

ଓଡ଼ିଆ ପଦ ପ୍ରକରଣର
ଏନୋଟେଟେଡ୍ କର୍ପସ୍



**ODIA PARTS OF SPEECH
ANNOTATED CORPUS**

ODIA 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:

***Subhashree Mohanty
Dr. Narayan Choudhary
Rajisha 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

Odia Parts of Speech Annotated Corpus

Authors: Subhashree Mohanty, Dr. Narayan Choudhary, Rajesha N., Manasa G

e-ISBN: 978-81-69175-88-3

CIIL Publication No.: 1695

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: Teena verma

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	Odia PoS Annotated Corpus	6
2.1	Grammatical Peculiarities of Odia	6
2.2	LDC-IL Customized BIS Tagset	7
2.3	Odia PoS Tagging	8
2.4	Summary of LDC-IL Odia PoS Data	14

Tables

Table 1-1: BIS Tagset	4
Table 1-2: LDC-IL PoS Status	5
Table 2-1: Odia Tagset	8
Table 2-3: Odia PoS Status	14
Table 2-4: Distinct Tag Count of Odia	15

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 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 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 metioned 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 ODIA POS ANNOTATED CORPUS

Subhashree Mohanty

2.1 GRAMMATICAL PECULIARITIES OF ODIA

Odia is a morphologically rich and structurally complex Indo-Aryan language distinguished by its productive word formation processes and dense grammatical encoding. The language makes extensive use of inflection and derivation, allowing a single lexical root to generate multiple word forms that express varied grammatical meanings. Its morphological system reflects both agglutinative and fusional characteristics, resulting in layered structures where several grammatical categories are embedded within a single token.

From a syntactic perspective, Odia predominantly follows a Subject–Object–Verb (SOV) word order, while also permitting a degree of flexibility influenced by discourse and emphasis. Grammatical relations are typically expressed through suffixation and postpositional markers rather than word order alone. The language systematically encodes categories such as number, case, tense, aspect, mood, person, and honorificity, creating a tightly integrated morpho-syntactic framework.

Furthermore, Odia demonstrates functional adaptability across lexical categories, where context plays a crucial role in determining grammatical class and interpretation. Its use of complex predicates, compounding, reduplication, and cliticization contributes to its expressive depth and structural richness. These grammatical peculiarities make Odia linguistically significant and pose important challenges for corpus development, linguistic description, and Part-of-Speech annotation.

Odia is a morphologically rich language characterized by extensive word formation processes. Words in Odia 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 Odia requires not only identifying the syntactic class of a token but also understanding the layered morphological information embedded within it.

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

2.2 LDC-IL CUSTOMIZED BIS TAGSET

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

Sl. No	Category			Label	Annotation Convention**	Examples
	Top level	Subtype (level 1)	Subtype (level 2)			
1	Noun			N	N	
1.1		Common		NN	N__NN	ଝିଅ
1.2		Proper		NNP	N__NNP	ଶୁଭଶ୍ରୀ
1.3		Nloc		NST	N__NST	ମଧୁସୂର
2	Pronoun			PR	PR	
2.1		Personal		PRP	PR__PRP	ତୁମ୍ଭେ
2.2		Reflexive		PRF	PR__PRF	ତାହାକୁ
2.3		Relative		PRL	PR__PRL	ଯିଏ, ଯାହାକୁ
2.4		Reciprocal		PRC	PR__PRC	ତୁମ୍ଭେ
2.5		Wh-word		PRQ	PR__PRQ	କ'ଣ ଚଳେ
3	Demonstrative			DM	DM	
3.1		Deictic		DMD	DM__DMD	ସେହି
3.2		Relative		DMR	DM__DMR	ଯେଉଁ
3.3		Wh-word		DMQ	DM__DMQ	କିଏ
4	Verb			V	V	
4.1		Main		VM	V__VM	ହଲେ
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		Auxiliary		VAUX	V__VAUX	ଅଛି
5	Adjective			JJ		ଅଦଭୁତ
6	Adverb			RB		ଆସ୍ତର
7	Postposition			PSP		ପାଇଁ
8	Conjunction			CC	CC	
8.1		Co-ordinator		CCD	CC__CCD	ଓ

