

ಕನ್ನಡ ಪದವರ್ಗಗಳ ಅನುಲೇಖಿತ ದತ್ತಾಂಶ



KANNADA
PARTS OF SPEECH
ANNOTATED CORPUS

KANNADA 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. Vijayalaxmi F Patil

Chetan Baji

Dr. Narayan Choudhary

Rajesh N.

**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

Kannada Parts of Speech Annotated Corpus

Authors: Dr. Vijayalaxmi F Patil, Chetan Baji, Dr. Narayan Choudhary, Rajesha N.

e-ISBN: 978-81-69175-32-6

CIIL Publication No.: 1702

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: Dr. Vijayalaxmi F Patil

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

Tables

Table 1-1: BIS Tagset	4
Table 1-2: LDC-IL PoS Status	5
Table 2-1: Kannada Tagset	8
Table 2-2: Pre-processed word list of Kannada with IPA	10
Table 2-3: Kannada PoS Status	15
Table 2-4: Distinct Tag Count of Kannada	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 hierarchichel tagset which has 3 level of tags. The first level represents the grammatical category, encompassing 11 primary tags. The second level 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 KANNADA POS ANNOTATED CORPUS

Dr. Vijayalaxmi F Patil

2.1 GRAMMATICAL PECULARITIES OF KANNADA

Kannada is a morphologically rich language characterized by extensive word formation processes. Words in Kannada 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 Kannada 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 Kannada, 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 Kannada. Therefore, LDC-IL customized the tagset to accomodate the specific needs of each language. The Kannada 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	ಮರ
1.2		Proper		NNP	N__NNP	ರಮೇಶ
1.3		Nloc		NST	N__NST	ಇಲ್ಲಿ, ಅಲ್ಲಿ
2	Pronoun			PR	PR	
2.1		Personal		PRP	PR__PRP	ನಾನು

2.2		Reflexive		PRF	PR__PRF	ತಾನೇ
2.3		Relative		PRL	PR__PRL	-
2.4		Reciprocal		PRC	PR__PRC	ಪರಸ್ಪರ
2.5		Wh-word		PRQ	PR__PRQ	ಏನು
3	Demonstrative			DM	DM	
3.1		Deictic		DMD	DM__DMD	ಆ, ಈ
3.2		Wh-word		DMQ	DM__DMQ	ಯಾವ
4	Verb			V	V	
4.1		Main		VM	V__VM	
4.1.1			Finite	VF	V__VM__VF	ಹೋದನು
4.1.2			Non-finite	VNF	V__VM__VNF	ಬಂದು
4.1.3			Infinitive	VINF	V__VM__VINF	ಹೇಳಲು
4.1.4			Gerund	VNG	V__VM__VNG	ತರುವುದು
4.2		Verbal Noun	Verbal noun	VNV	V__VM__VNV	ಎಂಬುವವನು
5	Adjective			JJ		ಒಳ್ಳೆಯ
6	Adverb			RB		ಬೇಗನೆ
7	Postposition			PSP		ಕುರಿತು
8	Conjunction			CC	CC	
8.1		Co-ordinator		CCD	CC__CCD	ಮತ್ತು
8.2		Subordinator		CCS	CC__CCS	ಏಕೆಂದರೆ
8.2.1			Quotative	UT	CC__CCS__UT	ಎಂದು
9	Particles			RP	RP	
9.1		Default		RPD	RP__RPD	ತಾನೇ
9.3		Interjection		INJ	RP__INJ	ಅಯ್ಯೋ
9.4		Intensifier		INTF	RP__INTF	ಬಹಳ

9.5		Negation		NEG	RP__NEG	ಅಲ್ಲ
10	Quantifiers			QT	QT	
10.1		General		QTF	QT__QTF	ತುಂಬ
10.2		Cardinals		QTC	QT__QTC	ಒಂದು
10.3		Ordinals		QTO	QT__QTO	ಒಂದನೆಯ
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	Sanskrit shlokas (in Kannada)
11.5		Echowords		ECH	RD__ECH	ಉಟ-ಗೀಟ

Table 2-1: Kannada 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 KANNADA PoS TAGGING

