Developing Resources for a Less-Resourced Language: Braj Bhasha

Mayank Jain[1], Yogesh Dawer[2], Nandini Chauhan[2], Anjali Gupta[2]
[1]Jawaharlal Nehru University, [2]Dr. Bhim Rao Ambedkar University
Delhi, Agra
{jnu.mayank, yogeshdawer, nandinipinki850, anjalisoniyagupta89}@gmail.com

Abstract

This paper gives the overview of the language resources developed for a less-resourced western Indo-Aryan language of India - Braj Bhasha. There is no language resource available for Braj Bhasha. The paper gives the detail of first-ever language resources developed for Braj Bhasha which are text corpus, BIS based POS tagset and annotation, Universal Dependency (UD) based morphological and dependency annotation. UD is a framework for cross-linguistically consistent grammatical annotation and an open community effort with contributors working on over 60 languages. The methodology used to develop corpus, tagset, and annotation can help in creating resources for other less-resourced languages. These resources would provide the opportunity for Braj Bhasha to develop NLP applications and to do research on various areas of linguistics - cognitive linguistics, comparative linguistics, typological and theoretical linguistics.

Keywords: Language Resources, Corpus, Tagset, Universal Dependency, Annotation, Braj Bhasha

1. Introduction

Braj or Braj Bhasha[1] is a Western Indo-Aryan language spoken mainly in the adjoining region spread over Uttar Pradesh and Rajasthan. In present times, when major developments in the field of computational linguistics and natural language processing are playing an integral role in empowering languages, it is essential to include as many languages as possible in this endeavour. It becomes more significant for Indian languages, which are far behind in this area. Following the poor situation of Indian languages in terms of computational resources and applications, there are no available language resources for Braj Bhasha. This is despite the large population of approximately 5 million native speakers. The present work is an attempt to create and develop of some of the basic resources for Braj Bhasha. The paper will focus on the creation and preparation of Text-Corpus, development of BIS based Braj POS Tagset, Universal dependency based annotation of the text corpus at POS level, Morphological features level and Syntactic level.

2. Corpus Creation

This section describes the first-ever annotated text corpus which has been created for Braj Bhasha. Kumar et al. (2016) mention about Braj corpus, however, the details of the corpus is not available.

2.1 Corpus Collection

Though there is a good amount of text written in Braj, these have limitations as almost no data is available in the digital/electronic format. The Braj data has been taken from offline sources which consist of various books and magazines. However, the books and magazines were not easily available. We had to make efforts to collect the text materials. The text we collected mainly belongs to religious texts, short stories, memoirs, culture, art, literary work.

2.2 Digitisation of Corpus

For digitising the available text, an Optical Character Recognition system was developed which makes use of Google OCR. The Google OCR gives a perfect result for Hindi texts written in Devanagari. Since the script of Braj Bhasha is Devanagari, the OCR tool gave quite a satisfactory result for Braj data. The process of digitization is a two-step procedure. First, the individual pages of books and magazines were scanned using a high-resolution scanner. Then, those scanned pages were converted to digital form with the help of the OCR system. At this stage, there were two issues which were needed to be addressed. The first was the cleaning and editing of digitized text because the OCR was meant for Hindi rather than Braj. It required manual editing of the OCRed text. The second issue was the abundance of poetry in some texts, it owes to the fact that Braj Bhasha had been predominantly used for writing poetry. Therefore, the digital data was cleaned to remove poetry text.

2.3 Corpus Statistics

At present, after cleaning, around 5000 pages of a raw unedited corpus is available. As of now, we have edited around 800 pages which consist of around 20,000 sentences and 300,000 tokens. The size of the corpus is being increased on regular basis as more and more data has been edited and cleaned regularly.

3. POS-Tagset and Annotation

POS annotated corpora is a basic resource for several NLP applications. The Braj corpus was annotated using BIS tagset which is shown in section 3.1. In section 3.2, the annotation guidelines for using Braj POS tagset have been discussed. Section 3.3 gives details of POS annotation.

3.1 Braj POS Tagset

There is no POS tagset available for Braj Bhasha as no work has been done on POS annotation of Braj. A Braj POS tagset has been developed for the current research which is based on the BIS[2] guidelines which are a national standard for Indian languages. The BIS tagset has been designed under the Bureau of Indian Standards by the Indian Languages Corpora Initiative (ILCI) group. This tagset takes care of linguistic characteristics of Indian languages. The main characteristic of this tagset is its

1ISO 639-3 Code: bra

2 http://tdil-dc.in/tdildcMain/articles/134692Draft%20POS%20Tag%20standard.pdf

hierarchical nature which takes into account the granularity of linguistic information. The categories at the level 1 are further divided into subtype level 2 and level 3. It is arranged in such a way that the categories at the higher level are more coarse whereas the categories at the lower level are more fine-grained in terms of linguistic/grammatical information. Since We are incorporating morphosyntactic features in UD based morphological and dependency annotation, the hierarchy of Braj POS tagset has been restricted to two levels only.

