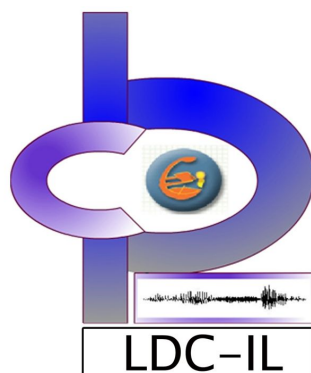


बर' सोदोबमा बुंफोरनाय रावथुन

BODO
PARTS OF SPEECH
ANNOTATED CORPUS

BODO 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:

*Hemlata Daimary
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

Bodo Parts of Speech Annotated Corpus

Authors: Hemlata Daimary, Dr. Narayan Choudhary, Rajesha N., Manasa G.

e-ISBN: 978-81-69175-92-0

CIIL Publication No.: 1705

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: Hemlata Daimary

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	Bodo PoS Annotated Corpus.....	6
2.1	Grammatical peculiarities of Bodo	6
2.2	LDC-IL Customized BIS Tagset.....	6
2.3	Bodo PoS Tagging	8
2.4	Summary of LDC-IL Bodo PoS Data	13

Tables

Table 1-1: BIS Tagset	4
Table 1-2: LDC-IL PoS Status.....	5
Table 2-1: Bodo Tagset.....	8
Table 2-2: Bodo PoS Status	13
Table 2-3: Distinct Tag Count of Bodo	14

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 annotations 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 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 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

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 Mishra 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 BODO POS ANNOTATED CORPUS

Hemlata Daimary

2.1 GRAMMATICAL PECULIARITIES OF BODO

Bodo is a morphologically rich language characterized by extensive word formation processes. Words in Bodo 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 Bodo 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

Some tags in the BIS framework, such as Finite, Non-finite, Infinitive verbs, Verbal Noun and Quotative Conjunction are rare or absent in grammatical structure of Bodo. Therefore, LDC-IL customized the tagset to accommodate the specific needs of each language. The Bodo customized tagset is as follows. 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)	Subtype (level 2)			
1	Noun			N	N	
1.1		Common		NN	N__NN	ओखाम, लामा
1.2		Proper		NNP	N__NNP	राजेन बर'
1.3		Nloc		NST	N__NST	सा, सिं
2	Pronoun			PR	PR	
2.1		Personal		PRP	PR__PRP	आं, नों, बि
2.2		Reflexive		PRF	PR__PRF	गाव

2.3		Relative		PRL	PR_PRL	जाय
2.4		Reciprocal		PRC	PR_PRC	गावजों गाव
2.5		Wh-word		PRQ	PR_PRQ	सोर, माब्ला, बबेयाव
2.6		Indefinite		PRI	PR_PRI	सोरबा
3	Demonstrative			DM	DM	
3.1		Deictic		DMD	DM__DMD	बे, बै
3.2		Relative		DMR	DM__DMR	जाय
3.3		Wh-word		DMQ	DM__DMQ	सोरनि, सोर
3.4		Indefinite		DM	DM__DMI	खायसे
4	Verb			V	V	
4.1		Main		VM	V__VM	फै, फैबाय, खालामबाय
4.1.1			Gerund	VNG	V__VM__VNG	जानाय, मावनाय
5	Adjective			JJ		मोजां, गोखा, गोखै
6	Adverb			RB		लासैयै, गोखैयै
7	Postposition			PSP		थाखाय, दारै
8	Conjunction			CC	CC	
8.1		Co-ordinator		CCD	CC__CCD	आरो, एबा
8.2		Subordinator		CCS	CC__CCS	बेनिखायनो
9	Particles			RP	RP	
9.1		Default		RPD	RP__RPD	खोमा, ना
9.2		Classifier		CL	RP__CL	सा, गां
9.3		Interjection		INJ	RP__INJ	आरे, आय!
9.4		Intensifier		INTF	RP__INTF	जोबोद
9.5		Negation		NEG	RP__NEG	दा
10	Quantifiers			QT	QT	
10.1		General		QTF	QT__QTF	इसे, गोबां
10.2		Cardinals		QTC	QT__QTC	से, नै, थाम
10.3		Ordinals		QTO	QT__QTO	सेथि, नैथि
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	दै-सै, हाथर-साथर

Table 2-1: Bodo 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 BODO POS TAGGING

Bodo PoS tagging is performed using the BIS tagset, with necessary adaptations from LDC-IL guidelines to suit Bodo's linguistic structure. Since Bodo is an agglutinative language, preprocessing steps such as word segmentation is carried out before tagging to fit in tagset framework. To improve tagging accuracy context-based tag assignment is carried out to ensure the consistency across the corpus.