8.2		Subordinator		CCS	CC__CCS	ତଥାପି
8.2.1			Quotative	UT	CC__CCS__UT	ବୋଲି
9	Particles			RP	RP	
9.1		Default		RPD	RP__RPD	ହଁ
9.3		Interjection		INJ	RP__INJ	ଆଲଟୋ
9.4		Classifier		CLS	RP__CL	ଜଣ
9.5		Intensifier		INTF	RP__INTF	ଖୁବ୍
9.6		Negation		NEG	RP__NEG	ନୁହେଁ
10	Quantifiers			QT	QT	
10.1		General		QTF	QT__QTF	କିଛି
10.2		Cardinals		QTC	QT__QTC	ସଇତିରିଶ
10.3		Ordinals		QTO	QT__QTO	ପ୍ରଥମ
11	Residuals			RD	RD	
11.1		Foreign word		RDF	RD__RDF	Dialects
11.2		Symbol		SYM	RD__SYM	%, +
11.3		Punctuation		PUNC	RD__PUNC	,
11.4		Unknown		UNK	RD__UNK	କୁହୁମ୍ବରକମ୍
11.5		Echowords		ECH	RD__ECH	ବୁଝୁଫୁଝୁ

Table 2-1: Odia Tagset

The type hierarchy's lowest level tag should be used for the annotation. The higher level tags ought to be automatically stored after the lower level tag has been chosen. The list was created during the PoS standards committee meeting that took place at IIIT Hyderabad on June 12, 2010.

2.3 ODIA POS TAGGING

Odia PoS tagging is performed using the BIS tagset, with necessary adaptations from LDC-IL guidelines to suit Odia's linguistic structure. Since Odia is an inflectional 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

- Odia PoS Tagging is done based on BIS tagset with customization to capture specific grammatical and morphological features.
- Compound words were examined carefully to determine appropriate tagging at the word level.

- Spelling mistakes rectified.
- Extra spaces were removed where tokens had been incorrectly split by spaces.
- Missing spaces were inserted where necessary.
- Validation has done by using ambiguous word list.
- Contextual validation was performed to assign the correct tag
- Manual verification was conducted to maintain data quality
- Disagreements were resolved through discussion

2.3.2 Tokenization

Tokenization is the process of segmenting a continuous text into smaller units called tokens, usually words or meaningful sub-units. It is a crucial preprocessing step in Natural Language Processing because the accuracy of Part-of-Speech (PoS) tagging depends entirely on correctly identified input tokens. PoS tagging assigns morpho-syntactic information—such as grammatical category, number, case, tense, and aspect - to each token, but it does not encode semantic or pragmatic information. If tokenization is incorrect, the PoS annotation will also be inaccurate.

In Odia, tokenization is especially important due to its rich morphology and agglutinative tendencies. Roots often combine with plural markers, case suffixes, clitics, and auxiliary elements to form morphologically complex words. Without proper segmentation, such complex forms may be assigned a single incorrect PoS tag. Therefore, careful preprocessing and language-specific tokenization rules are essential for accurate PoS tagging in Odia.

For example, ମୁଁ ଭାତ ଖାଉଛି will be divided into 3 tokens- ମୁଁ | ଭାତ | ଖାଉଛି। Through tokenization, it is quite obvious that one can provide a PoS tag to each separate token.

2.3.3 Selecting a Tag

A word has two important aspects: form and function. The function of a word provides crucial clues for further linguistic processing. Therefore, while assigning a Part-of-Speech (PoS) tag, a balanced and systematic approach is adopted that considers both form and function.

The category of a word is determined by examining its morphological, syntactic, and semantic properties. Although morphological features often make category identification straightforward, a word may function as a different category depending on the context. Such variations are resolved by observing syntactic clues.

For example, certain pronouns function as demonstratives when they occur before a noun.

Example:

1. ସେ ଘର ବାହାରେ ଠିଆ ହୋଇଛି । (He/She is standing outside the house.)
2. ସେ ପୁସ୍ତକଟି ଟେବୁଲ୍ ଉପରେ ରଖ । (Keep that book on the table.)

