

മലയാള ശബ്ദഭേദ വ്യാഖ്യാന
ദത്തശേഖരം

**MALAYALAM PARTS OF SPEECH
ANNOTATED CORPUS**

MALAYALAM 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. Rejitha K.S.

Dr. Saritha S. L.

Dr. Sajila S.

Dr. Narayan Choudhary

**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

Malayalam Parts of Speech Annotated Corpus

Authors: Dr. Rejitha K. S., Dr. Saritha S. L., Dr. Sajila S., Dr. Narayan Choudhary

e-ISBN: 978-81-69175-34-0

CIIL Publication No.: 1698

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	Malayalam PoS Annotated Corpus	6
2.1	Grammatical Peculiarities of Malayalam	6
2.2	LDC-IL Customized BIS Tagset	6
2.3	Malayalam PoS Tagging	8
2.4	Summary of LDC-IL Malayalam PoS Data	15

Tables

Table 1-1: BIS Tagset	4
Table 1-2: LDC-IL PoS Status	5
Table 2-1: Malayalam Tagset	8
Table 2-2: Pre-processed Word List of Malayalam with IPA	9
Table 2-3: Malayalam PoS Status	15
Table 2-4: Distinct Tag Count of Malayalam	16

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 MALAYALAM POS ANNOTATED CORPUS

Rejitha K S

2.1 GRAMMATICAL PECULIARITIES OF MALAYALAM

Malayalam is a morphologically rich language characterized by extensive word formation processes. Words in Malayalam 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 Malayalam 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 Malayalam, 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 Malayalam. Therefore, LDC-IL customized the tagset to accomodate the specific needs of each language. The Malayalam 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	മരം
1.2		Proper		NNP	N__NNP	രാമൻ
1.3		Nloc		NST	N__NST	താഴെ
2	Pronoun			PR	PR	
2.1		Personal		PRP	PR__PRP	അവൻ
2.2		Reflexive		PRF	PR__PRF	താൻ

2.3		Relative		PRL	PR_PRL	യാതൊരുവൻ
2.4		Reciprocal		PRC	PR_PRC	പരസ്പരം
2.5		Wh-word		PRQ	PR_PRQ	എന്ത്
3	Demonstrative			DM	DM	
3.1		Deictic		DMD	DM_DMD	ആ
3.2		Wh-word		DMQ	DM_DMQ	ഏത്
4	Verb			V	V	
4.1		Main		VM	V_VM	
4.1.1			Finite	VF	V_VM_VF	ചാടി
4.1.2			Non-finite	VNF	V_VM_VNF	ചാടുന്ന
4.1.3			Infinitive	VINF	V_VM_VINF	ചാടാൻ
4.1.4			Gerund	VNG	V_VM_VNG	ചാടുന്നത്
4.2		Verbal Noun	Verbal noun	VNV	V_VM_VNV	ചാട്ടം
5	Adjective			JJ		സുന്ദരമായ
6	Adverb			RB		വേഗത്തിൽ
7	Postposition			PSP		പോലെ
8	Conjunction			CC	CC	
8.1		Co-ordinator		CCD	CC_CCD	ഉം, അഥവാ
8.2		Subordinator		CCS	CC_CCS	എന്നാൽ
8.2.1			Quotative	UT	CC_CCS_UT	എന്ന്
9	Particles			RP	RP	
9.1		Default		RPD	RP_RPD	മാത്രം
9.3		Interjection		INJ	RP_INJ	അയ്യോ
9.4		Intensifier		INTF	RP_INTF	വളരെ
9.5		Negation		NEG	RP_NEG	അല്ല
10	Quantifiers			QT	QT	
10.1		General		QTF	QT_QTF	കുറച്ച്
10.2		Cardinals		QTC	QT_QTC	ഒന്ന്
10.3		Ordinals		QTO	QT_QTO	ഒന്നാം
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	എററററ

Table 2-1: Malayalam 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 MALAYALAM POS TAGGING

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

2.3.1 Annotation Guidelines

- Malayalam PoS Tagging is done based on BIS tagset with customization to capture specific grammatical and morphological features.
- Malayalam 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 semantic, discourse or pragmatic information. In agglutinative languages like Malayalam, Tamil, Telugu and Kannada two or three words are combined together to form a

complex word. Therefore, tokenization is needed because the accuracy of PoS tagged data depends on the context of the words. These words are from different categories or from same category. Such instances it is needed to pre-process the word to get desired category. Otherwise that complex word cannot be tagged with a single PoS tag.

