.g., *on, above, in, under, inside, outside*), many Indian languages use common nouns functioning as relational nouns to express spatial and temporal relations. For instance, Hindi *upara* and Manipuri মথক *məthək* express the meaning ‘*on/over/above*’, while Hindi *andara* and Manipuri মনুং *mənuŋ* correspond to ‘*in/inside*’.

These spatial nouns are treated as spatio-temporal relational nouns because they function as **two-**place predicates. One argument is realized by another NP in the [NP + GEN + NST] construction. For example, in Manipuri:

মহাক্সা য়ুমগী মথক্তা কারম্মী।

məhək-nə jum-gi mətʰək-tə kɑ-rəm-mi
he-NOM house-GEN above- LOC climb-EVI-DECL
‘He climbed above the house.’

In this sentence, মথক *məthək* “top/above” follows its possessor য়ুম *yum* “house”. Such relational nouns are semantically incomplete when used alone, since they always indicate the spatial

relation of one entity with respect to another. Thus, *jum* functions as the argument of the relation expressed by *məthək*.

Accordingly, a finite set of Manipuri relational nouns is consistently tagged as **NST**, including: *মমাং* *məmaŋ* ‘front’, *মতুং* *mətuŋ* ‘behind’, *মথক* *məthək* ‘above / upstairs’ *মখা* *məkha* ‘below / down’, *মনুং* *mənuŋ* ‘inside/within’ *মপান* *məpan* ‘outside’, etc.

These nouns are tagged as **NST** because they primarily denote location or time. However, the inclusion of lexical items such as *ঙসি* */ŋəsi/* ‘today’, *ঙৱাং* */ŋəraŋ/* ‘yesterday’, *হৌজিক* */həudʒik/* ‘now’, *হয়েং* */həjeŋ/* ‘tomorrow’, *কৈদৌঙে* */kəidəuŋe/* ‘when’, and *কৈদৌনুংদা* *kəidəunuŋdā* ‘never’ within this category remains uncertain and potentially confusing, as their status as NST elements is not clearly established.

2.3.4.5 Multiword

A sequence of words may function as a single linguistic unit, which often makes it difficult to assign a single grammatical category. This difficulty is particularly evident in Manipuri, an agglutinative language, where multiple morphemes are frequently combined to form a single token. In such languages, elements belonging to different grammatical categories can attach to one another within the same word.

For example, the token *চাউখৎপশিংদুখকতনি* */tʃaukʰətpə-siŋ-du-kʰak-tə-ni/* contains several morphemes that correspond to different grammatical categories: *V_VM_VGN* (plural) + *DM_DMD* + *RP_RPD* + *RP_RPD* + *V_VM_VF*.

In cases like this, where several functional and grammatical markers are fused into a single token, assigning a single Part-of-Speech (PoS) tag to the combined form becomes a significant challenge.

2.3.4.6 Compound Word

Two or more words may combine to form a single lexical unit, where the resulting token conveys a unified meaning. For example, *ফিশাৰা* */pʰisabə/* ‘weaver’ and *লাইনিংশং* */lainiŋsəŋ/* ‘prayer hall’ function as single semantic units despite being historically or structurally composed of multiple elements. In such cases, the forms are treated as lexicalized compounds and are assigned an appropriate Part-of-Speech (PoS) tag based on their overall grammatical function.

2.3.4.7 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 problem for the higher level of processing.

2.4 SUMMARY OF LDC-IL MANIPURI POS DATA

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

Dataset Summary	Aggregate Count
Total PoS Tags	543469
Total Sentences	44619
Total Word Count	449451
Total File Count	1012

Table 2-2: Manipuri PoS Status

Sl. No.	PoS Category	PoS Tag	Tag Count
1	Common Noun	N_NN	139009
2	Proper Noun	N_NNP	33796
3	Nloc Noun	N_NST	13301
4	Personal Pronoun	PR_PRP	27750
5	Reflexive Pronoun	PR_PRF	1203
6	Reciprocal Pronoun	PR_PRC	103
7	Wh-word Pronoun	PR_PRQ	3827
8	Indefinite Pronoun	PR_PRI	245
9	Deictic Demonstrative	DM_DMD	11835
10	Wh-word Demonstrative	DM_DMQ	576
11	Indefinite Demonstrative	DM_DMI	112
12	Finite Verb	V_VM_VF	54829
13	Non-finite Verb	V_VM_VNF	18159
14	Gerund	V_VM_VNG	70813
15	Adjective	JJ	4220
16	Adverb	RB	26215
17	Postposition	PSP	1244
18	Co-ordinator	CC_CCD	3443
19	Subordinator	CC_CCS	4625
20	Quotative	CC_CCS_UT	5
21	Particles	RP_RPD	4279
22	Interjection	RP_INJ	615
23	Intensifier	RP_INTF	2603
24	Quantifiers	QT_QTF	7171
25	Cardinals	QT_QTC	12645
26	Ordinals	QT_QTO	594
27	Foreign word	RD_RDF	2471
28	Symbol	RD_SYM	184
29	Punctuation	RD_PUNC	93834
30	Unknown	RD_UNK	1219

31	Echowords	RD_ECH	2544
----	-----------	--------	------

Table 2-3: Distinct Tag Count of Manipuri