

ગુજરાતી શબ્દભેદનું
એનોટેટેડ કોર્પસ

Gujarati Parts of Speech
Annotated Corpus

GUJARATI 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:

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

Gujarati Parts of Speech Annotated Corpus

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

e-ISBN: 978-81-69175-61-6

CIIL Publication No.: 1704

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: Rejitha K. S.

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	Gujarati PoS Annotated Corpus	6
2.1	Grammatical Peculiarities of Gujarati	6
2.2	LDC-IL Customized BIS Tagset	6
2.3	Gujarati PoS Tagging	8
2.4	Summary of LDC-IL Gujarati PoS Data	12

Tables

Table 1-1: BIS Tagset	4
Table 1-2: LDC-IL PoS Status	5
Table 2-1: Gujarati Tagset	8
Table 2-2: Gujarati PoS Status	12
Table 2-3: Distinct Tag Count of Gujarati	13

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 grammatical 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 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 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 GUJARATI POS ANNOTATED CORPUS

2.1 GRAMMATICAL PECULIARITIES OF GUJARATI

Gujarati is an Indo-Aryan language, follows a Subject–Object–Verb (SOV) word order. It has three grammatical genders such as masculine, feminine, and neuter. It has rich verb morphology. The features like tense, aspect, mood, person, number, and gender are embedded in verb. It uses an ergative case for subjects of transitive verbs in perfective tenses.

Gujarati grammar is characterized by its three-gender system, ergative constructions, postpositional structure, rich verbal morphology, and use of compound verbs. These features make it both structurally rich and typologically significant among Indo-Aryan languages.

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 Gujarati 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. To accommodate the specific features of Gujarati, 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 rare or absent in grammatical structure of Gujarati. Therefore, LDC-IL customized the tagset to accomodate the specific needs of each language. The Gujarati 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	વૃક્ષ
1.2		Proper		NNP	N__NNP	રામ
1.3		Nloc		NST	N__NST	જીવે
2	Pronoun			PR	PR	

2.1		Personal		PRP	PR__PRP	તે
2.2		Reflexive		PRF	PR__PRF	પોતે
2.3		Relative		PRL	PR__PRL	જે
2.4		Reciprocal		PRC	PR__PRC	પરસ્પર
2.5		Wh-word		PRQ	PR__PRQ	શું
2.6		Indefinite		PRI	PR__PRI	કોઈ
3	Demonstrative			DM	DM	
3.1		Deictic		DMD	DM__DMD	આ
3.2		Relative		DMR	DM__DMR	જેને
3.3		Wh-word		DMQ	DM__DMQ	કોણ
3.4		Indefinite		DMI	DM__DMI	કશું
4	Verb			V	V	
4.1		Main		VM	V__VM	ખાધું
4.2		Auxiliary		VAUX	V__VAUX	છે
4.2.1			Participle noun	VNP	V__VAUX_VNP	બેઠેલા
5	Adjective			JJ		સુંદર
6	Adverb			RB		ધીમે
7	Postposition			PSP		જેમ
8	Conjunction			CC	CC	
8.1		Co-ordinator		CCD	CC__CCD	અને
8.2		Subordinator		CCS	CC__CCS	પરંતુ
9	Particles			RP	RP	
9.1		Default		RPD	RP__RPD	ફક્ત
9.3		Interjection		INJ	RP__INJ	ઓ !!
9.4		Intensifier		INTF	RP__INTF	ઘણું
9.5		Negation		NEG	RP__NEG	ના
10	Quantifiers			QT	QT	
10.1		General		QTF	QT__QTF	કેટલાક
10.2		Cardinals		QTC	QT__QTC	એક
10.3		Ordinals		QTO	QT__QTO	પ્રથમ
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	૨૫-૨૫

Table 2-1: Gujarati 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 GUJARATI POS TAGGING

LDC-IL developed a customized Gujarati PoS tagset with the necessary adaptations to the BIS tagset to suit Gujarati's linguistic structure. To improve tagging accuracy context-based tag assignment is carried out to ensure the consistency across the corpus.

2.3.1 Annotation Guidelines

- Gujarati PoS Tagging is done based on BIS tagset with customization to capture specific grammatical and morphological features.
- 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 Gujarati, vocabulary item is considered as a token and each token is labeled according to the PoS category.

