
অসমীয়া পদ এন'টেটেড কৰ্পাচ



ASSAMESE
PARTS OF SPEECH
ANNOTATED CORPUS

ASSAMESE 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:

Syeda Mustafiza Tamim

Dr. Narayan Choudhary

Rajesh 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

Assamese Parts of Speech Annotated Corpus

Authors: Syeda Mustafiza Tamim, Dr. Narayan Choudhary, Rajesha N., Manasa G.

e-ISBN: 978-81-69175-06-7

CIIL Publication No.: 1707

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: Syeda Mustafiza Tamim, Teena Verma

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
2	ASSAMESE POS ANNOTATED CORPUS.....	6
2.1	GRAMMATICAL PECULARITIES OF ASSAMESE	6
2.2	LDC-IL CUSTOMIZED BIS TAGSET	6
2.3	ASSAMESE POS TAGGING	10
2.3.1	<i>Guidelines</i>	10
2.3.2	<i>Tokenization</i>	11
2.3.3	<i>Selecting a Tag</i>	11
2.3.4	<i>Observation on PoS Tagging of LDC-IL Assamese Text Corpora</i>	12
2.3.4.1	Inconsistency	13
2.3.4.2	Ambiguity	14
2.3.4.3	Abbreviation	16
2.3.4.4	Compound Word.....	16
2.3.4.5	Dialect Variation.....	17
2.3.4.6	Impact of Social Dialect.....	17
2.3.4.7	Verbal Noun.....	17
2.3.4.8	Unknown Words	17
2.4	SUMMARY OF LDC-IL ASSAMESE POS DATA	18

Tables

Table 1-1: BIS Tagset	4
Table 1-2: LDC-IL PoS Status.....	5
Table 2-1: Assamese Tagset	10
Table 2-2: Assamese PoS Status.....	18
Table 2-3: Distinct Tag Count of Assamese	19

1 POS ANNOTATED CORPUS BASED ON BIS TAGSET

Rejitha K .S., Narayan 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. This tagset is published on 2021. It defines tagsets for the Indian languages Bangla, Gujarati, Hindi, Kashmiri, Konkani, Maithili, Marathi, Punjabi, Urdu, Telugu, Kannada, Malayalam and Tamil.

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 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 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 VNF	
4.2.3			Infinitive	VINF	V VAUX VINF	
4.2.4			Gerund	VNG	V VAUX VNG	
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

LDC-IL 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 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 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 ASSAMESE POS ANNOTATED CORPUS

Syeda Mustafiza Tamim

2.1 GRAMMATICAL PECULIARITIES OF ASSAMESE

Assamese is an Indo-Aryan language with rich inflectional morphology and productive derivational processes. Assamese shows several distinctive grammatical characteristics that make it structurally rich and productive. Word formation in Assamese commonly occurs through the addition of suffixes, particles and classifiers to root words. Processes such as affixation, compounding and reduplication play an important role in expanding the vocabulary of the language. Another prominent feature of Assamese grammar is the extensive use of classifiers (e.g., জন, খন, টো), which are attached to nouns when counting or specifying objects and persons. Assamese vocabulary has also been influenced over time through contact with other languages such as Sanskrit, Bengali and English, which has introduced many borrowed words into everyday usage.

From a grammatical perspective, Assamese words can broadly be divided into inflecting and non-inflecting categories. Inflecting word classes include nouns, pronouns, adjectives and verbs, which may change their forms to express grammatical relations such as number, case, tense or aspect. Verbs in Assamese also form complex constructions with auxiliary verbs, which contribute to the expression of tense and aspect. In contrast, some categories remain relatively unchanged in form. These include conjunctions, particles, postpositions and certain adverbs that primarily serve grammatical or relational functions within a sentence. Together, these features contribute to the flexibility and richness of Assamese grammatical structure and make it suitable for detailed linguistic analysis in corpus-based studies.

The annotation follows the Bureau of Indian Standards (BIS) tagset (August 2010 onwards). The Assamese tag set contains 11 major categories with detailed subcategories (total of 36 tags).

2.2 LDC-IL CUSTOMIZED BIS TAGSET

Some tags in the BIS framework are not very common in the grammatical structure of Assamese, while certain features of Assamese require clearer representation in annotation. For example, Assamese shows specific usages of particles, postpositions, and verbal forms that may not always align perfectly with the general BIS categories. Therefore, while following the BIS framework, the Assamese PoS tagset used by LDC-IL includes certain adaptations to properly represent the grammatical characteristics of the language. The customized Assamese tagset is presented in Linguistic Standards for Indian Language (IL) Corpora Annotation: PoS Categories and Labels (June 12, 2010).

