Annotating Bhojpuri Corpus using BIS Scheme

Conference Paper · July 2018

CITATIONS
5

READS
31

2 authors, including:

Srishti Singh
Jawaharlal Nehru University
10 PUBLICATIONS **31** CITATIONS

SEE PROFILE

Annotating Bhojpuri Corpus using BIS Scheme

Srishti Singh and Esha Banerjee
Jawaharlal Nehru University
New Delhi, India
{singhsriss, esha.jnu}@gmail.com

Abstract

The present paper talks about the application of the Bureau of Indian Standards (BIS) scheme for one of the most widely spoken Indian languages 'Bhojpuri'. Bhojpuri has claimed for its inclusion in the Eighth Schedule of the Indian Constitution, where currently 22 major Indian languages are already enlisted. Recently through Indian government initiatives these scheduled languages have received the attention from Computational aspect, but unfortunately this non-scheduled language still lacks such attention for its development in the field of NLP. The present work is possibly the first of its kind.The BIS tagset is an Indian standard designed for tagging almost all the Indian languages. Annotated corpora in Bhojpuri and the simplified annotation guideline to this tagset will serve as an important tool for such well-known NLP tasks as POS- Tagger, Phrase Chunker, Parser, Structural Transfer, Word Sense Disambiguation (WSD), etc.

Keywords: Bhojpuri annotation, BIS, Classifiers, ergativity, and word formations.

1. INTRODUCTION

Bhojpuri is one of the major Indo-Aryan languages of north India which has been given code ISO 639-3among world languages. It is spoken in the Uttar Pradesh province of India mainly in Mirzapur, Ghazipur, Jaunpur, Ballia Gorakhpur Deoria, Basti, Azamagrh and Varanasi districts whereas in the Bihar province Rohtas, Eastern Champaran, Saran, Siwan, Ranchi and Bhojpur districts are major Bhojpuri speaking belt. Besides this, it is one of the official Languages in Nepal and Mauritius. It is also spoken in Guyana, Fiji, Uganda and in some parts of Burma. According to 2001 census the number of Bhojpuri speakers was 33,099,497 in India. This big population speaks the language with some regional differences as has been reported in Upadhyay, 1988:

1. More prestigious variety: Bhojpur and Rohtas in Bihar; and Ballia and Ghazipur in U.P.
2. Western Bhojpuri: Varanasi and nearby districts in U.P.
3. Madheshi spoken in Tihat and Gorakhpur.

The intelligibility ratio among different varieties of the language is very high (their syntactic structure and basic lexicons are common) but other elements such as affixes, auxiliaries, address terms, kingship terms and domain specific terms differs a lot.

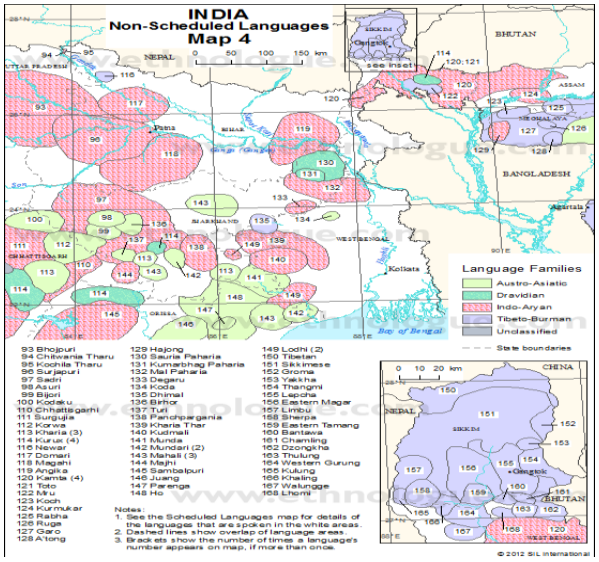

Figure 1: Map of non-schedule languages of Eastern India[1]

Bhojpuri entertains SOV word order, postpositions and final noun head; word formation may have 1 prefix, up to 5 suffixes; clause constituents indicated by both case-marking and word order; verbal affixation marks person, number and genders of subject and object. It is an ergative less non-tonal language with 34 consonant and 6 vowel phonemes, about 4 diphthongs. Bhojpuri writing system follows Devanagari and Kaithi script. Some magazines, newspapers, radio programs, dictionaries, are available in the language. Nowadays, local TV channels and films are the most popular resource for Bhojpuri. The domain specific Bhojpuri

[1] https://www.ethnologue.com/

corpus is also being developed for different fields especially media and entertainment. The Tagset of Bhojpuri will enable Bhojpuri corpus and other tools such as parser, chunker, morphological analyzer etc. to work better.