The Braj POS tagset contains eleven level 1 categories which are divided into 32 fine-grained categories at level 2 of the hierarchy. In the tagset, three level 1 categories: adjective, adverb and postpositions are not divided into sub-categories. Remaining 8 level 1 categories are further divided into sub-categories. The detailed tagset is given in table 1.

Sl. No	**Category**		**Label**	**Annotation Convention**	**Examples**
	Top level	Subtype (level 1)			
1	**Noun**		**N**	**N**	कृष्ण, विसैन,
1.1		Common	NN	N	शब्दन, ग्रंथ
		Proper	NNP	N__NNP	राधा, मथुरा
		Nloc	NST	N__NST	आगै, पीछै
2	**Pronoun**		**PR**	**PR**	मैं, वू, अपनौ
2.1		Personal	PRP	PR__PRP	मेरौ, तू, वू,
2.2		Reflexive	PRF	PR__PRF	अपनौ, अपन
2.3		Relative	PRL	PR__PRL	जो, जिस
2.4		Reciprocal	PRC	PR__PRC	आपस, परस्पर
2.5		Wh-word	PRQ	PR__PRQ	कौन, कित
2.6		Indefinite	PRI	PR_PRI	काऊ, कछू
3	**Demonstrative**		**DM**	**DM**	वू, जे, बिन
3.1		Deictic	DMD	DM__DMD	वे, वू, वा
3.2		Relative	DMR	DM__DMR	जे, बिन
3.3		Wh-word	DMQ	DM__DMQ	कौन,
3.4		Indefinite	DMI	DM__DMI	कछू, कहीं
4	**Verb**		**V**	**V**	है, लिखे, होय
4.1		Main	VM	V__VM	धर, रहतौ,
4.2		Auxiliary	VAUX	V__VAUX	है, हे, रहौ
5	**Adjective**		**JJ**	**JJ**	बड़ौ, सूधी,
6	**Adverb**		**RB**	**RB**	धीरें, जल्दी,
7	**Postposition**		**PSP**	**PSP**	पै, कूँ, कौ, सौं
8	**Conjunction**		**CC**	**CC**	पर, कै,
8.1		Co-ordinator	CCD	CC__CCD	अरु, पर
8.2		Subordinator	CCS	CC__CCS	तौ, कै,
9	**Particles**		**RP**	**RP**	तो, ही
9.1		Default	RPD	RP__RPD	भी, तो, ही
9.2		Interjection	INJ	RP__INJ	अरे, हे, ओ
9.3		Intensifier	INTF	RP__INTF	बहुतई, बेहद
9.4		Negation	NEG	RP__NEG	नाँय, न
10	**Quantifiers**		**QT**	**QT**	एक, थौड़ौ
10.1		General	QTF	QT__QTF	थौड़ौ, बहुत
10.2		Cardinals	QTC	QT__QTC	एक, दो,
10.3		Ordinals	QTO	QT__QTO	पहलौ, दूसरौ
11	**Residuals**		**RD**	**RD**	
11.1		Foreign word	RDF	RD__RDF	A word in a foreign script.
11.2		Symbol	SYM	RD__SYM	For symbols
11.3		Punctuation	PUNC	RD__PUNC	Only for punctuations
11.4		Unknown	UNK	RD__UNK	
11.5		Echo words	ECH	RD__ECH	

Table 1: Braj POS Tagset

3.2 Annotation Guidelines for Braj POS Tagset

The following description is the explanation of POS tags:

3.2.1 Noun (N)

The top-level category of the noun has three sub-categories which are as follows:

3.2.1.1 Common Noun (NN)

Words that belong to the types of common nouns, collective nouns, abstract nouns, countable and non-countable nouns. e.g. शब्दन, ग्रंथा

3.2.1.2 Proper Noun (NNP)

Words that denote the name of a person, place, day etc. e.g. राधा, मथुरा

3.2.1.3 Noun locative (NST)

These words can act as both location nouns and as a part of a complex postposition. e.g. आगै, पीछै

3.2.2. Pronoun (P)

The pronoun is divided into five sub-categories:

