

[illegible]

Nepali Parts of Speech Annotated Corpus

NEPALI 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:

*Umesh Chamling Rai,
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

Nepali Parts of Speech Annotated Corpus

Authors: Umesh Chamling Rai, Dr. Narayan Choudhary, Rajesha N., Prof. Shailendra Mohan

e-ISBN: 978-81-69175-03-6

CIIL Publication No.: 1696

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: Umesh Chamling Rai

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	Nepali PoS Annotated Corpus	6
2.1	Grammatical peculiarities of Nepali	6
2.2	LDC-IL Customized BIS Tagset	6
2.3	Nepali PoS Tagging	8
2.4	Summary of LDC-IL Nepali PoS Data	14

Tables

Table 1-1: BIS Tagset	4
Table 1-2: LDC-IL PoS Status	5
Table 2-1: Nepali Tagset	8
Table 2-2: Nepali PoS Status	14
Table 2-3: Distinct Tag Count of Nepali	15

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 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.

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 Nepali)

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 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 NEPALI POS ANNOTATED CORPUS

Umesh Chamling Rai

2.1 GRAMMATICAL PECULIARITIES OF NEPALI

Nepali is a morphologically rich language characterized by extensive word formation processes. Words in Nepali commonly undergo inflection, derivation, and compounding, resulting in an extremely large and virtually open-ended vocabulary. Because of these productive morphological operations, a single root or lemma can generate a wide range of word forms.

Grammatical productivity and morphological complexity pose significant challenges from the perspective of Part-of-Speech tagging. Multiple grammatical categories such as case, number, gender, tense, aspect, mood, person, and voice may be simultaneously labelled within a single word. As a result, PoS tagging in Nepali requires not only identifying the syntactic class of a token but also understanding the layered morphological information embedded within it.

A corpus effectively captures the linguistic nuances and unique characteristics of a language when it systematically represents the morphological, syntactic, and semantic variables. To accommodate the specific features of Nepali, the BIS tagset needs to be modified.

2.2 LDC-IL CUSTOMIZED BIS TAGSET

Some tags in the BIS framework, such as verbal nouns and quotatives, are rare or absent in the grammatical structure of Nepali. Similarly, only the main and auxiliary verbs have been included here. Therefore, LDC-IL customized the tagset to accomodate the specific needs of each language. The Nepali 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	भाइ
1.2		Proper		NNP	N__NNP	राम
1.3		Nloc		NST	N__NST	वहाँ
2	Pronoun			PR	PR	
2.1		Personal		PRP	PR__PRP	म
2.2		Reflexive		PRF	PR__PRF	आफू

2.3		Relative		PRL	PR__PRL	जुन
2.4		Reciprocal		PRC	PR__PRC	परस्पर
2.5		Wh-word		PRQ	PR__PRQ	को
2.6		Indefinite		PRI	PR__PRI	कोही
3	Demonstrative			DM	DM	
3.1		Deictic		DMD	DM__DMD	त्यो
3.2		Relative		DMR	DM__DMR	जो
3.3		Wh-word		DMQ	DM__DMQ	कस
3.4		Indefinite		DMI	DM__DMI	कुनै
4	Verb			V	V	
4.1		Main		VM	V__VM	गर
4.2		Auxiliary		VAUX	V__VAUX	छ
5	Adjective			JJ		रातो
6	Adverb			RB		झट्ट
7	Postposition			PSP		लागि
8	Conjunction			CC	CC	
8.1		Co-ordinator		CCD	CC__CCD	तर
8.2		Subordinator		CCS	CC__CCS	किनभने
9	Particles			RP	RP	
9.1		Default		RPD	RP__RPD	नै
9.2		classifier		CL	QT__CL	कोसा
9.3		Interjection		INJ	RP__INJ	अहो!
9.4		Intensifier		INTF	RP__INTF	साह्रै
9.5		Negation		NEG	RP__NEG	न....
10	Quantifiers			QT	QT	
10.1		General		QTF	QT__QTF	केही
10.2		Cardinals		QTC	QT__QTC	पाँच
10.3		Ordinals		QTO	QT__QTO	पहिलो
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	काम-साम