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

- Kannada PoS Tagging is done based on BIS tagset with customization to capture specific grammatical and morphological features.
- Kannada 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 Kannada, Tamil, Telugu 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. ಒಳಪಂಗಡಗಳುಂಟೆಂಬಂಶ 'o[lapaṅgaḍaga[luṅtɐmbaɱʃa'

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

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

Token in the original text	Pre processed word	tag	Pre processed word	tag	Pre processed word	tag
ವಿಗ್ರಹವೆಂದು vigrāhāvendu	ವಿಗ್ರಹ vigrāha	N_NN	ಎಂದು endu	CC_CCS_UT		
ಹೆಸರಿಲ್ಲದ hesarillaḍa	ಹೆಸರು hesaru	N_NN	ಇಲ್ಲದ illaḍa	V_VM_VNF		
ಉಳಿಗಾಲವಿಲ್ಲವೆಂದು uligā:lavuillaṇdu	ಉಳಿಗಾಲ uligā:la	N_NN	ಇಲ್ಲ illa	V_VM_VF	ಎಂದು endu	CC_CCS_UT
ಮುಖ್ಯವಲ್ಲ mukh'javalla	ಮುಖ್ಯ mukh'ja	JJ	ಅಲ್ಲ alla	V_VM_VF		
ಪ್ರತಿಭೆಯಿದ್ದರೂ pratibʰeɽiddaru:	ಪ್ರತಿಭೆ pratibʰe	N_NN	ಇದ್ದರೂ iddaru:	V_VM_VNF		
ಸಮತೋಲನವಿದೆ	ಸಮತೋಲನ	N_NN	ಇದೆ	V_VM_VF		

samaʈo:lanavide	samaʈo:lanā		ide			
ನಾನಿಲ್ಲಿದ್ದರೆ na:nilliddare	ನಾನು na:nu	PR_PRP	ಇಲ್ಲಿ illi	N_NST	ಇದ್ದರೆ iddare	V_VM_VNF

Table 2-2: Pre-processed word list of Kannada 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 Kannada word ಆಯ್ತು /a:jʈu/ is a main verb but sometimes it can act as other category. Following examples where /a:jʈu/ is act as verb and as particle.

In Kannada
ಅವರದು\PR_PRP ಊಟ\N_NN ಆಯ್ತು \V_VM_VF .RD_PUNC
ಆಯ್ತು \RP__RPD , .RD_PUNC ನೀವು\PR_PRP ಮನೆಗೆ\N_NN ಹೋಗಿ\V_VM_VF .RD_PUNC
In IPA
avaradu\PR_PRP u:ʈa\N_NN a:jʈu\V_VM_VF
a:jʈu\RP__RPD ni:vu\PR_PRP manege\N_NN ho:gi\V_VM_VF

2.3.4 Observation on PoS Tagging of LDC-IL Kannada 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 Kannada, 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. ಕನ್ನಡ\N_NNP ಭಾಷೆ \N_NNP, ನೇಪಾಳಿ\JJ ಹುಡುಗ \N_NN, ಪಂಜಾಬಿ\JJ ಊಟ\N_NN

Some of the Kannada words have single token with single tag like ಮಹಾರಾಜ \N_NN, but when the word ಮಹಾ occurs separately with another token the same word will be tagged as different category, for example – ಮಹಾ \RP_INTF ದಿಟ್ಟ\JJ ಮನುಷ್ಯ\N_NN

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

ಸಾರುಪ್ರಾಣಿ\N_NN - ಸಾರು\V_VM_VF ಪ್ರಾಣಿ\N_NN

ಇನ್ನೊಮ್ಮೆ\RB - ಇನ್ನೂ\RP__INTF ಒಮ್ಮೆ\RB ಹೇಳಿ\V_VM_VF

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

N__NN, V__VM__VF

In Kannada
ಬೆಳೆ\N__NN ಮಳೆಯಿಂದ\N__NN ನಾಶವಾಯಿತು\V__VM__VF .\RD_PUNC
ಮಕ್ಕಳಿಗೆ\N__NN ಪೌಷ್ಟಿಕ\JJ ಆಹಾರವನ್ನು\N__NN ನೀಡದಿದ್ದರೆ\V__VM__VNF ಸರಿಯಾಗಿ\RB ಬೆಳೆಯುವುದಿಲ್ಲ\V__VM__VF .\RD_PUNC
In IPA
beɭe\N__NN maɭejinda\N__NN na:ʃava:jiɽu\N__NN
makkalige\N__NN pavʃʰika\JJ a:ha:raɪvannu\N__NN ni:ɖaɖiddare\N__NN sarija:gi\RB beɭejuruɖilla\N__NN