For eg: the following word has three grammatical categories but it is a single token. അവരെപ്പോലെയായിരുന്നു ‘avareppo:leja:jirunnu’

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

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

Token in the original text	Pre processed word	tag	Pre processed word	tag	Pre processed word	tag
യന്ത്രഭാഗംപോലെ janṭrab ^h a:gampo:le	യന്ത്രഭാഗം janṭrab ^h a:gam	N_NN	പോലെ po:le	RP_RPD		
അനുഭവപ്പെട്ടെങ്കിലും anub ^h avappetteṅkilum	അനുഭവപ്പെട്ടു anub ^h avappettu	V_VM_VNF	എങ്കിലും eṅkilum	CC_CCS		
കിടക്കവിട്ടെഴുന്നേറ്റു kiṭakkaviṭṭe:ṭṭu	കിടക്ക kiṭakka	N_NN	വിട്ടു viṭṭa	V_VM_VNF	എഴുന്നേറ്റു e:ṭṭu	V_VM_VF
ആളല്ല a:lalla	ആൾ a:l	N_NN	അല്ല alla	RP_NEG		
സംഭവമുണ്ട് samb ^h avamunṭ	സംഭവം samb ^h avam	N_NN	ഉണ്ട് unṭa	V_VM_VF		
ധനികരെപ്പോലെയല്ല d ^h anikarepo:lejalla	ധനികരെ d ^h anikare	N_NN	പോലെ po:le	RP_RPD	അല്ല alla	RP_NEG
അസ്തിത്വമുണ്ടാകുക astit ^h pad ^h am ^h o:kke	അസ്തിത്വമുണ്ട് astit ^h pad ^h am ^h	N_NN	ഒക്കെ okke	RP_RPD		

Table 2-2: Pre-processed Word List of Malayalam 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 Malayalam word താൻ/ta:n/ is a reflexive

pronoun but sometimes it can act as other category. Following examples where /t̪a:n/ is act as particle and as reflexive pronoun.

In Malayalam			
ഇത്\PR_PRP	ബുദ്ധിയുടെ\N_NN	ഒരു\QT_QTF	തരം\N_NN
മരവിപ്പിന്\V_VM_VNV	വഴി\N_NN	തെളിയിക്കുന്നു\V_VM_VF	താനും\RP_RPD
താനും\PR_PRF പ്രാർത്ഥിക്കു\V_VM_VF .\RD_PUNC			
In IPA			
it̪\PR_PRP	budd̪hijute\N_NN	oru\QT_QTF	t̪aram\N_NN maraviɸɸina\V_VM_VNV
vaɸi\N_NN	t̪el̪ijikkunnu\V_VM_VF	t̪a:nuṁ\RP_RPD	
t̪a:nuṁ\PR_PRF pra:rt̪h̪ikkə\V_VM_VF .\RD_PUNC			

2.3.4 Observation on PoS Tagging of LDC-IL Malayalam 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 Malayalam, 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. മലയാളസാഹിത്യം /malaja:ɸasa:hiɸjam/ and മലയാള മലയാള /malaja:ɸa/ സാഹിത്യം /sa:hiɸjam/; നോക്കി /no:kki/ രസിക്കാൻ

/rasikka:n/ and നോക്കിരസിക്കാൻ /no:kkirasikka:n/ 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.

നോക്കി\V_VM_VNF രസിക്കാൻ\V_VM_VINF and നോക്കിരസിക്കാൻ V_VM_VINF

മലയാളം \J സാഹിത്യം\ N_NN and മലയാളസാഹിത്യം\ N_NN

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

V_VM_VNF & V_VM_VF

In Malayalam
കൂട്ടത്തോടെ\RB ചത്തു\V_VM_VNF വിറങ്ങലിച്ചിരിക്കുന്നു\V_VM_VF .\RD_PUNC
എത്ര\PR_PRQ പേര്\N_NN ചത്തു\V_VM_VF .\RD_PUNC
In IPA
ku:ʈʈatto:te\N_NN ʃattu\V_VM_VNF viranṅgalicʃirikkunnu\V_VM_VF .\RD_PUNC
eṭra\ PR_PRQ pe:r ə\N_NN ʃattu\V_VM_VF .\RD_PUNC

V_VM_VF& PSP

In Malayalam
ചെല്ലൻ\N_NNP മാനിനെ\N_NN പൊക്കിയെടുത്ത്\V_VM_VNF മുപ്പന്റെ\N_NN കാൽക്കൽ\N_NN വച്ചു\V_VM_VF .\RD_PUNC
ഒന്നേക്കാൽ\QT_QTC ഉറപ്പികയും\N_NN വെറ്റിലയടയ്ക്കയും\N_NN വച്ചു\PSP തൊഴണം\V_VM_VF .\RD_PUNC
In IPA
ʃellan\N_NNP ma:nine\N_NN pokkijetuttə\V_VM_VNF mu:ppanɽe\N_NN ka:lkkal\N_NN vatʃʃu\V_VM_VF .\RD_PUNC

onne:ka:l\QT_QTC	uruppikajum\N_NN	ver[ilaja]ajkkajum\N_NN	va[ŋ]u\PSP
tu[ŋ]am\V_VM_VF	. \RD_PUNC		

