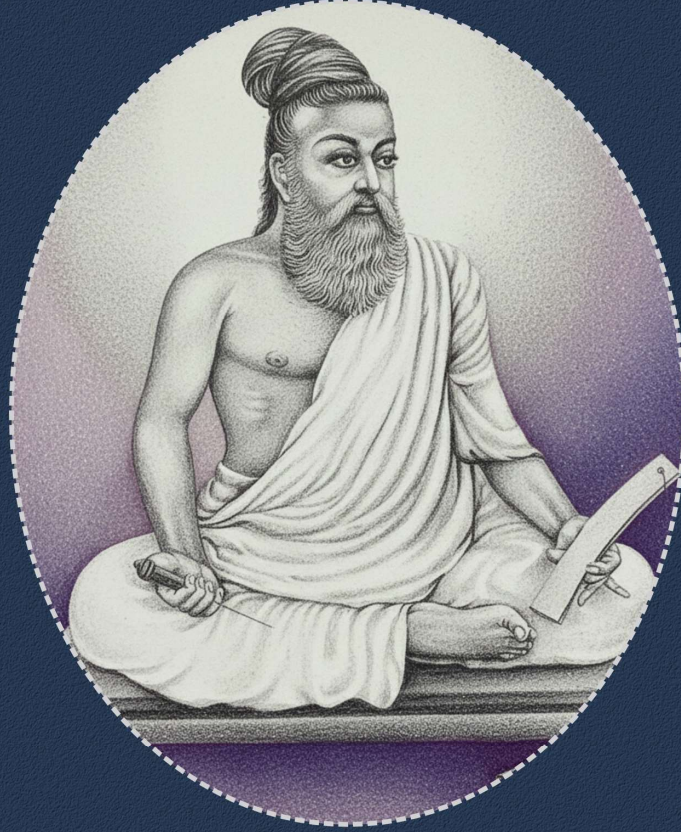


தமிழ்ச்
சொல்வகைப்பாட்டு
விளக்கத் தரவகம்



TAMIL PARTS OF SPEECH
ANNOTATED CORPUS

TAMIL 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. Amudha R

Dr. Kamaraj.S

Dr. Prem Kumar L. R

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

Tamil Parts of Speech Annotated Corpus

Authors: Dr. Amudha R, Dr. Kamaraj S, Dr. Prem Kumar L. R., Dr. Narayan Choudhary

e-ISBN: 978-81-69175-98-2

CIIL Publication No.: 1693

© 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: Amudha R.

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	Tamil PoS Annotated Corpus	Error! Bookmark not defined.
2.1	Grammatical peculiarities of Malayalam	Error! Bookmark not defined.
2.2	LDC-IL Customized BIS Tagset	7
2.3	Tamil PoS Tagging	10
2.4	Summary of LDC-IL Tamil PoS Data	18

Tables

Table 1-1: BIS Tagset	4
Table 1-2: LDC-IL PoS Status	5
Table 2-1: Tamil Tagset	7
Table 2-2: Pre-processed word list of Tamil with IPA	Error! Bookmark not defined.
Table 2-3: Tamil PoS Status	18
Table 2-4: Disntict Tag Count of Tamil	18

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 VNF	
4.2.3			Infinitive	VINF	V VAUX VINF	
4.2.4			Gerund	VNG	V VAUX VNG	
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 Mishra 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 213125 tags. The summary of LDC-IL parts of speech data is given below.

Dataset Summary	Aggregate Count
Total PoS Tags	2131256
Total Sentences	172089
Total Word Count	1750935
Total File Count	2934

Table 1-2: LDC-IL PoS Status

2 TAMIL PoS ANNOTATED CORPUS

Amudha R

2.1 GRAMMATICAL PECULIARITIES OF TAMIL

Tamil is one of the Dravidian languages and one of the longest surviving languages in the world. It has a great classical literature which has been documented for over 2000 years. Tamil is morphologically a very rich language. The morphological complexity and the complexity in the word structure make the tagging of grammatical information in the word tokens very complex. The word structure is generally a root with or without one or more affixes.

