

اردو پارٹس آف اسپچ اینوٹسڈ کارپس



Urdu Parts of Speech Annotated Corpus

URDU 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. Mansoor Khan

Dr. Shahnawaz Alam

Bi Bi Mariyam

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

Urdu Parts of Speech Annotated Corpus

Authors: Authors: Dr. Mansoor Khan, Dr. Shahnawaz Alam, Bi Bi Mariyam, Dr. Narayan Choudhary

e-ISBN: 978-81-69175-51-7

CIIL Publication No.: 1691

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: Shahnawaz Alam

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

Tables

Table 0-1: BIS Tagset	4
Table 2-1: Urdu Tagset	8
Table 2-2: Urdu PoS Status	13
Table 2-3: Distinct Tag Count of Urdu	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>partita, 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 0-1: 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 Bay, and Rajeev Sandal, 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's peculiar grammatical characteristics and data size are mentioned in the following chapters. LDC-IL PoS annotated data based on BIS framework has 12668303 million 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 URDU PoS ANNOTATED CORPUS

Dr. Mansoor Khan

2.1 GRAMMATICAL PECULAIRITIES OF URDU

Urdu 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. Urdu 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, Urdu also contains an intensive list of borrowed words.

Urdu 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, adjectives in Urdu have gender variations based on the noun they are attached to. On the contrary, the indeclinable category contains word classes which remain unchanged or static. These include conjunction, postposition, adverbs and participle.

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 Urdu texts, certain categories such as classifiers is not applied in Urdu tagging, as these features are either rare or not present in Urdu grammatical structure. Linguistic Standards for Indian Language (IL) Corpora Annotation: PoS Categories and Labels (June 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	کتاب، قلم
1.2		Proper		NNP	N__NNP	رام، رحيم
1.3		Nloc		NST	N__NST	اوپر، نيچے
2	Pronoun			PR	PR	
2.1		Personal		PRP	PR__PRP	میں، وہ

2.2		Reflexive		PRF	PR__PRF	خود، اپنا
2.3		Relative		PRL	PR__PRL	جو، جس
2.4		Reciprocal		PRC	PR__PRC	آپس، باہم
2.5		Wh-word		PRQ	PR__PRQ	کون، کہاں
2.6		Indefinite		PRI	PR__PRI	کوئی، کسی
3	Demonstrative			DM	DM	
3.1		Deictic		DMD	DM__DMD	یہ، وہ
3.2		Relative		DMR	DM__DMR	جو، جس
3.3		Wh-word		DMQ	DM__DMQ	کون، کس
3.4		Indefinite		DMI	DM__DMI	کوئی، کسی
4	Verb			V	V	
4.1		Main		VM	V__VM	گرا، گیا
4.2		Auxiliary		VAUX	V__VAUX	ہے، ہوا
5	Adjective			JJ		دلکش، سیاہ
6	Adverb			RB		تیز، جلد
7	Postposition			PSP		سے، نے
8	Conjunction			CC	CC	
8.1		Co-ordinator		CCD	CC__CCD	اور، بلکہ
8.2		Subordinator		CCS	CC__CCS	اگر، کیوں کہ
9	Particles			RP	RP	
9.1		Default		RPD	RP__RPD	ہی، بھی
9.2		Interjection		INJ	RP__INJ	اے، ارے
9.3		Intensifier		INTF	RP__INTF	بہت، بے حد
9.4		Negation		NEG	RP__NEG	نہ، نہیں
10	Quantifiers			QT	QT	
10.1		General		QTF	QT__QTF	چند، کثیر
10.2		Cardinals		QTC	QT__QTC	ایک، دو
10.3		Ordinals		QTO	QT__QTO	پہلا، دوسرا
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	دلہول، پیارہویار

Table 2-1: Urdu Tagset

The annotation is to be done using the lowest level tag of the type hierarchy. Once the lower level tag is selected, the higher level tags should be stored automatically. The list was prepared in the PoS standards committee meeting held on June 12, 2010 at IIIT, Hyderabad.

2.3 URDU POS TAGGING

Urdu PoS tagging is performed using the BIS tagset, with necessary adaptations from LDC-IL guidelines to suit Urdu's linguistic structure. Since Urdu is an agglutinative language, preprocessing steps such as word segmentation is carried out before tagging to fit in tagset framework. To improve tagging accuracy context-based tag assignment is carried out to ensure the consistency across the corpus.