Sl. No	Category	Label	Annotation	Assamese	Examples
--------	----------	-------	------------	----------	----------

					Convention* *		
	Top level	Subtype (level 1)	Subtype (level 2)				
1	Noun			N	N	বিশেষ্য	
1.1		Common		NN	N_NN	জাতিবাচক বিশেষ্য	সময়, শিক্ষক, জগত, জীৱন
1.3		Proper		NN P	N_NNP	ব্যক্তিবাচক বিশেষ্য	বৰুৱা, দাস, গুৱাহাটী, যোৰহাট
1.5		Locative		NS T	N_NST	স্থানবাচক বিশেষ্য	ইয়াত, তাত, ক'ত
2	Pronoun			PR	PR	সৰ্বনাম	
2.1		Personal		PRP	PR_PRP	ব্যক্তিবাচক সৰ্বনাম	তাই, মই, তেওঁ, তেখেত
2.2			Persona l+ suffix	*PR PP	PR_PRPP	ব্যক্তিবাচক সৰ্বনাম + পৰসৰ্গ	তাইৰ, মোৰ, তেওঁৰ, তেখেতৰ, আপোনাৰ
2.3		Reflexive		PRF	PR_PRF	আত্মবাচক সৰ্বনাম	নিজৰ, নিজকে, ইজনে-সিজনৰ
2.4		Relative		PRL	PR_PRL	সম্বন্ধয়বাচ ক সৰ্বনাম	যিজনে, যোনে, যোন
2.6		Reciprocal		PR C	PR_PRC	পাৰস্পৰিক সৰ্বনাম	নিজৰ মাজতে
2.8		Interrogative		PR Q	PR_PRQ	প্ৰশ্নবোধক সৰ্বনাম	কোন, কেতিয়া , কাৰ, কাক, কালৈ
2.9		Indefinite		PRI	PR_PRI	অনিশ্চয়বা চক সৰ্বনাম	কোনোবা, কো তিয়াবা, ক'ৰ বাত
3	Demonstrative			DM	DM	নিৰ্দেশবোধ ক	
3.1		Deictic		DM D	DM_DM D	প্ৰত্যক্ষ নিৰ্দেশক	এইখন/টো, সেইখন/টো,

							সোঁটো/খন,
3.2		Relative		DM R	DM_DM R	সমন্ধয়বাচ ক	যিখন,যোনটো /খন
3.3		Question		DM Q	DM_DM Q	প্ৰশ্নবোধক অব্যয়	কোনোবা, কালৈ,ক'লৈ
3.4		Indefinite		DM I	DM_DMI	অনিশ্চয়বা চক	কেনেবাকৈ, কোনোবা,কে তিয়াবা,
4	Verb			V	V	ক্ৰিয়া	
4.1		Main		VM	V_VM	মুখ্য ক্ৰিয়া	কৰ,ফুৰ,থা,পি ন্ধ
4.2			Finite	VF	V_VM_V F	পূৰ্ণ ক্ৰিয়া	গৈছিলো,থাই ছিলো,ফুৰিছি লো
4.3			Non- Finite	VN F	V_VM_V NF	অপূৰ্ণ ক্ৰিয়া	থাই, গৈ, উঠি, পৰি
4.5		Gerund		VN G	V_VM_V NG	ক্ৰিয়াবাচক বিশেষ্য	থাবলৈ, পহিবলৈ, যাবলৈ, থাওঁতে, যাওঁতে
4.6		Auxiliary		VA UX	VAUX	সহকাৰী ক্ৰিয়া	হৈছে, লৈছে, পাৰিলে,নোৱা ৰিলে
5	Adjective			JJ		বিশেষণ	অবুজ, বৰ, বেয়া, ভাল, ধুনীয়া, বেছি, কম
6	Adverb			RB		ক্ৰিয়া বিশেষণ	লাহে-লাহে, যল্দি, খৰকৈ, এতিয়া, তেতিয়া,