The process of POS tagging in such a morphologically rich language like Tamil is very complex. We can seek the guidance from traditional grammars like Tolkappiyam and Nunnulas well the modern linguistics based grammars for fixing the POS tag sets for Tamil. The POS tagging is necessary part of all the NLP applications. Part of Speech tagging is the processes of marking a word in a corpus or text with its grammatical category. The major grammatical categories in Tamil are noun, adjective, pronoun, verb, adverb, conjunction, and particle. The number of POS tags in a tag set depends on how fine grained we want the tags to be. To make a POS tagger for Tamil language is very challengeable. Tagging resolves the categorical ambiguity in the corpus or text. As we said already, tagging a Tamil corpus or text is a challenge. It expects a strong grammatical knowledge from the POS tag annotators.

The draft Bureau of Indian Standards (BIS) has recommended the use of a uniform tag set for the POS annotation of Indian languages. This tag set was created based on wider discussions by linguistics and language experts including computer scientist.

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

Some tags in the BIS framework, such as indefinite pronouns and classifiers, are rare or absent in grammatical structure of Tamil. Therefore, LDC-IL customized the tagset to accomodate the specific needs of each language. The Tamil 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	புத்தகம், மரம் puttakam, maram
1.2		Proper		NNP	N__NNP	அகரன், சென்னை akaran', cen'n'ai
1.3		Nloc		NST	N__NST	மேல், பக்கம் mEl, pakkam
2	Pronoun			PR	PR	
2.1		Personal		PRP	PR__PRP	அவர், நீ avar, nI
2.2		Reflexive		PRF	PR__PRF	தான் tAn'
2.3		Relative		PRL	PR__PRL	
2.4		Reciprocal		PRC	PR__PRC	ஒருவருக்கொரு வர் orugarukkoruvar
2.5		Wh-word		PRQ	PR__PRQ	யார். என்ன yAr. en'n'a
3	Demonstrative			DM	DM	
3.1		Deictic		DMD	DM__DMD	அந்த, இந்த anta, inta
3.2		Relative		DMR	DM__DMR	எந்த enta
3.3		Wh-word		DMQ	DM__DMQ	யாராவது, எதாவது

						yArAvatu, etAvatu
4	Verb			V	V	
4.1		Main		VM	V__VM	வா, சிரி, படி vA, ciri, paTi
4.1.1			Finite	VF	V__VM__VF	வந்தார், படித்தாள் vantAr, paTittAL
4.1.2			Non-finite	VNF	V__VM__VNF	வந்த, சென்றால் vanta, cen'RAI
4.1.3			Infinitive	VINF	V__VM__VINF	வர, போக vara, pOka
4.1.4			Gerund	VNG	V__VM__VNG	வருதல், ஓடுதல் varutal, OTutal
4.2		Verbal Noun	Verbal noun	VNV	V__VM__VNV	வருபவர்கள், படிப்பவர்கள் varupavarkaL, paTippavarkaL
4.3		Auxiliary		VAUX	V__VAUX	ஆகும், முடியும் Akum, muTiyum
			Finite	VF	V__AUX__VF	வேண்டும் vENTum
			Non-finite	VNF	V__VM__VNF	பார்த்துகொண்டு pArttukoNTu
			Infinitive	VINF	V__VM__VINF	எடுத்துப் பார்க்க வேண்டும் eTuttup pArkka vENTum
5	Adjective			JJ		நல்ல, அழகான nalla, aZakAn'a
6	Adverb			RB		வேகமாக, அவ்வாறு vEkamAka, avvARu
7	Postposition			PSP		பற்றி, குறித்து paRRi, kuRitt
8	Conjunction			CC	CC	
8.1		Co-ordinator		CCD	CC__CCD	மற்றும், அதனால்

