

Challenges in Automatic POS Tagging of
Indian Languages-
A Comparative Study of Hindi and Bhojpuri

*Dissertation Submitted to Jawaharlal Nehru University
in partial fulfillment of the requirements
for the award of the degree of*

MASTER OF PHILOSOPHY

SRISHTI SINGH



**Centre for Linguistics,
School of Language, Literature and Culture Studies,
Jawaharlal Nehru University,
New Delhi, India-110067**

2015

Dated: 24th July, 2015

DECLARATION BY THE CANDIDATE

This Thesis titled "Challenges in Automatic POS Tagging of Indian Languages - A Comparative Study of Hindi and Bhojpuri" submitted by me for the award of the degree of Master of Philosophy, is an original work and has not been submitted so far, in part or in full, for any other degree or diploma of any University or Institute.

Srishti Singh
(Srishti Singh)

M.Phil. student

Centre for Linguistics,

SLL&CS

JNU



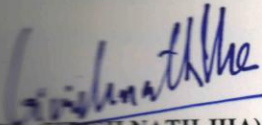
Centre for Linguistics
School of Language, Literature and Culture Studies
Jawaharlal Nehru University
New Delhi-110067, India

Dated: 24st July, 2015

CERTIFICATE

This thesis titled "Challenges in Automatic POS Tagging of Indian Languages - A Comparative Study of Hindi and Bhojpuri" submitted by Ms. Srishti Singh, Centre for Linguistics, School of Language, Literature & Culture Studies, Jawaharlal Nehru University, New Delhi, for the award of the degree of **Master of Philosophy**, is an original work and has not been submitted so far, in part or in full, for any other degree or diploma of any University or Institution.

This may be placed before the examiners for evaluation for the award of the degree of Master of Philosophy.


(DR. GIRISH NATH JHA)

SUPERVISOR
DR. GIRISH NATH JHA
Associate Professor
Centre for Sanskrit Studies
Jawaharlal Nehru University
New Delhi - 110067



(PROF. AYESHA KIDWAI)

CHAIRPERSON
Chairperson
Centre for Linguistics
School of Language, Literature & Culture Studies
Jawaharlal Nehru University, New Delhi-110067

Tel: +91-11-26704199 FAX: 91-011-26741586

DEDICATION

*to the service of
languages and communities,
socially abandoned
due to scarcity of technological resources*

ACKNOWLEDGEMENTS

This dissertation could probably not have been completed without the help, support, guidance and love of many people.

My foremost thanks and gratitude goes to my supervisor Dr. Girish Nath Jha who has been a constant support and encouragement ever since the beginning of this work. His faith in me and excellent guidance helped me finishing the task within the time frame. Since my inception at JNU, his dynamic personality, enthusiasm and attitude towards life has influenced me (both personally and professionally). Irrespective of the kind of problems that I came up with during these two years, he has always been very tolerant in resolving them.. I find myself blessed that out of many, I got this opportunity to work under his supervision. Here, I got the chance to see the practical aspect of Computational Linguistics of which earlier I was unaware. I extend my heartfelt gratitude to him. “I cannot simply ‘Thank’ you Sir, but can only give my regards to you, and I wish to continue working with you as a student with much more energy, productivity and dedication”.

I would also like to mention that it was due to his motivation that I find a platform like edDI for my professional and academic growth. Where I met another spark, Mr. Narayan Choudhary, who being a friend, senior and colleague, has always led me to the right direction with all homely warmth and bossy rules, at times. Not only this, this work is also proof read by him, and his valuable comments has added to its worth.

I owe my thanks to all the faculty members associated with Centre for Linguistics. Specially, Prof. Ayesha Kidwai, our current chairperson for her kind consideration. Despite heavy work load, she has always welcomed me and my discussions and has quickly responded to every issue that was raised even in the virtual space (e-mail). I thank Prof. Vaishna Narang, for her course on Ethics and the wonderful LEC workshop on English and Academic Writing, which was quite helpful in writing my chapters. My heartfelt thanks goes to Prof. Abbi for making us feel the value and power of endangered and lesser known languages which was yet another motivating factor for working on the technology development of **Bhojpuri**. I feel fortunate as my batch was the last one to have attended her course work before she got retired.

Here, a special thanks goes to my grany, Mrs. Shivvati Devi, who is behind all the motivation as my grany and as my participant (at one point of time). One of my earliest academic term papers on writing 'grammatical sketch of Benarasi' under Prof. Abbi's course designed for documentation, which later served as the data for pilot study, was the result of her stories which I am hearing since my childhood. And, perhaps the main reason for my inclination towards Bhojpuri to the extent that I could opt it as the field of research.

Thanks also go to all my colleagues at Indian Language Corpora Initiative (ILCI) for helping me out with various difficulties. Abhishek kumar for technical support, Abhishek Sharma and Abhishek Ranjan for programming and coding part, Rachita Sinha and Pitambar Behera with linguistic puzzles, Atma Prakash for making us smile in intense situations and Ranjit for his awesome tea. Special thanks to Atul Ojha, whose honest & tireless effort and guidance, consistently, right from setting up the tool to its implementation was very helpful.

Thanks are due to other friends from SCSS (Special centre for Sanskrit Studies) and CL (Centre for Linguistics), specially seniors (Esha banergee, Oinam Thoibi, and Harjit Singh, to name a few) for being available anytime I needed and helping me out even at the eleventh hour of deadlines.

I also thank the SCSS and CL family from library access to the office staffs for handling our odd visits to lab very calmly. Specially, Gopal Sir and Naveen Sir, for helping me with all the official matters since admission till submission.

I would like to take this opportunity to thank my parents, my family and close relatives. To all the family members for patiently listening to my bitter and sweet JNU stories and eagerly waiting for my come back on every occasion. Thank you mummy, papa and bhaiya (of course), for showing this much confidence in me, tolerating me and being with me at every step. Moreover, for making others understand about Linguistics and the nature of my work, when asked by others (a big deal).

I bid my thanks to my friend Aniruddha and Bhaskar for catching up with me a *number of times* just before the submission, and making me write happily. Last but never the least, my roommate Pushpa, I thank you for coordinating with my dancing schedule throughout, your *cooked food*

and rescuing me from the summer *lizard encounters* even at the middle of the night. I feel blessed to be with you all.

Table of Contents

| | | |
|---------|---|----|
| 1 | Introduction | 21 |
| 1.1 | Introduction | 21 |
| 1.2 | Method of Research | 22 |
| 1.3 | Statement of Problem | 23 |
| 1.4 | Research Question..... | 24 |
| 1.4.1 | Hypothesis..... | 25 |
| 1.5 | Background to Problem..... | 25 |
| 1.5.1 | Resources for Natural Language Processing | 26 |
| 1.5.1.1 | Corpus | 26 |
| 1.5.1.2 | Parts of Speech (POS) tagger | 26 |
| 1.5.1.3 | Chunker | 26 |
| 1.5.1.4 | Morphological analyser..... | 27 |
| 1.5.1.5 | Parser | 27 |
| 1.5.2 | Approaches to Stochastic Learning | 27 |
| 1.5.2.1 | Hidden Markov Model (HMM) | 27 |
| 1.5.2.2 | Conditional Random Fields (CRF) | 28 |
| 1.5.2.3 | Maximum Entropy Model (MEM)..... | 28 |
| 1.5.2.4 | Memory Based Learning..... | 28 |
| 1.5.3 | Support Vector Machines. | 29 |
| 1.5.3.1 | Application based SVM | 32 |
| 1.5.3.2 | Advantages of SVM Environment | 33 |
| 1.6 | Review of Literature..... | 34 |
| 1.6.1 | Bhojpuri | 34 |
| 1.6.2 | Historical Background | 34 |
| 1.6.3 | Status of Bhojpuri | 36 |
| 1.6.3.1 | Social Status | 36 |

| | | |
|---------|--|----|
| 1.6.3.2 | Educational Status | 36 |
| 1.6.3.3 | Political Status..... | 37 |
| 1.6.4 | Standardization of Bhojpuri | 38 |
| 1.6.5 | Existing endeavour in this field | 39 |
| 1.6.5.1 | Literature | 39 |
| 1.6.5.2 | Technological advancements | 42 |
| 1.6.6 | Existing Corpora in Indian Languages | 45 |
| 1.6.6.1 | EMILLE | 45 |
| 1.6.6.2 | Indian Languages Corpora, CIIL Mysore | 45 |
| 1.6.6.3 | LDC-IL..... | 45 |
| 1.6.6.4 | ILCI..... | 46 |
| 1.6.6.5 | Joshua..... | 46 |
| 1.6.7 | Parts of Speech Annotation Scheme | 47 |
| 1.6.7.1 | Penn Tagset | 47 |
| 1.6.7.2 | Annotation scheme for Indian Languages..... | 47 |
| 1.6.8 | Existing POS Taggers for Indian Languages..... | 49 |
| 1.6.8.1 | Bengali | 49 |
| 1.6.8.2 | Hindi..... | 49 |
| 1.6.8.3 | Malayalam..... | 50 |
| 1.6.8.4 | Punjabi..... | 50 |
| 1.6.8.5 | Sanskrit..... | 50 |
| 1.6.8.6 | Telugu..... | 51 |
| 1.7 | Relevance of this study | 51 |
| 2 | Corpus Creation for Bhojpuri | 52 |
| 2.1 | Research Methodology..... | 52 |
| 2.2 | Bhojpuri Corpus | 52 |
| 2.2.1 | Corpus Data | 53 |

| | |
|---------|---|
| | 10 |
| 2.2.2 | Source of Data..... 54 |
| 2.2.3 | Data Collection 56 |
| 2.2.3.1 | ILCrawler 57 |
| 2.2.4 | Corpus Cleaning..... 61 |
| 2.2.4.1 | ILSanitizer..... 61 |
| 2.2.5 | Data Management and File Format..... 61 |
| 2.3 | Features of Bhojpuri..... 64 |
| 2.3.1 | Issues in data collection 65 |
| 2.3.1.1 | Data from Other Languages 65 |
| 2.3.1.2 | Header on the webpage 67 |
| 2.3.1.3 | Some other common Crawler errors 69 |
| 2.3.2 | Corpus Validation 70 |
| 2.3.2.1 | Validation Challenges 70 |
| 3 | Annotated Corpus for Bhojpuri 77 |
| 3.1 | POS Tagging 77 |
| 3.2 | POS Annotation Scheme 77 |
| 3.2.1 | The BIS Scheme 78 |
| 3.2.2 | POS-Tagset for Bhojpuri 79 |
| 3.2.3 | A Preliminary Comparison of Hindi and Bhojpuri tagsets 80 |
| 3.2.4 | Revised Bhojpuri Tagset..... 80 |
| 3.3 | Data for Annotation..... 84 |
| 3.4 | Annotating Bhojpuri Corpus 85 |
| 3.4.1 | ILCIANN 85 |
| 3.4.1.1 | Online interface of the ILCIANN tool 85 |
| 3.4.2 | Bhojpuri POS annotation guideline 88 |
| 3.4.2.1 | NOUN (N)..... 88 |
| 3.4.2.2 | PRONOUN (PR)..... 90 |
| 3.4.2.3 | DEMONSTRATIVE (DM)..... 93 |
| 3.4.2.4 | VERBS (V) 95 |

| | | |
|----------|--|-----|
| 3.4.2.5 | ADJECTIVES (JJ) | 100 |
| 3.4.2.6 | ADVERB (RB) | 101 |
| 3.4.2.7 | POSTPOSITION (PSP) | 103 |
| 3.4.2.8 | CONJUNCTION (CC) | 104 |
| 3.4.2.9 | PARTICLES (RP) | 105 |
| 3.4.2.10 | QUANTIFIERS (QT) | 108 |
| 3.4.2.11 | RESIDUALS (RD) | 109 |
| 3.5 | Issues in POS tagging | 113 |
| 3.5.1 | Challenges in Manual Tagging | 113 |
| 3.5.1.1 | Unidentified Tokens and rare occurrences | 113 |
| 3.5.1.2 | Dialectal Variation | 115 |
| 3.5.1.3 | Inflected categories | 117 |
| 3.5.1.4 | Homophones | 120 |
| 3.5.1.5 | Different Realizations of one lexeme | 123 |
| 3.5.1.6 | Tagging inconsistency | 124 |
| 4 | SVM Based Bhojpuri Tagger | 125 |
| 4.1 | Bhojpuri and Support Vector Machine | 125 |
| 4.1.1 | Why SVM | 125 |
| 4.1.2 | Efficiency of the tagger | 125 |
| 4.2 | Description of the Tool | 126 |
| 4.2.1 | Property of the Tool | 126 |
| 4.2.2 | Design of the tagger | 127 |
| 4.2.3 | Tagger Models | 127 |
| 4.2.4 | Configuration | 128 |
| 4.2.5 | Format of the data | 128 |
| 4.2.6 | Tagger Architecture | 129 |
| 4.3 | Training the Tagger | 130 |
| 4.3.1 | Cautions for the training file | 131 |

| | | |
|---------|--|-----|
| 4.4 | Testing the Tagger..... | 132 |
| 4.5 | Evaluation of the Tagger | 133 |
| 4.5.1 | Gold corpus..... | 133 |
| 4.5.2 | SVMTeval..... | 134 |
| 4.5.2.1 | Accuracy as per class of ambiguity..... | 135 |
| 4.5.3 | Evaluation of the Tests..... | 136 |
| 4.5.3.1 | Evaluation of the first phase (model file with 30k tokens): | 136 |
| 4.5.3.2 | Evaluation of the current phase (model file with 90k tokens): | 139 |
| 4.5.3.3 | Results Obtained | 140 |
| 4.6 | Tagger based Tagging Issues | 141 |
| 4.6.1 | Pattern based errors..... | 141 |
| 4.6.1.1 | Error Pattern for Verb series | 141 |
| 4.6.1.2 | Error pattern for conjunct verbs | 142 |
| 4.6.2 | Errors due to ambiguous tokens..... | 143 |
| 4.6.3 | Random errors..... | 146 |
| 4.7 | Comparison and Performance of Hindi and Bhojpuri Tagger | 147 |
| 4.7.1 | Data for Comparison..... | 147 |
| 4.7.2 | Hindi Tagger | 147 |
| 4.7.3 | SVM performance on Hindi Tagger | 148 |
| 4.7.4 | SVM performance on Bhojpuri Tagger | 149 |
| 4.7.5 | Comparison of results | 149 |
| 4.7.5.1 | Error analysis of the Hindi tagger | 150 |
| 4.7.6 | Comparison Result..... | 155 |
| 4.8 | Development of the Bhojpuri tagger | 156 |
| 4.8.1 | Rules proposed for improvement of pattern errors for Bhojpuri tagger | 156 |
| 4.8.1.1 | Verbs | 156 |
| 4.8.1.2 | Adjectives..... | 157 |
| 5 | Discussion and Conclusion..... | 159 |
| 5.1 | Discussion | 159 |

| | | |
|-----|--|-----|
| 5.2 | Motivation of the Study..... | 159 |
| 5.3 | Stages of Development of Resource for Bhojpuri | 160 |
| 5.4 | Challenges Met in Corpus Creation and Annotation | 161 |
| 5.5 | Result of Comparative Analysis of Hindi and Bhojpuri | 163 |
| 5.6 | Final Result | 164 |
| 5.7 | Scope for Development..... | 164 |
| | References | 165 |