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 the same word can act as relative pronoun and relative demonstrative based on the context. જે functions as different grammatical categories depending on the context. The example is given below.

In Gujarati
ભગવાન\N_NN :RD_PUNC જે\PR_PRL સદા\N_NST રહે\V_VM એ\PR_PRP યથાર્થ\N_NN છે\V_VM .RD_PUNC
ફક્ત\RP_RPD એ\PR_PRP જ\RP_RPD એક\JJ છે\V_VM જે\DM_DMR વિચાર\N_NN પણ\CC_CCD કરે\V_VM છે\V_VAUX .RD_PUNC
In IPA
b ^h agava:nə\N_NN :RD_PUNC dʒe:\PR_PRL sadā:\N_NST rafie:\V_VM e:\PR_PRP ja ^h ṭha:r ^h ṭhə\N_NN ʃ ^h e:\V_VM
p ^h akaṭṭə\RP_RPD e:\PR_PRP dʒə\RP_RPD e:kə\JJ ʃ ^h e:\V_VM dʒe:\DM_DMR viʃa:rə\N_NN paŋə\CC_CCD kare:\V_VM ʃ ^h e:\V_VAUX

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

PoS tagging gives the lexical category and it gives some grammatical information. That means it taken care the form and function of the words. If it is possible to give the form and function 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 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 Inconsistency

The same word appeared differently in the corpus. For e.g. in Gujarati, 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. The same word can appear differently even though it gives the same meaning. Both combinations are exist in the corpus and it goes to different way while tagging.

The following examples shows the same elements can appear separately or as a single combined token with their respective PoS tags. The word is written as separately ‘શ્યામ સુંદર’ and combined ‘શ્યામસુંદર’.

In Gujarati
પીળાં\JJ વસ્ત્રોમાં\N_NN સુશોભિત\JJ શ્યામ\JJ સુંદર\JJ યુવક\N_NN ક્યાં\PR_PRQ છે\V_VM ?\RD_PUNC
મહાપ્રભુ\N_NNP જગન્નાથના\N_NNP મંદિરમાં\N_NN દરરોજ\N_NST જાય\N_NN છે\N_NN અને\CC_CCD જગન્નાથ\N_NNP તેમને\PR_PRP પોતાનું\PR_PRF સુંદરJJ શ્યામસુંદરJJ સ્વરૂપ\N_NN દેખાડે\N_NN છે\N_NN .\RD_PUNC
In IPA
pi:lɑ:m\JJ vɑstrɔ:mɑ:m\N_NN sufo:bʰitʰə\JJ ʃjɑ:mə\JJ sumd̪arə\JJ juvəkə\N_NN kja:m\PR_PRQ tʰe:\V_VM ?\RD_PUNC
mafia:prabʰu\N_NNP dʒaganna:tʰana:\N_NNP mamd̪irama:m\N_NN d̪araro:dʒə\N_NN dʒa:jə\N_NN tʰe:\V_VM VAUX ane:\CC_CCD dʒaganna:tʰə\N_NNP t̪e:mane:\PR_PRP po:t̪ɑ:nu\PR_PRF sumd̪arə\JJ ʃjɑ:masumd̪arə\JJ svaru:pə\N_NN d̪e:kʰa:de:\V_VM tʰe:\V_VM VAUX

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

V_VM & \V_VAUX

In Gujarati
ભગવાન\N_NN : \RD_PUNC એ\DM_DMD અહંકાર\N_NN છે\ V_VM . \RD_PUNC
એમાં\PR_PRL ભગવાનના\N_NN ઉપદેશો\N_NN સંક્ષેપમાં\N_NN આપ્યા\ V_VM છે\ V_VAUX . \RD_PUNC
In IPA
b ^h agava:nə\N_NN : \RD_PUNC e:\DM_DMD ahamka:rə\N_NN t̪ʰe:\ V_VM . \RD_PUNC
e:ma:m\PR_PRL b ^h agava:nana:\N_NN upaḍe:ʃo:\N_NN samkṣe:pama:m\N_NN a:pja:m\ V_VM t̪ʰe:\ V_VAUX . \RD_PUNC