2. BIS TAGSET

The BIS tagset is a national standard tagset for Indian languages that has been recently designed under the banner of Bureau of Indian Standards by the Indian Languages Corpora Initiative (ILCI) group. This is a hierarchical tagset and allows annotation of major categories along with their types and subtypes. In this framework the granularity of the POS has been kept at a coarser level. Thus, the hierarchy for most POS categories is only of two levels. The maximum depth for the POS tags is three levels so far. Most of the categories of this tagset seem to have been adapted either from the MSRI or the ILMT tagset. For morphological analysis it will take help from Morphological Analyzer, so morpho-syntactic features are not included in the tagset. The BIS scheme is comprehensive and extensible; captures appropriate linguistic information, and also ensures the sharing, interchangeability and reusability of linguistic resources (Gopal, 2012).

2.1 POS TAGGING

POS tagging (or morpho-syntactic tagging) is the process of assigning to each word in a running text a label which indicates the status of that word within some system of categorizing the words of that language according to their morphological and/or syntactic properties (Hardie, 2003). For natural language processing tasks, annotated corpus of a language has a great importance. Annotated corpora serve as an important tool for such well-known NLP tasks as POS-Tagger, Phrase Chunker, Parser, Structural Transfer, Word Sense Disambiguation (WSD), etc.

The description of the tagset is given in table 1.

Sl. No.	Category		Annotation Convention**	Examples	Remarks
	Top level	Subtype (level 1)			
1	Noun		N	Darwaza,	
1.1		Common	N_NN	DarwAzA,	
1.2		Proper	N_NNP	Bharbittan	
1.3		Nloc	N_NST	agwan,	
2	Pronoun		PR	jaun, je ke,	
2.1		Personal	PR_PRP	Okar, u, tu,	
2.2		Reflexive	PR_PRF	Apan, apan	
2.3		Relative	PR_PRL	Jaun, je ke	
2.4		Reciprocal	PR_PRC	Ek dusar, ,	
2.5		Wh-word	PR_PRQ	Kaun, kekar	
2.6		Indefinite	PR-PRI	Koi, kisi	
3	Demonstrative		DM	Ehar, hoary,	
3.1		Dictic	DM_DMD	Ehar, ohar,	
3.2		Relative	DM_DMR	Jaun, je	
3.3		Wh-word	DM_DMQ	kA, kaun	
3.4		Indefinite	DM_DMI	Koi, kisi	
4	Verb		V	roye, nahiwa	
4.1		Main	V_VM	roye, nahiwa	
4.2		Auxiliary	V_VAUX	bhael, lagal	
5	adjective		JJ	garib	
6	Adverbs		RB	bari , dhire	
7	prepositions		PSP	kA, ke, kar	
8	Conjunctions		CC	jabki, ki	
8.1		co-ordinator	CC_CCD	ya, balki	
8.2		Subordinator	CC_CCS	Magar, to,	
9	Particles		RP	To, hi, bhi,	
9.1		Classifier	RP_CL	Tho, The,	
9.2		Default	RP_RPD	To, hi, bhi,	
9.3		Interjection	RP_INJ	Are, he, e,	
9.4		Intensifier	RP_INTF	kHoob,	
9.5		Negation	RP_NEG	nAhi , nA	
10	Quantifiers		QT	Pura, sab,	
10.1		General	QT_QTF	purA, sab,	
10.2		Cardinals	QT_QTC	Ek, du	
10.3		Ordinals	QT_QTO	dusar, tisar	
11	Residuals		RD		
11.1		Foreign word	RD_FW		A word written in script other than the script of the original text
11.2		Symbol	RD_SYM	$, &, *, (,)	for symbols such as $, &, etc
11.3		Punctuation	RD_PUNC	।, ., ?, :, :	only for punctuations
11.4		Unknown	RD_UNK		
11.5		Echowords	RD_ECH	(chup-chAp)	

Table 1: Bhojpuri Tagset

2.2 Corpus and Data

The data for the present experiment is a collection of 9 folk stories currently having approximately 5,300 tagged words (3 stories). The data has been collected in spoken form and then transcribed. It includes two major dialects: 'Bhojpuri' spoken in Bhojpur and 'Benarasi' spoken in Varanasi. The examples below contain the interlinear glossing, free translation and PsOS tagged data (if required) of the sentences.