List of Tables

| | |
|---|-----|
| Table 1 Equations of SVM for resolved issues..... | 31 |
| Table 2 Concerned method and objectives of studies..... | 32 |
| Table 3 Methods of Multiclass problems..... | 33 |
| Table 4 Literary books/articles | 40 |
| Table 5 Contributing Doctoral Theses | 42 |
| Table 6 Corpus classification..... | 53 |
| Table 7 Major web links for data extraction..... | 55 |
| Table 8 Schedule languages of India | 60 |
| Table 9 Total files compiled for the present corpus | 62 |
| Table 10 Use of Determiners | 65 |
| Table 11 List of sentences from other languages, written in Devanagari..... | 66 |
| Table 12 List of Section headers found in the data..... | 68 |
| Table 13 List of other common crawler errors | 69 |
| Table 14 Discrepancy with genre specific sentences..... | 71 |
| Table 15 Examples of repeated phrases/words..... | 71 |
| Table 16 List of unaccepted terminology/ slangs | 72 |
| Table 17 Table listing the words/ phrases without space in between | 72 |
| Table 18 List of words with typing mistakes..... | 74 |
| Table 19 Fragmented words in the corpus | 74 |
| Table 20 List of miscellaneous errors | 75 |
| Table 21 Bhojpuri POS tagset (Revised) | 81 |
| Table 22 Composition of the Annotated corpus | 84 |
| Table 23 Reflexive Pronouns..... | 90 |
| Table 24 List of Wh-Pronouns..... | 92 |
| Table 25 Derived nouns from verbs (V<N)..... | 98 |
| Table 26 Annotation for other types of adverbs | 101 |
| Table 27 Echo words from the Corpus | 110 |
| Table 28 Echo Before words from the corpus | 112 |
| Table 29 List of unidentified words..... | 114 |
| Table 30 Word formations in Bhojpuri with inflected particles. | 118 |
| Table 31 Different Realizations of the Bhojpuri words..... | 123 |
| Table 32 Inconsistent tags..... | 124 |
| Table 33 Test result for seen data of 30k tokens (initial phase) | 136 |
| Table 34 Test result for unseen data of 10k token (initial phase)..... | 137 |
| Table 35 Test result for seen data of 10k tokens (current phase) | 139 |
| Table 36 Test result for seen data of 2k tokens (current phase) | 140 |
| Table 37 Error pattern for serial verbs..... | 142 |
| Table 38 Error pattern for conjunct verbs..... | 143 |

| | |
|---|-----|
| Table 39 Errors due to ambiguous tokens..... | 143 |
| Table 40 List of random errors | 146 |
| Table 41 Accuracy result of Hindi Tagger | 148 |
| Table 42 Comparative results of Hindi and Bhojpuri taggers | 150 |
| Table 43 Issues in Hindi tagger for tagging simple verb phrases | 150 |
| Table 44 Issues in Hindi tagger for tagging noun phrases | 152 |
| Table 45 Issues in detecting noun types in noun phrases | 153 |

List of Figures

| | |
|---|-----|
| Figure 1 Classification by SVM | 30 |
| Figure 2 Opus data and resources | 43 |
| Figure 3 Bhojpuri database | 44 |
| Figure 4 Procedure for creating the corpus | 53 |
| Figure 5 Screen shot of the crawled data | 59 |
| Figure 6 ILCIANN main page | 86 |
| Figure 7 ILCIANN annotation page (a) | 86 |
| Figure 8 ILCIANN annotation page (b) | 87 |
| Figure 9 Soft Margin vs. Hard Margin | 127 |
| Figure 10 Tagger Architecture | 129 |
| Figure 11 Training of the model file | 131 |
| Figure 12 Screen shot of the tagger generated tagging display | 133 |
| Figure 13 Screenshot of test on 30k seen data | 137 |
| Figure 14 Screenshot of test on 10k unseen data | 138 |
| Figure 15 Screenshot of test on 10k seen data | 139 |
| Figure 16 Screenshot of test on 2k unseen data | 140 |
| Figure 17 Screenshot of the evaluation for Hindi tagger | 149 |

List of abbreviations used

Abbreviations used for Technical Terms of Computational linguistics

| | |
|------|--------------------------------------|
| POS | Parts of Speech |
| WSD | Word Sense Disambiguation |
| NLP | Natural Language Processing |
| 90k | 90 Thousand |
| HMM | Hidden Markov Model |
| MEM | Maximum Entropy Model |
| CRF | Condition Random Field |
| SVM | Support Vector machine |
| IR | Information Retrieval |
| TTS | Text to Speech |
| STT | Speech to Text |
| ASR | Automatic Speech recognition |
| MT | Machine Translation |
| LFG | Lexical Functional grammar |
| GPSG | Generalized Phrase Structure Grammar |
| HPSG | head-Driven Phrase Structure grammar |
| EBMT | Example base Machine translation |
| NP | Nun Phrase |
| VP | Verb Phrase |
| NER | Named Entity Recognition |
| CPG | Computational Paninian Grammar |

Abbreviations used in the annotation of the corpus

| | |
|--------------|--------------------|
| <i>NN</i> | Common noun |
| <i>NST</i> | Spatio-temporal |
| <i>NNP</i> | Proper noun |
| <i>N</i> | Noun |
| <i>PR</i> | Pronoun |
| <i>PRP</i> | Personal Pronoun |
| <i>PRI</i> | Indefinite Pronoun |
| <i>PRC</i> | Reciprocal Pronoun |
| <i>PRF</i> | Reflexive Pronoun |
| <i>wh</i> | Question word |
| <i>PRQ</i> | wh-Pronoun |
| <i>DM</i> | Demonstrative |
| | Deictic |
| <i>DMD</i> | Demonstrative |
| | Relative |
| <i>DMR</i> | Demonstrative |
| | Indefinite |
| <i>DMI</i> | Demonstrative |
| <i>DMQ</i> | wh-Demonstrative |
| <i>V</i> | Verb |
| <i>VM</i> | Main Verb |
| <i>VAUX</i> | Auxiliary verb |
| <i>VF</i> | Verb Finite |
| <i>VNF</i> | Verb Infinitive |
| <i>VINF</i> | Verb Infinitive |
| <i>VNG</i> | Gerund |
| <i>JJ</i> | Adjective |
| <i>RB</i> | Adverb |
| <i>CC</i> | Conjunction |
| <i>CCD</i> | Coordinator |
| <i>CCS</i> | Subordinator |
| <i>Ech_b</i> | Echo before |
| <i>Ech</i> | Echo |
| <i>PSP</i> | Postposition |
| <i>RP</i> | Particles |
| <i>CI</i> | Classifier |
| <i>RPD</i> | Default |
| <i>INJ</i> | Interjection |
| <i>INTF</i> | Intensifier |
| <i>NEG</i> | Negation |

| | |
|------|---------------|
| QT | Quantifiers |
| QTC | Cardinals |
| QTF | Ordinals |
| QTO | General |
| RD | Residual |
| FW | Foreign Words |
| SYM | Symbols |
| PUNC | Punctuation |
| UNK | Unknown |
| misc | miscellaneous |
| pol | politics |
| ent | entertainment |
| lit | literature |
| spo | sports |
| EMPH | Emphatic |

List of symbols

| Devanagari | IPA | Transliteration | Devanagari | IPA | Transliteration |
|------------|-----------------|-----------------|------------|----------------|-----------------|
| क | k | k | ब | b | b |
| ख | k ^h | kh | भ | b ^h | bh |
| ग | g | g | म | m | m |
| घ | g ^h | gh | य | j | y |
| च | tʃ | c | र | r | r |
| छ | tʃ ^h | ch | ल | l | l |
| ज | dʒ | j | व | v | v |
| झ | dʒ ^h | jh | स | s | s |
| ट | t̪ | t | श | ʃ | sh |
| थ | t̪ ^h | th | ष | ʂ | Sh |
| द | ð | d | ह | h | h |
| ध | ð ^h | dh | अ | ə | a |
| न | n | n | आ | a | A |
| ट | t̪ | T | इ | ɪ | i |
| ठ | t̪ ^h | Th | ई | i | I |
| ड | d̪ | D | उ | ʊ | u |
| ढ | d̪ ^h | Dh | ऊ | u | U |
| ड़ | ɽ | Dx | ए | e | e |
| ढ़ | ɽ ^h | Dhx | ऐ | ɛ | ai |
| ण | ɳ | N | ओ | o | o |
| प | p | p | औ | ɔ | au |
| फ | p ^h | ph | अं | ŋ | M |

1 INTRODUCTION

1.1 INTRODUCTION

'Parts of Speech' is an important stage in any Linguistic analysis. The information that are provided by the web search engines are also inherited with POS information for the proper coding and decoding of information based on the user's input. A language learner first learns the vocabulary and the grammatical categorization for that language before he/she comes to analyze the basic sentence structure.

Parts-of-Speech (POS) annotation is one major building block for developing language technology. From the Computational Linguistic perspective, this stage of providing a text with the POS information is known as POS tagging. This is a process of assigning a word (token) with the tag (POS tag) which describes the grammatical category of that word in a given string. It might seem similar to the 'gloss' provided during the linguistic analysis.

POS annotation is a bridge between the corpus and other higher level language resources. The corpus of a language is the database or the documented piece of language in machine readable format. On the other hand, language resource is the resources helps in development of technological aspect of a language. These resources are chunker, parser, WSD (word sense disambiguator etc.). The annotation is an intermediate level which transfers the morpho-syntactic (POS) information to the corpus resulting in creating an annotated corpus of a language which is used for higher level NLP (natural language processing) research works.

The NLP work on Indian Language came rather late but by now many Indian languages have come up with different technologies for them. There are different consortia and research groups considering Indian languages for developing technologies.

The present work aims at developing such resource for Bhojpuri. The goal of the study is to develop a statistical POS tagger for Bhojpuri, as it is a less resource language and no such development has taken place in this regard. Support Vector Machine (SVM), serves as the learning model for the tagger as it is a less probabilistic and a pure classification model for both linear and non-linear data.

In due course, there are two other stages of developments involved, namely creating Bhojpuri corpus and the POS tagging of the corpus. For any analysis, first of all, a database of language is required; this has been fulfilled by creating a corpus for Bhojpuri under this experiment with approximately 1.69 lakh tokens. In the next stage, the corpus has been annotated for the training purpose and an annotated corpus of 90k tokens has been created. The tagger has trained in this endeavor shows the accuracy of 88.6% and rest part of tagger is under development.

The later section of the work deals with the comparative analysis of present Bhojpuri tagger with a contemporary Hindi tagger which is in being worked upon. The accuracies and the test results are evaluated for both the taggers. The some linguistic rules are proposed in the later part of the study, which could be applied at the post processing stage during the development of the tagger.

1.2 METHOD OF RESEARCH

There are many research groups like IIIT Hyderabad, IITs, JNU, UoHyd etc developing technologies for Indian Languages. Taggers have already been trained for languages like Bangla, Hindi, Malayalam, Telugu and Punjabi etc. with considerable results. Apart from rule based taggers, many statistical training models have also been employed for different languages. HMM, MEM, CRF, SVM are all tested and the accuracy of most of the tagger ranges between 86 -90%. Some have even higher accuracies.

Under this experiment a general domain corpus for Bhojpuri has been created following the standards for generating representative corpus and the source data was drawn from the web. The formatting and validation of the corpus has been done following the ILCI (Indian Languages Corpora Initiative) format and validation tool (ILCIANN) respectively.

For annotation purpose, the BIS (Bureau of Indian Standard) scheme for Indian languages has been adopted and the Bhojpuri POS tagset was devised following the same. The tagging has been done semi-automatically using the ILCIANN tool developed under the ILCI project. Lastly, the Bhojpuri tagger has been modelled on the SVM. The efficiency of the tagger has been compared with the Hindi tagger modelled on SVM. A comparative analysis of both the taggers has been done under controlled conditions where, size and domain of the test data and the learning model were kept constant. As a result of which we get that, although the taggers were calculated with approx 94% and 88% accuracies for Hindi and Bhojpuri, the results of the domain specific Hindi

tagger and the general domain Bhojpuri tagger reduces to one percent and the new accuracy was found to attain 93% and 87 %, respectively, when tested on the data from a new language domain. This result falsifies the hypothesis made in the beginning of the study which states that the result of a general domain tagger will surpass the domain specific tagger if the test data belongs to a specific domain, as the familiarity of the general domain tagger with the data will be higher than that of the domain specific tagger.

