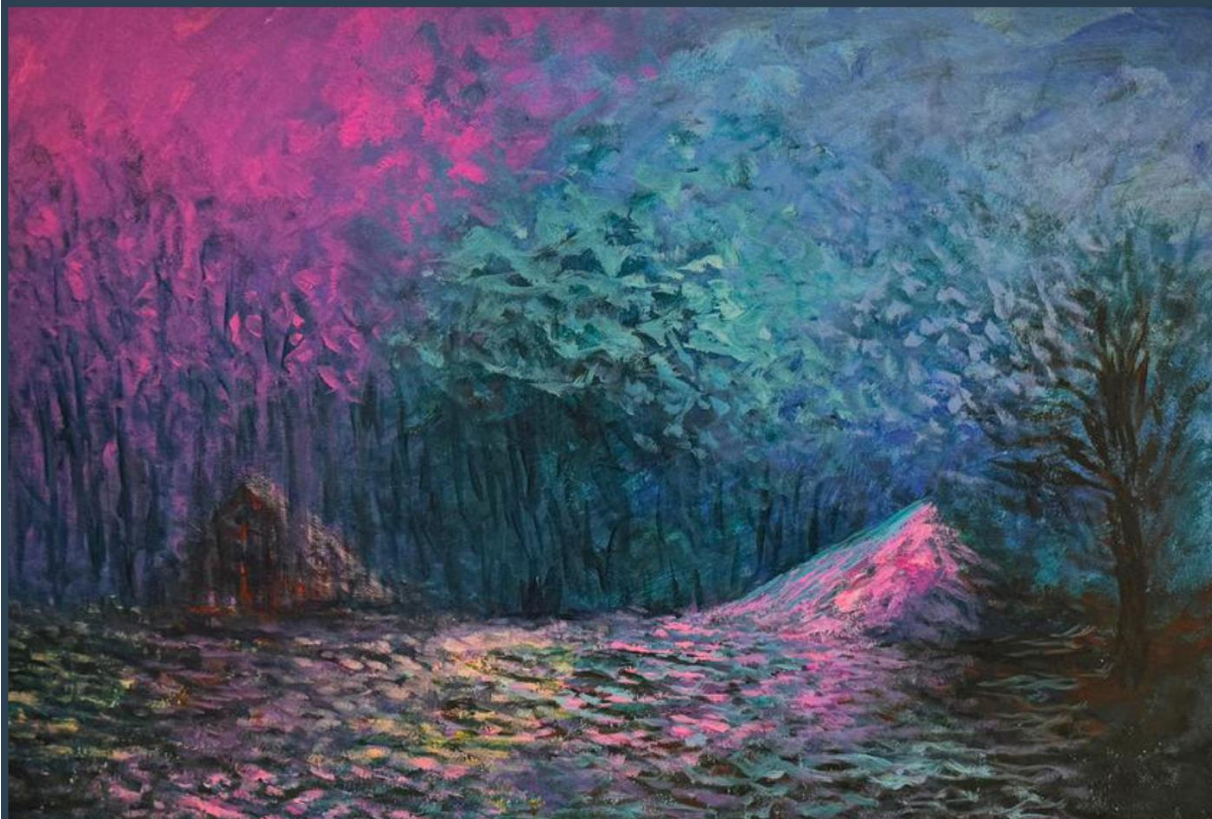


کأشرِ اجزائے کلامُک تحریری کارپس



KASHMIRI
PARTS OF SPEECH
ANNOTATED CORPUS

KASHMIRI 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. Zargar Adil Ahmad

Dr. Narayan Choudhary

Rajesh N.

Manasa G.

**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

Kashmiri Parts of Speech Annotated Corpus

Authors: Dr. Zargar Adil Ahmad, Dr. Narayan Choudhary, Rajesha N., Manasa G.

e-ISBN: 978-81-69175-77-7

CIIL Publication No.: 1701

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: Zargar Adil Ahmad

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	Kashmiri PoS Annotated Corpus	6
2.1	Grammatical peculiarities of Kashmiri	6
2.2	LDC-IL Customized BIS Tagset	6
2.3	Kashmiri PoS Tagging	8
2.4	Summary of LDC-IL Kashmiri PoS Data	13

Tables

Table 1-1: BIS Tagset	4
Table 1-2: LDC-IL PoS Status	5
Table 2-1: Kashmiri Tagset	8
Table 2-3: Kashmiri PoS Status	13
Table 2-4: Disntict Tag Count of Kashmiri	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, Kashmiri and Tamil.

