

తెలుగు భాషాభాగాల విశ్లేషిత దత్తాంశం



Telugu Parts of Speech
Annotated Corpus

TELUGU 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. Modugu Kasimbabu

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

Telugu Parts of Speech Annotated Corpus

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

e-ISBN: 978-81-69175-20-3

CIIL Publication No.: 1692

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: Modugu Kasimbabu

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	Telugu PoS Annotated Corpus	6
2.1	Grammatical peculiarities of Telugu	6
2.2	LDC-IL Customized BIS Tagset	6
2.3	Telugu PoS Tagging	8
2.4	Summary of LDC-IL Telugu PoS Data	16

Tables

Table 1-1: BIS Tagset	4
Table 1-2: LDC-IL PoS Status	5
Table 2-1: Telugu Tagset	8
Table 2-2: Pre-processed word list of Telugu with IPA	9
Table 2-3: Telugu PoS Status	16
Table 2-4: Disntict Tag Count of Telugu	17

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 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 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 TELUGU POS ANNOTATED CORPUS

Modugu Kasimbabu

2.1 GRAMMATICAL PECULARITIES OF TELUGU

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

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 Telugu 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 Telugu, 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 Telugu. Therefore, LDC-IL customized the tagset to accomodate the specific needs of each language. The Telugu 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	పుస్తకం
1.2		Proper		NNP	N__NNP	రాజు
1.3		Nloc		NST	N__NST	మునుపు
2	Pronoun			PR	PR	
2.1		Personal		PRP	PR__PRP	నేను, నువ్వు
2.2		Reflexive		PRF	PR__PRF	తననుతాను

2.3		Relative		PRL	PR__PRL	ఎవ్వడు
2.4		Reciprocal		PRC	PR__PRC	పరస్పరం
2.5		Wh-word		PRQ	PR__PRQ	ఎక్కడ
3	Demonstrative			DM	DM	
3.1		Deictic		DMD	DM__DMD	ఆ, ఈ
3.2		Relative		DMR	DM__DMR	ఏ
3.3		Wh-word		DMQ	DM__DMQ	ఏ
4	Verb			V	V	
4.1		Main		VM	V__VM	
4.1.1			Finite	VF	V__VM__VF	తిన్నాను
4.1.2			Non-finite	VNF	V__VM__VNF	తిని
4.1.3			Infinitive	VINF	V__VM__VINF	తిన
4.1.4			Gerund	VNG	V__VM__VNG	తినడం
4.2		Auxiliary		VAUX	V__VAUX	ఉంది
5	Adjective			JJ		తియ్యని
6	Adverb			RB		త్వరగా
7	Postposition			PSP		నుంచి
8	Conjunction			CC	CC	
8.1		Co-ordinator		CCD	CC__CCD	కానీ, ఇంకా
8.2		Subordinator		CCS	CC__CCS	అంటే
8.2.1			Quotative	UT	CC__CCS__UT	అని
			Introducer or Identifier		CC__CCS__ID	అంటూ
9	Particles			RP	RP	
9.1		Default		RPD	RP__RPD	అయితే
		Classifier		CL	RP__CL	మంది, గురు
9.3		Interjection		INJ	RP__INJ	అయ్యో, అమ్మో
9.4		Intensifier		INTF	RP__INTF	చాలా
9.5		Negation		NEG	RP__NEG	తప్ప
10	Quantifiers			QT	QT	
10.1		General		QTF	QT__QTF	కొంచెం
10.2		Cardinals		QTC	QT__QTC	ఒకటి, రెండు
10.3		Ordinals		QTO	QT__QTO	ఒకటవ, రెండవ
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	పులి-గిలి

Table 2-1: Telugu 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 Dravidian languages by Prof. Uma Maheswara Rao G and Dr. Sobha L on 12/07/2011

2.3 TELUGU POS TAGGING

LDC-IL developed a customized Telugu PoS tagset with the necessary adaptations to the BIS tagset to suit Telugu's linguistic structure. Since Telugu is an agglutinative 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 Annotation Guidelines

- Telugu PoS Tagging is done based on BIS tagset with customization to capture specific grammatical and morphological features.
- Telugu is an agglutinative language, pre-processing has been done because to fit the words according to the tags.
- Complex word was segmented to assign correct PoS categories.
- Compound words were examined carefully to determine appropriate tagging at the word level.
- Spelling mistakes rectified.
- Extra spaces were removed where tokens had been incorrectly split by spaces.
- Missing spaces were inserted where necessary.
- Validation 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 sematic, discourse or pragmatic information. In agglutinative languages like Telugu, Kannada, Tamil, and Malayalam 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.

