

বাংলা পদ-চিহ্নিত কর্পাস



Bengali Parts of Speech
Annotated Corpus

BENGALI 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:
Sonali Sutradhar,
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

Bengali Parts of Speech Annotated Corpus

Authors: Sonali Sutradhar, Dr. Narayan Choudhary, Rajesha N., Manasa G.

e-ISBN: 978-81-69175-00-5

CIIL Publication No.: 1706

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: Sonali Sutradhar

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

Tables

Table 1-1: BIS Tagset	4
Table 1-2: LDC-IL PoS Status	5
Table 2-1: Bengali Tagset	8
Table 2-2: Bengali PoS Status	13
Table 2-3: Distinct Tag Count of Bengali	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 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 and 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
	Top level	Subtype (level 1)	Subtype (level 2)			
1	Noun			N	1	Noun
1.1		Common		NN	1.1	
1.2		Proper		NNP	1.2	
1.3		Verbal		NNV	1.3	
1.4		Nloc		NST	1.4	

2	Pronoun			PR	2	Pronoun
2.1		Personal		PRP	2.1	
2.2		Reflexive		PRF	2.2	
2.3		Relative		PRL	2.3	
2.4		Reciprocal		PRC	2.4	
2.5		Wh-word		PRQ	2.5	
2.6		INDEFINITE		PRI	2.6	
3	Demonstrative			DM	3	Demonstrative
3.1		Deictic		DMD	3.1	
3.2		Relative		DMR	3.2	
3.3		Wh-word		DMQ	3.3	
3.4		Indefinite		DMI	3.4	
4	Verb			V	4	Verb
4.1		Main		VM	4.1	
4.1.1			Finite	VF	4.1.1	
4.1.2			Non-finite	VNF	4.1.2	
4.1.3			Infinitive	VINF	4.1.3	
4.1.4			Gerund	VNG	4.1.4	
4.2		Verbal		VN	4.2	
4.2		Auxiliary		VAUX	4.2	
4.2.1			Finite	VAUX	4.2.1	
4.2.2			Non-finite	VNF	4.2.2	
4.2.3			Infinitive	VINF	4.2.3	
4.2.4			Gerund	VNG	4.2.4	
4.2.5			Participle noun	VNP	4.2.5	
5	Adjective			JJ	5	Adjective
6	Adverb			RB	6	Adverb
7	Postposition			PSP	7	Postposition
8	Conjunction			CC	8	Conjunction
8.1		Co-ordinator		CCD	8.1	
8.2		Subordinator		CCS	8.2	
8.2.1			Quotative	UT	8.2.1	
9	Particles			RP	9	Particles

9.1		Default		RPD	9.1	
9.2		Classifier		CL	9.2	
9.3		Interjection		INJ	9.3	
9.4		Intensifier		INTF	9.4	
9.5		Negation		NEG	9.5	
10	Quantifiers			QT	10	Quantifiers
10.1		General		QTF	10.1	
10.2		Cardinals		QTC	10.2	
10.3		Ordinals		QTO	10.3	
11	Residuals			RD	11	Residuals
11.1		Foreign word		RDF	11.1	
11.2		Symbol		SYM	11.2	
11.3		Punctuation		PUNC	11.3	
11.4		Unknown		UNK	11.4	
11.5		Echowords		ECH	11.5	

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 own 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 BENGALI POS ANNOTATED CORPUS

Sonali Sutradhar

2.1 GRAMMATICAL PECULIARITIES OF BENGALI

Bengali is a morphologically rich language characterized by extensive word formation processes. Words in Bengali commonly undergo inflection, derivation, and compounding. Because of these productive morphological operations, a single root or lemma can generate a wide range of word forms. Bengali grammar is characterized by a Subject-Object-Verb (SOV) structure, lack of grammatical gender for nouns, and extensive use of postpositions rather than prepositions. It features honorific distinctions in pronouns and verbs, alongside a "classifier" system for counting nouns. Verbs are highly inflected for person, tense, and social status rather than number.

Grammatical productivity and morphological complexity pose significant challenges from the perspective of Part-of-Speech tagging. Multiple grammatical categories such as case, number, gender, tense, aspect, mood, person, and voice may be simultaneously labelled within a single word. As a result, PoS tagging in Bengali require 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. To accommodate the specific features of Bengali, the BIS tagset needs to be modified.

