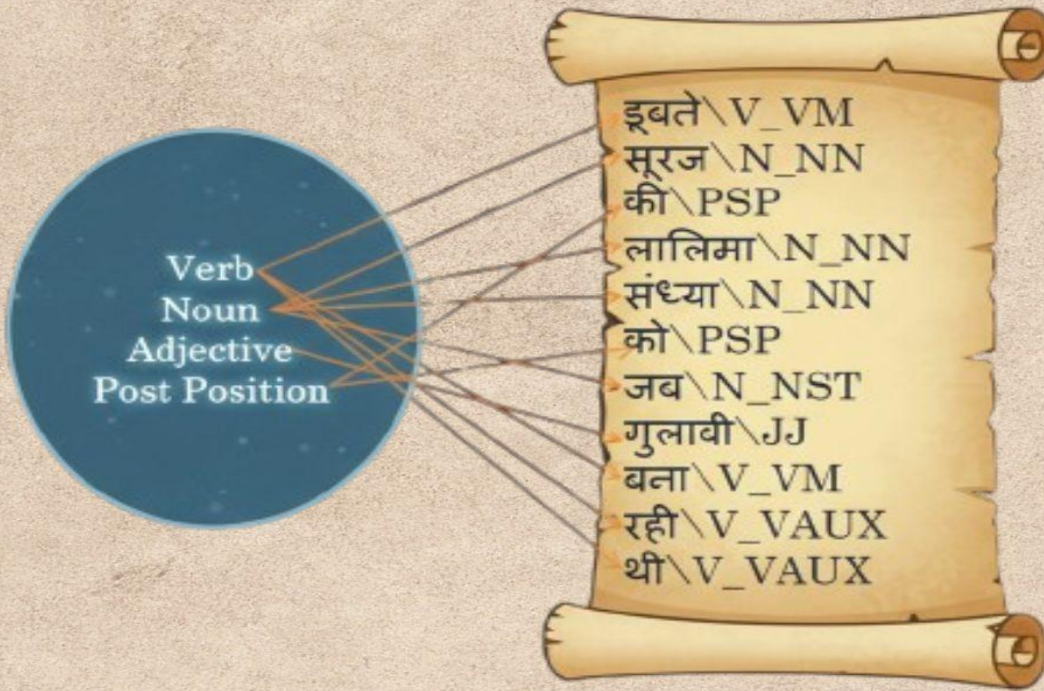


हिंदी शब्द-भेद एनोटेटेड कॉर्पस



Hindi
Parts of Speech
Annotated Corpus

HINDI 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. Satyaendra Kumar Awasthi

Dr. Narayan Choudhary

Rajesh N.

Prof. Shailendra Mohan

**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

Hindi Parts of Speech Annotated Corpus

Authors: Dr. Satyaendra Kumar Awasthi, Dr. Narayan Choudhary, Rajesha N., Prof. Shailendra Mohan

e-ISBN: 978-81-69175-44-9

CIIL Publication No.: 1703

First published: AD 2026 March :: Phalgunā 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

Table of 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	Hindi PoS Annotated Corpus	6
2.1	Grammatical peculiarities of Hindi	6
2.2	LDC-IL Customized BIS Tagset	6
2.3	Hindi PoS Tagging	7
2.4	Summary of LDC-IL Hindi PoS Data	13

Figures

No table of figures entries found.

Tables

Table 1-1: BIS Tagset	4
Table 1-2: LDC-IL PoS Status	5
Table 2-1: Hindi Tagset	7
Table 2-2: Hindi PoS Status	13
Table 2-3: Disntict Tag Count of Hindi	14

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 levels of tags. The first level represents the grammatical category, encompassing 11 primary tags. The second level is classified into 32 distinct subcategories, while the third level outlines 10 specialized categories. The different levels of tags minimize the redundancy by incorporating both lexical and syntactic information. Many PoS tags in the BIS are assigned specifically to individual lexical items.