For eg: the following word has three grammatical categories but it is a single token. మేమిచ్చినపని (me:mitʃ:inapani)

The grammatical combination is pronoun + adjective + noun. Without pre-processing it does not come under any lexical category.

Some pre-processed word list of Telugu with IPA is given below:

Token in the original text	Pre processed word	tag	Pre processed word	tag	Pre processed word	tag
పాటకూడా pa:ʈaku:ɖa:	పాట pa:ʈa	N_NN	కూడా ku:ɖa:	RP_RPD		
చేసినప్పటికీ tʃe:sinappaʈiki:	చేసిన tʃe:sina	V_VM_VNF	అప్పటికీ appaʈiki:	CC_CCD		
మాటవినివెళ్ళాను ma:ʈawiniwelʃa:nu	మాట ma:ʈa	N_NN	విని wini	V_VM_VNF	వెళ్ళాను welʃa:nu	V_VM_VF
పాటతప్ప pa:ʈaʈappa	పాట pa:ʈa	N_NN	తప్ప ʈappa	RP_NEG		
చేపల్లెవు tʃe:palle:wu	చేపలు tʃe:palu	N_NN	లేవు le:wu	V_VAUX_VF		
పువ్విచ్చింది puw:itʃʈiɳɖi	పువ్వు puw:u	N_NN	ఇచ్చింది itʃ:ɳɖi	V_VM_VF		
పనైతేతప్ప panajʈe:ʈappa	పని pani	N_NN	అయితే ajʈe:	RP_RPD	తప్ప ʈappa	RP_NEG
మాటమాత్రం ma:ʈama:ʈram	మాట ma:ʈa	N_NN	మాత్రం ma:ʈram	RP_RPD		

Table 2-2: Pre-processed word list of Telugu with IPA

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 noun may be an adjective if it occurs immediately before a noun. Semantic function of a word is also useful to determine the lexical category. The Telugu word చేసి/tʃe:si/ is a Non Finite Verb but sometimes it can act as other category. Following examples where చేసి/tʃe:si/ is act as Postposition and as Non Finite Verb.

In Telugu
రాధ\N_NNP వంట\N_NN చేసి\V_VM_VNF పంపింది\V_VM_VF
స్వామి\N_NN రావటం\V_VM_VNG జేసి\PSP భక్తుని\N_NN సంతోషం\N_NN
In IPA
ra:ɖʱa\N_NNP waɳɖa\N_NN tʃe:si\V_VM_VNF paɱpiɳɖi\V_VM_VF
swa:mi\N_NN ra:waɖaɱ\V_VM_VNG dʒe:si\PSP bʱakɖuni\N_NN saɱto:ʂaɱ\N_NN

2.3.4 Observation on PoS Tagging of LDC-IL Telugu 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 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 Inconsistency

The same word appeared differently in the corpus. For e.g. in Telugu, 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. Language names, while often tagged as proper nouns, can sometimes function as adjectives as well. నల్లగుర్రం\mallagurram and నల్ల\mallalla గుర్రం\gurram కేంద్రప్రభుత్వం\ke:ndraprab^hut^wwam and కేంద్ర\ke:ndra, ప్రభుత్వం\prab^hut^wwam are appear in corpus. Both combinations are exist in the corpus and it goes to different category while tagging.