							যেতিয়া, দৌৰা- দৌৰিকৈ
7	Postposit ion			PSP		অনুসৰ্গ	বাবে, লগত, পৰা, পৰাই
8	Conjunct ion			CC	CC	সংযোজক	
8.1		Coordina tor		CC D	CC_CCD	সমন্বয়ক	আৰু, কাৰণ, কিন্তু
8.2		Subordin ator		CC S	CC_CCS		যে, গতিকে, বাবে
9	Particles			RP	RP	আনুষাংগি ক অব্যয়	
9.1		Default		RP D	RP_RPD	সাধাৰণ	অকল, বিষয়ে, মাত্ৰ, নেকি, নিশ্চয়,
9.2		Classifier		CL	RP_CL	নিৰ্দিষ্টবাচক	টা, জন, টো, খন, দাল
9.3		Interjecti on		INJ	RP_INJ	বিষ্ময়বাচক	ইছ, আই'অ, উফ
9.4		Intensifie r		INT F	RP_INTF	তীব্ৰতাবোধ ক	একেবাৰে,
9.5		Negation		NE G	RP_NEG	নঞাৰ্থক	নকৰা, নালাগে, নোযোৱা
10	Quantifi ers			QT	QT	পৰিমাণবাচ ক	
10.1		General		QT F	QT_QTF	সাধাৰণ	এক, বহুত, কেইটামান, অজস্ৰ
10.2		Cardinals		QT	QT_QTC	সংখ্যাবাচক	এমাহ, এজনী,

				C			এগৰাকী, এথোজ
10.3		Ordinals		QT O	QT_QTO	ক্ৰমবাচক	প্ৰথমবাৰ,
11	Residual s			RD	RD	অৱশিষ্ট	
11.1		Foreign words		RD F	RD_RDF	বিদেশী শব্দ	
11.2		Symbol		SY M	RD_SYM	প্ৰতীক	@,\$
11.3		Punctuati on		PU NC	RD_PUN C	যতি চিহ্ন	, !
11.4		Unknown		UN K	RD_UNK	অজ্ঞাত	যিটো, যিবোৰ
11.5		Echo words		EC H	RD_ECH	ধ্বনাত্মক শব্দ	চাহ-তাহ, কৰি-মেলি
11.6		Reduplic ation		*R DP	RD_RDP	পুনৰাবৃত্তি	বহি-বহি, ৰৈ- ৰৈ,

Table 2-1: Assamese 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 ASSAMESE POS TAGGING

LDC-IL developed a customized Assamese PoS tagset with the necessary adaptations to the BIS tagset to suit the linguistic structure of Assamese. Since Assamese shows rich use of suffixes, particles, and postpositions, certain preprocessing steps such as proper word identification and segmentation are carried out before tagging so that the data fits well within the tagset framework. To improve tagging accuracy, context-based tag assignment is also applied to maintain consistency across the corpus.

2.3.1 Guidelines

- Assamese PoS Tagging is done based on BIS tagset with customization to capture specific grammatical and morphological features.
- Assamese is an agglutinative language, pre-processing has been done to fit the words according to the tags.

- Complex word was segmented to assign correct PoS categories.
- 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 was done by using an 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 dividing a sentence into its individual lexical units or tokens. Accurate tokenization is essential for effective PoS tagging, as it allows each word or unit to be assigned the correct morpho-syntactic category. While PoS tagging provides grammatical information about words, it does not capture semantic or discourse-level meaning. In Assamese, tokenization is particularly important due to features such as classifier attachment (জন, টো, খন, গৰাকী), compound verbs (খাই পেলাইছে, কৰি আছে, লৈ আহিল), postpositional constructions, and reduplication forms like বহি-বহি and বৈ-বৈ. For example, the sentence

মই ভাত খাই পেলাই আহিলো।

Tokenized as:

মই | ভাত | খাই | পেলাই | আহিলো

Tagged as:

মই/PR_PRP

ভাত/N_NN

খাই/V_VM_VNF

পেলাই/VAUX

আহিলো/V_VM_VF

Through tokenization, each lexical unit is assigned an appropriate PoS tag.

2.3.3 Selecting a Tag

In Assamese, a word has two important elements: its form (how it is built) and its function (how it is used). A balance is maintained between the two to ensure the tagging is consistent.

The category of a word is decided by looking at its grammar and its position in a sentence. While a word's category is often clear from its ending (suffix), it can sometimes act as a different

category depending on the context. We use syntactic clues to solve this. For example, some pronouns act as demonstratives if they appear right before a noun.

Example 1: Pronoun vs. Demonstrative

The word এই/ei/ changes its category based on whether it stands alone or modifies a noun.

এইটো মোৰ কলম।

— "This is my pen."