In the first sentence, the token “ସେ” functions as a **personal pronoun**. However, in the second sentence, it acts as a **demonstrative determiner** because it modifies the common noun ପୁସ୍ତକଟି.

In Odia, there are several words whose PoS category varies depending on the surrounding linguistic environment. For example:

- ଭଲ → Adjective (ଭଲ ଛାତ୍ର), Adverb (ଭଲ ପଢ଼େ)
- ଉପରେ → Noun (ଉପରେ ଥିବା ଅଂଶ), Postposition (ମନେ ଉପରେ)
- ଆସିଥିବା → Verb participle, Adjectival modifier

Thus, accurate PoS tagging in Odia requires careful analysis of contextual and syntactic behavior rather than relying solely on morphological form.

2.3.4 Observation on PoS Tagging of LDC-IL Odia 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 gives more than one sense and it can appear in different categories. If the contextual meaning and idiomatic expressions are taken into account the order of ambiguity will increases. Words remain untagged or are assigned an unknown tag due to a lack of information about the word or an inadequate definition of the tagset. This can occur with foreign words written in a foreign script, foreign words in transliterated form that are unfamiliar to the language, dialectal variations, and similar cases.

2.3.4.1 Inconsistency

In Odia PoS-tagged corpora, inconsistencies frequently arise due to the morpho-syntactic flexibility of the language. Although certain expressions share semantic similarity, they may display structural variation in written form. As discussed earlier, compound formations in Odia often appear in multiple orthographic realizations.

In Odia PoS-tagged corpora, inconsistencies also arise due to **conjunct verb constructions**, particularly in cases of single-token versus multi-token realizations. Odia frequently forms complex predicates by combining a main verb (often in non-finite form) with a vector or auxiliary verb. However, these combinations may appear either as a single orthographic unit or as separate tokens.

Example: Conjunct Verb Variation

(a) ସେ ପୁସ୍ତକଟି ନଢେଗଲା \V_VM ।

(b) ସେ ପୁସ୍ତକଟି ନଢେ \V_VM ଗଲା \V_VM ।

In sentence (a), the conjunct verb **ନିଲେଗଲା** appears as a single token and receives one PoS tag (V_VM). In sentence (b), the same semantic construction is split into two tokens: **ନିଲେ** (non-finite verb form) and **ଗଲା** (main/vector verb), each assigned separate PoS tags.

Although both sentences convey identical meaning (“He/She took the book away”), the difference in tokenization results in different tagging structures. This illustrates how single vs. multi-token verb constructions contribute to inconsistency in Odia PoS annotation.

Such variation highlights that inconsistencies in Odia corpora stem not only from annotation decisions but also from flexible orthographic and morpho-syntactic patterns inherent in the language. Standardized tokenization guidelines are therefore crucial for maintaining tagging consistency.

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

Adjective (JJ) vs Adverb (RB)

In Odia
ଭଲJJ ଛାତ୍ର\N_NN
ଭଲRB ପଢ଼େ\V_VM
In IPA
/bʱɔɭ/ \JJ /tʃʱa:ʈrɔ/ \N_NN
/bʱɔɭ/ \RB /pɔɽʱe/ \V_VM

Noun (N_NN) vs Adjective (JJ)

In Odia
ସୁନା\N_NN ମହଙ୍ଗାJJ ଅଟେ\V_VAUX ।\RD_PUNC
ସୁନାJJ ଅଳଙ୍କାର\N_NN ସୁନ୍ଦରJJ ଅଟେ\V_VAUX ।\RD_PUNC
In IPA
/suɳa:\N_NN /məɦɔŋga:\JJ /ɔte\V_VAUX /\RD_PUNC

/suṅa:/JJ /ɔɭɔŋka:ɾɔ/\N_NN /sundɔɾɔ/JJ /ɔte/\V_VAUX /.\RD_PUNC
--

Quantifier (QT_QTF) vs Adjective (JJ)

In Odia
କିଛିQT_QTF ପୁସ୍ତକ\N_NN ଟବ୍ଲେଟ୍\N_NST ଅଛି\V_VAUX ।\RD_PUNC
କିଛିJJ ଲବଙ୍ଗ\N_NN ଆସିଲେ\V_VM ।\RD_PUNC
In IPA
/ kichʰi\QT_QTF pustak\N_NN tebulɛ:\N_NST ɔchʰi:\V_VAUX .\RD_PUNC
/ kichʰi\JJ lo:kɔ\N_NN a:sile:\V_VM .\RD_PUNC