1.3 STATEMENT OF PROBLEM

Language resource is the primary asset of any language in this era of technology. Developing language resource is one of the primary objectives of software developers, computational linguist or language scientists. Though there is a lot of work done on different Indian languages and tools like POS tagger, chunker, shallow and full parsers, word net, sense disambiguators etc are available for all major Indian languages, mostly scheduled languages. But there are some more languages which in spite of having important role in the society, from linguistic point of view, struggle for their technological advancement. One such example is Bhojpuri, the language which has its existence all over the world, with more than 39.7 million speakers¹, in its spoken form, in writings, through researches, tourism and cinemas. But in this age of smart phones, if someone is looking for a single smart online app for Bhojpuri, there is none.

With the advent of computational linguistics it has now become easy to make a huge database of a language and create as many resources out of the one common robust collection of text, called corpus. When we say language resource, it means, the multi level analysis of language through systems and tools that are trained to study different language features and bring them to application level. Parts of Speech tagging or identification of word classes “is the first fundamental operation on raw text, and subsequent stages on NLP like phrase detection, parsing, semantic role labelling and so on follow after part-of-speech tagging”. (Bhattacharya P., 2010). Therefore, the very processing level demands lot of attention and consistency. This consistency is difficult to produce manually in a data of millions of word tokens. To achieve this we use data driven or statistical techniques.

¹ ethnologue.com/language/bho accessed 05/06/2015 time 01:25:08 am

The aim of this study is to create a stochastic POS (Parts of Speech) tagger for Bhojpuri taking SVM (Support Vector Machine) as the experiment model. The study will take off with building a general domain monolingual corpus for Bhojpuri, leading to the formulation of POS tagset for the language following the BIS (Bureau of Indian Standards) scheme for Indian languages, which is then followed by training of a statistical SVM based tagger. The study also aims at evaluating the suitability of the model based on the accuracy results obtained for Bhojpuri and the comparison with the accuracy found in pre-existing SVM tagger for Hindi, a genetically related language. Further, the scope of improvement by suggesting possible changes in the scheme and tagging technique and also by propounding rules for better applicability has been discussed.

This dissertation is, probably, one of the initial attempts for building language resource for Bhojpuri by introducing corpus with approximately 1.69 lakh word tokens along with the tagger rating 88.6 % accuracy, which, is quiet convincing achievement at the very first attempt.

1.4 RESEARCH QUESTION

The main focus of study is the creation of resources for Bhojpuri and comparing the results with that of a similar experiment done on Hindi. Each experiment undertaken is new to this field of study as well as to the researcher. Building a corpus is a rigorous process as it involves careful selection, dissemination and validation process. The corpus presented is the first corpus for Bhojpuri and so is the tagger. The output generated by the tagger is quite convincing and around the one generated by the Hindi tagger. But it is more interesting to see the annotation similarities and contrasts among Bhojpuri and Hindi.

Therefore, the present research will attempt to explain the following research questions:

1. Challenges for building POS annotated corpus for Bhojpuri.
2. Suitability of the SVM model for the language and in handling of issues/ambiguities.
3. Comparison with Hindi using the same model.
4. Evaluation of the taggers and scope for improvement.

1.4.1 Hypothesis

Apart from the resource generation, the study also involves a comparative analysis of two languages trained on the same model, namely Hindi and Bhojpuri. The comparison is made under some assumptions. First, Hindi and Bhojpuri are related languages, with quite similar syntax and wide range of vocabulary. Their word forms are also found quite different from each other. But the tagset evolved for these languages shares almost all tag categories except two, classifier and the echo_before. Based on this, if a comparison is made between the efficiency of the taggers, there should not be much difference in their accuracies and the error patterns.

Secondly, the Hindi tagger was being trained upon the ILCI corpus (as per the researcher's knowledge) which includes mainly two domains, health and tourism as noted from Nainwani (2011). And the Bhojpuri tagger presented here is a general domain tagger with the training corpus from different language domains. Based on this, another assumption was made which states that if the data from any domain (other than one found in Hindi tagger) is tested upon both the taggers, the accuracy result of Bhojpuri tagger must surpass the result generated by the Hindi tagger. Because the familiarity of the Hindi tagger with the domain of test corpus is less than that of the Bhojpuri tagger which includes data from varied domains.

1.5 BACKGROUND TO PROBLEM

Bhojpuri is among the major languages of the north India with 33,099,497 speakers². The parliament of India is now considering the inclusion of Bhojpuri in the Eighth Schedule of the Indian Constitution, with other 22 major Indian languages. Recently through Indian government initiatives all scheduled languages have received the attention from the Computational Linguistics community, but unfortunately this non-scheduled language still lacks such attention for its development in the field of NLP (Singh, S. 2014). By the time, Bhojpuri has developed its literature to a height. There are several online blogs, portals, newspaper and related web pages dealing in contents in various subject matters such as cinema, politics, sports, fun, chat, literature, criticism and so on. But, technically, only a few noticeable efforts have been made. The authentic web resources like online journals, newspapers and magazines are also very

² Census data 2001

limited. The need for development of technology for the language evokes the researcher to develop a POS (Parts-of-Speech) tagger for Bhojpuri using a stochastic approach.

1.5.1 Resources for Natural Language Processing

1.5.1.1 Corpus

Corpus is a large text of one or more languages in machine readable format. It is the collection of literary or non literary data from the natural language. Different types of corpus are used for different tasks, some require raw and some make use of annotated corpus. The usability of annotated corpus is quite high over raw un-annotated corpus as it serves as the primary resource data for other higher level NLP tasks. The corpora creation in Indian languages have started very late, yet have some major projects like MSRI, LDC-IL and ILCI, by the time, for creating corpus in almost all major and scheduled languages of India. The journey started with the advent of CII-EMILLE, which was also a huge general domain Hindi corpus with data from different language domains. More about corpus will be discussed under the section on review of Literature.

1.5.1.2 Parts of Speech (POS) tagger

Parts of speech tagging is one of the basic applications of NLP where each word token is given a definite tag based on the grammatical function of that word. The ‘tags’ are set of definite tokens formulated concerning the nature of corpus. The idea behind annotation is to devise the mapping of grammatical units within the human consciousness and its computer re-modelling. It can be done both ways, manually and automatically. Now a day, many automatic taggers have come up with convincing accuracy results. More on this is discussed in detail in the following section.

1.5.1.3 Chunker

Chunking is the task of grouping of phrases in a sentence as noun phrase (NP) and verb phrase (VP). It is also known as local word grouping. After POS annotation, it is the second level of linguistic annotation in NLP. Chunker is the tool used for marking the phrase boundaries for every word groups, automatically (fully or partially).

1.5.1.4 Morphological analyser

Analysers are tools for analysis of the internal structure of a word. They can either analyse the word in its inflectional and derivational components or used for generating the different possible morphological forms of a word from its root or word stem.

1.5.1.5 Parser

In general, syntactic parsing is the analysis of a string of words. It is a method of extracting the meaning of a sentence based on the theory of formal grammar. Parsing in computational linguistics is a similar process but this analysis results in a parse tree which shows the relation between the constituents in a string. This may also contain semantic information based on the nature of parsing, full or shallow. The tool responsible for processing the string in this field is known as parser.

1.5.2 Approaches to Stochastic Learning

A stochastic or statistical work includes frequency, statistics and probability. A frequency based method works on calculating the frequency of used tag for a word in annotated text and the other n-gram approach determines the tag by calculating the probability of the previous tag 'n'. This generally uses unigram, bigram and trigram models. Based on these, there are different statistical models with varying underlying principles and learning algorithms:

1.5.2.1 Hidden Markov Model (HMM)

Hidden Markov models (HMMs) is based on Bayesian inference. It was developed by Xie, Chang, Divakaran & Sun (2003). It has been the most popular tools in machine learning which considers all possible sequence of classes, for POS tagging. The goal is to try to establish a framework which supports both structure and event analysis and also serves as the building blocks for high level semantic analysis. There are two major distinctions between the concept of semantic space for semantic modelling of both mid and high level semantic structures and a new HMM with multi-faceted advantages for semantic analysis of videos (see Sober, 2010).

The HMM model is trained on annotated corpora to find out the transition and emission probabilities. For example, for a sequence of words w , HMM determines the sequence of tags t using the formula:

$$t = \operatorname{argmax} P(w, t)$$

Equation 1

1.5.2.2 Conditional Random Fields (CRF)

A Conditional Random Field (CRF) segments and labels a sequence of data on the basis of a probabilistic framework. A conditional model specifies the probabilities of possible label sequences given an observation sequence. This probability of the label sequence may depend on arbitrary or non-independent features of the observation sequence. The probability of a transition between labels may depend not only on the current observation, but also on past and future observations. The CRF model calculates the probability based on some features, which might include the suffix of the current word, the tags of previous and next words, the actual previous and next words etc.³

1.5.2.3 Maximum Entropy Model (MEM)

The Maximum Entropy Model is based on the principle of Maximum Entropy. This states that while selecting among a number of different probabilistic models for a set of data, the one which makes fewest arbitrary assumptions about the nature of the data is the most valid model.⁴ It was first introduced by Ratnaparkhi (1996) and McCallum et. al (2000). The model probability of history H with tags T is defined as:

$$p(h, t) = \pi \mu \prod_{j=1}^k \alpha_j^{f_j(h, t)}$$

Equation 2

In the above equation $\{a_1, \dots, a_k\}$ and $\{f_1, \dots, f_k\}$ are the positive model parameters and ‘features’, respectively.

where,

$f_j(h, t) = \{0, 1\}$ and parameter a_j corresponds to a feature f_j .

The chosen parameter to maximize the training data p is $\{a_1, \dots, a_k\}$ (see Karthik, K.).

1.5.2.4 Memory Based Learning

MBT is a memory-based tagger-generator and tagger in one. It is based on Memory Based Learning. The tagger tags new sequences and the tagger-generator generates a sequence tagger

³ Hasan, 2006.

⁴ MacKinlay, 2005.

on the basis of a training set of tagged data. MBT can be used as part-of-speech taggers or chunkers for NLP. It can also be used for domain specific named-entity recognition (NER) and information extraction and disfluency chunking in Speech to text.⁵

Other models

1. Stanford POS Tagger (Toutanova et al. 2003)
2. Multi-class SVM model which is trained by SVM- Multiclass (Joachims 1999)
3. Stochastic gradient descent (SGD) by Shalev-Shwartz et al. (2007).

1.5.3 Support Vector Machines.

Support Vector machine is a supervised learning model, generally defined as an optimal margin classifiers. It was developed by Vapnik in 1995 at COLT92 (Computational Learning Theory) conference, based on the statistical learning theory and well known for its generalization performance. SVM is used for analysing data, classification and pattern recognizing in indefinite feature space. Regression, here, means the extension of the machine to function approximation and the comparison of its performance with other function approximators (Kecman 2006).

In simple words, a Support Vector separates the two classes of variables by drawing hyperplane(s) in the features space. The data in each class are denoted with dots. The best hyperplane is the one which is at the farthest distance from the data point of any class, therefore, also called optimal margin classifier. Wider margin lowers the generalization error of the machine.

The reason for including this as the learning model for POS tagger is that it is used for classification and categorization of the input data in one of the two categories by training algorithm and making it a non-probabilistic linear classifier. Its extension to non-linear classifiers mapped on high dimension feature space is to fulfil the demand rose out of availability of very large no. of electronic data under process every minute. This mapping on infinite space is possible with the use of Kernel function $[k(x,y)]$.

A kernel function is used for calculating the dot product of two non linear vectors, mapped in the infinite feature space. Different kernel based methods derived in due course of time are namely- Kernel Least Square Method, Kernel Principle Component Analysis and Kernel Mahalanobis Distance.

⁵ <http://ilk.uvt.nl/mbt/>

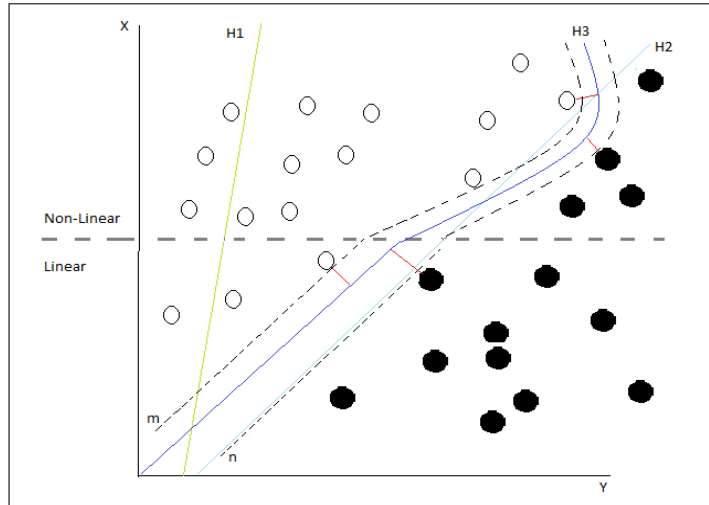


Figure 1 **Classification by SVM**

In the above figure 1, there are two sets of data and H1, H2 and H3 are three hyperplanes. The H1 hyperplane does not separate these classes, H2 does but the margin of separation are very less at different point and H3 separates the two classes with maximum margin for both linear and non-linear variables.

The set of n points can be calculated in a given training data D , as

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^p, y_i \in \{-1, 1\}\}_{i=1}^n \quad \text{Equation 3}$$

Where, X_i and Y_i belong to same class and Y_i is either 1 or -1. Each X_i is a p -dimensional real vector. The maximum-margin hyperplane will divide the points having $Y_i=1$ from $Y_i=-1$.

Therefore, set of points satisfying X can be written as the following hyperplane:

$$\mathbf{w} \cdot \mathbf{x} - \mathbf{b} = 0 \quad \text{Equation 4}$$

Where, w is the normal vector to the hyperplane.