2.3.1 Annotation Guidelines

- Bodo PoS Tagging is done based on BIS tagset with customization to capture specific grammatical and morphological features.
- Bodo is an agglutinative language. Words are formed by adding multiple affixes to a root, which can result in very long and detailed words.
- 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 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 important because accuracy of the PoS tagged data is depends on the input words. PoS tagging is incorporating the morpho-syntactic information to the lexical item. It does not insert any semantic, discourse or pragmatic information. In agglutinative languages like Bodo, Manipuri, Malayalam, Tamil, Telugu and Kannada two or three words are combined together to form a complex word. Therefore, tokenization is needed because the accuracy of PoS tagged data depends on the context of the words. These words are from different categories or

from same category. Such instances it is needed to pre-process the word to get desired category. Otherwise that complex word cannot be tagged with a single PoS tag.

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 addressed by looking at syntactic clues. For example some pronouns function as demonstratives if they occur before a noun.

बैयो ननि बाइजोआव गसनानै दङ। (He is standing outside the house.)

बै बिजाबखौ आरांगा सायाव दोन। (Keep that book on the table.)

In the first sentence, the token 'बैयो' serves as a personal pronoun. However, in the second instance, it works as a demonstrative because it modifies a common noun (बिजाबखौ). Similarly, indefinite pronouns such as जायखौब्लाबो and जायखिजाया also function as demonstratives when preceding a noun.

2.3.4 Observation on PoS Tagging of LDC-IL Bodo 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

The same word appeared differently in the corpus. For e.g. in Bodo, some words may appear as compound words in certain contexts, while the same words can also occur separately. Semantically both words give the same meaning but it will provide different morpho-syntactic

information. Fish name, while often tagged as common noun, can sometimes function as noun, verb as well. मासलांखार /maslaŋk^{har}/ and मास/mas/ लां/laŋ/ खार/k^{har}/

The following examples shows the same elements can appear separately or as a single combined token with their respective PoS tags.

मास\N_NN लां\V_VM खार\V_VM and मासलांखार\N_NN

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 Bodo PoS tagging. This was addressed with careful attention in LDC-IL Bodo PoS tagged text corpus.

V_VM & N_NN

In Bodo			
जिखौ\N_NN	दायो\N_NST	सु\V_VM	I\RD_PUNC
आंखौ\PR_PRP	सु\N_NN	हाबदौ\V_VM	I\RD_PUNC
In IPA			
zik ^h uu\N_NN	dajw\N_NST	su\V_VM	IRD_PUNC
aŋk ^h uu\N_NN	su\N_NN	habd ^u ŋ\V_VM	IRD_PUNC

JJ & RB

In Bodo			
बेयो\P_PRP	मोजां\JJ	खोलोम\N_NN	I\RD_PUNC
बेयो\P_PRP	जोबोद\RP_INTF	मोजां\RB	मोसायो\V_VM I\RD_PUNC
In IPA			
bejw\P_PRP	muzaŋ\JJ	k ^h u ^l wu ^m \N_NN	I\RD_PUNC
bejw\P_PRP	zwbwd\RP_INTF	muzaŋ\RB	musajw\V_VM I\RD_PUNC

N_NN & N_NNP

In Bodo	
सिखिरिया\N_NNP	बिबार\N_NN खादोंमोन\V_VM I\RD_PUNC
सिखिरिया\N_NN	बिबाराव\N_NN बिदै\N_NN सोबदोंमोन\V_VM I\RD_PUNC
In IPA	
sikʰirja\N_NNP	bibar\N_NN kʰadwɨmwɨn\V_VM I\RD_PUNC
sikʰirja\N_NN	bibaraw\N_NN bidwi\N_NN subdwɨmwɨn\V_VM I\RD_PUNC

N_NN & RP_INJ

In Bodo	
हाबाब\RP_INJ	सान्दुङा\N_NN जोबोद\RP_INTF गोसाथार\V_VM I\RD_PUNC
बिनि\PR_PRP	हाबाबखौ\N_NN खोनानानै\ V_VM बिमाया\N_NN मिनिदोंमोन\V_VM I\RD_PUNC
In IPA	
habab\RP_INJ	sandunja\N_NN zubwɨd\RP_INTF gwsatʰar\V_VM I\RD_PUNC
bini\PR_PRP	hababkʰu\N_NN kʰunananwi\ V_VM bimaja\N_NN minidwɨmwɨn\V_VM I\RD_PUNC

QT_QTO & N_NN

In Bodo	
नोंथाङा\PR_PRP	नैथि\QT_QTO सिरियाव\N_NN जिराय\V_VM I\RD_PUNC
बैयो\ PR_PRP	दिनै\N_NST गाइखेर\N_NN लोंदों\V_VM आरो\CC_CCD नैथिया\N_NN गांसो\N_NN जादोंमोन\V_VM I\RD_PUNC
In IPA	
nɨnʰtʰaŋa\PR_PRP	nwiθi\QT_QTO sirijaw\N_NN ziraj\V_VM I\RD_PUNC
bwiɟw\ PR_PRP	dinwi\N_NST gaikʰer\N_NN lɨŋdwɨ\V_VM arw\CC_CCD nwiθija\N_NN gaŋsw\N_NN zadwɨmwɨn\V_VM I\RD_PUNC

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 Bodo are- ना.गो., था.हि., आइ.पि.एल etc.