3.2.2.1 Personal Pronoun (PR)

These encode person feature in them. e.g. मेरौ, वू

3.2.2.2 Reflexive Pronoun (PRF)

It refers to a noun or pronoun which precedes it. e.g. अपनौ

3.2.2.3 Relative Pronoun (PRL)

It links two clauses in a single complex clause. e.g. जो, जि

3.2.2.4 Reciprocal Pronoun (PRC)
These words show reciprocity. e.g. आपस, परस्प्रा

3.2.2.5 Wh-word (PRQ)
Pronouns which falls into Wh-question category. e.g. कौन, कित

3.2.2.6 Indefinite Pronoun (PRI)
Words refer to something indefinite. e.g. काऊ, कछू

3.2.3. Demonstrative (DM)
Demonstratives have the same form as Pronouns. They are always followed by a noun which they modify. Whereas, a pronoun is used in place of a noun.

3.2.3.1 Deictic (DMD)
These are mainly personal pronouns like बे, बू, वा. But these must occur before a noun.

3.2.3.2 Relative (DMR)
It has the same form as a relative pronoun, but it should occur before the noun it modifies e.g. जे, बिन

3.2.3.3 Wh-word (DMQ)
It is same as Wh-pronoun and it should occur before the noun it modifies e.g. कौन

3.2.3.4 Indefinite (DMI)
It has the same form as an indefinite pronoun. It should occur before the noun it modifies. काऊ, कछू

3.2.4. Verb (V)
The verb is divided into two categories of Main and Auxiliary:

3.2.4.1 Main Verb (VM)
The main verb expresses the main predication of the sentence. It can be in the root form or one of its inflected form. A clause must have a main verb. e.g. धर, रहतौ

3.2.4.2 Auxiliary Verb (VAUX)
An auxiliary verb gives information about inflectional features like tense, aspect e.g. है, हे, रहौ

3.2.5. Adjective (JJ)
The adjectives fall into this category e.g. बड़ौ, लम्बी

3.2.6. Adverb (RB)
Only manner adverbs are annotated using this tag e.g. जल्दी, धीरे

3.2.7. Postposition (PSP)
Postpositions are tagged using this tag. e.g. नै, कौ, कूँ

3.2.8. Conjunction (CC)
A conjunction joins two phrases, clauses, noun, etc. These are divided into two sub-types:

3.2.8.1. Coordinate (CCD)
It joins two or more items of equal syntactic importance. e.g. अरू, पर

3.2.8.2 Subordinate (CCS)
It joins main clause with a dependent clause. It introduces dependent clause e.g. तौ, कै

3.2.9. Particles (RP)
Particles do not decline and they do not fall into any other categories mentioned here. These are divided into four sub-types:

3.2.9.1 Default (RPD)
The default particles are as follows: ही, तो, भी

3.2.9.2 Interjection (INJ)
Words which expresses emotions and gets the attention of people. e.g. अरे, हे, ओ

3.2.9.3 Intensifier (INTF)
Adverbial intensifiers fall under this category. e.g. बहुतई, बेहत

3.2.9.4 Negation (NEG)
The words which indicate negation. e.g. न, नाँय

3.2.10. Quantifiers (QT)
It quantifies the nouns. These are divided into three sub-types:

3.2.10.1 General (QTF)
These quantifiers do not indicate any precise quantity. e.g. थौड़ौ, बहुत

3.2.10.2 Cardinals (QTC)
These are absolute numbers either in digits or numbers e.g. 1, 3, एक, दो

3.2.10.3 Ordinals (QTO)
These include ordered part of the digits. e.g. पहलौ, दूसरौ

3.2.11. Residuals (RD)
These categories are the words which are not an intrinsic part of the language. These are divided into five sub-types:

3.2.11.1 Foreign Words (RDF)
The words are written in a non-Devanagari script. e.g. and

3.2.11.2. Symbol (SYM)
It is used for symbols like $, %. # etc

3.2.11.3 Punctuation (PUNC)
It is used for punctuations e.g. ( ) , " ।'

3.2.11.4. Unknown (UNK)
In this category, those words are kept whose annotation cannot be decided.

3.2.11.5 Echo words (ECH)

It is used for words formed by a morphological process known as Echo-formation. e.g. पानी-वानी

3.3 POS Annotation

For the POS annotation, the syntactic function of the word is given importance rather than its pure lexical category. It is necessary to take syntactic context into consideration so that the appropriate grammatical information of the word can be captured. Therefore, a word may change its lexical category depending on where it occurs in a sentence. Two POS annotated examples are given below:

(1) ब्रजभाषा\NNP कौ\PSP छेत्र\NN आज\NN हू\RPD भौत\INTF ब्यापक\JJ है\VM ।\PUNC

(2) पद्य\NN में\PSP तौ\RPD जे\DMD काम\NN भक्तिकाल\NN में\PSP ही\RP_RPD सम्पन्न\JJ है\V_VM. गयौ\V_VAUX हो\V_VAUX ।\RD_PUNC

3.4 Annotation Tool: WebAnno

An open source, general purpose web-based annotation tool, WebAnno[3] (Eckart de Castilho et al. 2016), is used for the annotation. One of the characteristic features of WebAnno is its suitability for a wide range of linguistic annotations including various layers of POS, morphological, syntactic, and semantic annotations. It allows adding custom annotation layer to facilitate the requirement of the user. It allows the distribution of annotations in various formats. WebAnno also allows multiple annotators to collaborate on a project. It is a flexible, easy-to-use annotation tool.

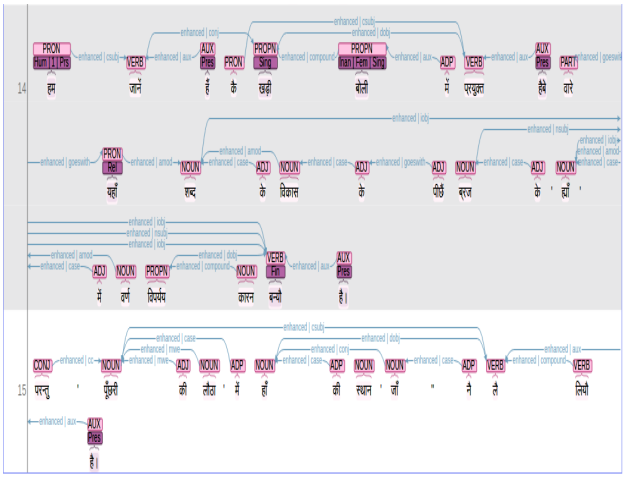

Figure 1. Annotation Examples in WebAnno

4. [4]Universal Dependency (UD) based Annotation

Nivre, J. et al. (2017) say "Universal Dependencies is a project that seeks to develop cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective."

UD framework addresses several NLP related issues. The problem of varied annotation schemes across languages has been addressed by providing cross-linguistically consistent grammatical annotation. "The annotation scheme is based on (universal) Stanford dependencies, Google universal part-of-speech tags, and the Interset interlingua for morphosyntactic tagsets." (Nivre, J. et al. 2017). The annotation scheme is based on existing standards.

One of the key features of UD framework is inclusivity. It provides universal taxonomy along with the scope to include language-specific extensions.

4.1 Basic Principles of UD

The UD annotation is based on the lexicalist view of syntax, which means that dependency relations hold between words. Words enter into syntactic relations. In lexicalist view, the basic annotation units are syntactic words. Words have morphological features encoded in them. Thus, words are not segmented into morphemes.

4.1.1 Morphological Annotation

The morphological specification of a (syntactic) word in the UD scheme consists of following three levels of representation[5]:

- A lemma representing the semantic content of the word.
- A part-of-speech tag a representing its grammatical class.
- A set of features representing lexical and grammatical properties of the lemma and particular word form.

One of the characteristics of the universal tags and features is that they do not include means to mark fusion words. Fusion words need to be split into syntactic word so that they will get POS tag and feature annotation.

4.1.2. Syntactic Annotation

There are three main ways in which syntactic dependency is marked in the UD framework:

1. The syntactic annotation in the UD scheme marks dependency relations between words.
2. The function words attach to the content words they modify and
3. The punctuation attaches to the head of the phrase or clause.

4.2 Morphological and Syntactic Annotation of Braj Bhasha in UD Framework

Under UD framework, there are over 200 contributors who are working on more than 100 treebanks in over 60 languages around the world. Six Indian languages (Hindi, Marathi, Sanskrit, Tamil, Telugu and Urdu) have been

3https://webanno.github.io/webanno/

4For details, refer to http://universaldependencies.org

5http://universaldependencies.org/u/overview/morphology.html

included in the UD framework. The present work attempts to incorporate the UD framework for annotating the Braj corpus. The following section describes it in detail.