From the above equation we get the imaginary hyperplanes m and n as

$$\mathbf{w} \cdot \mathbf{x} - \mathbf{b} = 1 \quad \text{and} \quad \text{Equation 5}$$

$$\mathbf{w} \cdot \mathbf{x} - \mathbf{b} = -1 \quad \text{Equation 6}$$

From geometry, the distance between the two hyperplanes comes out to be $2/\|\mathbf{w}\|$. By adding the constraint, we can prevent the data point falling into the margin as we get

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1$$

Equation 7

Support Vector has been applied to various problems in many fashions. In the due course, the SVM is found to come up with the following variants:

- (i) Least square SVM
- (ii) Linear programming SVM
- (iii) Robust SVM,
- (iv) Bayesian SVM
- (v) Committee machines.

Based on this basic principle of optimal margin, different problems have come up with various solutions. The equations derived for resolving the issues for different problems, with the help of SVM, are enlisted in the table 1 below:

Table 1 Equations of SVM for resolved issues

| | | |
|----|---|-----|
| a) | Quadratic programming optimization problem: | |
| | $\arg \min_{(\mathbf{w}, b)} \frac{1}{2} \ \mathbf{w}\ ^2$ | (a) |
| b) | Lagrange multipliers | |
| | $\arg \min_{\mathbf{w}, b} \max_{\alpha \geq 0} \left\{ \frac{1}{2} \ \mathbf{w}\ ^2 - \sum_{i=1}^n \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i - b) - 1] \right\}$ | (b) |
| c) | Karush-kunh-Tucker condition | |
| | $\mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i.$ | (c) |
| d) | Dual form | |
| | $\mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i.$ | (d) |
| | Where kernel is defined as $k(x_i, x_j) = x_i \cdot x_j$ | |
| e) | Soft margin principle for mislabelled examples by Corinna Cortes is resolve with | |
| | $\arg \min_{\mathbf{w}, \xi, b} \max_{\alpha, \beta} \left\{ \frac{1}{2} \ \mathbf{w}\ ^2 + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i - b) - 1 + \xi_i] - \sum_{i=1}^n \beta_i \xi_i \right\}$ | (e) |

1.5.3.1 Application based SVM

Since its inception, SVM has been studied by different scholars in different respects and for various kinds of applications as mentioned in the table 2 below:

Table 2 **Concerned method and objectives of studies**

| S. No. | Author | Concern | Objective |
|---------------|---------------------------------|---------------------------|-------------------------------|
| 1. | Bennett and Campbell | Geometrical Point of view | |
| 2. | Hebrick | Bayesian rule | exposition of kernel method |
| 3. | Hastie, Tibshirani and Friedman | | Statistical literature on SVM |

Vpjisav Kecman's (2006) review of the book "Support Vector Machine for pattern classification" by Shigeo Abe, tells us that SVM is an easy learning machine and efficient data mining tool. This books talks about classifiers, matrices, quadratic programming and semi definite kernels.

It talks about direct and indirect decision functions in which the boundary of class are given by the curves but in direct the function is not present whereas in indirect , both the functions receives the same value (Kecman, 2006). Hard margin for linear classification and two types of soft margin for overlapping is discussed. These are overlapping and L1 soft margin for linear sum and L2 soft margin for square sum slack variables are required for training data.

Methods for multiclass problems are one against-all SVMs (each class is separated from the remaining classes), pair wise SVMs (one class is separated from another class), the use of error-correcting output codes for resolving unclassifiable regions, and all-at-once SVMs (in which decision functions for all classes are determined at once). Their properties are mentioned in the table 3 below:

Table 3 **Methods of Multiclass problems**⁶

| Method | Property |
|-------------------------------|---|
| One against –all SVM | Every class is separated from the remaining cases |
| Pair wise SVMs | One class is separated from the other |
| Error correcting output codes | For resolution of unclassifiable regions |
| All-at-once SVM | Decision functions are determined at a time for all classes |

1.5.3.2 Advantages of SVM Environment

Support Vector Machines has proved out to be excellent in classification and regression analysis. The non probabilistic design targets at selecting the best possible vector out of the array of vectors. It provides several facilities to its user by being a flexible, portable and an efficient performer. The SVM get set with an easy installation process and the tagging speed if the tool is very high that it tags several thousand data in just seconds. The flexibility can be measured in this term that despite having language with varied language features, the tagger should be capable of handling the information and context, wisely. Because languages with richer morphology has bigger feature pattern and the ambiguity level also fluctuates, accordingly. These are taken care well by the SVM.

One of the major advantages of the machine is that it converts the linear model into a non linear one without any clustering because the input space automatically gets clustered in the feature space after mapping. This is helpful for non linear separation of the data class using kernel based method.

⁶ Abe (2005)

1.6 REVIEW OF LITERATURE

1.6.1 Bhojpuri

Bhojpuri is one of the prominent languages of Northern India belonging to Indo-Aryan language family. The ISO code given to the Bhojpuri language is 639-3 and it is spoken by 33,099,497 speakers as recorded in Census data 2001. It is also known as Hindustani, Pashcimas, Deshwali, Khotta, Bihari, Bhojpuriya and Purabiya. The major Bhojpuri speaking belts in India are Bihar and some part of Uttar Pradesh. Upadhyay H. S. (1988) in his book, Bhojpuri folksongs from Ballia, has listed out Mirzapur, Ghazipur, Jaunpur, Ballia, Gorakhpur, Deoria, Basti, Azamagrh and Varanasi districts in Uttar Pradesh and Rohtas, Eastern Champaran, Saran, Siwan, Ranchi and Bhojpur districts in Bihar as the major Bhojpuri speaking areas of the Northern belt. Besides this, he also mentions that it is accepted as one of the official Languages of Nepal and Mauritius, spoken widely in Guyana, and Fiji. He further classifies Bhojpuri into its ‘standard variety’ spoken mainly in Bhojpur and Rohtas, its ‘western variety’ spoken in Varanasi and nearby districts and the ‘Madheshi’, a variety spoken in Tihat and Gorakhpur. On the other, Udaya Narayan Tiwari (1960) classifies Magahi, Maithili and Bhojpuri as three dialects of the Magadhan group of which Bhojpuri is the western most variety. He tries to show some noteworthy differences in these varieties in form of verb morphology and the availability of literature. According to him Maithili is found to have much older literature as compared to Magahi and Bhojpuri which are applauded for their folk songs and poetries now developing into literature since the 20th century.

1.6.2 Historical Background

Bhojpuri is considered as the western most and the eldest variety of language in the Magadhan group. The other two are Magahi and Maithili. Udaya Narayan Tiwari (1960) tries to show some noteworthy differences in these varieties in form of verb morphology and the availability of literature. According to him the Maithili is found to have much older literature as compared to Magahi and Bhojpuri which being oral tradition for a long time, are praised for their folk songs and poetries and now developing into literature since the twentieth century.

The name 'Bhojpuri' was derived from the older name 'Bhojpuriya' which was kept after the name of Bhojpur district in Bihar. According to Tiwari⁷ (1960), the place holds great importance in the medieval history of India because of the kings of Ujjain (Bhojas). From the history we came to know that the name 'Bhojpur' was the result of the conquest between Cheros tribe and Bhojas, the descendents of the famous Raja Bhoj. It is interesting to see, from a city 'Bhojpur', the name spread over and became popular as a district. Similarly, the language 'Bhojpuri' is also becoming popular worldwide retaining the name from where it originated.

With time the names of places also undergo modification as language itself. And each name has some relevance to the history of that place. Mishra (2003) and Tiwari (1960) give an account of how a place gets a name which undergoes changes with time and the relevance of the name in the history of the language. According to Tiwari 'Pashcimas' and 'Hindustani' was used in Bengal, 'Deshwali', 'Mulki' and 'Khotta' were the names given to migrants engaged with lower jobs like gardening or vending. They address each other as Mulki as they both belonged to the same nation. The variety spoken in 'Chapra' was later propounded as Chhapariya, 'Purbi' and 'Purabiya' is the name of the variety spoken in other districts in the east of Bihar, Benaras in chief. The ideas behind the names are pointed out in Mishra . For example, Ghazipur is named after Guru Maharshi Vishwamitra who was the son of King Gadhi . Bharat, the son of king Duṣyant, is responsible for the coinage of Bharatvarsha. The language change is not a recent phenomenon and so is the change in the names. With time, the linguistic change brought changes in the names of ancient cities Araṇya, Saraṇya and Champa raṇya to modern Arr ah, Saran and Champaran. And their records can be dated back to centuries. Tiwari (1960) in his book *Origin and development of Bhojpuri*⁸ regards to 'The Bengal Atlas', 1971 by James Rennell as the first map of Bhojpuri where 'Bhojpur' and 'Shahabad' districts are mentioned with a different spelling as 'Bhoujpour' and 'Shawabad'. Not only this, old sayings in Bhojpuri, even dates it back to the late Mughal period but the first use of the term 'Bhojpuri' is found in the Grierson's *Linguistic Survey of India (1903)*.

⁷ Tiwari (1960)

⁸ Ibid

1.6.3 Status of Bhojpuri

1.6.3.1 Social Status

The social structure and system of Bhojpuri community can be understood from the following formula.

Understanding class, caste, dialect = understanding of Bhojpuri Speech Community (i)

Equation (i) is the Mishra's (2003) understanding of the notion of class (social status of a language), caste (the religion hierarchy within a society) and dialect (the speech variety of the society) can be very well understood with ones understanding of the Bhojpuri Speech Community. He describes main components of this society are mainly the Hindus and the Muslims. The *chaturvarna* caste system divided among the *Brahmins*, *Kshatriyas*, *Vaishyas* and *Shudras*; still prevails in the society. Culturally speaking, the large number of celebrations in form of festival, the devotion to follow the customs, rituals offered at every phase of life as *samskara*, ashram system, *purushartha* etc are believed to be the integral part of the living of the people. Lastly, the Panchayat system, joint family system, their seating arrangements and settlement patterns are the major pillars of the hierarchical social system. Many scholars believe Bhojpuri to be a very polite language, Tiwari also opines that the '*rauwa*' dialect of eastern Bihar should be chosen as the standard variety because of the sweetness of the tongue.

1.6.3.2 Educational Status

A number of significant efforts have been made since last century for recognition of Bhojpuri as a national language. Some universities in Bihar have initiated courses in Bhojpuri at the graduate level. In spite of facing difficulties like the lack of teaching posts, universities like Patna University, Magadh University, Vir Kunwar Singh University and Jai Prakash Narayan University, Bihar University and Ambedkar University are constantly trying to bring it to academics. Many work on collecting and editing the scattered folk literature currently running. Followed by this, a number of conferences like All India Bhojpuri Conference at national level and other small conferences are being organized in Mauritius, Siwan, Deoria, New Delhi and other parts of the world. The proceedings of the conferences are regularly published in Bhojpuri journals. One central figure to all these activities is Shri. Pandey Kapil, who is a retired Deputy Director of Bihar, is now a day, taking lead in the editing and publishing aspect of the affairs.

Along with the academia, Bhojpuri is also expanding in the technical sphere as there are Bhojpuri newspapers, magazines, literary and entertainment resources available on internet. The advent of Unicode has also paved way for languages to evolve them with the digitization. The resources like corpus etc can be easily worked upon installing the peculiar sounds of the language into script.

1.6.3.3 Political Status

Bhojpuri has been a powerful language since very beginning. It is believed that the bravery of Kunwar Singh and Mangal Pandey made the British seek Bhojpuri men in the army only after realising the power of the Bhojpuri rulers. The languages are playing potential role in nation building. After independence, the rising public demand for the recognition of their speech community compelled the Indian constitution to promise to divide states along their linguistic boundaries, soon after independence. This raised the wave of language movement, nationwide. Different language movements headed by scholars took place. The Annamalai E. Book by Ajit Jogi for recognition of Chattisgarhi, Movements on the linguistic purism for Tamil, recognition of Sindhi, establishment of Urdu as result of the Iranianization of Hindwi, and many more were raised and fought. Konkani and Punjabi also rest their case recently, in this regard. In this regards, Mishra is right in saying

“Bhojpuri is a practical language of an energetic race which is ever ready to accommodate itself to circumstances and which has made its influence felt all over India” (Mishra, 2003:1).

There are different foundations like Akhil Bharati Sahitya Sammelan 1973 and Akhil Bharatiya Bhojpuri Parishad, UP which are consequent to the growing consciousness among people towards the upliftment of their language, politically. But the recognition of Bhojpuri as a linguistic state is still a controversial debate. Bhojpuri has its spread both in UP and Bihar, and some outer regions, therefore, the idea of developing a hypothetical Bhojpur as the linguistic region for the language is argued by scholars Srivastava and Choudhary. Where, one claims that the development of the language is detrimental to Hindi and other in supports to the remedies for decreasing index of regional varieties. They suggest Hindi to remain the official language but the rise in functionality of the regional languages should be promoted.

1.6.4 Standardization of Bhojpuri

“Dialects are capable of producing classics”⁹

History is evident of the fact that dialects are linguistically equipped enough to function as an independent language. From the works of great authors like Kabir, Tulsī, Jayasī, it is evident that dialects are equipped enough to act as an independent language. This fact is also supported by Haugen’s statement “The History of language demonstrates convincingly that there is no such thing as inherently handicapped language. All the great languages of today were once underdeveloped”¹⁰

It is the time, society and the applicability of the language which makes it standard one. There are some stages, Mishra mentions, for standardizing a language, namely- Selection, Codification, Propagation, Elaboration, and Recognition. Putting Bhojpuri into this frame, linguistic uniformity is the first key measure to be taken for Bhojpuri. There should be a common speech irrespective of the class. The lack of uniformity will hamper the perfect communication throughout. But the selection process should not be so restricted that the language loses its essence. The omission of the unaccepted vocabulary is necessary but the elimination of rusticity; decency of Bhojpuri will change the language on the whole. With Unicode, the graphical representation is no more complex. And the use of lexical databases like dictionaries helps keeping it constant throughout. But evidently, only “grammar and dictionaries are not enough unless linguistic description is accepted by the user”. Because the reason being taken as a local language of the place, speakers often feel shameful to accept it as their first language. Once this is coped, the propagation and elaboration for Bhojpuri is ready to bloom as already discussed under previous section.