Tag: PR_PRP (Personal Pronoun)

Reason: Here, 'এইটো' acts as the subject of the sentence. It represents the object directly without being followed by a noun.

এই কলমটো মোৰ।

— "This pen is mine."

Tag: DM_DMD (Deictic Demonstrative)

Reason: Here, 'এই' modifies the noun 'কলমটো' (pen). It functions as a pointer (demonstrative) identifying which pen is being discussed.

Example 2: Noun vs. Adjective

Some words change their role based on the neighboring words. Consider the word পাচত /pasɔt/:

ঘৰৰ পাচত এখন বাৰী আছে।

— "There is a garden at the back of the house."

Tag: N_NST (Spatio-Temporal Noun)

Reason: It refers to a physical space or location (the backyard).

সি পাচৰ দুৱাৰেদি ওলাই গ'ল।

— "He went out through the back door."

Tag: JJ (Adjective)

Reason: Here, it describes the noun 'দুৱাৰ' (door), acting as a qualifier to specify which door was used.

By looking at the context, the annotator decides the correct tag for such ambiguous words.

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

PoS tagging assigns the lexical category of each token while also reflecting limited contextual information. In Assamese annotation, both the morphological form and the functional role of a word are considered to ensure systematic and consistent tagging. When formal and functional

properties are properly represented, the PoS layer becomes highly useful for advanced linguistic processing and NLP applications.

The annotated data can generally be divided into three groups:

- Correctly tagged words
- Ambiguously tagged words
- Unknown or untagged words

In the LDC-IL Assamese corpus, most nouns (N_NN, N_NNP), pronouns (PR_PRP, PR_PRF), adjectives (JJ), finite verbs (V_VM_VF), non-finite verbs (V_VM_VNF), adverbs (RB), and other major categories are accurately tagged following BIS guidelines.

However, lexical ambiguity is a natural characteristic of Assamese, as many words carry multiple senses and may function under different grammatical categories depending on context. The effectiveness of a PoS tagger is largely determined by its ability to resolve such ambiguity. Contextual meaning, idiomatic usage, compound constructions, and discourse functions often increase the level of ambiguity.

Certain words remain untagged or are marked as RD_UNK due to insufficient lexical information, limitations in tagset coverage, foreign words written in non-Assamese script, unfamiliar transliterated forms, dialectal variations, or orthographic inconsistencies. These factors present challenges in maintaining complete tagging consistency across the corpus.

2.3.4.1 Inconsistency

In an Assamese PoS-tagged corpus, inconsistencies frequently arise due to the morpho-syntactic characteristics of the language. Even when expressions convey similar meanings, they may appear in different structural forms. One major source of inconsistency in Assamese is variation in compound formation and tokenization. Certain compound nouns may occur either as a single orthographic unit or as separate tokens, which directly affects their PoS tagging.

For example, consider the following forms:

- ঘৰদুৱাৰ\N_NN and ঘৰ\N_NN দুৱাৰ\N_NN
- ৰাজপথ\N_NN and ৰাজ\N_NN পথ\N_NN
- জনসাধাৰণ\N_NN and জন\N_NN সাধাৰণJJ

In these cases, the first form is treated as a single lexical unit and is assigned one PoS tag (N_NN), since it functions as a compound noun with a unified meaning. For example, ঘৰদুৱাৰ refers collectively to the household, ৰাজপথ denotes a main/public road, and জনসাধাৰণ refers to the general public as a single conceptual entity.

However, when these expressions are tokenized separately (e.g., ঘৰ + দুৱাৰ, ৰাজ + পথ, জন + সাধাৰণ), each component receives an individual PoS tag based on its independent lexical

category. In some cases, this may even change the grammatical interpretation—for instance, জন (N_NN) + সাধাৰণ (JJ) represents a noun followed by an adjective, rather than a fixed compound noun.

Thus, although the semantic interpretation may remain closely related, variation in tokenisation leads to differences in PoS assignment. This demonstrates how orthographic and morphological flexibility in Assamese contributes to tagging inconsistency in annotated corpora

2.3.4.2 Ambiguity

Natural language is inherently ambiguous. Ambiguity arises when a single word or a sequence of words is associated with more than one meaning in the language system (Chierchia and McConnell-Ginet 1990:32). Generally, lexical ambiguity can be divided into two types: categorical (or syntactic) ambiguity and semantic lexical ambiguity. Categorical ambiguity occurs when the grammatical category of a word is uncertain, for example whether it functions as a noun, verb, adjective, or particle. Semantic lexical ambiguity arises when a word carries multiple meanings.