4.2.1. Morphological Features:
Morphological Features are additional lexical and grammatical properties of the word which are not covered by universal POS tags. The format in which a feature is used is Name=Value. A word can have any number of features separated by the vertical bar. For example, Number=Sing|Person=3

The following are some of the morphological features used for Braj Bhasha:

AdpType | AdvType | Animacy | Aspect | Case | Definite | Degree and Polarity | Echo | Foreign | Gender | Gender [psor] | Mood | Number | Number [psor] | NumType | Person | Polite | Poss | PronType | Tense | VerbForm | Voice |

An example of morphological features is given below:

(3) ब्रजभाषा\NNP.Inan.Acc.Fem.Sing कौ\PSP.Gen.Masc.Sing छेत्र\NN.Inan.Acc.Masc.Sing.3 आज\NN.Nom.Masc.Sing.3 हू\RPD भौत\INTF.Deg ब्यापक\JJ हैं\VM.Ind.Sing.3.Pres.Fin.Act. ।\PUNC

4.2.2. Syntactic dependency:
Syntactic annotation in the UD scheme consists of typed dependency relations between words. The basic dependency representation forms a tree, where exactly one word is the head of the sentence, dependent on a notional ROOT and all other words are dependent on another word in the sentence.
Apart from the basic dependency which is obligatory for all the syntactic annotation, an additional enhanced dependency representation can be incorporated which gives a complete basis of semantic interpretation.
The following are some of the dependency relations used for Braj Bhasha:

acl | advmod | aux | case | cc | ccomp | compound | conj | cop | det | dobj | iobj | mark | mwe | nmod | nsubj | obj | obl | punct | root | xcomp |

An example of dependency relations is given in Table 2:
(4) ब्रजभाषा कौ छेत्र आज हू भौत ब्यापक है ।

S.N.	Token	POS. Morph Features	Dependent on S.N.	Dependency relation
1	ब्रजभाषा	NNP.Inan.Acc.Fem.Sing.3	3	nmod
2	कौ	PSP.Gen.Masc.Sing	1	case
3	छेत्र	NN.Inan.Acc.Masc.Sing.3	7	nsubj
4	आज	NN.Nom.Masc.Sing.3	7	nmod
5	हू	RPD\INTF.Deg	4	dep
6	भौत	INTF.Deg	7	advmod
7	ब्यापक	JJ	0	root
8	है	VM.Ind.Sing.3.Pres.Fin.Act	7	cop
9	।	Punc	7	punct

Table 2: UD based dependency for sentence no. (4)

4.2.3 Annotation of Braj using UD

The morphological features and dependency relations, which are given in the previous section, have been used to annotate Braj corpus. At present, we have completed annotation of about 500 sentences. More data is being annotated on regular basis. Once, we have enough data, we would use machine learning approach to train the system. At present, we don't enough data for training purpose. The idea is to create comprehensive resources of Braj so that modern NLP applications can be developed for it.

5. Conclusion

The present research focuses on developing various resources and tools for Braj which does not have any of such resources. There has been some encouraging progress in this regard, as we have been able to create a first-ever digital corpus for Braj. Although it is a raw corpus, it was quite difficult to collect and create the corpus. Further work is in progress, where the digitised corpus is being annotated at different levels of linguistic annotation – POS, morphological features, and syntactic dependencies. Some important and essential resources - annotation guidelines, tagsets, etc - have been created. We are also experimenting with developing automatic tools by using machine learning approach. We have been making progress and hope to present some reliable results during our presentation. Along with these results, we would also discuss the issues and challenges which were faced during the progress the work.
We hope that our work would contribute towards building essential resources of Braj and our methodology would encourage such work for other less-resourced languages.

6. Acknowledgements

A special thank should go to Dr Ritesh Kumar for providing us with various tools to help us in the development of the corpus. A special acknowledgement should be made for Prof Girish Nath Jha for providing us encouragement and opportunity to work on Braj language. We would also like to acknowledge the open platform provided by UD team which allowed to use UD framework for our work.

7. Bibliographical References

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S.M., Hartmann, S., Gurevych, I., Frank, A. and Biemann, C. (2016). A Web-based Tool for the Integrated Annotation of Semantic and Syntactic

Structures. In Proceedings of the LT4DH workshop at COLING 2016, Osaka, Japan.

Kumar, R., Ojha, A. K., Lahiri, B., and Alok, D. (2016). Developing Resources and Tools for some Lesser-known Languages of India. Presented in Regional ICON(regICON) 2016, IIT-BHU, Varanasi, India.

Ojha, A. K., Behera P., Singh S., and Jha, G.N. (2015). Training & Evaluation of POS Taggers in Indo-Aryan Languages: A Case of Hindi, Odia and Bhojpuri. In the proceedings of 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. Pages 524-529. Poznań, Poland.

Nivre, J., Agić, Ž, and Ahrenberg, L. et al. (2017). Universal Dependencies 2.1, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-2515.

https://www.ethnologue.com/language/bra Accessed on 11-01-2018