Noun (N_NN) vs Verb (V_VM)

In Odia
ସେ\PR_PRP ଖଳେ\N_NN ଦେଖୁଛି\V_VM ।\RD_PUNC
ତୁମ୍ଭେ\PR_PRP ସବୁଦିନ\RB ସକାଳୁ\RB ଖଳେ\V_VM ।\RD_PUNC
In IPA
/se\PR_PRP kʰelɔ\N_NN dekʰutʰi\V_VM .\RD_PUNC/
/tume\PR_PRP sobudin\RB sɔka:lʊ\RB kʰelɔ\V_VM .\RD_PUNC/

Pronoun vs Adverb

In Odia
ତୁମ୍ଭେ\PR_PRP ଏଠା\RB କାହିଁକି\RB କରୁଛ\V_VM ?\RD_PUNC
ତାଙ୍କ\PR_PRP ଭିତର\PSP ଏଠା\PN ଭାବ\N_NN କାହିଁକି\RB ଆସିଛି\V_VM ?\RD_PUNC
In IPA
/tume\PR_PRP epa\RB ka:ɪki\RB kɔrutʰɔ\V_VM?\RD_PUNC/
/taŋkɔ\PR_PRP bʱitɔre\PSP epa\PN bʱa:bɔ\N_NN ka:ɪki\RB a:sitʰi\V_VM_VF ?\RD_PUNC/

Pronoun (P_PN) vs Quantifier (Q_Q)

In Odia
କେତେ/Q_Q ଲବଙ୍ଗ\N_NN ଆସିଛନ୍ତି\V_VM
କେତେ/P_PN ଜଣ/Q_Q ଆସିଛନ୍ତି\V_VM
In IPA
/keɽe\Q_Q lokɔ\N_NN a:sitʰɔntʰi\V_VM/
/keɽe\P_PN dʒɔŋɔ\Q_Q a:sitʰɔntʰi\V_VM/

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 Odia are ଇ.ସି.ଜି and ଏ.ଏସ୍.ପି

Odia	IPA
ଇ.ସି.ଜି Stands for Electro Cardio Gram	/i.si.dʒi/ means /ɪ.lɛk.tro: 'kɑ:r.dʒi.o: græm/
ଏ.ଏସ୍.ପି Stands for Additional superintendant of Police.	/e:.es.pi:/ means /ə'dɪʃənəl su:pərɪn'tendənt əv pə'li:s/

2.3.4.4 Compound Word

Compounding in Odia is a productive morphological process in which two or more morphemes combine to form a single lexical unit that conveys a unified meaning. In corpus annotation under the BIS tagset framework, orthographic representation plays a crucial role in tokenisation decisions. Hyphenated compounds such as ଦିନ-ରାତି\N_NN, ସୁଖ-ଦୁଃଖ\N_NN, ଭାଇ-ଭଉଣୀ\N_NN, and ଛଟା-ଟା-ବଡ଼\JJ are preferably treated as single tokens and assigned one primary PoS tag, as the hyphen signals semantic and morphological cohesion.

In contrast, non-hyphenated expressions written with a space, such as ସମୟ\N_NN ସୀମା\N_NN, ରାଜ୍ୟ\N_NN ସରକାର\N_NN, and ନୀଳ\JJ ଆକାଶ\N_NN, are tagged separately according to their individual grammatical categories. However, when the same forms occur without spacing (e.g., ସମୟସୀମା\N_NN, ରାଜ୍ୟସରକାର\N_NN, ନୀଳାକାଶ\N_NN), they are treated as single tokens. Therefore, consistent preprocessing and careful examination of orthographic patterns are essential for accurate PoS tagging and reliable corpus annotation in Odia.