2.2 LDC-IL CUSTOMIZED BIS TAGSET

Some tags in the BIS framework, such as indefinite pronouns and classifiers, are very common feature in grammatical structure of Bengali. Therefore, LDC-IL customized the tagset to accommodate the specific needs of each language. The Bengali 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	কলম
1.2		Proper		NNP	N__NNP	কলকাতা
1.3		Nloc		NST	N__NST	উপরে

2	Pronoun			PR	PR	
2.1		Personal		PRP	PR__PRP	আমরা
2.2		Reflexive		PRF	PR__PRF	নিজের
2.3		Relative		PRL	PR__PRL	যে
2.4		Reciprocal		PRC	PR__PRC	পরস্পর
2.5		Wh-word		PRQ	PR__PRQ	কে
2.6		Indefinite		PRI	PR__PRI	কেউ
3	Demonstrative			DM	DM	
3.1		Deictic		DMD	DM__DMD	এই
3.2		Relative		DMR	DM__DMR	যে
3.3		Wh-word		DMQ	DM__DMQ	কোন
3.4		Indefinite		DMI	DM__DMI	কোনো
4	Verb			V	V	
4.1		Main		VM	V__VM	করে
4.3		Auxiliary		VAUX	V__VAUX	ছিল
5	Adjective			JJ		সুন্দর
6	Adverb			RB		তাড়াতাড়ি
7	Postposition			PSP		দিয়ে
8	Conjunction			CC	CC	
8.1		Co-ordinator		CCD	CC__CCD	আর
8.2		Subordinator		CCS	CC__CCS	কিন্তু
9	Particles			RP	RP	
9.1		Default		RPD	RP__RPD	তো
9.2		Classifier		CL	RP__CL	টা
9.3		Interjection		INJ	RP__INJ	হায়
9.4		Intensifier		INTF	RP__INTF	খুব
9.5		Negation		NEG	RP__NEG	না
10	Quantifiers			QT	QT	
10.1		General		QTF	QT__QTF	অনেক
10.2		Cardinals		QTC	QT__QTC	এক
10.3		Ordinals		QTO	QT__QTO	প্রথম
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	জল টল

Table 2-1: Bengali 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. Finalized the tags for Indo-Aryan languages by Prof. Deepti Mishra Sharma and Dr. Soma Paul on 12/07/2011

2.3 BENGALI POS TAGGING

LDC-IL developed a customized Bengali PoS tagset with the necessary adaptations to the BIS tagset to suit Bengali's linguistic structure. Since Bengali is a morphologically rich language, pre-processing 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 Guidelines

- Bengali PoS Tagging is done based on BIS tagset with customization to capture specific grammatical and morphological features.
- 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 discussions.

2.3.2 Tokenization

Tokenization prepares the raw text so the POS tagger can analyze each unit. Tokenization is important because accuracy of the PoS tagged data is depending on the input word. PoS tagging is incorporating the morpho-syntactic information to the lexical item. It does not insert any semantic, discourse or pragmatic information. Depending on the language and the complexity of the model, tokenizers use different strategies. In the traditional approach, it splits text into individual words based on spaces and punctuation. In Bengali sometimes **apostrophe** (') is used

as a character eliminator inside the word, as “দুইটি” [ɽuiʈi] become “দু’টি” [ɽuʈi]. So the tokenizer or the automatic POS tagger has to be trained enough to handle this kind of issue. In the case of proper noun and common noun sometimes case marker is added with a **hyphene (-)**. In such cases, case marker has to be considered with the previous word and should be merged with the previous word.

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 a proper noun can function as a common noun. Semantic function of a word is also useful to determine the lexical category. The Bengali word আর [ar] is a conjunction but sometimes it can act as other category. Following examples where [ar] is act as conjunction and as adverb.

In Bengali
সোমা\N_NNP আর\ CC_CCD রমা\N_NNP আসবে\ V_VM I\RD_PUNC
সে\ PR_PRP আর\RB আসবে\ V_VM না\ RP_NEG I\RD_PUNC
In IPA
ʃoma\N_NNP ar\ CC_CCD rɔma\N_NNP aʃbe V_VM
ʃe\ PR_PRP ar\RB aʃbe\ V_VM na\ RP_NEG

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

PoS tagging gives the lexical category and it gives some grammatical information. That means, it takes care of the form and function of each token. If it is possible to give the form and function features then it will be very useful to higher the 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 contextual meaning and idiomatic expressions are taken into account, the degree of ambiguity

will increase. 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 Ambiguity

Natural language is ambiguous from its basic nature. Ambiguity occurs when a single word or a sequence of words is associated within a language with more than one meaning. 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 Malayalam PoS tagging. This was addressed with careful attention in LDC-IL Bengali PoS tagged text corpus.