Keeping the expansion and growing exposure and popularity of the language, the Government of India is considering including it in the eight schedule. The day is not far when Bhojpuri will be among all other official language of India.

⁹ Mishra, 2003

¹⁰ Mishra, 2003, (as quoted by Srivastava).

1.6.5 Existing endeavour in this field

1.6.5.1 Literature

Bhojpuri is as rich as any other tradition. The richness of the culture can be seen through the ages old repository of the language in the form of special occasional songs, poems, festivals and the retention of old religious customs till date. Being an oral tradition till late seventeen century it has a wider speaking community. Bhojpuri was able to hold much of its tradition alive which later contributed to the formation of literature of the language. The collection of folk songs, poems, short stories, mythological stories, narratives is largely written in Bhojpuri. Works in different disciplines are going on in the language, such as:

1.6.5.1.1 Folk and comparative literature

Bhojpuri folk literature has been portrayed in the works of Upadhyaya (2008)¹¹ and Upadhyay (1988)¹². *Bhojpuri folksongs from Ballia* by H.S. Upadhyay is a collection of more than one thousand folksongs collected from the Ballia district in Uttar Pradesh. The book depicts the richness of culture and language both. The songs for various ceremonies, festivals and incarnations show the sensitivity of people to their root and the versatile poetic genre, their linguistic empowerment. Whereas, *Bhojpuri Loksahitya* is a compact compilation of history and development of Bhojpuri literature by Krishna Dev Upadhyay. The Indian literary tradition of Bhojpuri folk art, its classification, forms and values are presented with exemplary folk songs, stories and idioms. The comparative literatures for Bhojpuri with other developing or established languages have also being looked upon. The history and poetic developments in Bhojpuri like compositions and its various types in an oral tradition, modern poets and inspirations of Avadhi and Bhojpuri languages are listed by Shaligram Shukla 'Neer', 1984 in his *Adhunik Avadhi and Bhojpuri: Etahas aur Kavya*.

1.6.5.1.2 Socio-linguistically oriented works

The journey starts with Grierson's *Linguistic Survey of India*, in which he dedicated one whole section in the second volume on the language of Bihar named as 'Bihar and Oriya' and collected

¹¹ Krishna Dev Upadhyay (2008)

¹² H. S. Upadhyay (1988)

language samples from different dialects of the Bihar district by giving them one umbrella term *Bihari*. He also explored the ‘Seven Grammar of the dialects and sub-dialects of Bihari language’, life of Bihar and collections of Bhojpuri folk literature. In this stream the ‘*Notes on the Bhojpuri Dialect in Hindi Spoken in Western Bihar*’ by John Beams is another landmark in the linguistic analysis of Bhojpuri. It includes study on phonology, noun and pronoun declension, verb conjugation and postposition.

Some of the major contributors of *Bhojpuri Sahitya* (literature) are listed in the table 4 below:

Table 4 Literary books/articles

| S. No. | Authors | Year | Title | Publication | Work Concerns |
|--------|---------------------------|---------|--|-------------------------------|--|
| 1 | A.B. Singh | 1976 | Auxiliary verbs in Bhojpuri | | Article |
| 2 | Durgashankar Prasad Singh | 1959 | Bhojpuri ke kavi aur Kavya | Patna | Culture and social concerns |
| 3 | Indradev | 1957-58 | The Sociology of Bhojpuri Literature | Lucknow University | Culture and social concerns |
| 4 | Krishna Dev Upadhyay | 1960 | Bhojpuri Loksahitya ka Adhyayan | Varanasi | Culture and social concerns |
| 5 | R. B. Mishra | 2003 | Socology of Bhojpuri Language | Swasti Publication, Varanasi | Culture and Social Concerns |
| 6 | S Ojha | 1956-63 | Bhojpuri Lokoktiyo ke Sanskritik Paksh ka Adhyayan | Rachi | Folk tales, songs and poetry |
| 7 | Satyavrat Sinha | 1957 | Bhojpuri Lokgatha | Hindustani Akadami, Allahabad | Culture and social concerns |
| 8 | Shashi Shekhar Tiwari | 1970 | Bhojpuri Lokoktiyã | Bihar Rashtrabhasha Parishad | Folk tales, songs and poetry |
| 9 | Sridhar Mishra | 1971 | Bhojpuri Loksahitya | - | modified Ph. D thesis (Dept of Hindi, BHU) |

| | | | | | |
|----|---------------------|------|--|--|----------|
| 10 | Sukhdev Singh | 1968 | Bhojpuri aur Hindi ka Tulatmak Adhyayan | Nilabh Prakashan, Allahabad | Academic |
| 11 | Surendra K. Gambhir | 1983 | A Case Study of Guyanese Bhojpuri and Standard Hindi | Trusties of Indian University on the behalf of Anthropological Linguistics | |

1.6.5.1.3 Linguistically oriented works

Udaya Narayan Tiwari has made his valuable contribution by exploring the history and geography of Bhojpuri along with the language description in terms of its phonology and morphology in his writing *Bhojpuri Bhasha aur Sahitya* and an exclusive trace of development of the language in his *The Origin and Development of Bhojpuri, 1960*. This book focuses on the phonological and morphological aspects of the language. The sound system of Bhojpuri, the change in sounds over a period of time, the retention and undergoing phonological alterations and emerging forms and different process of sound change from old to middle and from middle to new Indo Aryan is very well described by the author. Whereas the morphology section is a comparative study of word formation processes, in different ages. This includes the description of formative affixes, declension of nouns, and formation of different grammatical categories in varied dialects. This book also gives a clear picture of the Bhojpuri sound system and word formation process along with the sound changes along time.

Shukdev Singh's recent *Bhojpuri aur Hindi, 2009* covers all the three above mentioned categories. It is a fine grained comparison of sounds, sound changes, grammatical categories and their word forms and cases (Karakas) found in Bhojpuri and Hindi. The book also gives the description of related languages like Maithili, Magahi, Vajjika and Angika in a comparative analysis with Bhojpuri. And summarizes the literature from different genre like idioms and phrases used, terms from agriculture, *Biraha*, (*song of lamentation*) and *ghatahi boli* (the speech variety spoken around the ghats of Varanasi).

Kripa Shankar Singh's *Bhasha Vigyan aur Bhojpuri, 1973*, which is a modified version of his Ph.D thesis, gives the description of *Tagmemic principles* for Bhojpuri and Mahendra Nath Dubey's discussion on dialectal differences found in Bhojpuri spoken in the western, middle and eastern regions of Azamgarh district entitles *Bhojpuri is not Homogeneous even in a Single*

District, 1966 are fine examples of their contribution by presenting their Doctoral works, later on. Some other linguistic works in for of Ph.D thesis is presented in the table 5 below.

Table 5 Contributing Doctoral Theses

| S. No. | Authors | Year | Title | Publication |
|--------|--------------------------|---------|--|-------------------------------------|
| 1 | S Ojha | 1956-63 | Bhojpuri Lokoktiyo ke Sanskritik Paksh ka Adhyayan | Rachi |
| 2 | Ramnath Sharma | 1962 | Bhojpuri tatha Hindi Vibhakti-Prakriyaon ka Tulnatmak Adhyayan | Agra |
| 3 | Shaligram Shukla | 1960 | Bhojpuri Syntax | Cornell |
| 4 | Mahanth Mishra | 1972 | Bhojpuri Kriyapadon ka Vivranatmak Adhyayan | Patna |
| 5 | Arun Chaturvedi | 1979 | Bhojpuri Ka Kriya Sanrachana | Agra |
| 6 | Brija Nand Singh | 1971 | Bhojpuri Sangya Shabdō k a Gathnatmak ewam Arthparak Adhyayan | Jabalpur |
| 7 | Sarit Kishori Srivastava | 1995 | Varanasi ke sthan Namon ka Sanskritik Adhyayan | Vishwavidyalaya Prakashan, Varanasi |
| 8 | Vishwanath Prasad | 1950 | Early Phonetics and Phonology | University of London |
| 9 | A. B. Singh | 1960 | Study of the Speech of Awadhi and Bhojpuri Border Dialects | University of Allahabad |
| 10 | Mahendra Nath Dubey | 1966 | Bhojpuri is not Homogeneous even in a Single District | Banaras Hindu University |

1.6.5.2 Technological advancements

As already mentioned, technological developments for Bhojpuri are too few as mentioned below:

1.6.5.2.1 Opus Corpus

Opus is an acronym made out of open parallel corpus. it is an open source parallel corpus for different languages of the world developed by Open Linguistics Working Group¹³ (OWLG) as part of Linguistically/NLP-relevant dataset collection. It is a compilation of tools for automatic collecting, aligning and annotating online data. Some other related resources are corpus for medical documents, European constitution and parliament proceedings, open subtitles, KDEdoc and PHP monolingual corpus, Setimes and SPC parallel corpus as shown in figure 2 below.

The screenshot shows the 'Data and Resources' section of the Opus website. On the left, there is a 'Social' navigation bar with icons for Google+, Twitter, and Facebook. Below this is a section for 'OWLG' (Open Linguistics Working Group) with a description and a 'read more' link. The main content area is a list of data resources, each with a 'DATA' icon, a title, a brief description, and a 'More information' button. The resources listed are: EMEA - European Medicines Agency documents, EUconst - The European constitution, EUROPARL - European Parliament Proceedings, OO - the OpenOffice, OpenSubs - the opensubtitles, KDE4 - KDE4 localization files (v.2), KDEdoc - the KDE manual corpus, PHP - the PHP manual corpus, SETIMES - A parallel corpus of the Balkan languages, and SPC - Stockholm Parallel Corpora.

Figure 2 Opus data and resources¹⁴

The interface design gives information about the tools, links to those tools, sub-corpus and the browsing keywords in the very home page. The Bhojpuri corpus feed into the system has a total of two documents and the sentence count 19, as updated on its website.

¹³ <http://linguistics.okfn.org>

¹⁴ <http://datahub.io/dataset/opus>

OPUS ... the open parallel corpus

OPUS is a growing collection of translated texts from the web. In the OPUS project we try to convert and align free online data, to add linguistic annotation, and to provide the community with a publicly available parallel corpus. OPUS is based on open source products and the corpus is also delivered as an open content package. We used several tools to compile the current collection. All pre-processing is done automatically. No manual corrections have been carried out.

The OPUS collection is growing! Check this page from time to time to see new data arriving ...
Contributions are very welcome! Please contact <jorg.tiedemann@lingfil.uu.se >

Search & download resources: | |

Language resources: click on [tmx | mooses | xces | lang-id] to download the data! (raw = untokenized, true = truecasermodel, TM = phrase-based translation model)

| corpus | doc's | sent's | src tokens | trg tokens | XCES/XML | raw | TMX | Moses | mono | raw | true | TM | dic | freq | Browse Files |
|--------------|----------|-----------|-------------|--------------|---------------|-----------|-------|-----------|--------|--------|------|----|-----|------|----------------------------|
| Ubuntu | 2 | 19 | 0.1k | 0.1k | [xces bho en] | [bho en] | [tmx] | [mooses] | bho en | bho en | | | | | [sample] [xml bho][xml en] |
| total | 2 | 19 | 0.3k | 26.0k | | 19 | | 19 | | | | | | | |

Search & Browse

- OPUS multilingual search interface
- Europarl v7 search interface
- Europarl v3 search interface
- OpenSubtitles search interface
- EUconst search interface
- Word Alignment Database

Sub-corpora (downloads & infos):

- Books - A collection of translated literature (DOC)
- DGT - A collection of EU Translation Memories
- DOGC - Documents from the Catalan Government
- ECB - European Central Bank corpus
- EMEA - European Medicines Agency documents
- The EU bookshop corpus (EUbookshop/EUbooks)
- EUconst - The European constitution (EUconst0..)

Figure 3 Bhojpuri database¹⁵

1.6.5.2.2 Bhojpuri Tagset

The first Bhojpuri tagset was recently introduced in Language Resource and Evaluation conference, 2014. The paper entitled ‘*Annotating Bhojpuri Corpus using BIS Scheme*’¹⁶ (Singh & Banerjee, 2014) is a part of the present work and presents the first tagset for Bhojpuri. The tagset is based on the BIS¹⁷ scheme for Indian languages. The tagset is a two layered hierarchical tagset with 33 tags under 11 major categories. The language feature and the annotation challenges are discussed in the paper (More on this will be discussed in chapter 2).

Some more works in this domain is being done at MGIHU, Wardha and IIT, BHU. ‘*Initial Experiments in POS Tagging of Bhojpuri*’¹⁸ from IIT, BHU was presented at regICON: Regional Symposium on Natural Language Processing. The paper was an attempt of devising an SVM based POS tagger for Bhojpuri with the corpus length of 2000 sentences and 4880 word tokens. The average accuracy with this data was achieved around 65%. But the work for triggering the technological development of a language should be of more significant nature.

¹⁵ <http://opus.lingfil.uu.se/>

¹⁶ <http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-WILDRE%20Proceedings.pdf>

¹⁷ Bureau of Indian Standards, devised as part of ILCI project, at SCSS, JNU.

¹⁸ <http://regicon2015.sanchay.co.in/files/schedule.pdf>

1.6.6 Existing Corpora in Indian Languages

The Corpus technology in India started very late. Indian languages started gaining attention for their resource generation since the second half of the 20th century and another reason for this shift of focus was morphological equity of the languages. A number of significant works have been done in this sphere. There are programmes and consortiums which are constantly up to creating corpora and other value resources for different Indian languages from different domains and purpose. Some major projects are mentioned below.

1.6.6.1 EMILLE

Emille was the first corpus for Hindi with over 2 million texts from different genres developed by CIIL and CIIL-KHS-EMILLE¹⁹ combined was built under aegis. Some part of the corpus was also annotated following the older scheme of POS tagging but that was internal usage only.