To accommodate the diverse range of Indian languages within a single tagset, the BIS developed a superset of PoS tags specifically designed for Indian languages. All categories or subcategories within the BIS PoS tagset may not be applicable to every Indian language. Due to the linguistic diversity and structural variances among Indian languages, some categories may be relevant to certain languages while others might not be, depending on their distinctive grammatical features and constructions.

Sl. No	Category			Label	Annotation Convention**	Remarks
	Top level	Subtype (level 1)	Subtype (level 2)			
1	Noun			N	N	
1.1		Common		NN	N NN	
1.2		Proper		NNP	N NNP	
1.3		Verbal		NNV	N NNV	The verbal noun type is only for languages such as Tamil and Malayalam)

1.4		Nloc		NST	N NST	
2	Pronoun			PR	PR	
2.1		Personal		PRP	PR PRP	
2.2		Reflexive		PRF	PR PRF	
2.3		Relative		PRL	PR PRL	
2.4		Reciprocal		PRC	PR PRC	
2.5		Wh-word		PRQ	PR PRQ	
2.6		Indefinite		PRI	PR PRI	
3	Demonstrative			DM	DM	
3.1		Deictic		DMD	DM DMD	
3.2		Relative		DMR	DM DMR	
3.3		Wh-word		DMQ	DM DMQ	
3.4		Indefinite		DMI	DM DMI	
4	Verb			V	V	
4.1		Main		VM	V VM	
4.1.1			Finite	VF	V VM VF	
4.1.2			Non-finite	VNF	V VM VNF	
4.1.3			Infinitive	VINF	V VM VINF	
4.1.4			Gerund	VNG	V VM VNG	
4.2		Verbal		VN	V VN	<i>paTittam, naTattam, naTanam</i>
4.2		Auxiliary		VAUX	V VAUX	
4.2.1			Finite	VAUX	V VAUX VF	
4.2.2			Non-finite	VNF	V VAUX VN F	
4.2.3			Infinitive	VINF	V VAUX VI NF	
4.2.4			Gerund	VNG	V VAUX VN G	
4.2.5			Participle noun	VNP	V_VAUX_VNP	
5	Adjective			JJ		
6	Adverb			RB		Only manner adverbs
7	Postposition			PSP		
8	Conjunction			CC	CC	
8.1		Co-ordinator		CCD	CC CCD	
8.2		Subordinator		CCS	CC CCS	

8.2.1			Quotative	UT	CC CCS UT	
9	Particles			RP	RP	
9.1		Default		RPD	RP RPD	
9.2		Classifier		CL	RP CL	
9.3		Interjection		INJ	RP INJ	
9.4		Intensifier		INTF	RP INTF	
9.5		Negation		NEG	RP NEG	
10	Quantifiers			QT	QT	
10.1		General		QTF	QT QTF	
10.2		Cardinals		QTC	QT QTC	
10.3		Ordinals		QTO	QT QTO	
11	Residuals			RD	RD	
11.1		Foreign word		RDF	RD RDF	A word written in script other than the script of the original text
11.2		Symbol		SYM	RD SYM	For symbols such as \$, & etc.
11.3		Punctuation		PUNC	RD PUNC	Only for punctuations
11.4		Unknown		UNK	RD UNK	
11.5		Echowords		ECH	RD ECH	

Table 1-1: BIS Tagset

1.4 PoS ANNOTATION IN LDC-IL DATA

LDC-IL works on building annotated linguistic corpora for Indian languages. These corpora are automatically tagged and then verified by linguistic experts to ensure accuracy and consistency.

LDC-IL has developed the PoS tagged datasets in multiple scheduled Indian languages, including Assamese, Bengali, Bodo, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Nepali, Odia, Punjabi, Tamil, Telugu, and Urdu. This wide coverage highlights LDC-IL's efforts toward building a rich and standardized linguistic resource for diverse Indian languages.