V_VM & V_VAUX

In Bengali
কল্পনা\N_NNP কথাটা\N_NN বলে\ V_VM চলে\ V_VM গেল\ V_VAUX I\RD_PUNC
কল্পনা\N_NNP বাড়ি\N_NN গেল\ V_VM I\RD_PUNC
In IPA
kɔlpona\N_NNP kɔtʰat̪a\N_NN bole\ v_vm tʃole\ V_VM gelo\ V_VAUX
kɔlpona\N_NNP baʈi\N_NN gelo \ v_vm

V_VM & PSP

In Bengali
ভাত\N_NN দিয়ে\ V_VM দাও\ V_VAUX I I\RD_PUNC
হাত\N_NN দিয়ে\ PSP ভাত\N_NN খাও\ V_VM I\RD_PUNC
In IPA
bʱat̪\N_NN d̪ie\ V_VM dao\ V_VAUX
ʱat̪\N_NN d̪ie\ PSP bʱat̪\N_NN kʰao\ V_VM

RP_RPD & PR_PRL

In Bengali
সে\PR_PRP বলল\ V_VM যে\RP_RPD তোমাকে\PR_PRP আসতেই\ V_VM হবে\ V_VAUX I\RD_PUNC
যে\PR_PRL জোরে\RB দৌড়ায়\ V_VM সেই\PR_PRL খেলায়\N_NN প্রথম\QT_QTO হয়\ V_VAUX I\RD_PUNC
In IPA

ʃe\PR_PRP bo1:o\V_VM dʒe\RP_RPD tomake\PR_PRP aʃtei\V_VM hobe\V__VAUX
dʒe\PR_PRL dʒore\RB douɾaj\V_VM ʃei\PR_PRP kʰelaj\N_NN protʰom\QT_QTO hoj\V__VAUX

N_NN & N_NNP

In Bengali
আমি\PR_PRP কখনও\RB সাগর\N_NN দেখিনি\V_VM ।\RD_PUNC
সাগর\N_NNP খুব\RP_INTF ভালো\JJ ফুটবল\N_NN খেলে\V_VM ।\RD_PUNC
In IPA
ami\PR_PRP kəkʰono\RB ʃagor\N_NN dekʰini\V_VM
ʃagor\N_NNP kʰub\RP_INTF bʰalo\JJ phutboɪ\N_NN kʰæle\V_VM

CC_CCD & RB

In Bengali
দীনত\N_NN আর\CC_CCD দারিদ্র্য\N_NN দুইই\QT_QTC ছেয়ে\V_VM আছে\V_VAUX মালতীলতাকে\N_NNP ।\RD_PUNC
কেউ\PR_PRI আর\RB ডাকছে\V_VM না\RP_NEG ।\RD_PUNC
In IPA
di:nota\N_NN ar\CC_CCD ɖarid̪:ro\N_NN ɖui:\QT_QTC tʃeje\V_VM atʃe\V_VAUX malot̪ilotaʃe\N_NNP
keu\PR_PRI ar\RB daktʃe\V_VM na\RP_NEG

RP_NEG & RP_RPD

In Bengali
সাধারণত\RB পকেটে\N_NN সিগারেট\N_NN রাখি\V_VM না\RP_NEG ।\RD_PUNC
কাজটা\N_NN করে\V_VM দে\V_VAUX না\RP_RPD ।\RD_PUNC
In IPA
ʃadharonoto\RB pokete\N_NN sigaret\N_NN rakʰi\V_VM na\RP_NEG
kadzta\N_NN kore\V_VM de\V_VAUX na\RP_RPD

N_NNP & JJ

In Bengali	
রূপবতী\N_NNP	বলেছে\V_VM ,\RD_PUNC তুই\PR_PRP অখিলদার\N_NNP আইনসম্মত\JJ উত্তরাধিকারী\N_NN \RD_PUNC
তোমার\PR_PRP মেয়ে\N_NN খুবই\RP_INTF রূপবতী\JJ \RD_PUNC	
In IPA	
rupoboti\N_NNP boletʃe\V_VM tui\PR_PRP okʰildar\N_NNP ain-ʃommoʃo\JJ uʈːoradhikari\N_NN	
ʈomar\PR_PRP meje\N_NN kʰub-i\RP_INTF rupoboti\JJ	

2.3.4.2 Abbreviation

Abbreviations are shortened forms of words or phrases which are widely used in both written and spoken language. Some examples of abbreviations in Bengali are- বি.এ., এম.এ. etc.