BIS tagset is a hierarchical tagset which has 3 level of tags. The first level represents the grammatical category, encompassing 11 primary tags. The second level is classified into 32 distinct subcategories, while the third level outlines 10 specialized categories. The different levels of tags minimize the redundancy by incorporating both lexical and syntactic information. Many PoS tags in the BIS are assigned specifically to individual lexical items.

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

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

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, Kashmiri, 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, Kashmiri, 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 KASHMIRI POS ANNOTATED CORPUS

Zargar Adil Ahmad

2.1 GRAMMATICAL PECULIARITIES OF KASHMIRI

Kashmiri is a morphologically rich Indo-Aryan language, exhibits a high degree of inflection and derivation, allowing for extensive word formation and grammatical marking. Much of its grammatical meaning is expressed through the attachment of suffixes to lexical roots. Kashmiri operates through several productive structural processes, including affixation, compounding, reduplication, and derivation. Because of these productive morphological operations, a single root or lemma can generate a wide range of wordforms, each encoding distinct grammatical functions.

From a Part-of-Speech (POS) analysis perspective, the rich morphological structure of the language introduces significant challenges in accurate categorization and tagging. Kashmiri distinguishes between declinable and indeclinable words. Declinable categories include nouns, pronouns, adjectives, and verbs. These word classes or Parts of Speech undergo inflectional changes to express agreement in gender, number, person, and case. For example, adjectives agree with the noun they modify and may exhibit gender-based variation. Indeclinable categories include word classes such as conjunctions, postpositions, particles, certain adverbs, and some participial forms that do not undergo inflectional change.

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 Kashmiri, the BIS tagset needs to be modified.

2.2 LDC-IL CUSTOMIZED BIS TAGSET

Based on the BIS framework, LDC-IL customized the tagset to accommodate the specific needs of each language. The Kashmiri 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	گُڙ
1.2		Proper		NNP	N__NNP	بَشپَر
1.3		Nloc		NST	N__NST	بُون

2	Pronoun			PR	PR	
2.1		Personal		PRP	PR__PRP	بہ
2.2		Reflexive		PRF	PR__PRF	بینو
2.3		Relative		PRL	PR__PRL	ہیں
2.4		Reciprocal		PRC	PR__PRC	پانہ وانہ
2.5		Wh-word		PRQ	PR__PRQ	کس
2.6		Indefinite		PRI	PR__PRI	کانبہ
3	Demonstrative			DM	DM	
3.1		Deictic		DMD	DM__DMD	سہ
3.2		Relative		DMR	DM__DMR	ہیں
3.3		Wh-word		DMQ	DM__DMQ	کس
3.4		Indefinite		DMI	DM__DMI	کینہہ
4	Verb			V	V	
4.1		Main		VM	V__VM	دوان
4.1.1		Auxiliary		VAUX	V__VAUX	چہ
5	Adjective			JJ		بوڈ
6	Adverb			RB		یکدم
7	Postposition			PSP		پیٹھ
8	Conjunction			CC	CC	
8.1		Co-ordinator		CCD	CC__CCD	تہ
8.2		Subordinator		CCS	CC__CCS	زہ
9	Particles			RP	RP	
9.1		Default		RPD	RP__RPD	صرف
9.3		Interjection		INJ	RP__INJ	ہائے
9.4		Intensifier		INTF	RP__INTF	واریاہ
9.5		Negation		NEG	RP__NEG	نہ
10	Quantifiers			QT	QT	
10.1		General		QTF	QT__QTF	زیادہ
10.2		Cardinals		QTC	QT__QTC	زہ
10.3		Ordinals		QTO	QT__QTO	دوئیم
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	بڻ وٽ

Table 2-1: Kashmiri 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.

2.3 KASHMIRI POS TAGGING

Kashmiri PoS tagging is performed using the BIS tagset, with necessary adaptations from LDC-IL guidelines to suit Kashmiri’s linguistic structure. Data comprising detailed structural and grammatical features of Kashmiri were supplied to the system in order to enhance Part-of-Speech tagging accuracy. To further improve tagging accuracy context-based tag assignment is carried out to ensure the consistency across the corpus.