2.3 About TAGSET

The present tagset includes 33 categories divided into eleven major Parts of Speech catagories which are further sub divided among some lower level catagories designed in accordance to the utility of the tool. A word belonging to a particular lexical category may function differently in a given context. Besides adjective, adverb and preposition all other categories have some further distributions. For guidelines see appendices.

3. Characteristics of Bhojpuri

This section discusses the characteristics of Bhojpuri in detail with examples:

3.1 Classifiers

Like Bangla and Maithali, Bhojpuri also heavily use classifiers with numerals. The marker for classifiers in different dialects of language is different as: 'Tho'**,** 'go', 'The', Kho etc

1. There were two brothers.

du	go	bhAI	rahasan
Two	CL	brother	live.PST.HON

This confirms to the shifting paradigm among Indian languages where there is Awadhi on one hand (classifier less language) and Bengali (classifier rich language) on the other hand giving place to Bhojpuri somewhere in between.

3.2 Ergative

In Bhojpuri, there is no overt ergative case marker available with the nouns but the inflection of the verb represents the perfective aspect of the sentence which makes Bhojpuri a Eragtive less Language like Awadhi.This construction is quite different from Hindi which is a Ergative language In some dialects of Bhojpuri the Perfective is marked with the –l- or –les- forms. Following are examples :

2. Ganesu said that he would do everything.

ganesu	kahales
Ganesu.3MSg.	say.ERG.PST
Ham sab	kAm
1MSg all	work.ACC
Karab	
do.FUT.1MSg	

Tagged:

ganesu\N_NNP kahales\V_VM ham\PR_PRP sab\QT_QTF kAm\N_NN karab\V_VM

3. How much sugarcane did that old man eat.

kitnA	U	buddhA
How much	that	old-man
U.nkh	chuhales	
sugarcane	eat.ERG.PST	

Tagged:

kitnA\RP_INJ U\DM_DMD buDDHA\N_NN U.nkh\N_NN chuhales\V_VM ।

The ergativity in Bhijpuri is absent and the aspectual marker is there to make the construction sensible, therefore, we can not include it to any of the catagories at this level of annotation.

3.3 Word-Formation Process

With respect to Hindi morphology, Bhojpuri morphology differs greatly in certain categories. For instance, in Hindi the particle for emphatic 'hi' and 'bhi' are placed under Default Particle whereas in Bhojpuri these elements sometimes occur independently as a separate unit and most of the times these are merged with the host giving rise to a new subcategory under some major categories of the tagset.

For sample, some data paired with its respective variants are presented here to make the point more clear:

a.	biswAs	hI	→ biswAse
	Belief	EMPH	→ belief-EMPH
b.	koI	bhI	→ kauno
	Anybody	EMPH	→anybody-EMP
c.	tabhI	to	→tabbe/tabbae
	then	EMPH	→then-EMPH
d.	kabhI	to	→kabbo
	Sometime	EMPH	→sometime-EMPH

4. Nobody believed.

kaunoM	biswAse
Anybody.3MSg.EMPH	believe.EMPH
nahIM kareM	
not do.PRF.PST	

Tagged:

kaunoM\PR_PRI biswAse\N_NN nahIM\RP_NEG kareM\V_VM ।\RD_PUNC

5. They went and knocked the door still nobody woke up.

jA	ke	kiwADI

Go do door
khaT-khaTAwat hauan tabbo
knock.RDP aux.PRS.Pl still

Tagged:
jA\V_VM ke\V_VAUX kiwADI\N_NN khaT-khaTAwat\V_VM hauan\V_VAUX tabbo\RB nA\RP_NEG koI\PR_PRI utHat\V_VM hau\V_VAUX \RD_PUNC

3.4 Determiners

Determiners are such unique feature of the language which is not there in Hindi. These are basically the discourse particles but the present work is restricted to its syntactic utility only.