In Assamese, ambiguity in PoS tag assignment often occurs because certain lexical items can function as different grammatical categories depending on the syntactic context. However, the surrounding context usually helps resolve this ambiguity. The following examples illustrate how the same token may function as different parts of speech in Assamese.

N_NN & V_VM

In Assamese
তেওঁৰ খবৰ\N_NNটোৱে মোক দুখ দিলে।
মই তাৰ ভালকৈ খবৰ\V_VM ল'ম।
In IPA
t̪eɔ̌ɔ̌ ʰɔbɔɔɪ\N_NNtue mok dukh dile
moi taɪ bʰalɔkoi ʰɔbɔɔɪ\V_VM lom

N_NN & RP_NEG

In Assamese
এই কামটোত চৰকাৰৰ নিষেধ\N_NN আছে।
ইমান ৰাতি অকলে বাহিৰলৈ যোৱাটো নিষেধ\RP_NEG।
In IPA
ei kamtuɔ̌ sɔ̌kaɔ̌ɔ̌ nixedʰ\N_NN ase
iman ɾati ɔ̌kole bahiɔ̌loi jowatu nixedʰ\RP_NEG

JJ & RB

In Assamese
সি এজন ভালJJ ল'ৰা।
সি ভালRB গয়।
In IPA
xi ezɔn bʰalJJ lɔɾa।
xi bʰalRB gaj।

N_NN & N_NNP

In Assamese
পকাJJ আম মিঠা হয়।
আমবোৰ পকাV_VM।
In IPA
pɔkaJJ am mitʰa hoj।
ambuɾ pɔkaV_VM।

N_NN & RP_INJ

In Assamese
মোৰ আইN_NN ঘৰলৈ গৈছে।
অ' আইRP_INJ! ইমান ধুনীয়া দৃশ্য।
In IPA
moɾ aiN_NN gʱɔ.ɾɔloi goise।
ɔ aiRP_INJ! iman dʱunia dɾisjo।

QT_QTF – JJ

In Assamese
কমQT_QTF মানুহ উপস্থিত আছিল।

তেওঁৰ ধৈৰ্য্য কমJJ।
In IPA
kəm manuh upɔst ^h it asil।
teur d ^h ɔi.jɔ kəmJJ।

2.3.4.3 Abbreviation

Abbreviations are shortened forms of words or phrases that are commonly used in both written and spoken language. In Assamese as well, many words—especially names of degrees, institutions, titles, and organizations—are frequently written in abbreviated forms. The meaning of the abbreviation is usually understood from the context in which it appears.

Some examples of abbreviations used in Assamese are:

ড° (ডক্টৰ), শ্ৰী., বি.এ., এম.এ., ডি.চি., এছ.পি. etc.

Assamese	IPA
ড° ভূপেন হাজৰিকা এজন মহান শিল্পী আছিল।	dɔ bupen hadʒɔrika ɛdʒɔn mohan xilpi asil।
তেওঁ গুৱাহাটী বিশ্ববিদ্যালয়ৰ পৰা বি.এ. পাছ কৰিছে।	tɛu guwahati biswɔbiddalɔjɔ pɔɔ bi.e pas kɔɔise।
জিলাখনৰ ডি.চি. সভাত উপস্থিত আছিল।	zilak ^h ɔnɔ di.si. xɔb ^h at upɔst ^h it asil।

2.3.4.4 Compound Word

Compounding is the phenomenon wherein two or more morphemes combine to form a single lexical unit that conveys a complete meaning. In Assamese, compound words are frequently used in both written and spoken language. Examples include: শাক-পাচলি, কিতাপ পত্ৰ, ঘৰদুৱাৰ etc. These forms express a combined meaning and function together as one unit in a sentence

- In the case of hyphenated compounds, it is advisable to treat them as a single token and assign the PoS tag according to their overall grammatical category. For example: শাক-পাচলি\N_NN
- In the case of hyphenated compounds, it is advisable to treat them as a single token and assign the PoS tag according to their overall grammatical category. For example: কিতাপ\N_NN পত্ৰ\N_NN

- Conversely, when compound expressions appear with a space, each component should be tagged separately according to its own category. For example: ঘৰদুৱাৰ\N_NN