2.3.1 Annotation Guidelines

- Kashmiri PoS Tagging is done based on BIS tagset with customization to capture specific grammatical and morphological features.
- Kashmiri is an inflectional language; tokenization is essential because the accuracy of Part-of-Speech (POS) tagging largely depends on the contextual interpretation of word forms.
- 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 been 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 a stream of text into smaller units called tokens. It is important because accuracy of the PoS tagged data depends on the input words. PoS tagging is incorporating the morpho-syntactic information to the lexical item. It does not insert any semantic, discourse or pragmatic information. In Kashmiri sentences can be tokenized into its constituent parts at word level.

For example, The Kashmiri sentence “بہ چُہس بتہ کہپوان” will be tokenized into four distinct tokens- بہ| چُہس| بتہ| کہپوان, and each token can be assigned an individual Part-of-Speech (PoS) tag..

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. In the first sentence, the token ‘سو’ functions as a personal pronoun. However, in the second instance, it operates as a demonstrative determiner, as it modifies the common noun ‘کتاب’.

In Kashmiri
سو چہئے سکولس نیبر گنی کھڑا ۔
سو کتاب تھأوو کرسی پیٹھ۔
In IPA
sɔ \PR_PRP tʃe saku:las nebrɪ kənɪ kʰaɖa:
sɔ \DM_DMD kita:b tʰav kursi: peɽʰ

2.3.4 Observation on PoS Tagging of LDC-IL Kashmiri 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 Kashmiri, 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. As noted previously, compound formations in Kashmiri display variable instantiations. Expressions such as گڙ وول / گڙ وول and شری وول / شری وول, though semantically equivalent, differ in their segmentation at the tokenization stage. This variation directly impacts POS assignment, as the fused form may be treated as a single lexical unit, whereas the spaced form requires independent tagging of its constituent elements.

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

گڙ وول \ N_NN and وول \ RP_RPD گڙ \ N_NN or شری وول \ N_NN and وول \ RP_RPD شری \ N_NN

2.3.4.2 Ambiguity

Natural language is ambiguous from its basic nature. Ambiguity occurs when a single word or a sequence of words is associated within a language with more than one meaning. Generally lexical ambiguity is of two types; syntactic lexical ambiguity and semantic lexical ambiguity. Syntactic lexical ambiguity simply refers to ambiguity of category, e.g., noun or verb. Semantic lexical ambiguity means a word has several meanings.

The following are some examples of ambiguity observed in Kashmiri PoS tagging. This was addressed with careful attention in LDC-IL Kashmiri PoS tagged text corpus.

N_NN & RP_RPD

In Kashmiri
بس چہ آہڑ۔
بس، بہ گؤس تہ آس واپس۔
In IPA
bas\N_NN tʃhə a:mts
bas\RP_RPD bɪ govus tɪ a:s wa:pas

JJ, RP_RPD and RB

In Kashmiri
یہ چھ اصل پلو۔
اصل، سہ کر او؟
سہ چھ اصل گئیوان۔
In IPA
ʃi tʃʰu asɪl\JJ palav
asɪl \RP_RPD su kar a:v
su tʃʰu asɪl \RB geva:n

N_NN & JJ

In Kashmiri
مے گڑھ ژیر۔
مے گڑھ ژیر۔
In IPA
me gatsʰi tse:r\ N_NN
me gatsʰi tse:r\ JJ

N_NN & V_VM

In Kashmiri
ٹمی دیت مے اکھ ہار۔
ٹمی ہار نوؤس بہ لڑایہ منڑ۔
In IPA
təmj djut me akʰ ha:r\ N_NN
təmj ha:r\ V_VM novus bɪ laɖa:ʃi manz

QT_QTO & N_NN

In Kashmiri
یہ نہ، بیہ گمیز ہاؤ -
اکی تُل قلم، بیہ تُل شال۔
In IPA
ji nɪ beji \QT_QTO kəmi:z ha:v
əkɪ tul kalam \beji N_NN tul ʃa:l