RP_RPD& PR_PRF

In Malayalam			
ഇത്\PR_PRP	ബുദ്ധിയുടെ\N_NN	ഒരുതരം\QT_QTF	മരവിപ്പിന്\N_NN
വഴി\N_NN	തെളിയിക്കുന്നു\V_VM_VF	താനും\RP_RPD	. \RD_PUNC
താനും\PR_PRF	പ്രാർത്ഥിക്കു\V_VM_VF	. \RD_PUNC	
In IPA			
i[ɪ]tə\PR_PRP	buddʱijute\N_NN	oru t̪aram \QT_QTF	maravippinə\N_NN
va[ɪ]\N_NN	te[ɻ]ikkunnu\V_VM_VF	ta: num\RP_RPD	. \RD_PUNC
ta: num\PR_PRF	pra:rt̪hikkə\V_VM_VF	. \RD_PUNC	

N_NN & N_NNP

In Malayalam			
ഭൂമിയിൽ\N_NN	പൂക്കളും\N_NN	ഫലങ്ങളും\N_NN	കൊണ്ട്\PSP
കവിത\N_NN	രചിക്കുന്നു\V_VM_VNF	നിസാർ\N_NNP	അഹമ്മദിലൂടെ\N_NNP
ദേവദാസിനും\N_NNP	കവിത\N_NNP	പ്രിയ\JJ	പുത്രി\N_NN
ആണ്\V_VM_VF	. \RD_PUNC		
In IPA			
bʱu: miji\N_NN	pu: kka[ɻ]um\N_NN	pʱalaŋŋa[ɻ]um\N_NN	konɕə\PSP
kaviɕa\N_NN	raɕikkunna\V_VM_VNF	nisa: raɕi\N_NNP	pammaɖilu: te\N_NNP
de: vaɖa: sinum\N_NNP	kaviɕa\N_NNP	prija\JJ	puɕtri\N_NN
a: ɳə\V_VM_VF			

N_NST&RB

In Malayalam			
വീണ്ടും\N_NST	ഛാനമായ\JJ	നിശ്ശബ്ദം\N_NN	
എഴുതിയതു\V_VM_VNG	വീണ്ടും\RB	വായിച്ചുനോക്കി\V_VM_VNF	തിരുത്തണം\V_VM_VNF
എന്നു\RP_RPD	തോന്നിയാൽ\V_VM_VNF	തിരുത്തുകയും\V_VM_VINF	ചെയ്യും\V_VM_VF
. \RD_PUNC			
In IPA			
vi: ɳɕum\N_NST	m̪la: nama: ja\JJ	niɕeabɖaɕa\N_NN	

e:ɽɯɟɟatu\V_VM_VNG vi:ɳtʊm\RB va:ɟiɽɟɟuno:kki\V_VM_VNF ɽiruttan̩am\V_VM_VNF ennu\PRP ɽo:nnija:l\V_VM_VNF ɽiruttukajum\V_VM_VINF ɽejjum\V_VM_VF .\RD_PUNC

RP_NEG&RP_RPD

In Malayalam		
വാഹനത്തിന്റെ\N_NN	സഹായം\N_NN	ഒന്നും\QT_QTF
ഇല്ലാതെ\RP_NEG	മുന്നേമുക്കാൽ\QT_QTC	നാഴിക\N_NN
കൊണ്ടു\PSP		താണ്ടാൻ\V_VM_VINF
ദേവതക്കു\N_NN	സാധിക്കും\V_VM_VF	.\RD_PUNC
നിന്നെ\PR_PRP	പോലുള്ള\RP_RPD	ലക്ഷ്മിയെ\N_NNP
വിട്ടു\V_VM_VNF	എവിടെ\PR_PRQ	എന്നു\RP_RPD
തെണ്ടിത്തിരിയുന്നു\V_VM_VF		.\RD_PUNC
In IPA		
va:ɦanattin̩ɽe\N_NN	saha:jam\N_NN	onnum\QT_QTFilla:ɽe\RP_NEG
mu:nne:mukka:l\QT_QTC	na:ɽika\N_NN	kon̩ɽu\PSP ɽa:ɳɽa:n\V_VM_VINF
ɽe:vatakku\N_NN	sa:ɽɪkkum\V_VM_VF	.\RD_PUNC
ninne\PR_PRP	po:lulla\RP_RPD	lak̩ɽmije\N_NNP
enn\RP_RPD	illa:ɽe\RP_RPD	ɽen̩ɽittirijunnu\V_VM_VF
		.\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.