Table 2-1: Nepali 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 NEPALI POS TAGGING

LDC-IL developed a customized Nepali PoS tagset with the necessary adaptations to the BIS tagset to suit Nepali's linguistic structure. Hence, preprocessing steps such as word alignment/combine 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

- Nepali PoS Tagging is done based on BIS tagset with customization to capture specific grammatical and morphological features.
- In certain places, the words appeared separately (भइ लै); therefore, pre-processing has been done because to fit the words according to the tags.
- 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 important because accuracy of the PoS tagged data is 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 agglutinative languages like Bodo, Manipuri, Malayalam, Tamil, Telugu and Kannada two or three words are combined together to form a complex word. Therefore, tokenization is needed because the accuracy of PoS tagged data depends on the context of the words. These words are from different categories or from same category. Such instances it is needed to pre-process the word to get desired category. Otherwise that complex word cannot be tagged with a single PoS tag.

However, in a language like Nepali, which belongs to the Indo-Aryan language family, such issues are rarely observed; therefore, these processes have only been carried out where necessary.

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. Semantic function of a word is also useful to determine the lexical category. The Nepali word अन्त/अन्त/ is a noun location (NST) but sometimes it can act as other category. Following examples where /अन्त/ is act as particle and as noun location (NST).

In Nepali	
यहाँ\N_NST	मात्र\RB हेर्यौ\V_VM कि\CC_CCD अन्त\N_NST पनि\RP_RPD खोज्यौ\V_VM?\RD_SYM
उ\PR_PRP	मात्र\RB गयो\V_VM अन्त\RP_RPD !\RD_SYM
In IPA	
jʌɦã\N_NST	maʈɾʌ\RB ɦerjau\V_VM ki\CC_CCD\N_NST ʌɳtʌ\N_NST pʌɳi\RP_RPD kʰodzjau\V_VM?\RD_SYM
u\PR_PRP	maʈɾʌ\RB gʌjo\V_VM ʌɳtʌ\RP_RPD!\RD_SYM

2.3.4 Observation on PoS Tagging of LDC-IL Nepali 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.

The majority of nouns, pronouns, adjectives, verbs, adverbs etc. are correctly 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 Nepali, 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. Goddess name, while often tagged as proper nouns, can sometimes function as adjective quantifier and common noun as well. पञ्चकन्या/pʌŋtʂakʌŋja/ and पञ्च/pʌŋtʂʌ/ कन्या/kʌŋja/.

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

पञ्च \QT_QTF कन्या\ N_NN and पञ्चकन्या\ N_NNP

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 Nepali PoS tagging. This was addressed with careful attention in LDC-IL Nepali PoS tagged text corpus.

N_NN & N_NNP

In Nepali											
कमलले\	N_NNP	कमलको\	N_NN	फूलमा\	N_NN	मौरी\	N_NN	देख्यो\	V_VM	I\	RD_PUNC
आकाश\	N_NNP	आकाशतिर\	N_NN	हेर्छ\	V_VM	I\	RD_PUNC				
In IPA											
kʌmʌlʌle\	N_NN	kʌmʌlʌko\	N_NN	phʌlʌma\	N_NN	mauri\	N_NN				
dʌkʰjo\	V_VM	I\	RD_PUNC								
akʌʌ\	N_NNP	akʌʌtʃirʌ\	N_NN	hertʃʌ\	V_VM	I\	RD_PUNC				

N_NN & V_VM

In Nepali	
बस\N_NN	आइरहेको\V_VM छ\V_VAUX I\RD_PUNC
तिमी\PR_PRP	घरमा\N_NN बस\V_VM I\RD_PUNC
In IPA	
bASA\N_NN	airAheko\V_VM tshA\V_VAUX I\RD_PUNC
timi\PR_PRP	gʱArAmA\N_NN bASA\V_VM I\RD_PUNC