						maRRum, atan'Al
8.2		Subordinator		CCS	CC__CCS	ஆனால், என்ற An'Al, en'Ra
8.2.1			Quotative	UT	CC__CCS__UT	என்று en'Ru
9	Particles			RP	RP	
9.1		Default		RPD	RP__RPD	மட்டும், கூட maTTum, kUTa
9.2		Classifier		CL	RP CL	-
9.3		Interjection		INJ	RP__INJ	அய்யோ, ஆகா ayyO, AkA
9.4		Intensifier		INTF	RP__INTF	அதிக, மிக atika, mika
9.5		Negation		NEG	RP__NEG	இல்லை, அல்ல illai, alla
10	Quantifiers			QT	QT	
10.1		General		QTF	QT__QTF	கொஞ்சம், நிறைய konj'cam, niRaiya
10.2		Cardinals		QTC	QT__QTC	ஒன்று, இரண்டு on'Ru, iraNTu
10.3		Ordinals		QTO	QT__QTO	முதலாம், ஏழாம் mutalAm, EZAm
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	க்ரூஸ்விங் krUsving'
11.5		Echowords		ECH	RD__ECH	பாலுகீலு, மாடுகீடு pAlukIlLu, mATukITu
11.6		Reduplication		RDP		கடகட, பலபல kaTakaTa, palapala

Table 2-1: Tamil 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 TAMIL PoS TAGGING

LDC-IL developed a customized Malayalam PoS tagset with the necessary adaptations to the BIS tagset to suit Tamil's linguistic structure. Since Tamil 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

- Tamil PoS Tagging is done based on BIS tagset with customization to capture specific grammatical and morphological features.
- Tamil 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 Tamil, Malayalam, 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.
அவருக்காகக்காத்திருந்தவர்கள் /‘avarukkAkakkAttiruntavarkaL’/

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

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

Token in the original text	Pre processed word	tag	Pre processed word	tag	Pre processed word	tag
தமிழ்மொழி tamiZmoZi	தமிழ் tamiZ	N_NN	மொழி moZi	N_NN		
பெரியவீடு periyavITu	பெரிய periya	JJ	வீடு vITu	N_NN		
அந்தப்புத்தகம் antapputtakam	அந்த anta	DM_DMD	புத்தகம் puttakam	N_NN		
போகவேண்டும் pOkavENTum	போக pOka	V_VM_VINF	வேண்டும் vENTum	V_VAUX_VF		
பார்த்துபேசு	பார்த்து	V_VM_VNF	பேசு	V_VM_VF		
திரும்பிப்பார்த்துக்கொண்டு tirumpippArttukkoNTu	திரும்பிப் tirumpipp	V_VM_VNF	பார்த்துக் pArttuk	V_VM_VNF	கொண்டு koNTu	V_VAUX_VNF

அவன்கூடவந்தான் avan'kUTavantAn'	அவன் avan'	PR_PRP	கூட kUTa	RP_RPD	வந்தான் vantAn'	V_VM_VF
ஆதிரன்போல் Atiran'pOl	ஆதிரன் Atiran'	N_NNP	போல் pOl	PSP		
புத்தகம்மேலேயில்லை puttakammElEyillai	புத்தகம் puttakam	N_NN	மேலே mElE	N_NST	இல்லை illai	RP_NEG
அவளல்ல avaLalla	அவள் avaL	N_NN	அல்ல alla	RP_NEG		

Table 2-2: Pre-processed word list of Tamil 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 Tamil word தான்/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 Tamil	
அகரன்\N_NNP தான்\PR_PRF படித்தப்\V_VM_VNF பள்ளியில்\N_NN	
தான்\RP_RPD ஆசிரியராக\RB பணியாற்றுகிறான்\V_VM_VF	
தான்\PR_PRF பார்க்கவில்லை\V_VM_VF என்று\CC_CCS_UT	
கூறினான்\V_VM_VF	
In IPA	
akaran'\N_NNP tAn' \PR_PRF paTittap\V_VM_VNF paLLiyil\N_NN tAn'\RP_RPD	
AciriyarAka\RB paNiyARRukiRAn'\V_VM_VF	
tAn'\PR_PRF pArkkavillai\V_VM_VF en'Ru\CC_CCS_UT kURin'An'\V_VM_VF	

2.3.4 Observation on PoS Tagging of LDC-IL Tamil 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

The same word appeared differently in the corpus. For e.g. in Tamil, 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. தமிழ்க்கலைக்களஞ்சியம்\N_NNP 'tamiZkkalaikkaLanj'ciyam/

Tamil Encyclopedia and தமிழ்க் கலைக்களஞ்சியம்\N_NN '/tamiZk kalaikkaLanj'ciyam/' 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.