Like Maithili, Awadhi, and Bengali, Bhojpuri also gives a wide space for determiners to set in. Determiners in Bhojpuri can occur with almost all the common and proper nouns. Although these determiners are not found with honorific nouns, most of the address terms are suffixed heavily with Determiners. But like emphatic category in Bhojpuri the determiners also get merged with their host nouns and they need to be excluded from the main category of the tagset and also a new category under noun can be proposed where those nouns containing a demonstrative can be put.

Word final sounds	**-a/-A**	**-i/-I**	**-u**
Determiners	-vA	-jA	-A

From the tale above, we get the notion that the word final sound of a noun is responsible for the occurrence of determiners in Benarasi. A word ending with –a or –A sound will take –vA suffix, -i/-I will take –jA suffix and –u will take an –A suffix. Similar constructions are found in Maitheli, Magahi, Awadhi and other related languages.(Kachru, 1980)

Generally, such constructions determines which noun is talked about as infered from the examples below:

6. If Bharbittan would be there,he would have offered us abundance of guavas.
   bharbittanwA
   Bharbittan.3MSg.**DEM**
   bhAI rahat ta
   brother.3MSg/NOM be.PRF EMPH
   amrood kHoob toD-toD ke
   guava very pluck.RDP PP
   khiyAwat
   eat.CAUS.PST

Tagged:
bharbittanwA\N_NN bhAI\N_NN rahat\V_VM ta\CC_CCS amrood\N_NN khoob\RP_INTF toD – toD\V_VM ke\V_VAUX khiyAwat\V_VAUX

7. He had no money to buy the garland.
   paisawe nAhI rahal
   Money.DEM not live.PST
   ki mAlA kHarIde
   that garland buy.3MSg.PRS

Tagged:
paisawe\N_NN nAhI\RP_NEG rahal\V_VM ki\CC_CCS mAlA\N_NN kHarIde\V_VM

8. Jackal said to lamb to go to the well along to see who is fairer.
   siyarA lahalas
   Jackal.3MSg.DEM say.PST
   memanI se chalA-chalA
   lamb PP lets go.RDP

Tagged:
siyarA\N_NN lahalas\V_VM memanI\N_NN se\PSP chalA\V_VAUX -\RD_SYM CHALa\V_VAUX kua.n\N_NN meM\PSP dekHal\V_VM jAI\V_VAUX ke\PR_PRQ gor\JJ bA\V_VAUX ।

But the synthetic nature of the language makes it again difficult to give a separate tag for a bound morpheme. Therefore, this issue is presently left to be sorted out with the help of a more robust morphological analyzer and the Determiners are not given a separate category by the time.

3.5 Homophonous

In Bhojpuri homophonous cases are prevalent and this makes annotation task difficult. A human annotator needs to see the tokens in the given context and according to their function assign a proper tag. Some samples:

9. The lion started to roar loudly.
   ser mAre dahADe lagal
   Lion loudly roar started.aux.PST

Tagged:
ser\N_NN mAre\RP_INTF dahADe\V_VM lagal\V_VAUX

10. Then they all were about to beat him.
    ta sab mAre jAt
    CONJ all beat go.PST live
    rahalan
    aux.PROG.PST

Tagged:
ta\RP_RPD sab\N_NN mAre\V_VM jAt\V_VM rahalan\V_VAUX

11. He put the bundle on the ground with a thut.

U le jA ke
He.3MSg take go do
gaTHariyA paTak delas
bundle throw give.PRF.PST

Tagged:
U\PR_PRP le\V_VM ja\V_VM ke\V_VAUX
gaTHariyA\N_NN patak\V_VM delas\V_VM

12. He used to purchase the garlands with that money
ta U paisa ke
CONN. that money PP
rojAnA mAlA
daily garland.3FSg
kHaride
buy.PRF.PST
.