N_NN & RB

In Nepali	
पुस्तकमा\N_NN	पानीको\N_NN छिटो\N_NN पर्यो\N_NN I\RD_PUNC
केटो\N_NN	छिटो\RB छिटो\RB हिड्छ\N_NN I\RD_PUNC
In IPA	
puṣṭakAmA\N_NN	paɲiko\N_NN tshito\N_NN parjo\N_NN .\RD_PUNC
keto\N_NN	tshito\RB tshito\RB hɪdʃtsʰA\N_NN I\RD_PUNC

N_NST & RB

In Nepali	
तिमी\PR_PRP	भरै\N_NST जाऊ\N_NN I\RD_PUNC
उ\PR_PRP	भरै\RB भरै\RB गर्छ\N_NN I\RD_PUNC
In IPA	
timi\PR_PRP	bʱarai\N_NST dzau\N_NN I\RD_PUNC
u\PR_PRP	bʱarai\N_NST bʱarai\N_NST gaɽtsʰA\N_NN I\RD_PUNC

N_NN & RP_INJ

In Nepali	
नानीको\N_NN	उफ\N_NN सुनी\N_NN आमा\N_NN छक्क\RB परी\N_NN I\RD_PUNC
उफ\RP_INJ	यी\PR_PRP सबै\QT_QTF काम\N_NN मैले\PR_PRP नै\RD_RPD गर्नु\N_NN I\RD_PUNC
In IPA	

ṇaṇjiko\N_NN I\RD_PUNC	upʰa\N_NN	suṇj\V_VM	ama\N_NN	tʃʰak:ʌ\RB	pʌri\V_VM
upʰa\RP_INJ ṇai\RD_RPD	ji\PR_PRP gʌrṇu\ V_VM	sʌbai\ QT_QTF I\RD_PUNC	kamʌ\N_NN	maile\PR_PRP	

JJ & RB

In Nepali					
पुण्य\N_NNP	राम्रो\JJ	विद्यार्थी\N_NN	हो\V_VM	I\RD_PUNC	
छिरिङ्\N_NPP	साह्रै\RP_INTF	राम्रो\RB	बोल्छिन्\V_VM	I\RD_PUNC	
In IPA					
pʊɲjʌ\N_NNP	ramro\JJ	widjartʃi\N_NN	ɦo\V_VM	I\RD_PUNC	
tʃʰiriŋ\N_PP	safirai\RP_INTF	ramro\RB	bolʃʰiɳ\V_VM	I\RD_PUNC	

QT_QTO & N_NN

In Nepali					
तिमी\PR_PRP	दोस्रो\QT_QTO	लहरमा\N_NN	बस\V_VM	है\RD_RPD	I\RD_PUNC
आज\N_NST	यसले\PR_PRP	दूध\N_NN	खायो\V_VM	दोस्रोले\N_NN	घाँस\N_NN I\RD_PUNC
In IPA					
ʈimi\PR_PRP	ɖosro\QT_QTO	lʌɦarʌma\N_NN	bʌsʌ\V_VM	ɦai\RD_RPD	
adza\N_NST	jʌsʌle\PR_PRP	ɖuɖʱʌ\N_NN	kʰajo\V_VM	ɖosrole\N_NN	gʱãsʌ\N_NN

RP_NEG & RP_RPD

In Nepali					
न\RP_NEG	गए\V_VM	हुन्न\V_VM	र\RD_RPD	I\RD_PUNC	
तिमी\PR_PRP	काम\N_NN	गर\ V_VM	न\RD_RPD	I\RD_PUNC	
In IPA					
ɳʌ\RP_NEG	gʌe\V_VM	ɦuɳ:ʌ\V_VM	rʌ\RD_RPD	I\RD_PUNC	
ʈimi\PR_PRP	kamʌ\N_NN	gʌrʌ\V_VM	ɳʌ\RD_RPD	I\RD_PUNC	

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 Nepali are- ए.बी., आई.पी.एल., पी.एम. 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, 'हर-दिन', 'भित्र-बाहिर', 'तल-माथि', 'खाँदै-बोल्दै', 'गाँउ-घर' 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 'खर्च पानी', 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, 'सफामन'.