2.3.4.4 Multiword

A sequence of words can act as a single unit, making it difficult to assign a single category. Bodo is an agglutinative language, so multiple words are often joined together to form a single token. In Bodo, any category can attach to another category. For eg: जॉफोरनिसिमबो PR_PRP+PSP+PSP+N_NST+VM etc. In such cases, assigning an appropriate PoS tag to the combined word is challenging. Therefore, preprocessing is necessary to identify the correct category.

2.3.4.5 Compound Word

Compounding is the phenomenon wherein two or more morphemes combine to form a single unit and the whole token gives a meaning. For example, 'सुख-दुख', 'थैनाय-थानाय', 'थार-नंखाय', 'मोजा-गात्रि' etc. These words are taken as whole and assigned the respective appropriate tag. In the case of hyphenated compounds, it is advisable to treat them as a single token and thereby tag them using the main category. Conversely, in the case of non-hyphenated compounds such as 'थाखा फैसा', they should be tagged separately with their distinct categories. However, if they occur without any space, they can be considered to be a single token, for instance, 'फिसाजो'।

2.3.4.6 Dialect Variation

It is possible for a language to have regional dialectal variations. Similarly, the Bodo language also shows dialectal variations depending on location. If an annotator does not have proper knowledge of dialectal variation or cannot distinguish it correctly, it may lead to misinterpretation of meaning, or the annotator may simply tag it as “unknown” and leave it as such.

For example, in some regions where Bodo is spoken, 'Rice' is referred to as 'औखाम,' while in other regions it is called 'खाम' Likewise, in some places 'chili' is called 'बानलु,' whereas in other places it is known as 'आलौ' If an annotator is not familiar with such regional vocabulary, they may make errors during annotation.

2.3.4.7 Impact of Social Dialect

Dialect variation in Bodo is often shaped by social factors such as age, education, urban/rural setting, cultural exposure etc. Speakers from urban backgrounds often code-mix Bodo with English due to which a large amount of English words have entered Bodo corpus. Words such as 'प्लिज', 'केन्सेल', 'हेल', 'स्ल' etc. have become very common in day-to-day use. Similarly, instances can be found in Bodo corpus where expressions from other languages such as Assamese, Urdu are also present. For example, 'माति/माथि गाल्दाब = माथि-गाल्दाब' (Assamese) and 'लाइ जाम = लाइजाम' are in Assamese and Urdu language but written using Devanagari. Words from other languages such as साइखेल, सुफिन etc. can also be observed.

2.3.4.8 Classifier

In Bodo, classifiers are words used with numbers to count certain nouns, especially when referring to people, objects, or items in formal or careful speech. The most common general classifier is सा (sa)', used for counting human (e.g., सा-से मानसि – one man). Others are, e.g. मा, गां, फां, दां, बार etc.. These classifiers have been tagged according to their presence.

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 problem for the higher level of processing.

2.4 SUMMARY OF LDC-IL BODO POS DATA

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

Dataset Summary	Aggregate Count
Total PoS Tags	335987
Total Sentences	23824
Total Word Count	287722
Total File Count	377

Table 2-2: Bodo PoS Status

Sl. No.	PoS Category	PoS Tag	Tag Count
1	Common Noun	N_NN	105811
2	Proper Noun	N_NNP	22329
3	Nloc Noun	N_NST	9298
4	Personal Pronoun	PR_PRP	8736
5	Reflexive Pronoun	PR_PRF	1246
6	Relative Pronoun	PR_PRL	198
7	Reciprocal Pronoun	PR_PRC	61
8	Wh-word Pronoun	PR_PRQ	596
9	Indefinite Pronoun	PR_PRI	550
10	Deictic Demonstrative	DM_DMD	6887
11	Relative Demonstrative	DM_DMR	605
12	Wh-word Demonstrative	DM_DMQ	276
13	Indefinite Demonstrative	DMI_DMI	401
14	Main Verb	V_VM	45423
15	Gerund	V_VM_VNG	19987
16	Adjective	JJ	19000
17	Adverb	RB	6398

18	Postposition	PSP	3807
19	Co-ordinator	CC_CCD	10285
20	Subordinator	CC_CCS	1620
21	Particles	RP_RPD	2593
22	Classifier	RP_CL	291
23	Interjection	RP_INJ	431
24	Intensifier	RP_INTF	838
25	Negation	RP_NEG	18
26	Quantifiers	QT_QTF	4973
27	Cardinals	QT_QTC	8367
28	Ordinals	QT_QTO	374
29	Foreign word	RD_RDF	4934
30	Symbol	RD_SYM	592
31	Punctuation	RD_PUNC	47673
32	Unknown	RD_UNK	1263
33	Echowords	RD_ECH	126

Table 2-3: Distinct Tag Count of Bodo