

कोंकणी

शब्दजाती

भाश्यात (एन्नोटेटेड) कॉर्पस



KONKANI

PARTS OF SPEECH
ANNOTATED CORPUS

KONKANI 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:

Mr. Saurabh Varik

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

Kokani Parts of Speech Annotated Corpus

Authors: Mr. Saurabh Varik, Dr. Narayan Choudhary

e-ISBN: 978-81-69175-91-3

CIIL Publication No.: 1700

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: Ankita Tiwari

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	Konkani PoS Annotated Corpus	6
2.1	Grammatical peculiarities of Konkani	6
2.2	LDC-IL Customized BIS Tagset	6
2.3	Konkani PoS Tagging	10
2.4	Summary of LDC-IL Konkani PoS Data	19
2.5	References	20

Tables

Table 1-1: BIS Tagset	4
Table 1-2: LDC-IL PoS Status	5
Table 2-1: Konkani Tagset	9
Table 2-2: Konkani PoS Status	19
Table 2-3: Distinct Tag Count of Konkani	20

1 POS ANNOTATED CORPUS BASED ON BIS TAGSET

Rejitha K .S., Narayan Kumar 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 labeling 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)

Konkani Parts of Speech Annotated Corpus

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	
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 569408 tags. The summary of LDC-IL parts of speech data is given below.

Dataset Summary	Aggregate Count
Total PoS Tags	569408
Total Sentences	53161
Total Word Count	468967
Total File Count	877

Table 1-2: LDC-IL PoS Status

2 KONKANI POS ANNOTATED CORPUS

Saurabh Varik

2.1 GRAMMATICAL PECULIARITIES OF KONKANI

Konkani is an Indo-European (Indo-Aryan) language evolved from Sanskrit. It is a morphologically rich Indian language (Almeida, 1989). It is one of the twenty two languages included in the Eighth Schedule of the Indian Constitution. Konkani, an Indo-Aryan language spoken primarily along India’s western coast, is a linguistic mosaic, written in five scripts—Devanagari (Nagri), Roman (Romi), Kannada, Malayalam, and Persian-Arabic—and spoken in diverse dialects across Goa, Maharashtra, Karnataka, and Kerala. It is spoken by the people of the state of Goa, in some parts of Maharashtra, Karnataka and in some pockets of Kerala. It is influenced and enriched by various other languages like Marathi, Kannada, Malayalam, Hindi, Arabic, Persian, Portuguese and English. It is the official language of Goa with Devanagari as the officially recognized script. It is also written in Roman, Kannada, and Malayalam scripts (Walawalikar, et.al. 2010).

The BIS POS tagset is prepared keeping in view the comments of experts in the area of NLP and Language Technology (LT) for Indian languages. This tagset is an important step taken by DIT to ensure that NLP practitioners involved in tagging follow a common tagset while tagging various corpora. The tagset initially consisted of 38 tags. These tags were then modified after taking inputs from linguists, computer scientists and language experts. More details of the BIS tagset can be found in Chaudhary, 2010. The BIS tagset is a commendable effort. But the process of tagging was, at times, quite challenging as it was different from the conventional style of tagging (for example, the case of marking adverbs of place and location as NSTs (locative nouns) .

From a Part-of-Speech (POS) perspective, this morphological complexity poses significant challenges. Multiple grammatical categories—such as case, number, gender, tense, aspect, mood, person, and voice—may be simultaneously encoded within a single word. As a result, PoS tagging in Konkani 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.

2.2 LDC-IL CUSTOMIZED BIS TAGSET

Based on the BIS tagset and LDC-IL–annotated Konkani texts, certain categories—such as Finite, Non-Finite, Infinitive, Gerund verbs, and Quotative conjunctions—are not applied in Konkani part-of-speech tagging, as these features are either rare or not attested in the grammatical structure of Konkani. This approach follows the *Linguistic Standards for Indian Language (IL) Corpora Annotation: PoS Categories and Labels* (June 12, 2010).