2.3.4.5 Dialect Variation

In Dialect variation in Odia affects PoS tagging because taggers are typically trained on Standard Odia. Lexical differences (e.g., ଛୁଆ vs ପିଲା, କିଏ vs କିରା), morphological variations in verbs and auxiliaries (e.g., ଯାଉଛି vs ଯାଉଛୁ), phonological reductions (ଅଛି vs ଅତି), and pronoun differences (ତୁମେ vs ତୁ) can lead to misclassification errors. Syntactic changes in word order or particle usage further reduce tagging accuracy.

To mitigate this, dialect-aware corpora, normalization, and inclusion of regional variants in training data are essential for reliable PoS tagging in Odia. If the annotator is not familiar with such type of word then there is a chance to make mistake while assigning tag. So the annotator needs language expertise while tagging.

2.3.4.6 Impact of Social Dialect

Dialect variation in Odia is similarly influenced by social factors such as age, education, urban–rural background, and cultural exposure. Urban and educated speakers frequently code-mix Odia with English, resulting in the incorporation of English lexical items into the Odia corpus. Words such as ଉଲ୍ଲିଭୁ, କ୍ୱାନ୍ତାଲେ, ହଲେଟି, ଉଲ୍ଲେଟି, ଟାଉନ୍, ଅଫିସ୍ etc. are now widely used in everyday communication, often written in Odia script but retaining English pronunciation patterns. In addition, in multilingual and border regions of Odisha, expressions from languages such as Hindi, Telugu, Bengali, and tribal languages enter Odia discourse. These borrowed or regionally influenced forms may appear in Odia texts using Odia script, even though their origin lies in another language. For instance, dialectal forms like କନ୍ଦୋ (for Standard କନ୍ଦିତି), ତୁ (for ତୁମ୍ଭେ), or region-specific lexical items from Sambalpuri/Kosli varieties frequently occur in informal speech and social media writing. Such socially and regionally conditioned variations enrich the language but also introduce non-standard spellings and hybrid constructions, which can create challenges for corpus annotation and PoS tagging.

2.3.4.7 Unknown Words

Certain words in the text cannot be processed by the tagset. These words are regarded as unfamiliar terms, which can create issues for more advanced levels of processing. Unknown words in Odia text corpus often arise due to spelling variations, code-mixing, presence of dialectal forms, etc. For example, ଫେରିଲିକିRD_UNK, ଫାଇନ୍RD_UNK etc.

2.4 SUMMARY OF LDC-IL ODIA PoS DATA

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

Dataset Summary	Aggregate Count
Total PoS Tags	683112
Total Sentences	53288
Total Word Count	587653
Total File Count	953

Table 2-2: Odia PoS Status

Sl. No.	PoS Category	PoS Tag	Tag Count
1	Common Noun	N_NN	213917
2	Proper Noun	N_NNP	32288
3	Nloc Noun	N_NST	19558
4	Personal Pronoun	PR_PRP	34927
5	Reflexive Pronoun	PR_PRF	2857
6	Relative Pronoun	PR_PRL	1069
7	Reciprocal Pronoun	PR_PRC	688

8	Wh-word Pronoun	PR_PRQ	4518
9	Deictic Demonstrative	DM_DMD	10842
10	Wh-word Demonstrative	DM_DMQ	1175
11	Main Verb	V_VM	106524
12	Non-finite Verb	V_VM_VNF	0
13	Infinitive Verb	V_VM_VINF	0
14	Auxiliary Verb	V_VAUX	11122
15	Verbal Noun	V_VM_VNV	0
16	Adjective	JJ	49512
17	Adverb	RB	12743
18	Postposition	PSP	12103
19	Co-ordinator	CC_CCD	20069
20	Subordinator	CC_CCS	7854
21	Quotative	CC_CCS_UT	0
22	Particles	RP_RPD	10349
23	Interjection	RP_INJ	717
24	Intensifier	RP_INTF	894
25	Negation	RP_NEG	3223
26	Quantifiers	QT_QTF	11464
27	Cardinals	QT_QTC	7581
28	Ordinals	QT_QTO	1095
29	Foreign word	RD_RDF	2932
30	Symbol	RD_SYM	205
31	Punctuation	RD_PUNC	95254
32	Unknown	RD_UNK	1391
33	Echowords	RD_ECH	06

Table 2-3: Distinct Tag Count of Odia