In developing these corpora, LDC-IL adopted the Bureau of Indian Standards tagset as the underlying framework. Since Indian languages are structurally diverse and do not always share identical grammatical categories, the BIS tagset was customized to accommodate the unique linguistic features of each language. Certain morphological or syntactic categories may be present in one language but absent in another. For example, while some languages distinguish between honorific and non-honorific verb forms, others may not.

LDC-IL developed the PoS data in multiple languages based on the guidelines prepared by the Language Technologies Research Centre, IIIT Hyderabad. The document, [AnnCorra: Annotating Corpora – Guidelines for POS and Chunk Annotation for Indian Languages](#), was created by Akshar Bharati, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal, and released on 15-12-2006. These annotation standards define a PoS tagging scheme for Indian languages consisting of 26 part-of-speech tags.

LDC-IL has been able to achieve both standardizations 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 HINDI POS ANNOTATED CORPUS

Dr. Satyaendra Kumar Awasthi, Teena Verma

2.1 GRAMMATICAL PECULIARITIES OF HINDI

Hindi is a morphologically rich language with scope for extensive word formation due to its inflectional nature. It builds meaning by attaching distinct suffixes to a root word. Hindi possesses a diverse lexicon, largely shaped by structural processes such as affixation, reduplication, compounding etc. Furthermore, owing to social influences of neighboring languages such as Persian and English in the 20th century, Hindi also contains an intensive list of borrowed or foreign words.

Hindi morphology is anchored by a binary distinction between declinable and indeclinable words. Word classes or Parts of Speech that *decline* or undergo change to agree with gender, number or case fall under the declinable category. For example, nouns, adjectives, pronouns and verbs. On the contrary, the indeclinable category contains word classes which remain unchanged or static. These include conjunction, postposition, adverbs and participle.

A corpus acts as a digital warehouse which aids in identifying patterns, grammatical trends and linguistic nuances by providing data about the use of a language in real-world contexts.

2.2 LDC-IL CUSTOMIZED BIS TAGSET

Based on the BIS tagset and LDC-IL annotated Hindi texts, certain categories including sub-categories of verbs such as finite, non-finite, infinitive and gerund are not applied in Hindi tagging. 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	घर, लड़का
1.2		Proper		NNP	N_NNP	राम, सीता
1.3		Nloc		NST	N_NST	ऊपर, नीचे
2	Pronoun			PR	PR	
2.1		Personal		PRP	PR_PRP	मैं, तुम
2.2		Reflexive		PRF	PR_PRF	अपना, स्वयं
2.3		Relative		PRL	PR_PRL	जो, जिसे
2.4		Reciprocal		PRC	PR_PRC	आपस, परस्पर
2.5		Wh-word		PRQ	PR_PRQ	क्या, कौन
2.6		Indefinite		PRI	PR_PRI	किसी, कोई
3	Demonstrative			DM	DM	
3.1		Deictic		DMD	DM_DMD	यह, वह

3.2		Relative		DMR	DM_DMR	जो, जिसे
3.3		Wh-word		DMQ	DM_DMQ	किस, कौन
3.4		Indefinite		DMI	DM_DMI	किसी, कोई
4	Verb			V	V	
4.1		Main		VM	V_VM	खा, जाता
4.2		Auxiliary		VAUX	V_VAUX	है, हो, था
5	Adjective			JJ		अच्छा, मोटा
6	Adverb			RB		धीरे, जल्दी
7	Postposition			PSP		से, के लिए
8	Conjunction			CC	CC	
8.1		Co-ordinator		CCD	CC_CCD	और, तथा
8.2		Subordinator		CCS	CC_CCS	क्योंकि, अगर
9	Particles			RP	RP	
9.1		Default		RPD	QT	ही, भी
9.3		Interjection		INJ	RP_INJ	ओह, आह
9.4		Intensifier		INTF	RP_INTF	बहुत, अत्यंत
9.5		Negation		NEG	RP_NEG	न, नहीं, मत
10	Quantifiers			QT	QT	
10.1		General		QTF	QT_QTF	थोड़ा, कम
10.2		Cardinals		QTC	QT_QTC	एक, दो
10.3		Ordinals		QTO	QT_QTO	पहला, दूसरा
11	Residuals			RD	RD	
11.1		Foreign word		RDF	RD_RDF	A word written in script other than the script of the original text
11.2		Symbol		SYM	RD_SYM	\$. &
11.3		Punctuation		PUNC	RD_PUNC	, .
11.4		Unknown		UNK	RD_UNK	
11.5		Echowords		ECH	RD_ECH	पानी-वानी,