Tagged:
ta\RP_RPD U\PR_PRP paisa\N_NN ke\PSP
nA\RP_RPD ,\RD_PUNC rojAnA\N_NST
mAlA\N_NN-|RD_PUNC phool\N_NN kharide
\V_VM

3.5 Different Realizations in Spoken Bhojpuri

Since no natural language is free from ambiguities, Bhojpuri morphology also carries functional ambiguities where one form can be interpreted in more than one functions. It shows great variations in the spoken form of the language. A single entity is pronounced differently in different context and places of their occurances. Within a variety, these variations can be easily noticed at conjunction, particle and preposition's end. These make it difficult for the tool (a POS tagger or analyzer) to process efficiently on the data and sometimes gives bad results.
As explained below:

13. And he too ate.
aur U bhI Apan
And he EMPH he.REFL
khsayiles
seat.PRF.PST .

Tagged:
aur\CC_CCD U\PR_PRP bhI\RP_RPD Apan\RP_PRF
kHayiles\V_VM ।\RD_PUNC

14. And he climbed up.
Ta U chaDH gael .
And he.3MSg climb go.PRF.PST

Tagged:
a\CC_CCD U\PR_PRP chaDH\V_VM gael\V_VM

4. Annotation Challenges

The very first challenge with digitizing any language is to have the available data in the desired domain and format.The corpus data here is an extraction from the spoken data which is transcribed into written corpus, for the task. Bhjpuri is an Ergativeless and classifier rich language which places it somewhere in between the two langugages 'Hindi' and 'Awadhi'. The different realizations of the same lexicon and their correct categorization is as challangng as finding out the homophonous words in the data and differentiating their meanings contextually. Determiners inflected with the nouns and the floating particles occurring in between the proper nouns, though not so big a problem at this level, might seek significant attention at other levels of annotations.

5. Conclusion

Bhojpuri, being a developing language, requires more attention which can be attained by generating more NLP resources. However, the corpus annotated with this tagset would be more useful as it is tagged by a standard tagset/scheme. This will maximize the usage of sharing tagged data. The initiative for tagging less resourced Indian languages with the present standard tagset is a promising effort in this direction with the hope that all Indian languages corpora annotation programmes will follow these linguistic standards for enriching their linguistic resources.

6. References

Bhaskaran, S., Bali, K., Bhattacharya, T., Bhattacharya, P., Choudhury, M., Jha, G.N., S, Rajendran, K. Sravanan, L. Sobha and Subbarao KVS. (2008).A Common Parts-of-Speech Tagset Framework for Indian languages. In *Proceeding of 6th Language Resources and Evaluation Conference (LREC, '08)*.

Baskaran, S., Bali, K., Bhattacharya, T., Bhattacharya, P., Choudhury, M., Jha, G.N., Rajendran S., Sravanan, K., Sobha, L. and Subbarao K.V.S. (2008). Designing a common POS-Tagset Framework for Indian Languages. In *Proceeding of VI workshop on Asian Language Resources, IIIT, Hyderabad*.

Gopal, M. (2012). Annotating Bundeli Corpus Using the BIS POS Tagset. In *Proceeding of workshop on Indian Language Data: Resources and Evaluation, under (LREC'12)*, pp. 50-56.

Hardie, A. (2003). *The Computational Analysis of Morphosyntactic Categories in Urdu*. PhD Thesis Lancaster University.

Jha, G. N., Gopal, M. and Mishra, D. (2009). *Annoating Sanskrit Corpus: Adapting IL-POSTS*. Spriger Heidelberg Dordrecht London Ney York.

Kachru, Yamuna.(1980) *Aspects of Hindi Grmmar*. Manohar, niversty if Michigan.

Mitkov, R. (2003). *The Oxford Handbook of Computational Lingustics.* Oxford University Press, New York.

Upadhyay, H. S. (1988). *Bhojpuri folksongs from Ballia*. India Enterprises Encorporated.

http://www.ethnologue.com/language/bho

http://shiva.iiit,ac,in/SPASAL2007/iiit_tagset_guidelines.pdf

Appendix
Giudelines to Bhojpuri Tagset

1. NOUN (N)

The first major category in the Tagset is 'nouns'. Though the categorization is made keeping in view that it can broadly cover all the dialects and sub dialects of Bhojpuri, therefore it has three sub types in it.

1.1 Common noun (N_NN)

The nouns that simply function as noun and are content words should be marked as the common noun. This includes the general variety of all the nouns, e.g. darawAzA, samay, log, sAdhU etc.
For example:

1) Ab rAnI\N_NN tiyAr hot hain
   Now the queen is getting ready.

1.2 Proper noun (N_NNP)

Proper nouns are generally names that stands for some particular person or place. For example Bharbittan, GanesU, Chunamun etc.

2) Cunmun\N_NNP DerAyal
   Cunmun got afraid.