2.3.1 Guidelines

- Urdu PoS Tagging is 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 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. On the word-level, tokenization refers to breaking down of a sentence into the constituent words. For example, گانا گاتا ہوں میں will be tokenized into four tokens- گانا | میں | گاتا | ہوں

In inflectional languages such as Urdu, tokenization helps in breaking down the morphological and syntactic status of the words. Furthermore, in compound words, tokenization helps understand the layered meaning of a word.

Some compound words often pose a challenge to the tokenization process and need to be pre-processed. For example, the expressions such as غیر ذمہ دار ‘ye:r zım:e:ḡa:r’ (Negation+Noun+Suffix), خوش حال ی ‘xu:f.ɦa:li.’ (Adj+Noun+Suffix) have multiple grammatical categories but they function as a single token, tagged as a noun.

It is interesting to note that though compound words in Urdu are written using a hyphen, non-hyphenated instances can also be found. For eg: the following words have two instances of the same category (Noun). Though the grammatical combination is noun+noun, the words function as single tokens- خواب گاہ ‘xvɑ:bḡɑ:h’, سنگدل ‘səŋ.dɪl’.

2.3.3 Selecting a Tag

A word has two important elements; one is its form and other is its function. The function of a word offers significant clues for the subsequent stages of processing. So here adopt a balance between the form and the function of a word in a systematic and consistent way.

Category of a word is decided by utilizing the morphological, syntactic and semantic functions. A word’s category can be easily identified by morphological properties, but functionally it can act as another category due to context. This variation is extracted by syntactic clues. For example a noun may be an adjective if it occurs immediately before a noun.

وہ مندر کے اندر کھڑا ہے (He is standing inside the temple)

وہ قلم میز پر رکھ دو (keep that pen on the table)

In the first sentence, the token ‘وہ’ serves as a personal pronoun. However, in the second instance, it works as a demonstrative because it modifies a common noun (قلم). Similarly, indefinite pronouns such as کسی and کوئی also function as demonstratives when preceding a noun.

2.3.4 Observation on PoS Tagging of LDC-IL Urdu 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 give 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

In Urdu 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 `انٹے والا` and `کپڑے والا` or `انٹے والا` and `انٹے والا` convey similar meanings but they differ in tokenization and subsequent tagging.

`انٹے والا` \N_NN and `کپڑے والا` \N_NN or `انٹے والا` \N_NN and `انٹے والا` \N_NN

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

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 Urdu also often creates ambiguity in PoS tag assignment. However, examining the context solves this problem. Following are some examples of ambiguity observed in Urdu, demonstrating that the same token functions as a noun, particle, verb etc. depending upon the context.

RP_RPD, N_NN & V_VM

In Urdu
وہ بس میں چڑھ گیا۔
بس، میں ابھی گیا اور ابھی آیا۔
وہ امریکہ میں جا کر بس گئے۔
In IPA
vo bəs\N_NN mē: tʃəɾʰi gəja:
bəs\RP_RPD, mē: əbʰi gəja: ɔ:r əbʰi a:ja:
vo: əme:rika me:n dʒa: kər\V_VM bəs gəe

JJ, RP_RPD & RB

In Urdu
یہ گھر اچھا ہے۔
اچھا، وہ کیوں گیا تھا؟
وہ اچھا گاتا ہے۔
In IPA
je: ɡʱər ətʃʰa: \JJ hɛ:
ətʃʰa: \RP_RPD, vo: kjoː ɡəja: tʰa:
vo: ətʃʰa: \RB ɡaːtɑ: hɛ

N_NN & V_VM

In Urdu
سڑک پر گائے کھڑی ہے۔
اُس نے رفیع کی آواز میں گائے گائے۔
In IPA
səɾək pər ɡaː.e \N_NN kʰəɾi: hɛ
ʊs ne: rəfi: ki: aːvɑːz mɛː ɡaːne: ɡaː.e \V_VM

N_NN & RP_NEG

In Urdu
تمہاری مت ماری گئی ہے۔
اس موسم میں تم باہر مت کھیلو۔
In IPA
tʊmfia:ri: mət \N_NN ma:ri: ɡəi: ɦɛ
ɪs mʊ:səm mɛː tʊm baːɦər mət \RP_NEG kʰeːlo

N_NN & RP_INJ

In Urdu
رانی کی آہ سن کر راجہ دوڑا چلا آیا۔
آہ! آج بہت کام کرنا ہے۔
In IPA
ra:ni ki: a:h \N_NN sʊn kər raːdʒa ɖoːɾa: tʃʌla: aːja:
a:h! aːdʒ boːɦət ka:m kərna: hɛ