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

పట్టి\V_VM_VNF తెచ్చాను\V_VM_VF and పట్టితెచ్చాను\V_VM_VF

నడుస్తున్న\JJ చరిత్ర\ N_NN and నడుస్తున్నచరిత్ర\ N_NNP

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

V_VM_VNF & V_VM_VF

In Telugu
పాట\N_NN విని\V_VM_VNF కూర్చున్నది\V_VM_VF .\RD_PUNC
అలాగే\CC_CCS పడుకున్నది\V_VM_VF పాట\N_NN విని\V_VM_VF .\RD_PUNC
In IPA
pa:ʈa\N_NN vini\V_VM_VNF ku:rtfunnaði\V_VM_VF .\RD_PUNC
ala:ge:\CC_CCS paḍukunnaði\V_VM_VF pa:ʈa\N_NN vini\V_VM_VNF.\RD_PUNC

PR_PRP & DM_DMD

In Telugu
అది\PR_PRP నా\PR_PRP మరదలు\N_NN శ్రీలత\N_NNP
అది\DM_DMD మా\PR_PRP ఊరి\N_NN బస్సు\N_NN
In IPA
aɖi\PR_PRP na:\PR_PRP maraɖalu\N_NN ʃri:lata\N_NNP
aɖi\DM_DMD ma:\PR_PRP u:ri\N_NN bassu\N_NN

N_NN & N_NNP

In Telugu
నిన్న\N_NST పత్రికలో\N_NN ఒక\QT_QTO కవిత\N_NN చదివాను\V_VM_VF
కవిత\N_NNP నన్ను\PR_PRP చూసి\N_NN సంతోషించింది\N_NN
In IPA
ninna\N_NST pat̪rikalo:\N_NN oka\QT_QTO kawita\N_NN tʃaɖiwa:nu\N_NN
kawita\N_NNP nann:u\PR_PRP tʃu:si\N_NN sam̪to:ʃim̪tʃiɳɖi\N_NN

N_NST&RB

In Telugu
అప్పుడే\N_NST అందరూ\PR_PRF చూశారు\N_NN
అందరూ\PR_PRF అప్పుడే \RB వచ్చారు\N_NN
In IPA
appuɖe:\N_NST aɳɖaru:\PR_PRF tʃu:ʃa:ru\N_NN
aɳɖaru:\PR_PRF app:uɖe:\RB watʃtʃa:ru\N_NN

V_VM_VNF & JJ

In Telugu
నేను\PR_PRP అ\DM_DMD పని\N_NN చేసిన\V_VM_VNF తర్వాత\PSP ఊరికి\N_NN వెళ్తాను\V_VM_VF
నువ్వు\PR_PRP చేసిన\JJ వంట\N_NN బాగుంది\V_VM_VF
In IPA
ne:nu\PR_PRP a:\DM_DMD pani\N_NN tʃe:sina\V_VM_VNF t̪arwa:t̪a\PSP u:riki\N_NN weɭt̪a:nu\V_VM_VF
nuwwu\PR_PRP tʃe:sina\JJ want̪a\N_NN ba:ɡundi\V_VM_VF

V_VM_VF & N_NN

In Telugu
నువ్వు\PR_PRP పది\QT_QTC గంటల\N_NN వరకు\PSP చదువు\V_VM_VF
నేనే\PR_PRP అమెకు\PR_PRP చదువు\N_NN చెప్పించాను\V_VM_VF
In IPA
nuwwu\PR_PRP paɖi\QT_QTC ɡaɳt̪ala\N_NN waraku\PSP tʃaɖuʋu\V_VM_VF
ne:ne:\PR_PRP a:meku\PR_PRP tʃaɖuʋu\N_NN tʃeppint̪a:nu\V_VM_VF

RP_NEG&RP_RPD

In Telugu
కొంచెం\QT_QTF మాత్రమే\RP_NEG పచ్చరంగు\N_NNP మిగతాదంతా\QT_QTF ఎరుపురంగు\N_NNP
మీరు\PR_PRP మాత్రమే\RP_RPD మంచివారు\N_NN
In IPA

kɔntʃem\QT_QTF ma:ʈrame:\RP_NEG patʃarəngu\N_NNP migəʈa:ɖəntə:\QT_QTF erupurəngu\N_NNP
mi:ru\PR_PRP ma:ʈrame:\RP_RPD maɪtʃiwa:ru\N_NN

2.3.4.3 Abbreviation

Some people or institution is 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.