2.3.4.3 Abbreviation

Abbreviations are shortened forms of words or phrases that are widely used in both written and spoken language. In Kashmiri, abbreviations are commonly found in newspapers, academic writing, official documents, and digital communication. Some examples of abbreviations used in Kashmiri are: - کے جے کے، ایس ایم ایچ ایس، جے کے بی او ایس ای، جے کے - etc. بی بی سی، کے پی، جے کے، ایس ایم ایچ ایس، جے کے بی او ایس ای، جے کے کے -

2.3.4.4 Compound Word

Two or more morphemes may combine to form a single lexical unit where the entire token conveys a unified meaning. For Example, کلہ دود، دل ملول، برونہہ پهن گپن جاعے، etc. These words are treated as whole units and assigned the appropriate POS tag based on their overall grammatical function. In the case of hyphenated compounds, it is advisable to consider them as a single token and tag them according to their primary syntactic category. Conversely, in the case of non-hyphenated compounds written with a space, such as “لال بازر”، each component should be tagged separately according to its individual grammatical category. However, when compound elements occur without any space and form a single orthographic word, they should be treated as a single token and assigned one POS tag based on their overall function. For example, راج باغ.

2.3.4.5 Dialect Variation

Kashmiri exhibits significant dialectal variation across regions, characterized by differences in phonology, vocabulary, and morphology. The language displays regional stylistic variation across three principal dialect zones: mara:z (southern region), kamra:z (northern region), and yamra:z (central region). If the annotator is not familiar with such dialectal or region-specific lexical items, there is a high likelihood of errors in assigning the correct POS tag. For example وڊن، وڊان، برسُن، وڙه، ووٹ، پکن، پکان

2.3.4.6 Impact of Social Variation

Social variation in Kashmiri refers to differences in language use that arise due to social factors such as age, education, gender, religion, socioeconomic status, and urban–rural background rather than geographical location. Speakers from an urban background frequently engage in code-mixing, particularly between Kashmiri–English. As a result, a substantial number of English lexical items have entered the Kashmiri corpus. Words such as ‘پلېز’ (please), ‘ہیلو’ (hello), ‘سواری’ (sorry), ‘کرفیو’ (curfew) have become extremely common in everyday communication. Similarly, the corpus also reflects instances of Persian and Urdu lexical insertions within Kashmiri text corpus. Expressions such as ‘خودا’ (Almighty), ‘لباس’ (Clothes), ‘خواب’ (Dream), ‘انتظار’ (wait) occasionally appear in Kashmiri textual data. In case the annotator is unaware about the term it may tag as RD_UNK.

2.3.4.7 Unknown Words

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

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

Dataset Summary	Aggregate Count
Total PoS Tags	103488
Total Sentences	5331
Total Word Count	92746
Total File Count	49

Table 2-2: Kashmiri PoS Status

Sl. No.	PoS Category	PoS Tag	Tag Count
1	Common Noun	N_NN	26189
2	Proper Noun	N_NNP	6320
3	Nloc Noun	N_NST	1981
4	Personal Pronoun	PR_PRP	4677
5	Reflexive Pronoun	PR_PRF	688
6	Relative Pronoun	PR_PRL	688
7	Reciprocal Pronoun	PR_PRC	90
8	Wh-word Pronoun	PR_PRQ	130
9	Deictic Demonstrative	DM_DMD	1627
10	Wh-word Demonstrative	DM_DMQ	51

11	Main Verb	V_VM	13744
12	Auxillary Verb	V_VAUX	5907
16	Adjective	JJ	6473
17	Adverb	RB	930
18	Postposition	PSP	8652
19	Co-ordinator	CC_CCD	4502
20	Subordinator	CC_CCS	1297
22	Particles	RP_RPD	2263
23	Interjection	RP_INJ	89
24	Intensifier	RP_INTF	295
25	Negation	RP_NEG	493
26	Quantifiers	QT_QTF	1028
27	Cardinals	QT_QTC	1614
28	Ordinals	QT_QTO	227
29	Foreign word	RD_RDF	479
30	Symbol	RD_SYM	36
31	Punctuation	RD_PUNC	10706
32	Unknown	RD_UNK	1449
33	Echowords	RD_ECH	63

Table 2-3: Distinct Tag Count of Kashmiri