Konkani Parts of Speech Annotated Corpus

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	झाड, न्हय, मनीस, पुस्तक, रूख, आंबो, माड
1.2		Proper		NNP	N__NNP	गोंय, कोंकणी, शणै गोंयबाब, रामायण, बायबल, कुराण
1.3		Nloc		NST	N__NST	भायर, भितर, वयर, सकयल, काल, आयज,आज, पयलीं, मागीर, फाल्यां, लागींचें, उपरांत
2	Pronoun			PR	PR	
2.1		Personal		PRP	PR__PRP	हांव, तूं, तो, तें, ते, त्यो, तुमचें, आमचें, तांचें,
2.2		Reflexive		PRF	PR__PRF	आपूण, स्वता
2.3		Relative		PRL	PR__PRL	जातूंत, जो, जे, जाणें, जांकां, जाका
2.4		Reciprocal		PRC	PR__PRC	एकामेकाक, आपसांत
2.5		Wh-word		PRQ	PR__PRQ	कोण, कितें, खंयचो
2.6		Indefinite		PRI	PR__PRI	कोणूय, कितेंय, खंयचेंय, कोणूच, कितेंच, खंयचेंच
3	Demonstrative			DM	DM	
3.1		Deictic		DMD	DM__DMD	तो भुरगो, हो गांव, अशें चली

3.2		Relative		DMR	DM__DMR	जो सरकार, ज्या मनशान
3.3		Wh-word		DMQ	DM__DMQ	कोणें, कसलें
3.4		Indefinite		DMI	DM__DMI	कोणाचेंय, कसलेंय
4	Verb			V	V	
4.1		Main		VM	V__VM	येवप, आयलो, येता, येतलो, आयिल्लो, येतालो, आयला, येत, येतकच, येवन, आयिल्ल्यान, येवंक, करत, केल्लो, केल्यार, आसूं, व्हरूं, करूं, काडूं, खावप, खावपी, वाडप, वाडपी
4.3		Auxiliary		VAUX	V__VAUX	करूंक लागलो, केलां आसतलें, खालें जातलें, करता आसत, वता जायत, दिवंक लागत
5	Adjective			JJ		सोबीत गांव, दुश्ट मनीस, बरो भुरगो, काळी गाय, शाणें बाय
6	Adverb			RB		सवकास, परतो, हळू हळू, बेगीन बेगीन
7	Postposition			PSP		खातीर, पासत, बगर, कडेन, लागीं
8	Conjunction			CC	CC	
8.1		Co-ordinator		CCD	CC__CCD	आनी, वा
8.2		Subordinator		CCS	CC__CCS	जाल्यार, जर-तर, देखून, म्हणल्यार, पुणून, अशें, की
9	Particles			RP	RP	
9.1		Default		RPD	RP__RPD	बी, आदी, इत्यादी, मूं,

Konkani Parts of Speech Annotated Corpus

						व्हय, हय, वय, काय, खंय
9.3		Interjection		INJ	RP__INJ	आरे, चूप
9.4		Intensifier		INTF	RP__INTF	उपाट, भरपूर, सगल्यांत
9.5		Negation		NEG	RP__NEG	ना, न्हय
10	Quantifiers			QT	QT	
10.1		General		QTF	QT__QTF	थोडे, चड, कांय, खूब
10.2		Cardinals		QTC	QT__QTC	एक, दोन
10.3		Ordinals		QTO	QT__QTO	पयले, दुसरे
11	Residuals			RD	RD	
11.1		Foreign word		RDF	RD__RDF	A word written in script other than the script of the original text AND or IIT*
11.2		Symbol		SYM	RD__SYM	\$, &
11.3		Punctuation		PUNC	RD__PUNC	, . . , ? - /
11.4		Unknown		UNK	RD__UNK	चाबी, ऑफकोर्स, मिसिंग, ओव्हर, श्री
11.5		Echowords		ECH	RD__ECH	जेवण-बिवण (here बिवण is the Echo word) टिन्ही-शीन्ही, सोयरीं-धायरीं

Table 2-1: Konkani 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 during the PoS Standards Committee meeting held on June 12, 2010, at IIIT Hyderabad. The PoS tags for the Konkani language were finalized by Prof. Madhavi Sardesai and Mr. Damodar Ghanekar at the LDC-IL PoS workshop.