Table 2-1: Hindi 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 2010 at IIIT, Hyderabad.

2.3 HINDI POS TAGGING

Hindi PoS tagging is performed using the BIS tagset, with necessary adaptations from LDC-IL guidelines to suit Hindi's linguistic structure. Data containing information about the structural and grammatical features of Hindi was provided to the system to augment tag accuracy. To further improve tagging accuracy context-based tag assignment was carried out to ensure consistency across the corpus.

2.3.1 Guidelines

- Hindi PoS Tagging was done based on BIS tagset with customization to capture specific grammatical and morphological features.
- Compound words were examined carefully to determine appropriate tagging at the word level.
- Spelling mistakes were rectified.
- Extra spaces were removed where tokens had been incorrectly split by spaces.
- Missing spaces were inserted where necessary.
- Validation was done by using an 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 provide any semantic, discourse or pragmatic information. In Hindi, on the word-level, tokenization refers to segmenting a sentence into its individual components. For example, मैं खाना खाती हूँ will be divided into four tokens- मैं | खाना | खाती | हूँ. Through tokenization, it is quite obvious that one can provide a PoS tag to each separate token.

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. Hence, a balance is adopted between the form and the function of a word in a systematic and consistent way.

Category of a word is decided by observing the morphological, syntactic and semantic functions. Though a word's category can be easily identified by morphological properties, functionally it can act as another category due to context. This variation is addressed by looking at syntactic clues. For example some pronouns function as demonstratives if they occur before a noun.

वह घर के बाहर खड़ा है। (He is standing outside the house.)

वह किताब मेज पर रख दो। (Keep that book on the table.)

In the first sentence, the token 'वह' serves as a personal pronoun. However, in the second instance, it works as a demonstrative because it modifies a common noun (किताब). Similarly, indefinite pronouns such as किसी and कोई also function as demonstratives when preceding a noun.

In Hindi, there exists a plethora of words which differ in their function depending upon the neighboring linguistic environment such as ‘डाल’, ‘होनी’, ‘नाबालिग’ etc.

2.3.4 Observation on PoS Tagging of LDC-IL Hindi Text Corpora

PoS tagging provides the lexical category or word class along with contextual information. This is to say that it provides information about the form and function of words. Providing both the formal and functional features of the language would be highly desirable for advanced levels of processing. The PoS data can then be classified into accurately tagged words, ambiguously tagged words and untagged words.

The majority of nouns, pronouns, adjectives, verbs, adverbs etc. are correctly tagged. Ambiguity is a prevalent feature of natural language that often occurs at the lexical level. The competence and efficiency of a PoS tagger can be assessed by its ability to handle lexical ambiguity. Several lexical items present different meanings according to the context and hence can belong to varying categories. This ambiguity can be resolved by taking into account the contextual meaning and idiomatic expressions. Furthermore, due to insufficient information about a word, it can stay untagged or receive an unknown tag. This may happen with foreign words written in a different script, transliterated foreign lexical items, dialectical variations and other analogous instances.

2.3.4.1 Inconsistency

In Hindi 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 सब्जीवाला and सब्जी वाला or दूधवाला and दूध वाला convey similar meanings but they differ in tokenization and subsequent tagging-

सब्जीवाला\N_NN and सब्जी\N_NN वाला\RP_RPD or दूधवाला \N_NN and दूध\N_NN वाला\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 linguistic properties such as variable tokenization.