1.6.6.2 Indian Languages Corpora, CIIL Mysore

Indian Languages Corpora project²⁰, under TDIL of the Ministry of communication and information technology, running at CIIL claims to have developed a 45 million word corpora for all Scheduled Indian Languages. The recently added languages like Bodo, Dogri and Maithili are also included. The corpus types are text, parallel and spoken. The ISCII encoded data has now been converted into UNICODE. Text corpus of eleven Indian languages is accessible on the project website and the CILL²¹ Spoken Corpus for lesser known languages has also been introduced by CIIL.

1.6.6.3 LDC-IL

The LDC-IL (Linguistic Data Consortium for Indian Languages) was set up on the lines of LDC at the University of Pennsylvania. Originally initiated by CIIL, Mysore in collaboration with other institutions like IISC Bangalore, IIT Bombay, IIT Madras, IIT Hyderabad etc. It was supported by government of India and sponsored by DIETY. LDC-IL was introduced to create and manage larger databases of Indian Languages and also conducting forums for researchers

¹⁹ <http://www.lancaster.ac.uk/fass/projects/corpus/emille/>

²⁰ <http://www.ciilcorpora.net/>

²¹ <http://www.ciil-spokencorpus.net/>

working on Indian Languages from all over the world. Speech recognition and synthesis, character recognition and corpora creation were anticipated to be the direct beneficiary of the project²². POS Annotated corpora based on LDC-IL POS tagset in 12 Indian Languages has been updated with an average of 8000 token tagged in each language. And 12+3 major and minor schedule language data is tagged following LDC-IL LEX tagset. They are also keen on segmenting and annotating corpus on ISL (Indian Sign Languages)²³.

1.6.6.4 ILCI

Indian Languages Corpora Initiative (ILCI) project is building parallel aligned corpora for major Indian languages. It is sponsored by Technology Development for Indian Languages (TDIL) program of Ministry of Communication and Information Technology (MCIT). The project is currently running in its second phase and is funded by the Department of Information Technology (DIT), Govt. of India.

It has eleven consortium partners including the host institute Jawaharlal Nehru University (JNU). ILCI²⁴ project is headed by currently running in its second phase where 5 new Indian Languages has been introduced to the pre-existing 17 Indian Languages. Along with corpora creation and annotation at POS level, it is now heading towards the chunking. The POS and Chunk annotations are done semi-automatically using server based online ILCIANN tool for POS tagging and Chunking each of the languages.

1.6.6.5 Joshua

University of Edinburgh, as part of JOSHUA (machine translation Toolkit), initiated the task of building corpus for 6 Indian languages in 2012. Bengali, Hindi, Malayalam, Tamil, Telugu and Urdu are the participating languages to this project. The idea is to construct parallel corpus for these languages via crowd source²⁵ and their translation.

²² <http://www.ldcil.org/default.aspx>

²³ <http://www.ldcil.org/workInProgress.aspx>

²⁴ For more details, visit <http://sanskrit.jnu.ac.in/ilci/index.jsp>

²⁵ <http://aclweb.org/anthology/W/W12/W12-3152.pdf>

1.6.7 Parts of Speech Annotation Scheme

Annotation schemes are the detailed guidelines of the tagset used for the annotation of the corpus. There is an attached list of different schemes which are in developed so far and are some are still being used.

1.6.7.1 Penn Tagset

Penn tagset is most popular tagset for English, based on tagset for Brown corpus. It was developed by Santorini in 1990. Other tagsets earlier to this has long list of tags like LOB (Lancaster-Oslo/Bergen) had around 135 tags, Lancaster URCEL about 165 tags, London Lund Corpus (a spoken English corpus) used 197 tags which are overpowered by Penn tagset with a total of 36 simplified tags with both lexical and syntactic information.

1.6.7.2 Annotation scheme for Indian Languages

1.6.7.2.1 IIIT tagset

The first IIIT tagset for Indian Languages was based in Penn Treebank tagset. This tagset has a total of 26 lexical categories as listed in guidelines for annotating Indian languages (Bharat et al., 2006 as cited in Choudhary, N., 2011) developed at IIIT Hyderabad through workshops on NLP.

1.6.7.2.2 TDIL program

TDIL was started by MCIT in 1991 for building technology solutions for Indian languages which aimed at developing information processing tool; facilitate human-machine interaction and creating multilingual technology resource. The project was expanded to different research institute for several purposes e.g. English to Indian language MT system ran at CDAC Pune and IIT Kanpur, IIIT Hyderabad look after Indian to Indian Languages machine translation, HCU for Sanskrit-Hindi MT system, handwriting recognition by IISc Bangalore, Cross-lingual information access by IIT Bombay and corpora development to JNU (see Jha, 2010).

1.6.7.2.3 ILMT

Sampark was developed with the combined efforts of 11 consortium member under *Indian language to India Language Machine translation* (ILMT) project funded by TDIL program of Dept of Information Technology, Govt. of India.²⁶ ILMT project has developed a multipart

²⁶ <http://sampark.iiit.ac.in/sampark/web/index.php/registration/validuser>

machine translation system for 18 language pairs from 9 Indian languages as ‘Hindi and Urdu over Punjabi, Telugu, Bengali, Tamil, Marathi, Kannada’ and ‘Tamil and Malayalam over Telugu’. For languages analysis Sampark makes use of Computational Paninian Grammar (CPG) and hybrid approach (both rule-based and dictionary-based) with statistical learning. It applies transfer method of MT (i.e. analyze- transfer-generate paradigm). Their latest versions SSF (Shakti Standard Format) will provide inter operability between heterogeneous modules of NLP to help debugging.

1.6.7.2.4 IMS’s Tree Tagger

This is a language-independent POS tagger. It is available free for academic use and comes with free language models for approximately 10 languages. It is not an open source and its assessment is relatively poor. This is a HMM tagger using decision trees for smoothing. The tagger was developed as part of the project Text corpora and Erschließungswerkzeuge (1993-1996) at the IMS (same project as Corpus Workbench, CWB/CQP) and has not changed substantially since then.

1.6.7.2.5 IL-POST

IL-POST (Indian Language – Part of Speech Tagset) is a hierarchical tagset, following EAGLES guidelines. EAGLES guideline is hierarchical in nature representing features and values of the grammatical categories. This tagset was developed by Microsoft Research India, Bangalore in 2008 in the course of developing annotated corpus for Indian Languages. Baskaran et.al. (2008).

1.6.7.2.6 BIS scheme

There was no consensus on the POS tagging schema for the Indian Languages till very late. Though some POS taggers were developed for languages like Shrivastava, M. & Bhattacharya, P., for Hindi (2008), Avinesh, PVS & G. Karthick for Telugu (2007), Dandpat, S. et.al. for Bengali (2007) etc (as mentioned in Choudhary N. and Jha G.N., 2011). ILMT (Indian Language Macune Translation) also claims to have developed different taggers but they was not a big success. BIS (Bureau of Indian Standards) and ILCI made the first national standard POS annotation. This hierarchical tagset coarsely allows only two levels of POS categories (sometimes to three levels, at max), the major categories and sub-categories. MSRI and ILMT categories are included to a great extent, in the tagset. It is a comprehensive tagset with no

morpho-syntactic features. It is capable of sharing, reusing and interchanging the linguistic resources and captures appropriate linguistic information (Gopal, 2012).

1.6.8 Existing POS Taggers for Indian Languages

1.6.8.1 Bengali

Bengali is said to have originated from the eastern variety of the Magadhi Apabhramsa. It is a member of Indo-Aryan language family. Like other Indian languages, Bengali is also morphologically very rich. The language ranks seventh in terms of popularity across the world and second in India. Bengali also enjoys the authority of being the national language of Bangladesh (Kumar, 2010). Kumar in his survey enlisted all four models, namely HMM and ME, SVM and CRF where the result for HMM-ME comparative research showed that supervised learning models gives the best performance. The SVM model demonstrated the overall accuracy of 86.8% applied on a standard test set of 20K word forms. And CRF tagger was tested with 72,341 and 20K word forms considering NER, Lexicon and unknown words, achieved accuracy of 90.3 %.

1.6.8.2 Hindi

Hindi is the official language of India with more than 182 million speakers as noted in Dalal et.al. It is a morphologically rich language. Different tagging approaches have been proposed for Hindi including Morphology driven tagger, Maximum Entropy (ME) based tagger, HMM based tagger and CRF based tagger. The average accuracy achieved by ME after the 4-fold, cross validation of the first model was 93.5%, the second model reached 88.04 % over around 10 runs. EI-HMM was better than simple HMM model as the accuracy percentage reached 93.12% . CRF using CRF ++ was 82.67% with training data of 21,470 words and 78.66 % with test data of 4924 words.

Another tagger for Hindi based on SVM is under development at ILCI, JNU. This tagger is, to my knowledge, being trained upon the ILCI Hindi corpus for Health and tourism domain and the accuracy of the tagger has been calculated 94.2 % for the Hindi data. The tagger can be found at

the ILCI website²⁷. The later section of the fourth chapter includes a section on the comparison of Hindi and Bhojpuri tagger. The Bhojpuri tagger is the one presented in the present work and the Hindi tagger is the same as mentioned here. More on this will be dealt in Chapter 4.

1.6.8.3 Malayalam

Malayalam is one of the important Dravidian languages in terms of its literary tradition. Malayalam has found its own script, a syllabic alphabet consisting of independent consonant and vowel graphemes plus diacritics.²⁸ It has its separate script including a syllabic alphabet, independent consonants and vowels and diacritic. Malayalam POS tagger has been devised using HMM and SVM models. HMM model gave accuracy of 90% with 80% of the sequences generated automatically and the accuracy of SVM increased from 86 to 89% percent with increase in the size of the lexicon to 180,000.

1.6.8.4 Punjabi

Punjabi language is a member of the Indo-Aryan family, spoken in India, Pakistan, USA, Canada, England, and other countries with Punjabi immigrants. It is the official language of Punjab, written in ‘Gurumukhi’ script in eastern Punjab and ‘Shahmukhi’ script in western Punjab. A Rule Based tagger has been developed for Punjabi which was further used for grammar checking. The tagset contained 630 fine grained tags 8-million words corpus of Punjabi collected from different registers. The accuracy was measured in percentages of words which are accurately tagged by the tagger which was 80.29% and 88.86 % respectively on including and excluding the unknown words (see Kumar et.al).

1.6.8.5 Sanskrit

Rule based and statistical taggers have been devised for Sanskrit. The tagset for rule-based tagger has a total of 13 tags with 65 word class, 43 feature sub-class, 25 punctuation tags and one UN for unknown words. This is recently developed by R. Chandrashekhar (2007) as a part of his Ph.D thesis and the credit for statistical tagger processing un-pre-processed Sanskrit text goes to Oliver Hellwig²⁹. This tagger contains a manually annotated corpus of currently about 1,50,000

²⁷ <http://sanskrit.jnu.ac.in/pos/index.jsp>

²⁸ <http://en.wikipedia.org/wiki/Malayalam>

²⁹ http://www.indsenz.com/int/index.php?content=sanskrit_tagger

words and employs a Markov model for tokenization and part-of-speech tagging on a freeware under a permissive license and standalone application.

1.6.8.6 Telugu

In language classification Telugu has its place in Dravidian Language. It is official language of Andhra Pradesh and uses many morphological processes to join words together to form complex words. Three POS tagging approaches (1) Rule based, Transformation based and Maximum entropy Model has already been applied for Telugu which achieved accuracy of 98 %, 90 % and 81.7 % respectively with error rate reduces by 3 % for machine language technique and 7% for Rule Based Telugu Tagger.

1.7 RELEVANCE OF THIS STUDY

At present, the technology for Indian languages is advancing with leaps and bounds. Several language resource programs are making effort for developing by-products with better performance and more convincing results. Resource development is in progress for almost all schedule languages and major languages of India. Unfortunately, Bhojpuri, despite being a major language of the India, is deprived in this regards. The present study aims at developing a language resource for Bhojpuri. The whole experiment has been conducted in three stages- Creating corpus for Bhojpuri, POS annotation of the Bhojpuri corpus and training the SVM based statistical tagger for Bhojpuri. The annotated corpus generated in this endeavour is a collection of 192,000 tokens from the general domain capturing six most popular sub domains and the tagger has been efficiently trained with an accuracy reaching 94.24 % for the general domain data.

This study is an initiative taken for settling up the technological ground for Bhojpuri and for its recognition, by developing resources in the language. The corpus built is likely to serve as base for many other resources and application waiting to launch.

2 CORPUS CREATION FOR BHOJPURI

2.1 RESEARCH METHODOLOGY

The study aims at generating the POS tagger for *Bhojpuri*, a lesser resource language. The experiment has been conducted in three consecutive stages namely

- a) Creating Corpus for Bhojpuri
- b) POS tagging/annotation of the Bhojpuri corpus developed as part of the study and
- c) Training of a statistical tagger for Bhojpuri modelled on Support Vector Machine

The present chapter is the description of the corpus created for Bhojpuri. This Corpus of a language serves as the database for the further linguistic processing and here, also form a ground for Parts of Speech annotation to proceed. This is the pathway to arrive at the answer to the very first research question as discussed in the earlier chapter. This chapter talks about the process of corpus building for Bhojpuri and its variants like data collection and cleaning, management and dissemination. The challenges met in validating the corpus of a language which is exposed to such platform for the first time is also dealt.

The present discussion will be carried forward in the next chapter which deals chiefly with the corpus annotation. Annotation scheme, language specific concerns and improvising, Bhojpuri tagset along with the extensive guideline for tagging the corpus will be focused upon. This section will explain the motivation behind the generated POS tagger.

2.2 BHOJPURI CORPUS

Corpus is an electronic collection of text data that is representative of language in its complete form. It is one of the primary resources of language because it is repository of the time, culture, tradition, people, occupation, and other peculiarities of the community through the language.