JJ & V_VM

In Urdu
اُستاد نے مجھے ہرا کپڑا دیا۔
میں نے اسے جنگ میں ہرا دیا۔
In IPA
ʊst̪ɑːd neː mʊdʒʰeː h̪əɾɑː\JJ kəp̪ɾɑː d̪jaː
mẽ ne use dʒəŋg mẽ h̪əɾɑː\V_VM d̪jaː

QT_QTO & N_NN

In Urdu
پہلا نہیں دوسرا قلم دکھاؤ۔
ایک لڑکا آیا، دوسرا چلا گیا۔
In IPA
p̪ɛɦlaː nəɦĩː duːsɾɑː\QT_QTO qələm d̪ɪkʰɑːoː
eːk ləɾkɑː aːjaː duːsɾɑː\N_NN tʃəlaː gəjaː

2.3.4.3 Abbreviation

Some people are known by the abbreviation of their name. Based on the context it can identify as proper noun or any other category. It is tagged accordingly.

Urdu	IPA
اے پی جے۔ عبدالکلام	/eː piː dʒeː əbd̪ʊl kəlaːm/
اے۔ آر۔ رحمان	/eɪ aːr rəɦmaːn/

2.3.4.4 Compound Word

Two or more word combined to form a single unit and the sole token gives a meaning. For eg: سنگدل/səŋg.dɪl/, دلفریب/dɪl.fəriːb/. These types of words are taken as such and assigned the appropriate tag.

2.3.4.5 Dialect Variation

Urdu exhibits significant dialect variations across regions which are characterized by differences in phonology, vocabulary and morphology. Varieties such as dakhni, Brajbhasha, Dehelvi etc. though often grouped under the Urdu umbrella, display very unique features setting them apart from Standard Urdu. This dialect 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 dialect vocabulary into the Urdu corpora. Some expressions from regional dialects of Urdu found in the corpus are ڈھیٹ ڈھیٹہ، بھابی بھابی، بچوں بچو، etc.

2.3.4.6 Impact of Social Dialect

Dialect variation in Urdu is often shaped by factors such as age, education, urban/rural setting, cultural exposure etc. Speakers from urban background often code-mix Urdu with English due to which a large amount of English words have entered Urdu corpus. Words such as 'کونسل', 'پلیز', 'بلیک بورڈ', 'ٹیل' etc. have become very common in day-to-day use.

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 a problem for the higher level of processing. For example, RD_UNK\شورٹ, RD_UNK\مینیسٹو, RD_UNK\اندرد etc.

2.4 SUMMARY OF LDC-IL URDU PoS DATA

Dataset Summary	Aggregate Count
Total PoS Tags	1502213
Total Sentences	63184
Total Word Count	1374380
Total File Count	755

Table 2-2: Urdu PoS Status

Sl. No.	PoS Category	PoS Tag	Tag Count
1	Common Noun	N_NN	346024
2	Proper Noun	N_NNP	83041
3	Nloc Noun	N_NST	21218
4	Personal Pronoun	PR_PRP	54014
5	Reflexive Pronoun	PR_PRF	10308
6	Relative Pronoun	PR_PRL	6487
7	Reciprocal Pronoun	PR_PRC	252
8	Wh-word Pronoun	PR_PRQ	2759
9	Indefinite Pronoun	PR_PRI	1220
10	Deictic Demonstrative	DM_DMD	31515
11	Relative Demonstrative	DM_DMR	6077
12	Wh-word Demonstrative	DM_DMQ	1383
13	Indefinite Demonstrative	DM_DMI	5317
14	Main Verb	V_VM	160564
15	Auxiliary Verb	V_VAUX	98376
16	Adjective	JJ	96328
17	Adverb	RB	14120

18	Postposition	PSP	248955
19	Co-ordinator	CC_CCD	58095
20	Subordinator	CC_CCS	26717
21	Particles	RP_RPD	31263
22	Interjection	RP_INJ	834
23	Intensifier	RP_INTF	4631
24	Negation	RP_NEG	15600
25	Quantifiers	QT_QTF	20425
26	Cardinals	QT_QTC	18775
27	Ordinals	QT_QTO	6570
28	Foreign word	RD_RDF	2515
29	Symbol	RD_SYM	1028
30	Punctuation	RD_PUNC	126805
31	Unknown	RD_UNK	964
32	Echowords	RD_ECH	33

Table 2-3: Distinct Tag Count of Urdu