2.3.4.2 Ambiguity

Natural language is ambiguous from its basic nature. Ambiguity arises 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; categorical (or syntactic) ambiguity and semantic lexical ambiguity. Categorical ambiguity refers to the uncertainty about the grammatical category of a word e.g., noun or verb. On the other hand, semantic lexical ambiguity arises when a word has several meanings.

The phenomenon of zero derivation in Hindi also often creates ambiguity in PoS tag assignment. However, examining the context solves this problem. The following examples illustrate ambiguity in Hindi, demonstrating that the same token functions as a noun, particle, verb etc. depending upon the context.

N_NN, RP_RPD & V_VM

In Hindi
वह बस\N_NN में बैठ गयी।
बस\RP_RPD, मैं अभी गया और अभी आया।
वे अमरीका में बस\V_VM गए हैं।
In IPA
ʋəɦ bəs\N_NN mẽ beṭʰ gəji:
bəs\RP_RPD, mɛ̃ əbʰi :gəjə ɔr əbʰi :ajə
ʋe əməri:ka mẽ bəs\V_VM gəe hɛ̃

JJ, RP_RPD & RB

In Hindi
यह कपड़ा अच्छा\JJ है।
अच्छा\RP_RPD, वो क्यों आई थी?
वह अच्छा\RB नाचता है।
In IPA
jəɦ kəpəɽə əccʰa\JJ hɛ
əccʰa\RP_RPD, ʋo kjõ ai :tʰi:
ʋəɦ əccʰa\RB nacəṭə hɛ

N_NN & V_VM

In Hindi
मुझे खाना\N_NN खाना है।
मुझे खाना खाना\V_VM है।
In IPA
muʃʰe kʰana\N_NN kʰana hɛ
muʃʰe kʰana kʰana\V_VM hɛ

N_NN & RP_NEG

In Hindi
मुश्ताक़ जी मत\N_NN देने गए हैं।
इस हालत में तुम बहार मत\RP_NEG जाना।
In IPA
muʃtaq ʒi: mət\N_NN dene gae hẽ
is halət meẽ ʈum bæhar mət\RP_NEG jana

N_NN & RP_INJ

In Hindi
सीमा की उफ़\N_NN सुनकर राजीव भागा चला आया।
उफ़\RP_INJ ! आज बहुत सारा काम करना है।
In IPA
si:ma ki :uʃ\N_NN sunəkər rəʒi:v bʱaga çəla aja
uʃ\RP_INJ ! aʒ bæhʊt sara kam kərəna hẽ

JJ & V_VM

In Hindi
शिक्षिका ने मुझे हरा\JJ कागज़ दिया।
उसने मुझे लड़ाई में हरा\V_VM दिया।
In IPA
ʃikʂika ne muʒʱe hərə\JJ kagəz d̪ija
usəne muʒʱe ləɽai :meẽ hərə\V_VM d̪ija

QT_QTO & N_NN

In Hindi
यह नहीं, दूसरी\QT_QTO साड़ी दिखाओ।
एक लड़की ने कलम ली, दूसरी\N_NN ने किताब।
In IPA
jəh nəhiː, du:səri:\QT_QTO saɽi :dikʰao

ek ləʔəki :ne kələm li:, d̪u:səri:\N_NN ne kiʈab

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 Hindi are- के.बी., एल.एल.बी., बी.डी.ओ, बी.ए. etc.

2.3.4.4 Compound Word

Compounding is the phenomenon wherein two or more morphemes combine to form a single unit and the whole token gives a meaning. For example, ‘रात-भर’, ‘रंग-बिरंगे’, ‘उल्टी-सीधी’, ‘आते-जाते’, ‘दिल-दिमाग’ etc. These words are taken as whole and assigned the respective appropriate tag. In the case of hyphenated compounds, it is advisable to treat them as a single token and thereby tag them using the main category. Conversely, in the case of non-hyphenated compounds such as ‘समय सीमा’, they should be tagged separately with their distinct categories. However, if they occur without any space, they can be considered to be a single token, for instance, ‘पत्थरदिल’.