அவன்\PR_PRP பள்ளியை\N_NN நோக்கி\PSP நடக்கிறான்\V_VM_VF
.\RD_PUNC. /avan'\PR_PRP paLLiyai\N_NN nOkki\PSP naTakkiRAn'\V_VM_VF
.\RD_PUNC./ and

அவன்\PR_PRP வரைபடத்தைக்\N_NN கூர்ந்து\V_VM_VNF நோக்கினான்.
/avan'\PR_PRP varaipaTattaik\N_NN kUrntu\V_VM_VNF nOkkin'An'\V_VM_VF
.\RD_PUNC.

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

N_NN & N_NNP

In Tamil
இது\PR_PRP வெள்ளி\N_NN மோதிரம்\N_NN .\RD_PUNC
இன்று\N_NST வெள்ளி\N_NNP கிழமை\N_NNP .\RD_PUNC
In IPA
itu\PR_PRP veLLi\N_NN mOtiram\N_NN .\RD_PUNC
in'Ru\N_NST veLLi\N_NNP kiZamai\N_NNP .\RD_PUNC

N_NN & V_VM_VF

In Tamil
வீட்டின்\N_NN படி\N_NN உயரமாக\RB உள்ளது\V_VM_VF .\RD_PUNC

பாடத்தை\N_NN படி\V_VM_VF .RD_PUNC
In IPA
vITTin\N_NN paTi\N_NN uyaramAka\RB uLLatu\V_VM_VF .RD_PUNC
pATattai\N_NN paTi\V_VM_VF .RD_PUNC

RP_RPD & PR_PRF

In Tamil
இன்று\N_NST தான்\RP_RPD மழை\N_NN பெய்தது\V_VM_VF .RD_PUNC
அமுதா\N_NNP தான்\PR_PRF வாங்கிய\V_VM_VNF புத்தகத்தைப்\N_NN படித்தாள் V_VM_VF .RD_PUNC
In IPA
in'Ru\N_NST tAn\RP_RPD maZai\N_NN peytatu\V_VM_VF .RD_PUNC
amutA\N_NNP tAn\PR_PRF vAng'kiya\V_VM_VNF puttakattaip\N_NN paTittAL\V_VM_VF .RD_PUNC

N_NN & V_VM_VF

In Tamil
நீ\PR_PRP அங்கே\N_NST ஆடு\V_VM_VF .RD_PUNC
ஆடு\N_NN அங்கே\N_NST இருக்கிறது\V_VM_VF .RD_PUNC
In IPA
n\PR_PRP ang'kE\N_NST ATu\V_VM_VF .RD_PUNC
ATu\N_NN ang'kE\N_NST irukkiRatu\V_VM_VF .RD_PUNC

N_NST & RB

In Tamil
மேஜையின்\N_NN மேல்\N_NST புத்தகம்\N_NN உள்ளது\V_VM_VF .RD_PUNC
புத்தகம்\N_NN பத்துக்கு\QT_QTC மேல்\RB கொடு\V_VM_VF .RD_PUNC
In IPA

mEjaiyin\N_NN mEl\N_NST puttakam\N_NN uLLatu\V_VM_VF .\RD_PUNC
puttakam\N_NN pattukku\QT_QTC mEl\RB koTu\V_VM_VF .\RD_PUNC

PSP & V_VM_VNF

In Tamil			
காட்டில்\N_NN	விலங்குகள்\N_NN	தண்ணீர்காக\N_NN	
அலைந்து\V_VM_VNF	சுற்றி\V_VM_VNF	திரிந்தன\V_VM_VF	.\RD_PUNC
கோட்டையைச்\N_NN	சுற்றி	மதில்சுவர்\N_NN	உள்ளன\V_VM_VF .\RD_PUNC
In IPA			
kATTi\N_NN	vilang'kukaL\N_NN	taNNIrka\N_NN	alaintu\V_VM_VNF
cuRRi\V_VM_VNF	tirintan'a\V_VM_VF	.\RD_PUNC	
kOTTaiyaic\N_NN	cuRRi matilcuvar\N_NN	uLLan'a\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.