1.3 Spatio-temporal noun (N_NST)

There are a specific set of words that functions both as preposition and argument of a verb.Such words are marked as spatio- temporal irrespective of their function in a given context. Some of them are agaweM, pacHaweM, upaM, nichaweM etc.

3) Phir Age\N_NST Gael
   Then he went forward

2. PRONOUNS (PR)

The category of pronoun has been divided into six sub-categories. These include personal, reflexive, relative, reciprocal, wh-word and indefinite. These categories should be self-explanatory and follows the same definitions as posited in common linguistic literature.

2.1 Personal Pronouns (PR_PRP)

Personal pronouns cover all the pronouns that denotes a person, place or thing. This includeS all their cases as well for example: okar, U, toke, tU, ham etc.

4) ser kahalA - ab maiM\PR_PRP tumko khAungA
   Lion said, now I will eat you.

2.2 Reflexive Pronouns (PR_PRF)

Reflexive pronouns are the ones that denote to ownership to its antecedent which can be either a noun or a pronoun. There are only a few words in this category, namely Apan, apan, khud etc.

5) tab tU Apan/PR_PRF dHolak le ke bhAg jAe
   Then you run away with your Dholak.

2.3 Relative Pronouns (PR_PRL)

The relative pronouns are those pronouns whose antecedent can be either a noun or a pronoun. However, these pronouns do not make any difference in number or gender as in the case of personal pronouns. The relative pronoun in Bhojpuri are jaun, jeke, jahAM etc.

2.4 Recipocal Pronouns (PR_PRC)

Reciprocal pronouns denote some reciprocity. This is commonly denoted by ek dusar, Apas, apne meM etc.

2.5 Wh-pronouns (PR_PRQ)

The wh-word pronouns are typically the pronouns that are used to ask questions. These words are kaun/ke, kab, kehar etc.

2.6 Indefinite Pronouns (PR_PRI)

The indefinite pronouns refer to unspecified objects, places or things. These words are koi, kisietc.

3. DEMONSTRATIVES (DM)

The category of demonstrative has been separated from the category of pronouns as the demonstratives mainly indicate about a noun and does not act as anaphora. The demonstratives have been sub-categorized into four divisions- deictic, relative, wh-words and indefinite.

3.1 Deictic (DM_DMD)

The deictic demonstratives are default demonstratives that demonstrate the noun it modifies. The deictic demonstratives in Hindi are typically I, U, ehar, ohar, je etc. They generally occur before a noun.

3.2 Relative Demonstrative (DM_DMR)

The relative demonstrative occur in the same form as the relative pronoun. The difference is only that these relatives are always followed by a noun that it modifies.For example jaun, je etc.

3.3 Wh-Word Demonstrative (DM_DMQ)

The wh-demonstratives are the same question words as wh-pronouns. The difference is that in their demonstrative function they do not ask question, rather only demonstrates. The wh-word demonstratives in Bhojpuri are kA, kaun etc.

3.4 Indefinite Demonstratives (DM_DMI)

Like indefinite pronouns, the indefinite demonstratives refer to unspecified objects, places or things. These words are koi, kisi etc.

4. VERBS (V)

The verbs are sub- divided into two only- Main (V_VM) and Auxiliary(V_VAUX) Verbs. While the

auxiliary verb is a closed set of verb, the main verb can be anything from a root verb to any of its inflected forms. Each sentence or clause must have a main verb. A sentence can have one more auxiliary verbs.

6) kaunoM biswAse nahiM kareM\V_VM
Nobody believed.

5. ADJECTIVES (JJ)

Adjective is a single whole category. There is one definition for an adjective which is self-explanatory. These are mostly attributive adjectives.

7) bahut garIb\JJ rahal
He was very poor.

6. ADVERB (RB)

Adverb also is mono-category part-of-speech. The standards document says that the category of adverb (RB) is only for manner adverbs. For example, words like chAhe jaise bhi, tabhiM, bArI, dhIre etc.

8) bAbA ke saNge , jaise\RB tU log tAs khelalA waise\RB U pAsA kheleM
The way you play cards, he used to pay 'pasa'.

7. POSTPOSITION (PSP)

Postpositions are all the parts-of-speech that work as case marker. Words like meM, se, kA, ke, kar etc. are examples of postposition.

9) hAr, bhagwAn jI ka\PSP, herA gael
The god's necklace was lost.