Telugu	IPA
జి. ఉమామహేశ్వరరావు; జి means గారపాటి	dʒi\ uma:mahe:ʃwarara:wu\ dʒi means ga:rapa:ʈi
బి.హెచ్.కె. means భద్రరాజు కృష్ణమూర్తి	bi\hetʃke\ means bʰaɖrira:dʒu\ kruʃɳamu:rtʃi

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. But these words can come along with noun, verb and gerunds as well.

తినకముందు/ ʈinakamundu/ the word ‘ముందు’ mundu/ means ‘before’ come with verbs so there is a confusion that whether to split the word or consider it as single token. If it separated ‘ముందు’ is fit for N_NST tag and if it is together verb category can accommodate this word.

అప్పుడు/appudu/ is a word which come under N_NST according to BIS tagset. It comes with both noun and verb and if it comes with noun it is preprocessed, but if it comes with verb the whole token is considered as verb category. For eg:

Telugu
నేను\PR_PRP భోజనం\N_NN చేస్తున్నప్పుడు\ V_VM_VNF వచ్చారు\ V_VM_VF
IPA
ne:nu\PR_PRP bʰo:dʒanam\N_NN tʃe:stʌnnappudu\ V_VM_VNF watʃtʃa:ru\V_VM_VF

2.3.4.5 Multiword

A sequence of words can act as a single unit, making it difficult to assign a single category. Telugu is an agglutinative language, so multiple words are often joined together to form a single token. In Telugu, any category can attach to another category. For eg: VERB+AUX + N_NST+QT_QTF, NOUN+RP_RPD+AUX etc. In such cases, assigning an appropriate PoS tag to the combined word is challenging. Therefore, preprocessing is necessary to identify the correct category. A pre-processed multiword is given below:

Telugu	
తింటున్నంతసేపువుంది → తింటు\V_VM_VNF ఉన్న\V_VAUX అంత\QT_QTF సేపు\N_NN ఉంది\V_VAUX	
IPA	
ṭɪntʊnnant̪sɛ:puvundi → ṭɪntu\V_VM_VNF unna:\V_VAUX ant̪a\QT_QTF se:pu\N_NN undi\V_VAUX	

2.3.4.6 Compound Word

Two or more word combined to form a single unit and the sole token gives a meaning. For eg: తిక్కనభారతం\ṭikṇanabʰa:raṭam\, తెలుగుభూమిపత్రిక\ṭelugubʰu:mipaṭṛika\, జేబుదొంగ\dʒɛ:budɔŋga\, పెద్దమట్టిగుర్రంబండి\peḍḍamaṭṭigurrambaṇḍi\ these types of words are taken as such and assigned the appropriate tag.

2.3.4.7 Dialect Variation

If the annotator is unaware about the dialect variation it may lead to incorrect tagging. In Telugu, the word కైకిలి /kaikili/ ‘coolie’ is using in Telangana region. Its standard form is ‘కూలీ’/ku:li:/. The word ‘సోల’ /so:la/ means ‘A certain dry measure’ some of the Rayalaseema region people they don’t use this term. ‘A father’s sister’ called ‘మేనత్త’ /me:natt̪a/ in Telugu, and in Telangana region they called like ‘బాప్ప’ /ba:ppa/. The word ‘quickly’ is called ‘త్వరగా’ /ṭwaraɡa:/ some of the Rayalaseema speakers they use to call ‘బిన్నా’ /binna:/. The word ‘తక్కాలి’ /ṭakka:li/ is used in Rayalaseema region, but in the other regions of Telugu they can say టమాట /ṭama:ṭa/. So there are many words like this, if the annotator is not familiar with such type of words then there is a chance to make mistake while assigning tag. So the annotator needs language expertise while tagging.

2.3.4.8 Impact of Social Dialect

The word సిఫారసు/sip^ha:rasu/ means ‘recommendation’ it is an Hindustani word and it is nativised, so this word tagged as N_NN category. Same as ‘కితాబు’/kiṭa:bu means ‘A title of honor’ it is an Hindustani word. ‘దవఖాన’/ḍawak^ha:na/ means ‘Hospital’, తకరారు /ṭakara:ru/ means ‘A dispute’; ‘నామోషి’ /na:mo:ṣi/ means ‘shyness’; ‘జల్దీ’ /dʒalḍi:/ means ‘quickly’; ‘వగైరా’ /wagaira:/ means ‘Etcetera’; ‘హుషారు’ /huṣa:ru/ means ‘carefulness’; ‘పితూరీ’ /piṭu:ri:/ means ‘An intrigue’; ‘పరాకు’ /para:ku/ means ‘Carelessness’; ‘మతలబు’ /maṭalabu/ means ‘meaning’. These are all we are localized and using our language. In case the annotator unaware about the term it may tag as RD_UNK.