Bhojpuri, despite being widely spoken language, falls under less resourced languages. By the time, there are abundance web sources with Bhojpuri content and information but no full-fledged platform has been offered to bring all into a single whole. This language of millions of people does not have a single good application. Mishra (2003) also points out that “Bhojpuri differs

widely in pronunciation, grammar and vocabulary and does not have a well established orthography....It remains so far mostly unexplored and hence needs thorough exploration in respect of all aspects of its linguistics complexities.”

The creation of the corpus involves the following procedures as shown in the following figure 4.

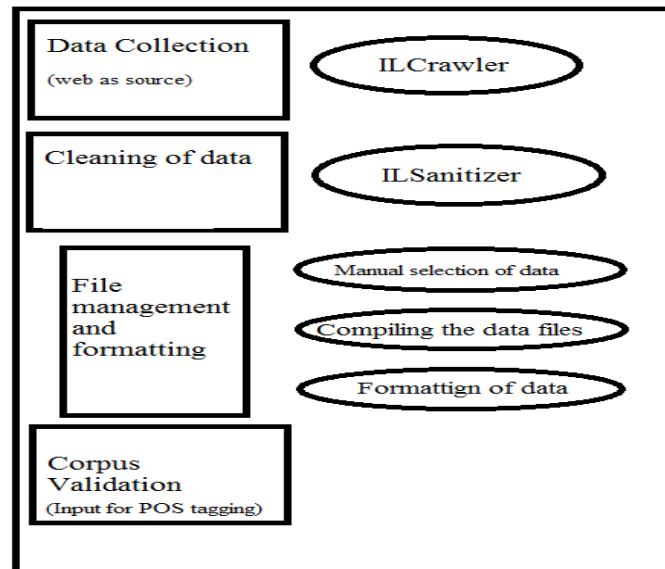


Figure 4 Procedure for creating the corpus

2.2.1 Corpus Data

Corpus is considered to be the basic building block for Language Technology. Being electronically retrievable, it becomes as easy source for the developer/researcher/practitioner to take this as base for further data processing. The Bhojpuri corpus, presented here, is a general domain monolingual annotated corpus built on the amalgamation of six sub-domains, namely blogs, entertainment, literature, politics, sports and miscellaneous. The sub-domains chosen qualify for both quality and quantity, considering the representativeness of the corpus. The length of corpus is 9,019 sentences with 1,69,275 word tokens. Each sub-domain contributes to the corpus, in the following ratio a shown in table 6:

Table 6 Corpus classification

| S. No | Sub-domains | No. of sentences | Total no. of word tokens |
|-------|-------------|------------------|--------------------------|
| 1 | Blogs | 3,597 | 60,814 |

| | | | |
|---|---------------|-------|----------|
| 2 | Entertainment | 2,569 | 54,369 |
| 3 | Literature | 272 | 5,493 |
| 4 | Politics | 961 | 18,971 |
| 5 | Sports | 19 | 443 |
| 6 | Miscellaneous | 1,601 | 29,185 |
| | Total | 9,019 | 1,69,275 |

From, table 6. it is evident that blogs and entertainment are the major contributing sections in the corpus and the sports is less opulent. The smaller sports data is due the availability of less sports related data on web and some parts of the sports data is also included under blogs and miscellaneous sections. Blogs and miscellaneous domains contain data from varied discipline irrespective of the domains specified above. These two sub-domains are shaped this way for the training purpose of the tagger. Broader the area of coverage of the training file, the better will be the performance of the tagger. These two files were initial inputs for training the model file. More on this will be discussed in another chapter on tagger.

2.2.2 Source of Data

The authenticity of the corpus also depends upon the source where the input data comes from. In this respect, the verifiability and representativeness of the corpus is also importance. Concerning this fact, websites for Bhojpuri newspapers, literary sites and only authentic reports were chosen as the source for retrieving data. The major websites accessed for the data crawling were:

- <http://www.thesundayindian.com/>
- <http://www.bhojpurika.com/>
- <http://trendsarrived.com/>
- <http://anjoria.wordpress.com/>
- <http://tatkhakhabar.com/>

Blogs and literary columns are also drawn from ‘<http://norivers.org/>’ which is a literary website. The list of the web links provided are tabulated below:

Table 7 Major web links for data extraction

| S. No. | Major Web links | Access Date |
|--------|---|-------------|
| 1 | http://www.thesundayindian.com/bh/category/4/ | 20150107 |
| 2 | http://www.thesundayindian.com/bh/category/खेल/4/1 | 20150107 |
| 3 | http://www.thesundayindian.com/bh/category/2/ | 20150107 |
| 4 | http://www.thesundayindian.com/bh/category/ | 20150107 |
| 5 | http://www.bhojpurika.com/category/concerns/राजनीति/ | 20150107 |
| 6 | http://www.thesundayindian.com/bh/blogs/ | 20150107 |
| 7 | http://www.bhojpurika.com/category/entertainment/ | 20150107 |
| 8 | http://bhojpurika.com/wp-content/uploads/2014/06/HelloBhojpuri-June14.pdf | 20150107 |
| 9 | http://www.bhojpurika.com/category/literature/ | 20150107 |
| 10 | http://www.bhojpurika.com/category/columns/ | 20150107 |
| 11 | http://www.bhojpurika.com/category/blogs/ | 20150107 |
| 12 | http://www.bhojpurika.com/category/concerns/ | 20150107 |
| 13 | http://www.facebook.com/ApniBhojpuri/posts/438608122886529 | 20150415 |
| 14 | http://www.facebook.com/permalink.php?id=331015513661458&story_fbid=641342285962111 | 20150415 |
| 15 | http://norivers.org/amazing/box/0SbjEn0vZ3k/.html | 20150415 |

| | | |
|----|---|--------------|
| 16 | http://trendsarrived.com/files/box/BPZwrf4M7Bk/तर-मई-क-द.html | 2015041 5 |
| 17 | http://www.facebook.com/JayBhojpuri/posts/682974015065210 | 2015041 5 |
| 18 | http://anjoria.wordpress.com/tag/कइसन-पियवा-के-चरित्तर-बा/ | 2015041 5 |
| 19 | http://anjoria.wordpress.com | 2015041 5 |
| 20 | http://tatkakhabar.com/?cat=5 | 2015041 5 |
| 21 | http://tatkakhabar.com/?s=जनवरी%2C+2015 | 2015041 5 |
| 22 | http://tatkakhabar.com/?s=मार्च%2C+2015 | 2015041 5 |
| 23 | http://tatkakhabar.com/?s=अप्रैल%2C+2015 | 2015041 5 |
| 24 | http://tatkakhabar.com/?s=दिसम्बर%2C+2014 | 2015041 5 |

Though the copying of data from web can become a major legal issue if redistributed any further same as the copyright issue as discussed in Hardie (2012, 57-60). But most of the data are freely available on internet are open source already and the author is being paid for every visit made on his page. Still there are legal concerns. One possibility is mentioned in Hardie is to use the data for research and development without sharing it with any other research group or publically, to avoid the breach of copyright. Similar approach has been incorporated for the present experiment.

2.2.3 Data Collection

The data for the present corpus has been borrowed from different Bhojpuri websites. The main reason for the web drawn corpus was to maintain the uniformity of the writing style. Though Bhojpuri does not have any standard dialect therefore different Bhojpuri writers do write in the variety used in their community. Web is found to be a platform where the uniformity is maintained to a large extent as the websites for news like 'tataka khabar' and the Sunday Indian

have such criteria that no other variety of speech should be included. The data was collected and cleaned for some noises using the tools namely- ILCrawler and ILSanitizer. The tools and their utility have been discussed in the sections below:

2.2.3.1 ILCrawler

Unicode has made the resources so easily and readily available that just some keywords serve with some hundreds of pages of related information. In this high tide of data, it is difficult to go through each source independently. Most part of this can be checked manually with huge cost and labour, certainly not all. Therefore, an application is designed to reduce the human pressure of copying, editing and formatting millions of data with just some mouse clicks, is called Crawler.

Crawler is data mining software which helps in extracting all the available text material of a language, from the web, and collecting them all at one place. ILCrawler is a java based application which is capable of digging out the contents from all the sub-links under the main webpage. ILCrawler and ILSanitizer used in this experiment were developed as part of computational linguistics research group at JNU

2.2.3.1.1 Design of Crawler

The crawler is divided into three major components:

- a) **Language** – This folder has two sub-sections, one for the specific language and its domain and another for the selection of language.

The first option is called ‘language -1’ (roughly) contains the list of codes of all the languages. The user needs to select the language of their choice. The second option is ‘language’ it has the information of the language for which the tool is being applied. The researcher/programmer has to go to select the respective language name for which the data has to be driven and the domain names into which the data has to be categorized. For example- In this case the language was Bhojpuri and the domain names contained all six aforementioned corpus domains.

- b) **Links** – This section contains the input web links for crawling or extracting data. The advantage of a crawler here is that the user does not require feed the links for every web

page, manually. The tool itself derives all the related and sub-branching links of the main webpage and extracts information from each of these sub-links.

- c) **Output** – The crawled data is stored in this section with the files named after the source link it was extracted from. This way the metadata information link source links, parent website etc are stored here, which can be used later on for keeping records.

2.2.3.1.2 Features of a ILCrawler

- a) It is a very useful technique for the practitioners and linguists because from huge pile up of different varieties of columns in different languages on a single page, it drives out only the data from the prescribed languages and relevant discipline (if configured as such).
- b) The script of a language is the key component to a crawler. This application is able to detect the language on the basis of its script. Suppose, the app is running on a Hindi webpage which has section heading in one language (suppose Hindi) and main heading or buttons in other language (perhaps, English) or its translation. The content of the whole page is more or less similar. This is why; the crawler is featured with the ability of differentiating among different scripts. From the same page, the tool will extract information's for Hindi from the text written in Devanagari and for English, by selecting the roman script only.

2.2.3.1.3 Advantages of ILCrawler

The crawler can be set up on a device/computer. The crawler is found to overpower the following problems of manual collection of data:

- a) The advantage of such software over manual work is great as it selects all and only data from the language concerned. There is very low (almost 0 %) chances of having a foreign language data or script, as it is an automated process.
- b) Crawling down the data from web is a very cost efficient process and saves a lot of researchers time spent over the extremely labour intensive manual collection (Choudhary, N., 2011).

- c) The data is collected at one place with each file named after the sublinks they belong to. These links serves as metadata information for future application and verification or documentation of the corpus see the figure 5 below.

| Name | Date modified | Type | Size |
|---|------------------|----------|-------|
| thesundayindian.com_5d519ec3-7184-4c92-a47c-e2985239b71_09-01-2015 | 09-01-2015 02:49 | TXT File | 20 KB |
| thesundayindian.com_0e8c5d82-b0c0-4e7e-9a95-3def038fc45e_09-01-2015 | 09-01-2015 02:51 | TXT File | 9 KB |
| thesundayindian.com_0f32b62d-edd9-4d83-ba56-4b479aacc47b_09-01-2015 | 09-01-2015 02:50 | TXT File | 5 KB |
| thesundayindian.com_1c32a33-2091-4c95-a17d-eeb1648d1ce4_09-01-2015 | 09-01-2015 02:50 | TXT File | 24 KB |
| thesundayindian.com_1d2d97b7-d4d3-40e3-ac9b-f3cf79a7b877_09-01-2015 | 09-01-2015 02:49 | TXT File | 15 KB |
| thesundayindian.com_1d9b550d-2015 | 09-01-2015 02:50 | TXT File | 10 KB |
| thesundayindian.com_2a7941f-b-2015 | 09-01-2015 02:51 | TXT File | 2 KB |
| thesundayindian.com_5b71be98-2015 | 09-01-2015 02:49 | TXT File | 4 KB |
| thesundayindian.com_06ef7ed2-2d99-421b-a86e-5743094cede5_09-01-2015 | 09-01-2015 02:50 | TXT File | 13 KB |
| thesundayindian.com_6d42aa4c-8ecf-454b-b4f5-d973af73da56_09-01-2015 | 09-01-2015 02:52 | TXT File | 2 KB |
| thesundayindian.com_7eb7d7a5-7dea-4c0f-824b-5e64891793f5_09-01-2015 | 09-01-2015 02:51 | TXT File | 8 KB |
| thesundayindian.com_8ade2656-171e-4181-ac54-971ed2875acb_09-01-2015 | 09-01-2015 02:50 | TXT File | 2 KB |
| thesundayindian.com_8ce25011-242b-45bd-83ee-cb7585ec53af_09-01-2015 | 09-01-2015 02:50 | TXT File | 15 KB |
| thesundayindian.com_9af65778-8ef0-49c1-8973-e66279367eed_09-01-2015 | 09-01-2015 02:49 | TXT File | 2 KB |
| thesundayindian.com_9c15851e-8a30-4194-bf09-07cedf523d91_09-01-2015 | 09-01-2015 02:50 | TXT File | 2 KB |
| thesundayindian.com_9ecd1413-2668-46e7-9260-f50d28d026b4_09-01-2015 | 09-01-2015 02:51 | TXT File | 20 KB |
| thesundayindian.com_18ace91d-2642-42db-af4b-1b12b850a819_09-01-2015 | 09-01-2015 02:50 | TXT File | 9 KB |
| thesundayindian.com_18e4527c-70ee-4a1e-b64b-bc439f7bb721_09-01-2015 | 09-01-2015 02:52 | TXT File | 2 KB |
| thesundayindian.com_25ad92d4-b134-4412-ad59-181c3bfddca_09-01-2015 | 09-01-2015 02:50 | TXT File | 11 KB |
| thesundayindian.com_28b2b23d-9866-40a9-b187-8789a189f94_09-01-2015 | 09-01-2015 02:51 | TXT File | 12 KB |
| thesundayindian.com_28ca7dde-1675-42cc-818b-e5ades7e2a16_09-01-2015 | 09-01-2015 02:48 | TXT File | 13 KB |
| thesundayindian.com_31c1ca1c-8c13-4817-8635-da66d2e2c831_09-01-2015 | 09-01-2015 02:52 | TXT File | 3 KB |
| thesundayindian.com_40a25379-7b37-4e23-b84f-344c0156b018_09-01-2015 | 09-01-2015 02:52 | TXT File | 4 KB |
| thesundayindian.com_44eafe46-dfad-4446-a62e-778293e14316_09-01-2015 | 09-01-2015 02:50 | TXT File | 19 KB |
| thesundayindian.com_0628cd8b-248b-4586-b3b9-17805cc3099b_09-01-2015 | 09-01-2015 02:49 | TXT File | 16 KB |
| thesundayindian.com_74e79a4-d988-4cd7-b113-80e90f9979a_09-01-2015 | 09-01-2015 02:50 | TXT File | 82 KB |
| thesundayindian.com_0079f580-5088-4dd1-e14e-f58b20ee55cd_09-01-2015 | 09-01-2015 02:50 | TXT File | 13 KB |

Figure 5 Screen shot of the crawled data

Fig. 5 captures the image of the crawled data from ‘the Sunday Indian’ website. The leftmost column lists the web links containing all politics data and the column next to it mentions the date on which it has been collected. By giving the link from the main page (let say, link to the heading ‘politics’ on the Sunday Indian page) will make the crawler visit each and every sub-links inside the given link. These internally processed sub links are collected as separate files with original content of the page. This data is further sent for the cleaning process.