CC_CCS & RP_RPD

In Gujarati
ભગવાને\N_NN કહ્યું\ V_VM કે\CC_CCS સો\N_NNP સૈયદે\N_NNP એ\PR_PRP વાંચ્યો\ V_VM છે\ V_VAUX . \RD_PUNC
ઊઘ\N_NN બરાબર\JJ આવે\ V_VM છે\ V_VAUX કે\RP_RPD ? \RD_PUNC
In IPA
b ^h agava:ne:\N_NN kaɦjum\ V_VM ke:\CC_CCS so.\N_NNP saijade:\N_NNP e:\PR_PRP va:m̪t̪ʃo:\ V_VM t̪ʰe:\ V_VAUX . \RD_PUNC
u:mg̃hə\N_NN bara:barə\JJ a:ve:\ V_VM t̪ʰe:\ V_VAUX ke:\RP_RPD ? \RD_PUNC

2.3.4.3 Abbreviation

Some people or institution are known by the abbreviation of their name. Based on the context it can identify as proper noun or any other category. It is tagged accordingly.

Gujarati	IPA
એસ.એ.બી.ટી\N_NNP	/ e:sə.e:.bi:.t̪i:/ \N_NNP

2.3.4.4 Confusion in the definition of Nloc

Space and time denoting words tagged as N_NST tag which is under Noun category and Nloc is its subtype. ‘ત્યારે’ /t̪ja:re:/ means ‘then’ and functions as an adverb of time. It is tagged under the noun category, which creates confusion for the annotator.

2.3.4.5 Compound Word

Two or more word combined to form a single unit and the sole token gives a meaning. For eg: સમીજજીવન /sama:ɖʒadʒi:vʌnane:/. This type of words are taken as such and assigned the appropriate tag.

2.3.4.6 Dialect Variation

If the annonator is unaware about the dialect variation it may lead to incorrect tagging. 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.

2.3.4.7 Unknown Words

There will be some words in the text which cannot be handled by the tagset. Words from other languages written in the Gujarati script that are not part of the Gujarati lexicon are tagged as unknown. Such words are considered as unknown words and it will pose problem for the higher level of processing.

2.4 SUMMARY OF LDC-IL GUJARATI POS DATA

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

Dataset Summary	Aggregate Count
Total PoS Tags	759604
Total Sentences	48598
Total Word Count	655358
Total File Count	552

Table 2-2: Gujarati PoS Status

Sl. No.	PoS Category	PoS Tag	Tag Count
1	Common Noun	N_NN	180787
2	Proper Noun	N_NNP	44718
3	Nloc Noun	N_NST	17550
4	Personal Pronoun	PR_PRP	42458
5	Reflexive Pronoun	PR_PRF	3023
6	Relative Pronoun	PR_PRL	3399
7	Reciprocal Pronoun	PR_PRC	118
8	Wh-word Pronoun	PR_PRQ	2790
9	Indefinite Pronoun	PR_PRI	1592
10	Relative Demonstrative	DM_DMR	3249
11	Deictic Demonstrative	DM_DMD	11032

12	Wh-word Demonstrative	DM_DMQ	524
13	Indefinite Demonstrative	DM_DMI	2449
14	Main Verb	V_VM	116353
15	Auxiliary Verb	V_VAUX	48996
16	Participle Noun	V_VAUX_VNP	26
17	Adjective	JJ	47729
18	Adverb	RB	6406
19	Postposition	PSP	21439
20	Co-ordinator	CC_CCD	19809
21	Subordinator	CC_CCS	6582
22	Particles	RP_RPD	36881
23	Interjection	RP_INJ	765
24	Intensifier	RP_INTF	1881
25	Negation	RP_NEG	6093
26	Quantifiers	QT_QTF	11211
27	Cardinals	QT_QTC	11964
28	Ordinals	QT_QTO	2497
29	Foreign word	RD_RDF	1418
30	Symbol	RD_SYM	299
31	Punctuation	RD_PUNC	103947
32	Unknown	RD_UNK	1615
33	Echowords	RD_ECH	4

Table 2-3: Distinct Tag Count of Gujarati