2.3 KONKANI POS TAGGING

Konkani PoS tagging is performed using the BIS tagset, with necessary adaptations from LDC-IL guidelines to suit Konkani’s linguistic structure. Since Konkani is primarily a highly inflectional (synthetic) language that also displays strong agglutinative features., It inflects nouns for case, gender, and number, and conjugates verbs for tense, aspect, and mood. It acts agglutinatively by stacking suffixes onto a stem to modify meaning. With morphemes combining to form words that inflect for case, number, gender (nouns), and tense, aspect, mood, and agreement (verbs). For instance, in Goan Konkani, the word /gharâk/ (“to the house”) segments into /ghar/ (free morpheme, “house”) and /âk/ (bound morpheme, dative case). Similarly, /kha-tâ/ (“are eating”) divides into /kha/ (free, “eat”) and /tâ/ (bound, present progressive, 3rd person plural). These morphemes are consistent across dialects, but their phonological realization and orthographic representation vary by script and community.

2.3.1 Guidelines

- Konkani PoS Tagging is done based on BIS tagset with customization to capture specific grammatical and morphological features.
- Konkani is primarily a highly inflectional (synthetic) language that also displays strong agglutinative features.
- 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 the process of breaking down a sentence into its constituents. It is important because accuracy of the PoS tagged data depends on the input words. PoS tagging incorporates the morpho-syntactic information to the lexical item. It does not insert any semantic, discourse or pragmatic information. In Hindi, on the word-level, tokenisation refers to breaking down a sentence into the respective constituents. For example, *हांव आंबो खायत आसां* will be tokenized into 4 tokens- *हांव* | *आंबो* | *खायत* | *आसां*. Through tokenisation, it is quite obvious that one can provide a PoS tag to each separate token.

In the old Konkani corpus, some **postpositions were written attached to the main word** such as nouns, pronouns, and verbs. For example: *रामाखातीर, ताकालागून, खावपाकम्हूण, चेतनलेगीत*. However, according to the **BIS PoS tagging guidelines**, these postpositions should be **separated**

from the main lexical item and tagged independently based on their grammatical category. Therefore, during annotation we split these forms into two units so that the noun/pronoun/verb receives its appropriate tag and the postposition or as per category is tagged separately.

Examples:

1. रमाखातीर → रामा + खातीर
IPA: / rəma:/ /kʰa:ti:r/
2. ताकालागून → ताका + लागून
IPA: /ta:ka/ /la:gu:n/
3. महाकाव्यलेगीत → महाकाव्य + लेगीत
IPA: /məɦa:ka:vjə/ /le:gi:t/
4. चेतनलेगीत → चेतन + लेगीत
IPA: /tʃe:tən/ /le:gi:t/

Thus, although these forms appear as **single orthographic words in the older corpus**, in **BIS-level POS annotation** they are **segmented and tagged separately** to maintain consistent morpho-syntactic analysis.

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 some pronouns function as demonstratives if they occur before a noun.

तो घरा भायर उबो आसा. (He is standing outside the house.) - - तो **PR_PRP**

तो भुरगो घरा भायर उबो आसा. (The boy is standing outside the house..) - तो **DM_DMD**

In the first sentence, the token 'तो' serves as a personal pronoun. However, in the second instance, it works as a demonstrative because it modifies a common noun (भुरगो). Similarly, indefinite pronouns such as कोण, कितें, खंयचो and कोणे also function as demonstratives when preceding a noun.

Ambiguous Words: ते and तेवूय/तेंवूय: The word (ते) has different grammatical categories like pronoun (they), demonstrator (that) and conjunction (to).

Examples:

1. 30 ते 50 (30 to 50) The word (ते) lexically and functionally acts as conjunction, hence to be tagged as CCD.
2. ते म्हण्टाले (they said) - Here word (ते) acts as personal pronoun, hence to be tagged as PR_PRP
3. ते खंय आसात? (where are they?): Here word (ते) acts as relative demonstrator, hence to be tagged as DM_DMR
4. रमेशान पोलीसांक फोन केलो आनी ते दोनूय चोर सापडले. (Ramesh called police and those two thieves got caught).:Here, word (ते) is modifying its succeeding noun (thief) so it is Deictic demonstrator, hence to be tagged as DM_DMD.
5. तांकां पसंत नाशिल्लें, तांच्या भुरग्यान संगीत शिकचें आनी तेंवूय नृत्य (He never wanted his son to learn music and that too the dance form) Here, the word (तेंवूय) is an ambiguous word. It is modifying succeeding noun or previous context. Here, bound morpheme and conveys the meaning 'also'. Therefore word (तेंवूय) should be tagged as DM_DMR.