Tamil	IPA
ரூ. means ரூபாய்	/rU./ means /rUpAy/
டாக்டர்.ஏபிஜே அப்பதுல்கலாம்	means /TAKTar.EpijE/ means /TAKTar.EpijE. appatulkalAm/

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.

வெற்றியடையுமுன்பே /veRRiyaTaiyumun'pE/ the word ‘முன்பே’ /mun'pE/ 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 ‘முன்பே’ is fit for N_NST tag and if it is together verb category can accommodate this word.

அப்புறம் /appuRam/ 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:

Tamil
வீட்டின்\N_NN மேல்\N_NST கோழி\N_NN கூவியது\N_NN VF
IPA
vITTin'\N_NN mEl\N_NST kOZi\N_NN kUviyatu\N_NN VF

2.3.4.5 Multiword

A sequence of words can act as a single unit, making it difficult to assign a single category. Tamil is an agglutinative language, so multiple words are often joined together to form a single token. In Tamil, any category can attach to another category. For eg: N_NN + V_VM_VNV + V_VM_VNF + RP_RPD, V_VM_VNF + V_VAUX_VINF + V_VAUX_VNF 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:

Tamil
சுயாதீனத்தன்மையற்றவர்களாயிருக்கின்றபடியால் → சுயாதீனத்தன்மை \N_NN யற்றவர்களா\N_NN V_VM_VNV யிருக்கின்ற\N_NN V_VM_VNF படியால்\RP_RPD
cuyAtIn'attan'maiyaRRavarkaLAyirukkin'RapaTiyAl -cuyAtIn'attan'mai \N_NN yaRRavarkaLA\N_NN V_VM_VNV yirukkin'Ra\N_NN V_VM_VNF paTiyAl\RP_RPD
நினைத்துப்பார்க்கமுடியாதவைகளாகிவிட்டதனால் → நினைத்துப்\N_NN V_VM_VNF பார்க்க\N_NN V_VAUX_VINF முடியாதவைகளாகிவிட்டதனால்\N_NN V_VAUX_VNF
nin'aittuppArkkamuTiyAtavaikaLAKiviTTatan'Al → nin'aittup\N_NN V_VM_VNF pArkka\N_NN V_VAUX_VINF muTiyAtavaikaLAKiviTTatan'Al\N_NN V_VAUX_VNF
படிக்கவைக்கப்பட்டிருக்கமாட்டார்களாலும்கூட → படிக்க\N_NN V_VM_VNF வைக்கப் \N_NN V_VAUX_VINF பட்டிருக்க \N_NN V_VAUX_VNF மாட்டார்களாலும்\N_NN V_VM_VNV கூட\RP_RPD
paTikkavaikkappaTTirukkamATTArkaLAlumkUTa → paTikka\N_NN V_VM_VNF vaikkap \N_NN V_VAUX_VINF paTTirukka \N_NN V_VAUX_VINF mATTArkaLAlum\N_NN V_VM_VNV

kUTa\RP_RPD

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: **மாட்டுவண்டி** /mATTuvaNTi/, **விளையாட்டுமைதானம்** /viLaiyATTumaitAn'am/. These types of words are taken as such and assigned the appropriate tag.