2.3.4.9 Classifier

Telugu has a classifier marker. However, it is used only to count humans. For example, ముగ్గురు/mugguru/, నలుగురు/naluguru/ means ‘Four persons’, పదకొండుమంది/paḍakonḍumandī/ means ‘Eleven persons, పన్నెండుమంది/pannenḍumandī/ means ‘Twelve persons’. Here, గురు /guru/ and మంది /mandi/ are classifier markers. Telugu don’t have non-human classifier markers. In the words గురు and మంది above, ‘Guru’ is used for three, four, up to ten. From eleven, the mandi marker is used. Such as పదకొండుమంది, పన్నెండుమంది. Mandi is also used for the words ఎనిమిది/enimidi/, తొమ్మిది/tommiḍi/, పది/paḍi/, and also నూరు/nu:ru/ means hundred. However, in the case of Guru, Mandi is gradually lost and Mandi is established.

2.3.4.10 Unknown Words

There will be some words in the text which cannot be handled by the tagset. కాంట్రవర్సిటీలు\ka:nṭrawarsiti:lu, పొరప్లాట్లు\porapla:ṭlu, ఖాతార్నాక్\k^ha:ṭa:rna:k, వైచిపీ\waitʃi:pi:, పాలిట్రక్స్\pa:liṭriks, కట్ కటా\kaṭ kaṭa:. Such words are considered as unknown words and it will pose problem for the higher level of processing.

2.4 SUMMARY OF LDC-IL TELUGU POS DATA

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

Dataset Summary	Aggregate Count
Total PoS Tags	37840
Total Sentences	3992
Total Word Count	30119
Total File Count	31

Table 2-3: Telugu PoS Status

Sl. No.	PoS Category	PoS Tag	Tag Count
Sl. No.	PoS Category	PoS Tag	Tag Count
1	Common Noun	N_NN	7248
2	Proper Noun	N_NNP	3591
3	Nloc Noun	N_NST	1365
4	Personal Pronoun	PR_PRP	1170
5	Reflexive Pronoun	PR_PRF	535
6	Relative Pronoun	PR_PRL	190
7	Reciprocal Pronoun	PR_PRC	5
8	Wh-word Pronoun	PR_PRQ	112
9	Deictic Demonstrative	DM_DMD	681
10	Wh-word Demonstrative	DM_DMQ	195
11	Relative Demonstrative	DM_DMR	91
12	Finite Verb	V_VM_VF	3002
13	Non-finite Verb	V_VM_VNF	2042
14	Infinitive Verb	V_VM_VINF	614
15	Gerund	V_VM_VNG	499
16	Auxiliary Finite Verb	V_VAUX_VF	235
17	Auxiliary Non-finite Verb	V_VAUX_VNF	44
18	Auxiliary Infinitive Verb	V_VAUX_VINF	24
19	Adjective	JJ	3676
20	Adverb	RB	1772
21	Postposition	PSP	484
22	Co-ordinator	CC_CCD	335
23	Subordinator	CC_CCS	279
24	Quotative	CC_CCS_UT	153
25	Particles	RP_RPD	88
26	Interjection	RP_INJ	18
27	Intensifier	RP_INTF	191
28	Negation	RP_NEG	89
29	Classifier	RP_CL	26
30	Quantifiers	QT_QTF	480
31	Cardinals	QT_QTC	595
32	Ordinals	QT_QTO	241
33	Foreign word	RD_RDF	3
34	Symbol	RD_SYM	2
35	Punctuation	RD_PUNC	7719
36	Unknown	RD_UNK	12
37	Echowords	RD_ECH	14

Table 2-4: Distinct Tag Count of Telugu