8. CONJUNCTION (CC)

Conjunctions words act as joiners of phrases or clauses within a sentence. The category of conjunction has been divided into two sub-categories of coordinator and subordinator.

8.1 Co-ordinating conjunctions (CC_CCD)

Coordinators are typically the words that join two phrases(noun or verb), of the same category or a clause.Some common conjunctions are aur, par, yA, balkI etc.

10) Amrood khUb toD –toD ke khAe aur\CC_CCD Apan sab bhaiyan ke delas
Plucking the guavas he ate it and gave to all his brothers.

8.2 Subordinating conjunctions (CC_CCS)

Subordinator typically conjoins two clauses and the second clause is subordinated. Some of the subordinate conjunctions are magar, to ki etc

11) ta dhobi kahe **ki**\CC_CCD tU kAm bahut kaile hauA hamAr
The washer man said **that** you have worked a lot for me.

9. PARTICLES (RP)

Particles are words that do not decline and also do not fall into any other categories described above and elsewhere. Bhojpuri particles includes following five sub-categories

9.1 Classifiers (RP_CL)

A classifier sometimes called a **measure word** is a word or morpheme used in some languages to classify the referent of a countable noun according to its meaning.

12) dU\QT_QTC go\RP_CL bhaI rahasan
There were two brothers.

9.2 Default Particle (RP_RPD)

Default Particle is a category that includes all those element of the language which though do not have any lexical important but are auspecious functionally. Some of the Bhojpuri default particles are to hI bhI, nA, jI etc.

13) to ganes jI\RP_RPD ek The bAlak kA rUp rakh ke ayilan
Then lord Ganesa appeared in the form of a boy.

9.3 Interjection (RP_INJ)

Interjections are particles which denote exclamation utterances. The common exclamatory marks in Bhojpuri are अरे, हे, ए , हो etc.

14) are\RP_INJ ! sab tarapf se band hai
It is closed from all sides

9.4 Intensifier (RP_INTF)

Intensifiers are words that intensify the adjectives or adverbs. The common intensifiers in Bhojpuri are arkhoob, itnA, bahut, mAre, ittI etc.

15) aum bahut\RP_INTF cHoTe\JJ hokahAM jaoge
You are too young to go with them.

9.5 Negation (RP_NEG)

The negation particles are the words that indicate negation. These include nAhi, nA, mat, binA, bagair etc.

16) ab mat\RP_NEG Aye, nAhiM\RP_NEG ta bahut mArab
Now don't follow or I will beat you.

10. QUANTIFIERS (QT)

Quantifiers are the words that indicate quantity and modify nouns or adjectives. These have been sub-categorized into three parts- general, cardinals and ordinals.

10.1 General (QT_QTF)

The general quantifiers do not indicate any precise quantity, e.g, purA, sab, ek etc.

17) Rahat ta khUb\QT_QTF seb khiyAwat
If he would be there, must have brought us a lot of apples.

10.2 Cardinals (QT_QTC)

The cardinal quantifiers are absolute numbers, either in digits or in words such as 1, 2, 3, ek, do, tIn etc.

18) ek\QT_QTC The dhobi ke ghar rahal
There was a washerman's house.

10.3 Ordinals (QT_QTO)

The ordinals denote the order part of the digits such as pahilA, dusar, tIsar etc.

19) duno\QT_QTO log gayilan khet meM
Both went to the fields.

11. RESIDUALS (RD)

The category of residuals has been demarcated for the words that are usually not intrinsic part of the language/speech. Divided into five parts, these include foreign words, symbols, punctuations, unknown words and echo-words.

11.1 Foreign Words (RD_RDF)

The foreign words are all the words that are not written in the Devanagari script.

11.2 Symbols (RD_SYM)

The symbols are the characters that are not part of the regular Devanagari script such as *, @, #, $, % etc.

11.3 Punctuations (RD_PUNC)

Punctuations include the characters that are considered as the regular punctuation marks in Hindi, e.g. (,),,,!,?,- etc.

11.4Unknown (RD_UNK)

Unknown words would the words for which a category cannot be decided by the annotator. These may include words from phrases or sentences from a foreign language written in Devanagari.

11.5 Echo-Words (RD_ECH)

The echo-words are the words that are formed by the morphological process known as echo-formation e.g. (chup-chaap), (sach-much), (kAT-kUT) etc.