

ਪੰਜਾਬੀ
ਸ਼ਬਦ ਭੇਦ ਦਾ
ਐਨੋਟੇਟਿਡ
ਕੋਰਪਸ



PUNJABI
PARTS OF SPEECH
ANNOTATED CORPUS



PUNJABI 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. Shalinder Singh
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

Punjabi Parts of Speech Annotated Corpus

Authors: Dr. Shalinder Singh, Dr. Narayan Choudhary, Rajesha N., Manasa G.

e-ISBN: 978-81-69175-64-7

CIIL Publication No.: 1694

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: Shalinder Singh

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	Punjabi PoS Annotated Corpus	6
2.1	Grammatical peculiarities of Punjabi	6
2.2	LDC-IL Customized BIS Tagset	6
2.3	Punjabi PoS Tagging	7
2.4	Summary of LDC-IL Punjabi PoS Data	12

Tables

Table 1-1: BIS Tagset	4
Table 1-2: LDC-IL PoS Status	5
Table 2-1: Punjabi Tagset	7
Table 2-2: Punjabi PoS Status	13
Table 2-3: Distinct Tag Count of Punjabi	13

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 annotations 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.

S l. N o	Category			Label	Annotati on Conventi on**	Remarks
	Top level	Subty pe (level 1)	Subtyp e (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 VNF	
4.2.3			Infinitive	VINF	V VAUX VINF	
4.2.4			Gerund	VNG	V VAUX VNG	
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

LDCIL 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 been 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 on 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 own 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 mentioned 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 PUNJABI POS ANNOTATED CORPUS

Shalinder Singh

2.1 GRAMMATICAL PECULIARITIES OF PUNJABI

Punjabi 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. Punjabi 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, Punjabi also contains an intensive list of borrowed or foreign words.

Punjabi 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 Punjabi texts, certain categories including sub-categories of verbs such as finite, non-finite, infinitive and gerund are not applied in Punjabi 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	ਘਰ
1.2		Proper		NNP	N_NNP	ਰਾਮ
1.3		Nloc		NST	N_NST	ਉੱਪਰ
2	Pronoun			PR	PR	
2.1		Personal		PRP	PR_PRP	ਮੈਂ
2.2		Reflexive		PRF	PR_PRF	ਆਪਣਾ
2.3		Relative		PRL	PR_PRL	ਜੋ
2.4		Reciprocal		PRC	PR_PRC	ਆਪਸ
2.5		Wh-word		PRQ	PR_PRQ	ਕੀ
2.6		Indefinite		PRI	PR_PRI	ਕਸਿ
3	Demonstrative			DM	DM	

3.1		Deictic		DMD	DM_DMD	ਇਹ
3.2		Relative		DMR	DM_DMR	ਜੋ
3.3		Wh-word		DMQ	DM_DMQ	ਕਸਿ
3.4		Indefinite		DMI	DM_DMI	ਕਸਿ
4	Verb			V	V	
4.1		Main		VM	V_VM	ਖਾ
4.3		Auxiliary		VAUX	V_VAUX	ਹੈ
5	Adjective			JJ		ਚੰਗਾ
6	Adverb			RB		ਹੌਲੀ
7	Postposition			PSP		ਲਈ
8	Conjunction			CC	CC	
8.1		Co-ordinator		CCD	CC_CCD	ਅਤੇ
8.2		Subordinator		CCS	CC_CCS	ਜੇ
9	Particles			RP	RP	
9.1		Default		RPD	QT	ਵੀ
9.3		Interjection		INJ	RP_INJ	ਉਹ
9.4		Intensifier		INTF	RP_INTF	ਬਹੁਤ
9.5		Negation		NEG	RP_NEG	ਨਹੀ
10	Quantifiers			QT	QT	
10.1		General		QTF	QT_QTF	ਬੋੜ੍ਹਾ
10.2		Cardinals		QTC	QT_QTC	ਇੱਕ
10.3		Ordinals		QTO	QT_QTO	ਪਹਿਲਾ
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	ਪਾਣੀ-ਪਾਣੀ

Table 2-1: Punjabi 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 PUNJABI POS TAGGING

Punjabi PoS tagging is performed using the BIS tagset, with necessary adaptations from LDC-IL guidelines to suit Punjabi's linguistic structure. Data containing information about the structural and grammatical features of Punjabi 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

- Punjabi 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 Punjabi, on the word-level, tokenization refers to segmenting a sentence into its individual components. For example, ਮੈਂ ਖਾਣਾ ਖਾਂਦੀ ਹਾਂ will be divided into 4 tokens- ਮੈਂ | ਖਾਣਾ | ਖਾਂਦੀ | ਹਾਂ. 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.