2.3.4 Observation on PoS Tagging of LDC-IL Konkani 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 lexical item may occur in different grammatical roles within the corpus. For instance, in Konkani, a word may function as an adjective in certain contexts and as a noun in others. Although the semantic interpretation remains unchanged, each occurrence conveys distinct morpho-syntactic information. Additionally, nouns may also function in adjectival positions depending on the syntactic environment.

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

-कुटुंबांतल्या जाण्ट्या\JJ मनशांनीं मतदान केलें (all the aged members of the family voted)

- सगळ्या जाण्ट्यांनी\N_NN मतदान केलें. (all the aged people voted)

Here the word (जाण्टो or जाण्टो) lexically acts as an adjective as well as a noun. At the syntactic level, in the first example, it is functioning as adjective hence to be tagged as JJ, while in the second example it is functioning as a noun hence to be tagged as N_NN.

In Konkani PoS tagged corpus, inconsistencies often arise due to the morpho-syntactic flexibility present in the language where though expressions share semantic similarity, they exhibit structural differences. As discussed earlier, compound forms have varying instantiations. Notably, forms such as रिक्शावालो and रिक्शा वालो or लोलयेंकार and लोलयें कार convey similar meanings but they differ in tokenisation and subsequent tagging-

रिक्शावालो\N_NN and रिक्शा\N_NN वालो\RP_RPD or लोलयेंकार\N_NN and लोलयें\N_NNP कार\RP_RPD

Though this results in inconsistent tags for the same token, these variations highlight that inconsistencies in Hindi POS-tagged corpora stem not only from annotation practices but also the rich linguistic properties such as variable tokenisation.

2.3.4.2 Ambiguity

Natural language is ambiguous from its basic nature. Ambiguity arise when a single word or string of words is associated in the language system with more than one meaning (Chierchia and McConnell-Ginet 1990:32). 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 phenomenon of zero derivation in Konkani also often creates ambiguity in PoS tag assignment. However, examining the context solves this problem. Here are some ambiguous Konkani words that can function as more than one part of speech, with clear contextual examples. I've kept the explanation academic and corpus-friendly.

Ambiguous POS: Noun and Conjunction

Example: word कारण kAraNa, reason; because).

At semantic level, the word कारण (kAraNa) has two meanings, one is 'a reason' which acts as a noun and another is 'because' which acts as a conjunction.

Word Sense Disambiguation Challenges

The hyphen(-) is marked as punctuation in the BIS tagset, but in Konkani it conveys different information on different occasions. In some cases, a pair of two words joined by a hyphen is a

compound word whereas in other cases it may be just a noun phrase. Various senses conveyed by the hyphen need to be separated and structurally identified as specific compounds. This would facilitate word sense disambiguation.

बरे-वायट bareM-vAyaT ('good-bad'). In this example, the hyphen marks a pair of antonyms. In isolation, the words in this pair function as adjectives whereas the pair as a whole functions as a noun.

For example, in **तेँ एक बरेँ चली** teM eka bareM chali ('she is a good girl'), and **तेँ एक वायट चली** teM eka vAyaT chali ('she is a bad girl', **बरेँ** bareM 'good' and **वायट** vAyaT 'bad' both function as adjectives. However, in **ताका बरेँ-वायट समजना** tAkA bareM-vAyaT samajanA ('he does not know what is good and what is bad' i.e. 'he does not know his good'), the pair **बरेँ-वायट** bareM-vAyaT functions as noun.

The hyphen in the pair **बायल-मनीस** bAyal manIsa ('woman-human') marks a specifier-specified relationship. In Konkani, **मनीस** manIsa 'human being' is a gender neutral term. The following sentences illustrate this point:

1. **तो एक बरो मनीस** to ek baro manIs ('he is a good human being')
2. **ती एक बरी मनीस** ti ek bari manIs ('she is a good human being')

We see that in both the above sentences **मनीस** 'human being' remains the same. However, the term can be specified by other gender specific nouns such as **दादलो** dAdlo (masculine) 'man', **चली** chali (feminine/neuter) 'girl/daughter', **बायल** bAyl (feminine) 'woman' when it occurs in compounds such as **दादलो मनीस** ('man-human'), **बायल मनीस** ('woman-human'), ('girl-human'). For example, in **दादल्या मनशाक ते कळचेंना** dAdlya manashAka teM kaLacheMnA ('a male human (i.e. a man) will not understand this'), dAdlo 'man' specifies the masculine gender of **मनीस** manIsa 'human being'. Thus the role of hyphen cannot be ignored in the compounds of the above type. **भाट-बेंस** bhAT- beMsa ('property-assets'), **भांगर-शिंगर**, **धन-दौलत**, **साडे-सात** sADe-sAta (half past seven), **सवाय-आठ** - savAya-ATh quarter past eight), **ताका मारपाक सात-आठ मनीस आशिल्ले** (tAkA mArApAka sAta-ATh manIs Ashille) 'to hit him seven-eight people were there' (i.e. 'around seven to eight people were present to hit him'), the phrase - **सात-आठ** is joined by the hyphen. Here, the hyphen expresses a sense of 'unsurety' within the mind of the speaker about number of people that were present before him.