2.3.4.3 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, ‘হাত-পা’, ‘দিন-রাত’, ‘স্বর্গ-নরক’, ‘মা-বাবা’ 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. However, if they occur without any space, they can be considered to be a single token, for instance, ‘রথযাত্রা’.

2.3.4.4 Dialect Variation

Bengali exhibits significant dialectal variations across regions which is characterised by differences in phonology, vocabulary etc. Varieties such as Rarhi, Bangali, Varendri, Rajbangshi etc. though often grouped under the Bengali umbrella, display very unique features setting them apart from Standard Bengali. This dialectal 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 dialectal vocabulary into the Bengali corpora. Some expressions from regional dialects of Bengali found in the corpus are লবন, মরিচ, পানি etc.

2.3.4.5 Impact of Social Dialect

Dialect variation in Bengali is often shaped by social factors such as age, education, urban/rural setting, cultural exposure etc. Speakers from urban backgrounds often code-mix Bengali with English due to which a large amount of English words have entered Bengali corpus. Words such as ‘চেক করো’, ‘সরি’, ‘স্লো’ etc. have become very common in day-to-day use.

2.3.4.6 Classifier

Classifiers are affixes which categorize nouns into noun classes on the basis of shared, usually perceived characteristic properties of objects they denote. In Bengali, these affixes make a semantic classification of entities based on personality, shape, and size. Classifiers are basically a closed class morphemes which are attached postpositionally with nouns, pronouns and sometime with adjective. They are divided into groups on the basis of the personality, shape, and size of entities they classify. For example, টি, টা, থানা, থানি, জন are some important classifiers of Bengali language.

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. Unknown words in Bengali text corpus often arise due to spelling variations, code-mixing, presence of dialectal forms etc. For example, রহিস\RD_UNK, গ্রাউণ্ড\RD_UNK, ডুয়েল\RD_UNK etc.

2.4 SUMMARY OF LDC-IL BENGALI POS DATA

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

Dataset Summary	Aggregate Count
Total PoS Tags	407958
Total Sentences	38510
Total Word Count	338790
Total File Count	240

Table 2-2: Bengali PoS Status

Sl. No.	PoS Category	PoS Tag	Tag Count
1	Common Noun	N_NN	102404
2	Proper Noun	N_NNP	22785
3	Nloc Noun	N_NST	10134
4	Personal Pronoun	PR_PRP	21594

5	Reflexive Pronoun	PR_PRF	1121
6	Relative Pronoun	PR_PRL	981
7	Reciprocal Pronoun	PR_PRC	22
8	Wh-word Pronoun	PR_PRQ	3398
9	Indefinite Pronoun	PR_PRI	1756
10	Deictic Demonstrative	DM_DMD	5123
11	Relative Demonstrative	DM_DMR	785
12	Wh-word Demonstrative	DM_DMQ	69
13	Indefinite Demonstrative	DM_DMI	1413
12	Main Verb	V_VM	66029
13	Auxiliary Verb	V_VAUX	19299
14	Adjective	JJ	20240
15	Adverb	RB	8405
16	Postposition	PSP	9869
17	Co-ordinator	CC_CCD	4954
18	Subordinator	CC_CCS	2941
19	Particles (Default)	RP_RPD	7980
20	Classifier	RP_CL	57
21	Interjection	RP_INJ	548
22	Intensifier	RP_INTF	1650
23	Negation	RP_NEG	5992
24	Quantifiers (General)	QT_QTF	7329
25	Cardinals	QT_QTC	6663
26	Ordinals	QT_QTO	396
27	Foreign word	RD_RDF	1280
28	Symbol	RD_SYM	78
29	Punctuation	RD_PUNC	69090
30	Unknown	RD_UNK	3377
31	Echowords	RD_ECH	193

Table 2-3: Distinct Tag Count of Bengali