Malayalam	IPA
ജി means ജി.ശങ്കരക്കുറുപ്പ്	/ɽɟi/ means /ɽɟi.ɕaŋkarakkuɽup̩p̩/
ഇ.എം.എസ് means	/i.em.esə/ means
ഇ.എം .ശങ്കരൻനമ്പൂതിരിപ്പാട്	/i.em.ɕaŋkarannampu:ɽirippa:ɽə/

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.

സഫലമാകുമുമ്പേ /sap̩ɦalama:kummumpe:/ the word ‘മുമ്പേ’ /mumpe:/ means ‘before’ comes with verbs so there is a confusion that whether to split the word or consider it as single token. If it separated ‘മുമ്പേ’ is fit for N_NST tag and if it is together verb category can accommodate this word.

അപ്പോൾ /appo:l/ 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:

Malayalam			
അവൾ\PR_PRP	അദ്ഭുതം\N_NN	പ്രകടിച്ചപ്പോൾ\V_VM_VNF	വ്യക്തമായി\V_VM_VF .\RD_PUNC
IPA			
ava\PR_PRP	adbʱutam\N_NN	prakatijʃfappo:l\V_VM_VNF	vjaktama:ji\V_VM_VF .\RD_PUNC

2.3.4.5 Multiword

A sequence of words can act as a single unit, making it difficult to assign a single category. Malayalam is an agglutinative language, so multiple words are often joined together to form a single token. In Malayalam, 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:

Malayalam	
കഴിഞ്ഞകാര്യത്തെക്കുറിച്ചാർത്ത് → കഴിഞ്ഞJJകാര്യത്തെ\N_NN	കുറിച്ച്\PSP ഓർത്ത്\V_VM_VNF
IPA	
ka:ɻɪnnaka:ɻattekku:ɻiʃfo:rtʃə → ka:ɻɪnnaJJka:ɻatte\N_NN ku:ɻiʃʃə\PSP o:rtʃə\V_VM_VNF	

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: കാറ്റടിച്ചു /ka:ɻatʃiʃʃu/, മഴനനഞ്ഞു /ma:ʃananaɻɳu/. 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 Malayalam, the word അക്ക് /ɳakkə/ ‘to you’ is using in Malabar region. Its standard form is ‘നിങ്ങൾക്ക്’ /niɳaɻkkə/. The words ഓൻ /oon/ ‘he’ and ഓൾ /ool/ ‘she’ are using instead of ‘avan’ and ‘avaL’. 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.

2.3.4.8 Impact of Social Dialect

The word ‘മുസിബത്ത’ /musi:battə/ means ‘trouble’; it is an Arabic word and it is nativised by Muslims so this word tagged as N_NN category. In case the annotator unaware about the term it may tag as RD_UNK.

2.3.4.9 Verbal Noun

The word which denoting result of action is tagged as verbal noun category other type of verbal nouns are tagged as noun.

2.3.4.10 Unknown Words

There will be some words in the text which cannot be handled by the tagset. Words from other languages written in the Malayalam script that are not part of the Malayalam 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 MALAYALAM POS DATA

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

Dataset Summary	Aggregate Count
Total PoS Tags	1322728
Total Sentences	149781
Total Word Count	1089199
Total File Count	1622

Table 2-3: Malayalam PoS Status

Sl. No.	PoS Category	PoS Tag	Tag Count
1	Common Noun	N_NN	355148
2	Proper Noun	N_NNP	74854
3	Nloc Noun	N_NST	34012
4	Personal Pronoun	PR_PRP	79945
5	Reflexive Pronoun	PR_PRF	6600
6	Relative Pronoun	PR_PRL	8
7	Reciprocal Pronoun	PR_PRC	1108
8	Wh-word Pronoun	PR_PRQ	7224
9	Deictic Demonstrative	DM_DMD	14365
10	Wh-word Demonstrative	DM_DMQ	2221
11	Finite Verb	V_VM_VF	140246
12	Non-finite Verb	V_VM_VNF	114840

13	Infinitive Verb	V_VM_VINF	22921
14	Gerund	V_VM_VNG	26396
15	Verbal Noun	V_VM_VNV	120
16	Adjective	JJ	47217
17	Adverb	RB	21906
18	Postposition	PSP	25712
19	Co-ordinator	CC_CCD	3998
20	Subordinator	CC_CCS	12647
21	Quotative	CC_CCS_UT	427
22	Particles	RP_RPD	29673
23	Interjection	RP_INJ	908
24	Intensifier	RP_INTF	1747
25	Negation	RP_NEG	1698
26	Quantifiers	QT_QTF	42946
27	Cardinals	QT_QTC	11758
28	Ordinals	QT_QTO	2107
29	Foreign word	RD_RDF	3772
30	Symbol	RD_SYM	429
31	Punctuation	RD_PUNC	233100
32	Unknown	RD_UNK	2621
33	Echowords	RD_ECH	54

Table 2-4: Distinct Tag Count of Malayalam