Punctuation

Marking an inverted comma as punctuation is misleading in some cases. For example, in **'काँग्रेस पार्टी' चो** kA.NgresA pArTI cha.o ('of Congress Party'), **चो** is a suffix of the noun phrase 'Congress Party'. This phrase is put in inverted commas as it is a foreign language phrase. Thus marking inverted commas in such cases as punctuation is not proper.

Tagging of spatio-temporal adverbs as NST

All nouns in Konkani undergo oblique formation before they take a case suffix. राम rAm becomes रामा rAmA before it takes any suffix like -क ka, -न na, -चो cho, etc. However, the spatio-temporal adverbs take suffixes before undergoing oblique formation. E.g. भायर+लो bhAyara + lo> bhAyalo ‘of outside’, सकयल+च्यान sakayala +chyAna > सकयलच्यान sakayalyAna ‘from below’. Moreover, these adverbs take only the above mentioned suffixes. Combinations like भायर+क bhAyara +ka ‘to the one outside’ भायर+आंत bhAyara+AMta ‘in the one outside’ are not permitted in the language. If these adverbs are tagged as nouns (NSTs) then they must be treated as a special category of nouns.

Recognition of frozen expressions

Some expressions in language function in a certain way. खरें म्हणल्यार khareM mhaNalyAra (Truly/Really speaking) खरें KhareM in Konkani means ‘truth’ whereas म्हणल्यार mhaNalyAra means ‘having told’ which is a verb/quotative. However, when these two words come together, they always function like an adverb. Tagging these words individually would serve no purpose. Some more examples in the same category are listed below - सांगपाचें म्हणल्यार sAMgapAcheM mhaNalyAra “Of-telling if told” means ‘actually’. तशें पळेल्यार tasheM paLayalyAra “that way if seen” also means ‘actually’.

Negation

1. ना. हांव येवचो ना nA. hAMva yevaMcho nA (‘No. I will **not** come’)
 2. ना. आमी येवंचे नात nA. Aml yevaMche nAt (‘No. We will **not** come’)
- In Hindi, negation occurs at all places as a particle (in the sentences 1. and 2.), whereas in Konkani the first word in 1. and 2. is a particle and the second one in 1. and 2. is a finite verb but as per BIS guidelines, all these forms are tagged as *Negation*.

JJ, RP_RPD and RB

तें चादर बरें JJ आसा.

बरेँ\RP_RPD, तूं म्हजें काम केन्ना मेरेन करता?

तें बरेँ\RB गायता.

N_NN & RP_INJ

सिमान म्हळ्ळें आवयस\N_NN आयकून राजीव धांवत आयलो.

आवयस\RP_INJ ! ते दोंगुल्लेर आमी कशे वचपाचे?

N_NN & JJ

मोटें\N_NN आयज शाळेंत आयले ना. (motẽ a:j d̪ɜ̃ ʃa:l̪et̪ a:j|e: na:)(fatty didn't come to school today.)

मोटें\JJ भुरगें रडटालें. (motẽ bʱurgẽ rəɖt̪a:l̪ẽ) (The fatty child was crying.)

N_NNP & JJ

गुलाबी\N_NNP आयज मडगांव वतलें.

गुलाबी\JJ रंग म्हज्या आवडिचो.

CC_CCD & RP_RPD

Ambiguous word: तर (tara)

Examples:

Conjunction: जर हांव वेळार थंय पावना तर ताचें वाचप शक्य नाशिल्लें (ɖʒərə hã:uə ue:l̪a:rə t̪ʰəjə pa:uəɳa: t̪ərə t̪a:l̪i: uə:l̪əpə ʃəkjə ɳa:ʃɪll̪ẽ:, if I had not gone on time he would have not survived).