2.3.4.7 Dialect Variation

If the annonator is unaware about the dialect variation it may lead to incorrect tagging. In Tamil, the word **‘நீங்க எங்குட்டு போறீங்க?’** /nIng'ka eng'yUTTtu pORIng'ka?/ ‘Where are you going?’ is using in Kongu region. Its standard form is **‘நீங்கள் எங்கே போகிறீர்கள்?’** /nIng'kaL eng'kE pOkIRIrkaL?/. The words **நீங்க** /nIng'ka/ ‘you’, **எங்குட்டு** /eng'yUTTtu/ ‘Where’ and **போறீங்க?** /pORIng'ka?/ ‘going?’ are using. 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

In Tamil, a social dialect (sociolect) isn't just about different words; it involves distinct shifts in phonology (sound), morphology (word structure), and vocabulary based on class, caste, or urban-rural divide. "While the standard Tamil phrase **என்ன செய்கிறாய்?** / en'n'a ceykiRAY/ translates to 'What are you doing?', shifts into the Chennai sociolect as **‘என்ன பா பண்ணுற?’** /en'n'a pA paNNuRa?/. This colloquialism incorporates the particle 'ba'—a local adaptation of the Urdu/Hindustani term *bhai*—and is uniquely classified as a common noun (N_NN) in specific linguistic tagging frameworks, despite its functional role as a full interrogative

2.3.4.9 Verbal Noun

Verbal Nouns are derived from verbs. Basically, they function as a common noun and they can be modified by an adjective. Earlier Verbal Nouns are tagged under the category of Verbs. In Tamil, the words such as **படிப்பவர்கள்** /paTippavarkaL/, **வருபவர்கள்** /varupavarkaL/ are considered as Verbal Nouns.

- **அலை அடித்தபோது ஆரம்பமானது** /alai aTittappOtu ‘Arampan-
NN An'a-JJ ‘tu’/
‘It started when the wave hit’

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 Tamil script that are not part of the Tamil lexicon are tagged as unknown. Such words are considered as unknown words and it will pose problem for the higher level of processing. க்ருஸ்விங் /krUsving/

2.4 SUMMARY OF LDC-IL TAMIL PoS DATA

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

Dataset Summary	Aggregate Count
Total PoS Tags	2131256
Total Sentences	172089
Total Word Count	1750935
Total File Count	2934

Table 2-3: Tamil PoS Status

Sl. No.	PoS Category	PoS Tag	Tag Count
1	Common Noun	N_NN	660814
2	Proper Noun	N_NNP	145852
3	Nloc Noun	N_NST	36251
4	Personal Pronoun	PR_PRP	106685
5	Reflexive Pronoun	PR_PRF	3720
6	Relative Pronoun	PR_PRL	148
7	Reciprocal Pronoun	PR_PRC	290
8	Wh-word Pronoun	PR_PRQ	13439
9	Deictic Demonstrative	DM_DMD	16796
10	Wh-word Demonstrative	DM_DMQ	1702
11	Finite Verb	V_VM_VF	164455
12	Non-finite Verb	V_VM_VNF	199552
13	Infinitive Verb	V_VM_VINF	35449
14	Gerund	V_VM_VNG	7519
15	Verbal Noun	V_VM_VNV	43439
16	Auxiliary Finite Verb	V_VAUX_VF	24746
17	Auxiliary Non-finite Verb	V_VAUX_VNF	8994
18	Auxiliary Infinitive Verb	V_VAUX_VINF	602
19	Adjective	JJ	48390
20	Adverb	RB	45859
21	Postposition	PSP	23022
22	Co-ordinator	CC_CCD	11907

23	Subordinator	CC CCS	36808
24	Quotative	CC CCS UT	11187
25	Particles	RP RPD	15209
26	Interjection	RP INJ	758
27	Intensifier	RP INTF	4090
28	Negation	RP NEG	5907
29	Quantifiers	QT QTF	15752
30	Cardinals	QT QTC	44740
31	Ordinals	QT QTO	6162
32	Foreign word	RD RDF	9532
33	Symbol	RD SYM	5177
34	Punctuation	RD PUNC	375144
35	Unknown	RD UNK	832
36	Echowords	RD ECH	327

Table 0-1: distinct Tag Count of Tamil