2.3.4.5 Dialect Variation

It is possible for a language to have regional dialectal variations. Similarly, the Nepali language also shows dialectal variations depending on location. If an annotator does not have proper knowledge of dialectal variation or cannot distinguish it correctly, it may lead to misinterpretation of meaning, or the annotator may simply tag it as “unknown” and leave it as such.

For example, in some regions where Nepali is spoken, 'cauliflower' is referred to as 'बन्दकोबी,' while in other regions it is called 'काउली.' Likewise, in some places 'round chili' is called 'भ्यान्टे,' whereas in other places it is known as 'डल्ले.' If an annotator is not familiar with such regional vocabulary, they may make errors during annotation.

2.3.4.6 Impact of Social Dialect

Dialect variation in Nepali is often shaped by social factors such as age, education, urban/rural setting, cultural exposure etc. Speakers from urban backgrounds often code-mix Nepali with English due to which a large amount of English words have entered Nepali corpus. Words such as 'प्लीज', 'सरी', 'हेलो', 'स्लो' etc. have become very common in day-to-day use. Similarly, instances can be found in Nepali corpus where expressions from other languages such as Hindi, Bengali are also present. For example, 'अच्छा गर्यो' (Hindi) and 'वर्णाली दी' are in Hindi and Bengali language but written using Devanagari. Words from other languages such as भाल, खुरूमजरी etc. can also be observed.

2.3.4.7 Classifier

In Nepali, classifiers are words used with numbers to count certain nouns, especially when referring to people, objects, or items in formal or careful speech. The most common general classifier is 'वटा (wota)', used for counting things (e.g., दुईवटा किताब – two books). For people,

‘जना (jana)’ is used (e.g., तीनजना विद्यार्थी – three students). These words have been tagged according to their presence.

2.3.4.8 Unknown Words

There will be some words in the text which cannot be handled by the tagset. Such words are considered as unknown words and it will pose problem for the higher level of processing. Unknown words in Nepali text corpus often arise due to spelling variations, code-mixing, presence of dialectal forms etc. For example, बारोमोइ, वनङ्ककम\RD_UNK, गासोडी\RD_UNK etc.

2.4 SUMMARY OF LDC-IL NEPALI POS DATA

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

Dataset Summary	Aggregate Count
Total PoS Tags	820475
Total Sentences	54540
Total Word Count	682565
Total File Count	1483

Table 2-2: Nepali PoS Status

I. No.	PoS Category	PoS Tag	Tag Count
1	Common Noun	N_NN	222130
2	Proper Noun	N_NNP	37109
3	Nloc Noun	N_NST	22208
4	Personal Pronoun	PR_PRP	48041
5	Reflexive Pronoun	PR_PRF	4926
6	Relative Pronoun	PR_PRL	1506
7	Reciprocal Pronoun	PR_PRC	160
8	Wh-word Pronoun	PR_PRQ	1682
9	Deictic Demonstrative	DM_DMD	4481
10	Wh-word Demonstrative	DM_DMQ	14
11	Verb	V_VM	157248
12	Axulary Verb	V_VAUX	17497
13	Adjective	JJ	46532
14	Adverb	RB	27285
15	Postposition	PSP	5807
16	Co-ordinator	CC_CCD	26641
17	Subordinator	CC_CCS	1469

18	Particles	RP_RPD	20629
19	Interjection	RP_INJ	319
20	Intensifier	RP_INTF	978
21	Negation	RP_NEG	62
22	Quantifiers	QT_QTF	20018
23	Cardinals	QT_QTC	3066
24	Ordinals	QT_QTO	1027
25	Foreign word	RD_RDF	1160
26	Symbol	RD_SYM	82
27	Punctuation	RD_PUNC	137828
28	Unknown	RD_UNK	3304
29	Echowords	RD_ECH	2297

Table 2-3: Distinct Tag Count of Nepali