Particle: 'बरेँ! येतां तर हांव!' ('bərẽ:!! je:t̪ã: t̪ərə hã:uəl', 'yes! now I am leaving.

In the above sentences, the word तर (təṛə) is used as a supplementary or stressable word so somewhat special as to give meaning in the sentence. Hence it should be treated as CCD.

तुमी तर लाखांनी रुपया मागतात आनी हांव तर गरीब मनीस tumi: tər lakʰā̃ni: rupeja magəṭat ani: hamw tər gəri:b məni:s you are asking for lakh rupees and I am a poor person). In this sentence, the word तर (tara) indicates opposition with respect to meaning between two connected sentences. Hence, it should be treated as a RPD

QT_QTO & N_NN

ही न्हय, दुसरी\QT_QTO दुसरी गाडी दाखयात.
एका चलयेन पेन घेतलें, दुसऱ्यान\N_NN पुस्तक.

2.3.4.3 Abbreviation

Abbreviations are shortened forms of words or phrases which are widely used in both written and spoken language. Some examples of abbreviations in Konkani are- के.बी., एल.एल.बी., बी.डी.ओ, बी.ए. etc.

2.3.4.4 Compound Word

The hyphen sometimes functions as an integral part of compounds and phrases. Marking it as a punctuation mark will mean ignoring the subtle information that it conveys in these occurrences. We feel that the role of the hyphen has to be recognized as it would not only help in word sense disambiguation (as in the case of बरें-वायट bareM-vAyaT ‘good-bad’) but also bring to light certain peculiarities of Konkani (as in the case of भाट-बेंस bhAT- beMsa ‘property-assets’ where the synonymic compound is formed with a word from Portuguese).

Frozen expressions, for example can be marked with a special tag at POS level itself so that there is no wastage of time in unnecessary work, later. For example, in खरें-म्हणल्यार KhareM mhaNalyAra ‘really/truly speaking’ tagging individual words would serve no purpose. Instead the token ‘ ’ could be marked as FE i.e. frozen expression. The appropriate category (i.e. adverb) could be marked at the next level of NLP.

2.3.4.5 Dialect Variation

Dialect variation in Konkani means that the language changes a bit from place to place—mainly in pronunciation, vocabulary, and sometimes grammar. This happens because Konkani speakers

live across different regions and have long contact with other local languages. Konkani is spoken in Goa, coastal Karnataka, and Maharashtra. Over time, interaction with Marathi, Kannada, and Portuguese influenced local speech, creating regional varieties. Major Konkani dialects (with examples)

1. Goan Konkani (Antruzi / Bardeshi): Spoken mainly in Goa; considered the standard form in education and media.

Sentence: तूं किदें करता? tũ: kiḍẽ: kəɾəɽɑ:?

Meaning: What are you doing?

2. Malvani Konkani: Spoken in southern Maharashtra and northern Goa; heavily influenced by Marathi.

Sentence: तुं काय करता? tũ ka:jə kəɾəɽɑ:?

Meaning: What are you doing?

(Notice “काय” ka:jə instead of “किदें” kiḍẽ:)

3. Canara / Karnataka Konkani: Spoken in coastal Karnataka; influenced by Kannada.

Sentence: तूं किदें करताय? tũ: kiḍẽ: kəɾəɽɑ:jə?

Meaning: What are you doing?

(Different verb ending)

4. Saraswat Konkani: Used by Saraswat communities across regions; shows lexical differences.

Word for water:

Goan: उदक uḍəkə | Karnataka: नीर ɲi:rə

Simple comparison (one idea, different forms)

Meaning	Goan	Malvani	Karnataka
What	किदें (kiḍẽ:)	काय (ka:jə)	किदें (kiḍẽ:)
Water	उदक/उदीक (uḍəkə/uḍi:kə)	पाणी (pa:ɲi:)	नीर (ɲi:rə)
Where	खंय (kʰəjə)	कुडे (kude:)	खंय (kʰəjə)

2.3.4.6 Impact of Social Dialect

Konkani is a highly diverse Indo-Aryan language that has been significantly shaped by its geographical location, history of migration, and foreign colonization, leading to heavy influences from Portuguese, Kannada, Marathi, and Malayalam. Due to these, Konkani is spoken in multiple dialects and written in five different scripts (Devanagari, Roman, Kannada, Malayalam, and Perso-Arabic). **Portuguese Influence** - साबांव Sabão (Soap), पाव Pão (Bread), कुस्तार Kustar