2.3.4.5 Dialect Variation

Hindi exhibits significant dialectal variations across regions which is characterized by differences in phonology, vocabulary etc. Varieties such as Awadhi, Brajbhasha, Bundeli etc. though often grouped under the Hindi umbrella, display very unique features setting them apart from Standard Hindi. This dialectal variation greatly influences the composition and annotation of linguistic corpora as speakers often incorporate regional variants into their language. This results in the seeping in of dialectal vocabulary into the Hindi corpora. Some expressions from regional dialects of Hindi found in the corpus are नून, मनई, ढीठ, मांड-भात, लुकाइ, रोकत etc.

2.3.4.6 Impact of Social variation

Social variation in Hindi is often shaped by factors such as age, education, urban/rural setting, cultural exposure etc. Speakers from urban backgrounds often code-mix Hindi with English due to which a large amount of English words have entered Hindi corpus. Words such as ‘प्लीज’, ‘कैंसिल’, ‘हलो’, ‘स्लो’ etc. have become very common in day-to-day use. Similarly, instances can be found in Hindi corpus where expressions from other languages such as सुण, रज्जों, पैरीं etc. can also be observed. Words and phrases from Kashmiri are likewise present. For example, ‘कूरी कूरी तबाशीरी’ and ‘पननि कूरी-लग्अयय’ are in Kashmiri language but written using Devanagari.

2.3.4.7 Unknown Words

There may be some words in the corpus that cannot be handled by the tagset. Such words are considered unknown and can pose challenges for higher levels of processing. Unknown words in Hindi text corpus often arise due to spelling variations, code-mixing, presence of dialectal forms etc. For example, किमानंदो\RD_UNK, अनालयी\RD_UNK etc.

2.4 SUMMARY OF LDC-IL HINDI PoS DATA

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

Dataset Summary	Aggregate Count
Total PoS Tags	770004
Total Sentences	47212
Total Word Count	663282
Total File Count	260

Table 2-2: Hindi PoS Status

Sl. No.	PoS Category	PoS Tag	Tag Count
1	Common Noun	N_NN	143831
2	Proper Noun	N_NNP	30728
3	Nloc Noun	N_NST	14592
4	Personal Pronoun	PR_PRP	45193
5	Reflexive Pronoun	PR_PRF	5586
6	Relative Pronoun	PR_PRL	2402
7	Reciprocal Pronoun	PR_PRC	204
8	Wh-word Pronoun	PR_PRQ	4656
9	Indefinite Pronoun	PR_PRI	4371
10	Deictic Demonstrative	DM_DMD	8741
11	Relative Demonstrative	DM_DMR	817
12	Wh-word Demonstrative	DM_DMQ	603
13	Indefinite Demonstrative	DM_DMI	2783
14	Main Verb	V_VM	109140
15	Auxiliary Verb	V_VAUX	64974
16	Adjective	JJ	33089
17	Adverb	RB	9273
18	Postposition	PSP	97308

19	Co-ordinator	CC_CCD	16494
20	Subordinator	CC_CCS	11398
21	Particles	RP_RPD	23144
22	Interjection	RP_INJ	842
23	Intensifier	RP_INTF	2121
24	Negation	RP_NEG	12183
25	Quantifiers	QT_QTF	7689
26	Cardinals	QT_QTC	8081
27	Ordinals	QT_QTO	1439
28	Foreign word	RD_RDF	230
29	Symbol	RD_SYM	43
30	Punctuation	RD_PUNC	106679
31	Unknown	RD_UNK	1339
32	Echowords	RD_ECH	31

Table 2-3:Distinct Tag Count of Hindi