2.3.4.5 Dialect Variation

Assamese exhibits considerable dialectal variation across regions, particularly among varieties such as Eastern Assamese (Standard), Kamrupi, and Goalpariya. These dialects differ in phonology, vocabulary, and certain grammatical features, although they remain largely mutually intelligible. Such variation also affects the composition and annotation of linguistic corpora, as speakers often incorporate regional lexical items into their language. Consequently, dialectal vocabulary frequently appears alongside standard forms in Assamese corpora. For example, the Standard Assamese form গ'ল corresponds to গেছী in Kamrupi and হইল in Goalpariya. Similarly, কেতিয়া (Std.) appears as কিতৈ in Kamrupi and কোনবেলা in Goalpariya; ল'ৰা-ছোৱালী (Std.) corresponds to ছলি-পলি in Xualkusi and ছাৱা in Goalpariya; and আলহী (Std.) corresponds to অতিথি in Kamrupi and গুষ্টি in Goalpariya. These examples illustrate how regional vocabulary from different Assamese dialects may enter linguistic corpora and influence the annotation process.

2.3.4.6 Impact of Social Dialect

Social dialect variation in Assamese is influenced by education, profession, and exposure to other languages. In urban areas, many speakers mix Assamese with English, and several English words appear in Assamese script in everyday communication. Words such as বাছ (bus), টিকেট (ticket), ডাক্তৰ (doctor), and মোবাইল (mobile) are commonly used in daily speech and are treated as regular Assamese words in many contexts. Additionally, due to regional contact, Assamese also includes words borrowed from neighbouring languages. For instance, words like মজদুৰ (labourer) and বজাৰ (market) have entered Assamese through contact with Hindi/Urdu and are used in certain contexts. These variations reflect how social interaction and language contact shape vocabulary within different Assamese-speaking communities.

2.3.4.7 Verbal Noun

The word which denotes the result of action is tagged as a verbal noun category; other types of verbal nouns are tagged as nouns.

2.3.4.8 Unknown Words

There may be some words in the corpus that cannot be handled by the tagset. Such words are considered unknown and can pose challenges for higher levels of processing. Unknown words in Assamese text corpus often arise due to spelling variations, code-mixing, presence of dialectal forms etc. For example, ছিছম\RD_UNK, বনি\RD_UNK etc.

2.4 SUMMARY OF LDC-IL ASSAMESE POS DATA

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

Dataset Summary	Aggregate Count
Total PoS Tags	185522
Total Sentences	13467
Total Word Count	159138
Total File Count	404

Table 2-2: Assamese PoS Status

Sl. No.	PoS Category	PoS Tag	Tag Count
1	Common Noun	N_NN	59506
2	Proper Noun	N_NNP	11303
3	Nloc Noun	N_NST	3065
4	Personal Pronoun	PR_PRP	6556
5	Reflexive Pronoun	PR_PRF	480
6	Relative Pronoun	PR_PRL	404
7	Reciprocal Pronoun	PR_PRC	33
8	Wh-word Pronoun	PR_PRQ	786
9	Pronoun Interrogative	PR_PRI	347
10	Deictic Demonstrative	DM_DMD	3681
11	Wh-word Demonstrative	DM_DMQ	0
12	Demonstrative Relative	DM_DMR	319
13	Demonstrative Marker Indefinite	DM_DMI	210
14	Verb Main Verb	V_VM	21760
15	Finite Verb	V_VM_VF	2378
16	Non-finite Verb	V_VM_VNF	1197
17	Infinitive Verb	V_VM_VINF	144
18	Gerund	V_VM_VNG	0
19	Verbal Noun	V_VM_VNV	0
20	Verb Auxiliary	V_VAUX	6177
21	Adjective	JJ	15169
22	Adverb	RB	3979
23	Postposition	PSP	3260
24	Co-ordinator	CC_CCD	4878
25	Subordinator	CC_CCS	1508
26	Quotative	CC_CCS_UT	0
27	Particles	RP_RPD	2650
28	Interjection	RP_INJ	129

29	Intensifier	RP_INTF	658
30	Negation	RP_NEG	6
31	Quantifiers	QT_QTF	3458
32	Cardinals	QT_QTC	2356
33	Ordinals	QT_QTO	408
34	Foreign word	RD_RDF	1664
35	Symbol	RD_SYM	543
36	Punctuation	RD_PUNC	25841
37	Unknown	RD_UNK	587
38	Echowords	RD_ECH	82

Table 2-3: Distinct Tag Count of Assamese