(Costly - from *Custar*), कदेल *Kadeir* (Chair - from *Cadeira*), मेज *Meiz* (Table - from *Mesa*), जनेल *Janel* (Window - from *Janela*), **Kannada words** - बागील *Bāgil* (Door), replacing the generic दार *Dār*, कोरता आसा *korta asa* - is doing. **Marathi Influence** (Northern Konkani/Malvani) नंतर *Nantar* (Afterwards) or कधी *Kadhi* (When). **Malayalam/Tulu Influence** - सिडी *Sidi* (Ladder - from *Malayalam*), चेट्टू *Chettu* (Coconut tree - from *Malayalam*)

2.3.4.7 Unknown Words

There will be some words in the text which cannot be handled by the tagset. Such words are considered as unknown words and it will pose problem for the higher level of processing. For example, क्या (*kja:*)\RD_UNK, है (*hai*)\RD_UNK, नहीं (*nāhi:*)\RD_UNK, सारी (*sa:ri*)\RD_UNK, थ्येक्यू (*θæŋkju:*)\RD_UNK etc.

2.4 SUMMARY OF LDC-IL KONKANI POS DATA

The status of the LDC-IL Konkani 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 2-2: Konkani PoS Status

Sl. No.	PoS Category	PoS Tag	Tag Count
1	Common Noun	N_NN	136905
2	Proper Noun	N_NNP	25404
3	Nloc Noun	N_NST	19950
4	Personal Pronoun	PR_PRP	50267
5	Reflexive Pronoun	PR_PRF	3328
6	Relative Pronoun	PR_PRL	625
7	Reciprocal Pronoun	PR_PRC	194
8	Wh-word Pronoun	PR_PRQ	3279
9	Indefinite Pronoun	PR_PRI	2404
10	Deictic Demonstrative	DM_DMD	9026
11	Relative Demonstrative	DM_DMR	622
12	Wh-word Demonstrative	DM_DMQ	345
13	Indefinite Demonstrative	DM_DMI	837
14	Main Verb	V_VM	99798

15	Auxiliary Verb	V_VAUX	14673
16	Adjective	JJ	26004
17	Adverb	RB	10530
18	Postposition	PSP	4693
19	Co-ordinator	CC_CCD	14940
20	Subordinator	CC_CCS	6934
21	Particles	RP_RPD	5322
22	Interjection	RP_INJ	1244
23	Intensifier	RP_INTF	1577
24	Negation	RP_NEG	5974
25	Quantifiers	QT_QTF	8421
26	Cardinals	QT_QTC	10501
27	Ordinals	QT_QTO	1019
28	Foreign word	RD_RDF	1109
29	Symbol	RD_SYM	127
30	Punctuation	RD_PUNC	100314
31	Unknown	RD_UNK	2961
32	Echowords	RD_ECH	22

Table 2-3: Distinct Tag Count of Konkani

2.5 REFERENCES

- Almeida, Matthew. (1989). *A Description of Konkani*. Miramar, Goa: Thomas Stephens Konkani Center.
- BIS Annotation Standards With Reference to Konkani Language, Edna Vaz, Shantaram V. Walawalikar, Dr.Jyoti Pawar, Dr.Madhavi Sardesai, Goa University, Taleigao-Goa.
- Part-of-Speech Annotation Challenges in Marathi, Gajanan Rane, Nilesh Joshi, Geetanjali Rane, Hanumant Redkar, Malhar Kulkarni and Pushpak Bhattacharyya, Center For Indian Language Technology (CFILT) Indian Institute of Technology Bombay.
- ILCI Parts of Speech (PoS) Annotation Guidelines by Narayan Chaudhary, Pinkey Nainwani, Ritesh Kumar, Esha Banerjee, ILCI group, JNU
- Miranda, Rocky V. 2003. Konkani. In George Cardona & Dhanesh Jain (eds.), *The Indo-Aryan Languages*. London: Routledge, pp. 729-765.
- Miranda, Rocky V. 2019. Konkani Language., In G. N. Devy (ed.), *The Languages of Goa – Peoples, Linguistic Survey of India*, Vol. 8 Part II.
- Katre, S. M. 1966., *The Formation of Konkani.*, Poona: Deccan College Post-Graduate and Research Institute.