ਉਹ ਘਰ ਦੇ ਬਾਹਰ ਖੜਾ ਹੈ। (He is standing outside the house.)

ਉਹ ਕਤਾਬ ਮੇਜ਼ ਤੇ ਰੱਖ ਦਿਓ। (Keep that book on the table.)

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

In Punjabi, there exists a plethora of words which differ in their function depending upon the neighboring linguistic environment such as ‘ਡਾਲ’ , ‘ਚੋਟੀ’ , ‘ਨਾਬਾਲਗਿ’ etc.

2.3.4 Observation on PoS Tagging of LDC-IL Punjabi 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 Punjabi 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 ਸਬਜ਼ੀਵਾਲਾ and ਸਬਜ਼ੀ ਵਾਲਾ or ਦੁੱਧਵਾਲਾ and ਦੁੱਧ ਵਾਲਾ convey similar meanings but they differ in tokenisation and subsequent tagging-

ਸਬਜ਼ੀਵਾਲਾ\N_NN and ਸਬਜ਼ੀ\N_NN ਵਾਲਾ\RP_RPD or ਦੁੱਧਵਾਲਾ \N_NN and ਦੁੱਧ\N_NN ਵਾਲਾ \RP_RPD

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

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

N_NN, RP_RPD & V_VM

In Punjabi
ਉਹ ਬੱਸ\N_NN ਵੱਚਿ ਬੈਠ ਗਈ।
ਬੱਸ\RP_RPD, ਬੱਸ, ਮੈਂ ਹੁਣ ਗਿਆ ਤੇ ਹੁਣ ਆਇਆ।
ਉਹ ਅਮਰੀਕਾ ਵੱਚਿ ਬੱਸ \V_VM ਗਏ ਹਨ।
In IPA
ʊh bəs ^h \N_NN vɪtʃ bəɪ̯t̪ ɡəɪ̯
bəs ^h \RP_RPD bəs ^h mɛ̃ hʊɳ ɡəɪ̯a te hʊɳ aɪ̯a
ʊh əmɾiːkaː vɪtʃ bəs ^h \V_VM ɡəɪ̯ hən

JJ and RB

In Punjabi
ਇਹ ਕਪੜਾ ਚੰਗਾ\JJ ਹੈ।
ਉਹ ਚੰਗਾ\RB ਨੱਚਦਾ ਹੈ।
In IPA
ih kəpɾaː t͡ʃʌŋɡaː\JJ hɛː
ʊh t͡ʃʌŋɡaː\RB nət͡ʃːdaː hɛː

N_NN & V_VM

In Punjabi
ਸੜਕ ਉੱਤੇ ਲੜਕੇ\N_NN ਦੌੜ ਰਹੇ ਹਨ।
ਦੋਨੋਂ ਲੜਕੇ\V_VM ਆ ਗਏ ਹਨ।
In IPA
səɾək utte ɭəɾkɛ\N_NN d̪əɾ rəhe hən.
donõ ɭəɾkɛ\V_VM a ɡəe hən.

N_NN & RP_INJ

In Punjabi
ਸੀਮਾ ਦੀ ਉਫ਼ \N_NN ਸੁਣ ਕੇ ਰਾਜੀਵ ਦੌੜਦਾ ਆ ਗਿਆ।
ਉਫ਼! \RP_INJ ! ਅੱਜ ਬਹੁਤ ਸਾਰਾ ਕੰਮ ਕਰਨਾ ਹੈ।
In IPA
si:ma: di: ʊf\N_NN suṇ ke: ra:ɖʒi:v dɔ:ɽda: a: gɪ:a:
ʊf! \RP_INJ ədʒ bəɦʊt sa:ra: kəm kərna: hɛ:

JJ & V_VM

In Punjabi
ਅਧਿਆਪਕਿ ਨੇ ਮੈਨੂੰ ਹਰਾ\JJ ਕਾਗਜ਼ ਦੱਤਾ।
ਉਸਨੇ ਮੈਨੂੰ ਲੜਾਈ ਵੱਚਿ ਹਰਾ\V_VM ਦੱਤਾ।
In IPA
əɖɦia:pika: ne: mə:ɳu: hərə:\JJ ka:gəz dɪtta:
usne: mə:ɳu: ləɽa:i: vɪtʃ hərə:\V_VM dɪtta:

QT_QTO & N_NN

In Punjabi
ਇਹ ਨਹੀ, ਦੂਜੀ\QT_QTO ਸਾੜੀ ਦਖਾਓ।
ਇੱਕ ਕੁੜੀ ਨੇ ਕਲਮ ਲਈ, ਦੂਜੀ\N_NN ਨੇ ਕਤਿਬ।
In IPA
ɪh nəhi:, du:dʒi:\QT_QTO sa:ɽi: dɪkʰa:o:
ɪk kuɽi: ne kələm ləi:, du:dʒi:\N_NN ne kɪta:b

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 Punjabi are- ਕੇ.ਬੀ., ਐਲ.ਐਲ.ਬੀ., ਬੀ.ਡੀ.ਓ, ਬੀ.ਏ. 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, ‘ਰਾਤ-ਭਰ’, ‘ਰੰਗ-ਬਰੰਗੇ’, ‘ਉਲਟੀ-ਸਾਧੀ’, ‘ਆਉਦੇ-ਜਾਂਦੇ’, ‘ਦਲਿ-ਦਮਾਗ’ 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 ‘ਸਮੇ ਸੀਮਾ’, 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, ‘ਪੱਥਰਦਲਿ’.

2.3.4.5 Dialect Variation

Punjabi exhibits significant dialectal variations across regions which is characterised by differences in phonology, vocabulary etc. Varieties such as Haryanvi Punjabi, Rajisthani Punjabi, Delhi Punjabi, Himachali Punjabi, Canadian Punjabi etc. though often grouped under the Punjabi umbrella, display very unique features setting them apart from Standard Punjabi. 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 Punjabi corpora. Some expressions from regional dialects of Punjabi found in the corpus are ਚਲਾ ਗਿਆ (ਪੰਜਾਬ) _ ਬਗ ਗਿਆ (ਹਰਿਆਣਾ) etc.

2.3.4.6 Impact of Social Dialect

Dialect variation in Punjabi is often shaped by social factors such as age, education, urban/rural setting, cultural exposure etc. Speakers from urban backgrounds often code-mix Punjabi with English due to which a large amount of English words have entered Punjabi corpus. Words such as ‘ਪਲੀਜ਼’, ‘ਕੈਸਲ’, ‘ਹੈਲੋ’, ‘ਸਲੋ’ etc. have become very common in day-to-day use.

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 Punjabi text corpus often arise due to spelling variations, code-mixing, presence of dialectal forms etc. For example, ਫਤੁਮ \RD_UNK, ਮੀਚਾਓ \RD_UNK etc.

2.4 SUMMARY OF LDC-IL PUNJABI POS DATA

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

Dataset Summary	Aggregate Count
Total PoS Tags	1298034
Total Sentences	65452
Total Word Count	1150325
Total File Count	502

Table 2-2: Punjabi PoS Status

I. No.	PoS Category	PoS Tag	Tag Count
1	Common Noun	N_NN	278358
2	Proper Noun	N_NNP	55784
3	Nloc Noun	N_NST	22999
4	Personal Pronoun	PR_PRP	64390
5	Reflexive Pronoun	PR_PRF	9956
6	Relative Pronoun	PR_PRL	5793
7	Reciprocal Pronoun	PR_PRC	530
8	Wh-word Pronoun	PR_PRQ	2482
9	Deictic Demonstrative	DM_DMD	16787
10	Wh-word Demonstrative	DM_DMQ	895
11	Verb	V_VM	162319
12	Axulary Verb	V_VAUX	108494
13	Adjective	JJ	73753
14	Adverb	RB	13700
15	Postposition	PSP	181762
16	Co-ordinator	CC_CCD	30766
17	Subordinator	CC_CCS	18735
18	Particles	RP_RPD	36766
19	Interjection	RP_INJ	817
20	Intensifier	RP_INTF	3694
21	Negation	RP_NEG	15402
22	Quantifiers	QT_QTF	13723
23	Cardinals	QT_QTC	16177
24	Ordinals	QT_QTO	2804
25	Foreign word	RD_RDF	1694
26	Symbol	RD_SYM	1266
27	Punctuation	RD_PUNC	146443
28	Unknown	RD_UNK	1726
29	Echowords	RD_ECH	143

Table 2-3: Distinct Tag Count of Punjabi