N_NNP & N__NN

In Kannada
ಕಮಲ\N_NNP ಶಾಲೆಗೆ\N__NN ಹೋಗಿದ್ದಾಳೆ\N__NN
ಲಕ್ಷ್ಮಿ ದೇವಿಗೆ\N_NNP ಕಮಲ\N__NN ಶ್ರೇಷ್ಠ\JJ
In IPA
kamala\N_NNP ʃa:lege\N__NN ho:giɖɖa:ɭe\N__NN
lakʃmiɖe:viɖe\N_NNP kamala\N__NP ʃre:ʃʰa\N__NN

PSP & QT__QTF

In Kannada
ಮನೆಯ\N__NN ತುಂಬಾ\N__NN ಸಾಮಾನುಗಳು\N__NN ಇವೆ\N__NN
ಅಲ್ಲಿ\N__NST ತುಂಬಾ\JJ ಸೊಳ್ಳೆಗಳು\N__NN ಇವೆ\N__NN
In IPA
maneja\N__NN ɽumba:ʌPSP sa:ma:nugaɭu\N__NN ive\N__NN
alli\N__NST ɽumba:ʌJJ solɭegaɭu\N__NN ive\N__NN

N_NNP & RB

In Kannada
ಸುಂದರ\N__NN ಏಕೆ\PR__PRQ ಇನ್ನೂ\N__NST ಬಂದಿಲ್ಲ\N__NN ?\RD_PUNC

ಅ\DM__DMD ಮನೆಯನ್ನು\N_NN ಸುಂದರವಾಗಿ\RB ಕಟ್ಟಿಸಿದ್ದಾರೆ\V_VM_VF .\RD_PUNC
In IPA
sundara\N_NN e:ke\PR__PRQ innu:\N_NST bandilla\V_VM_VF
a:\DM__DMD manejanu\N_NN sundarava:gi\RB kattisidda:re\V_VM_VF

RB & RP__RPD

In Kannada
ಮಗು\N_NN ಸರಿಯಾಗಿ\RB ಬರೆ\V_VM_VF .\RD_PUNC
ಸರಿ\RP__RPD ,\RD_PUNC ಈಗ\N_NST ನೀವು\PR__PRP ಹೋಗಿ\V_VM_VF .\RD_PUNC
In IPA
magu\N_NN sarija:gi\RB bare\V_VM_VF .\RD_PUNC
sari\RP__RPD ,\RD_PUNC i:ga\N_NST ni:vu\PR__PRP ho:gi\V_VM_VF

JJ & RB

In Kannada
ಅ\DM__DMD ಹುಡುಗ\N_NN ಮಂದ\N_NN ಬುದ್ಧಿಯವನು\N_NN .\RD_PUNC
ಮಜ್ಜಿಗೆ\N_NN ಮಂದವಾಗಿ\RB ಇದೆ .\RD_PUNC
In IPA
a:\DM__DMD huduga\N_NN manda\N_NN budd ^h ijavanU\N_NN
madʒdʒigE\N_NN mandava:gi\RB ide\V_VM_VF

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.

Kannada	IPA
ಡಾ. ಗಂಗೂಬಾಯಿ ಹಾನಗಲ್	/d̪aː gaŋguːbaːji haːnaɡal/
ಟಿ.ಬಿ. ಚಂದ್ರ	/ʈi bi tʃand̪ra/

2.3.4.4 Multiword

A sequence of words can act as a single unit, making it difficult to assign a single category. Kannada is an agglutinative language, so multiple words are often joined together to form a single token. In Kannada, any category can attach to another category. For eg: ADJECTIVE+NOUN+VERB, NOUN+VERB+QUOTATIVE, NOUN+PSP+VERB 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:

Kannada
ಸ್ವಾತಂತ್ರ್ಯವೆಷ್ಟಿರಬೇಕಿತ್ತೆಂಬುದರನುಗುಣವಾಗಿ → ಸ್ವಾತಂತ್ರ್ಯ \N_NN ಎಷ್ಟು\PR_PRQ ಇರಬೇಕಿತ್ತು\V_VM_VF ಎಂಬುದರ\V_VM_VNF ಅನುಗುಣವಾಗಿ\PSP
IPA
svaːt̪ant̪r̪javeʃʈirabeːkitt̪embuɖaranuɡuṇavaːgi → svaːt̪ant̪r̪ja\N_NN eʃʈu\PR_PRQ irabeːkitt̪u\V_VM_VF embuɖara\V_VM_VNF anuɡuṇavaːgi\PSP

2.3.4.5 Compound Word

Two or more word combined to form a single unit and the sole token gives a meaning. For eg: ಕರಿಹಲಗೆ /karihalage/, ನೀಲಕಂಠ /niːlakaṇṭʰa/. These types of words are taken as such and assigned the appropriate tag.

2.3.4.6 Dialect Variation

If the annotator is unaware about the dialect variation it may lead to incorrect tagging. In Kannada, the word ಬ್ಯಾಳಿ /bjaːli/ ‘pulses’ is using in Dharwad region. Its standard form is ‘ಬೇಳೆ’ /beːle/. ಒಯ್ಯಿದೆ /oj̪ɪɖide/ ‘is going’ is using in Mysore but in standard it ಹೋಗುತ್ತಿದೆ /hoːɡuɪt̪ɪɖide/. ‘ಹೋಕು’ /hoːku/ ‘go’ is used in the place of ಹೋಗಬೇಕು /hoːgabeːku/ by Coastal Kannada speakers.

If the annotator is not familiar with such type of word then there is a chance to make mistake while assigning tag. So the annotator needs language expertise while tagging.

2.3.4.7 Impact of Social Dialect

The word ‘ಮರ್ಜಿ’ /mardzi/ means ‘wish’; it is an Urdu word and it is nativised by some specific Kannada region people. so this word tagged as N_NN category. In case the annotator unaware about the term it may tag as RD_UNK.

2.3.4.8 Verbal Noun

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

2.3.4.9 Unknown Words

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

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

Dataset Summary	Aggregate Count
Total PoS Tags	769767
Total Sentences	66113
Total Word Count	642885
Total File Count	1682

Table 2-3: Kannada PoS Status

Sl. No.	PoS Category	PoS Tag	Tag Count
1	Common Noun	N_NN	225396
2	Proper Noun	N_NNP	61765
3	Nloc Noun	N_NST	19406
4	Personal Pronoun	PR_PRP	42302
5	Reflexive Pronoun	PR_PRF	5377
7	Reciprocal Pronoun	PR_PRC	241
8	Wh-word Pronoun	PR_PRQ	5277
9	Deictic Demonstrative	DM_DMD	10646
10	Wh-word Demonstrative	DM_DMQ	1162
11	Finite Verb	V_VM_VF	69815
12	Non-finite Verb	V_VM_VNF	61877

13	Infinitive Verb	V_VM_VINF	5094
14	Gerund	V_VM_VNG	7568
15	Verbal Noun	V_VM_VNV	2289
16	Auxiliary Finite Verb	V_VAUX_VF	1229
17	Auxiliary Non-finite Verb	V_VAUX_VNF	410
18	Auxiliary Infinitive Verb	V_AUX_INF	30
19	Adjective	JJ	35828
20	Adverb	RB	15887
21	Postposition	PSP	10026
22	Co-ordinator	CC_CCD	9954
23	Subordinator	CC_CCS	2676
24	Quotative	CC_CCS_UT	11288
25	Particles	RP_RPD	5722
26	Interjection	RP_INJ	558
27	Intensifier	RP_INTF	2140
28	Negation	RP_NEG	344
29	Quantifiers	QT_QTF	8560
30	Cardinals	QT_QTC	13236
31	Ordinals	QT_QTO	2237
32	Foreign word	RD_RDF	2690
33	Symbol	RD_SYM	2064
34	Punctuation	RD_PUNC	124818
35	Unknown	RD_UNK	1839
36	Echowords	RD_ECH	16

Table 2-4: Distinct Tag Count of Kannada