2.2.3.1.4 Restrictions of ILCrawler

Though the present crawler is one of the initial applications developed and employed by the research group, it has some restrictions which are being worked upon. These are:

- a) There are about 100 languages spoken in India, from among which 22 are listed as schedule languages in the Indian constitution and 12 scripts for all (See table 8). The first ten languages in the table shares one common script i.e. Devanagari, the next three are written in Bengali and last two have adopted Perso-Arabic. Only seven languages from serial 14 to 20 have their separate writing system. Identification of the language only on the basis of its script is quite helpful for languages with different scripts (like 14-20 i.e. Gujarati, Kannada, Malayalam, Oriya, Punjabi, Tamil and Telugu). For the languages sharing common script, if there is a foreign language data using same script is present within the text, the crawler will not be able to differentiate between both the languages. Such examples are found in this experiment also where there are several occurrences of Hindi and Sanskrit sentences; even Marathi (a few sentences) was found within the blog and miscellaneous domain. More on this is dealt under the validation issues.

Table 8 Schedule languages of India

| Sl. No. | Language | Script |
|---------|-----------|----------------------|
| 1. | Hindi | Devanagari |
| 2. | Sanskrit | Devanagari |
| 3. | Marathi | Devanagari |
| 4. | Konkani | Devanagari |
| 5. | Nepali | Devanagari |
| 6. | Maithili | Devanagari |
| 7. | Sindhi | Devanagari |
| 8. | Bodo | Devanagari |
| 9. | Dogri | Devanagari |
| 10. | Santhali | Devanagari, Ol Chiki |
| 11. | Bengali | Bengali |
| 12. | Assamese | Bengali |
| 13. | Manipuri | Bengali, Meithei |
| 14. | Gujarati | Gujarati |
| 15. | Kannada | Kannada |
| 16. | Malayalam | Malayalam |
| 17. | Oriya | Oriya |
| 18. | Punjabi | Gurmukhi |
| 19. | Tamil | Tamil |
| 20. | Telugu | Telugu |
| 21. | Urdu | Perso-Arabic |
| 22. | Kashmiri | Perso-Arabic |

These sentences had to be manually removed from the corpus during the validation process as it cannot be checked by the automatic cleaner. Some set of files has considerable number of such sentences, although the ratio of their overall occurrence is not very high (discussed in detail under corpus validation).

- b) In contradiction to the above section, the language data available in roman script, like blogs or poetry, cannot be identified by the crawler. That is a difficult task for even for a linguist to include it in the data because then the whole text is to be transcribed first, into the original script and then bring into use.
- c) The crawler is restricted to look for all and only Devanagari alphabets, which makes it to exclude some useful information like dates, numbers (number of session, money amount, and even some relevant information like the medicine or herb names, names of some event or process written in Roman for the focus of the reader.

2.2.4 Corpus Cleaning

Cleaning is the process of reconstructing the hampered text. This includes all the possible discrepancies found in a running text while copying it down like spelling error, redundancy, omitted letters or word(s), improper sentence break etc. It also includes the error related to grammar and other languages, for example, in many cases the online Hindi data is found with period (.) as the sentence boundary marker. Corpus cleaning also involves the fill ups of the missed and omitted elements from the primary data source. The complete check of such noises and reducing them to the least is the objective of cleaning, which is a necessary component of building an error free corpus.

Data sanitation and selection are the two major processes of cleaning a corpus. The

2.2.4.1 ILSanitizer

The sanitation is performed automatically with the help of ILSanitizer. As already mentioned above, Sanitizer is a program developed for the automatic checking of noises found in the data collected from the web via. Crawler. The ILSanitizer has three main functions:

- a) Compiling the data files
- b) Data Cleaning
- c) Shuffling of the data

2.2.5 Data Management and File Format

The crawled data has been compiled into files with a maximum limit of one thousand sentences. The ILCI file format was followed which contains two columns of Sentence ID and the sentences. Sentence ID is a unique index for every sentence which follows a particular naming

convention of each sentence ID. It includes the first three letters from the language name followed by the serial no. beginning with the set number to which the file belongs. For example, the first sentence of the third set for Bhojpuri will have the sentence ID ‘BHO3001’ where *BHO* stands for Bhojpuri, 3 stands for the third set with sentences ranging between 3000-3999 followed by the serial no. of sentence in that file. The file is converted to UTF 8 format.

2.2.5.1.1 Compiling the data files

The crawled data is list of scattered and non-uniform data of varied length. The Sanitizer is programmed such that it is able to take all the data from each domain and convert them into separate files. In this process the Sanitizer also breaks each sentence into a new line and decides the length of each file in word tokens.

Each file was estimated to have 1000 sentences each for the ease of manual selection and restriction process and the average word length of sentences in the corpus was found to be 16. Therefore, the sanitizer was set to compile files of 100,000 word tokens each.

Table 9 Total files compiled for the present corpus

| S. No. | File names | No. of Sentences | No. of words/tokens |
|--------|-------------|------------------|---------------------|
| 1 | Blogs_set 1 | 917 | 15,192 |
| 2 | Blogs_set 2 | 936 | 15,200 |
| 3 | Blogs_set 3 | 933 | 16,032 |
| 4 | Blogs_set 4 | 811 | 14,390 |
| 5 | Ent_set 1 | 795 | 16,059 |
| 6 | Ent_set 2 | 884 | 17,522 |
| 7 | Ent_set 3 | 861 | 18,400 |
| 8 | Ent_set 4 | 29 | 2,388 |
| 9 | Lit_set 1 | 272 | 5,453 |
| 10 | Pol_set 1 | 961 | 18,971 |
| 11 | Spo_set 1 | 19 | 443 |
| 12 | Misc_set 1 | 1,066 | 19,322 |
| 13 | Misc_set 2 | 535 | 9,863 |
| | Total | 9,019 | 16,9275 |

From table 9, the total no. of files compiled were 13 with four sets from blogs (all types of data) and entertainment, two sets from miscellaneous (collective data of all genre, including technical sphere) and one set from politics, literature and sports. The number of sentences and word length of each file is same as presented in the table above.

2.2.5.1.2 Data Cleaning

In general, cleaning of data involves elimination, assimilation and edition in the text under process. Here, in terms of Sanitizer, cleaning is restricted to the elimination of the undesired elements. In the process of crawling, sometimes fragments of a sentences written in another script are left, the sanitizer will eliminate put those fragments and other leftovers like meaningless words, scattered letters, and non grammatical sentences.

There are cases when same input link is provided twice or the data is overwritten over the previous output. In such cases, the chances of duplicity are high. The sanitizer looks after this and efficiently checks for the duplicity by removing duplicate sentences.

2.2.5.1.3 Features of Sanitizer

- a) ILSantizer is tested for eliminating any sentence found with three or less than three words with random selection.
- b) It does not allow any scattered or left-over material to dwell in the corpus, any further.
- c) Sanitizer is very helpful in separating all the sentences from a paragraph into new lines.
- d) The metadata information can also be stored along with the text as the sanitizer offers to keep or remove the source link for each sentence.
- e) It also helps deciding the maximum length o the sentences. This is not applicable for the present corpus as the sentences are manually edited for its length and fixed to an average of 16 tokens per sentence.

2.2.5.1.4 Shuffling of the data

All the three phenomenon from compiling to cleaning and shuffling are internal process of the sanitizer. Unlike compilation, data shuffling is an optional feature, this purely depends upon the requirement of the researcher and the nature of the corpus, whether the data is to be preserved in actual order or should be reordered. The shuffling is simply the random ordering of the data within or across files.

2.3 FEATURES OF BHOJPURI

3.1 Ergativity

Like Awadhi, Bhojpuri is also an ergative less language. The ergative case marker (e.g. *ne* in Hindi) is not present in Bhojpuri. The absence of this feature is replaced by the contrasting perfective aspect of the sentence which makes it grammatically correct. This construction is quite different from Hindi as Hindi makes use of adjectives. In some dialects of Bhojpuri the Perfective is marked with *-l* or *-les* suffixes. Following are examples :

[1] गनेसू कहलेस हम सब काम करब³⁰

Gloss: ganesu kəhəles həmə səbə kamə kərəbə

Translation: Ganesu said that he would do everything.

3.4 Determiners

A determiner is one that determines a quantity of a noun. Determiners are similar to the classifiers in function but the main difference is that in Bhojpuri the determiners and classifiers determine the number of nouns and numerals respectively. This unique feature is not found in Hindi. It is basically a discourse particle which comes automatically, out of the speaker, with narration something in flow.

Maithili, Awadhi, and Bengali also make use of determiners. Determiners might found in Hindi spoken in contact areas of the Bhojpuri speaking region. Determiners can be attested in Bhojpuri with all nouns except with honorific. This also justifies that there are two sets of reference, one honorific which does not make use of determiner at all, spoken either in formal speech and to the elders of the family and second, with determiners, which can be used with friends, among members of same age group and is basically informal.

Terms of address in Bhojpuri are rich in determiner use. These words are glued with the host nouns, same as an emphatic which does not allow it to make a separate tag level category in the

³⁰ This example has been adopted from Singh and Banerjee, 2014.

tagset. The use of determiners as explained by Kachru, (2006) also applies to the syntax of Bhojpuri, as shown in the given table 10:

Table 10 Use of Determiners

| Word final sounds | -ə/-a | -i/-i | -u |
|-------------------|-------|-------|----|
| Determiners | -va | -ja | -a |

The above table shows that the words ending are the responsible carriers of determiners. For example, if a word ends with the final sound –ə or –a will take –va suffix as the determiner, words with –i/-i ending will take –ja and words with final sound –u will take an –a suffix. Similar constructions are found in Maithili, Magahi, Awadhi and other related languages.(Kachru, 1980) Generally, such constructions determine which noun is talked about as inferred from the examples below:

[2] भर्बित्तन भाई रहत त अमरूद खूब तोड़ –तोड़ के खियावत³¹

Gloss: b^hərəbittənə b^hai rəhət t̃ əmərud̃ə x̃ub̃ t̃oḍ̃ –t̃oḍ̃ ke k^hiavət̃

Translation: If Bharbittan would be there, he would have offered us with abundance of guavas.

2.3.1 Issues in data collection

2.3.1.1 Data from Other Languages

The ILCrawler is restricted to the extraction of Devanagari script only. This facilitates the linguist to directly visit the desired website and crawl down the required data. Any extra or unnecessary information, like advertisements on the page, descriptions, keyword, date etc are not copied. But it is not able to restrict the extraction of other languages written in Devanagari. The statements or speech made in Sanskrit, exemplified under the Bhojpuri blog, will be crawled as part of the Bhojpuri text. These sentences are eliminated from the corpus while validation for the string of words is unknown to the tagger. Some similar examples are listed in the table 11 below:

³¹This example has been adopted from Singh and Banerjee, 2014.

Table 11 List of sentences from other languages, written in Devanagari

| S. No. | Data from other languages | Languages |
|--------|---|----------------|
| 1 | इमे भोजा अंगिरसो विरूपा दिवस्पुत्रासो असुरस्य वीरा | Sanskrit |
| 2 | श्रीकाशी विश्वनाथो विजयतेतराम् | Sanskrit |
| 3 | श्वेते वृषे समारूढा श्वेताम्बरधरा शुचिः महागौरी शुभं दद्यान्महादेवप्रमोददा | Sanskrit |
| 4 | चन्द्रहासोज्ज्वलकरा शार्दूलवरवाहना कात्यायनी शुभं दद्याद्देवी दानवघातिनी | Sanskrit |
| 5 | एकवेणी जपाकर्णपूरा नग्रा खरास्थिता लम्बोष्ठी कर्णिकाकर्णी तैलाभ्यक्तशरीरिणी वामपादोल्लसल्लोहलताकण्टकभूषणा | Sanskrit |
| 6 | वर्धनमूर्धध्वजा कृष्णा कालरात्रिर्भयङ्करी | Sanskrit |
| 7 | सिंहासनगता नित्यं पद्माश्रितकरद्वया शुभदास्तु | Sanskrit |
| 8 | त्वदीयं वस्तु गोविन्दम् तुभ्यमेव समर्पयामि | Sanskrit |
| 9 | लास्ट सेवन डेज | English |
| 10 | आल इज वेल | English |
| 11 | टेकेन फार ग्रान्टेड | English |
| 12 | मुंबईतील बॉम्बस्फोट विश्वविजेत्यांचा सन्मान अण्णा हजारेंचा | Other language |
| 13 | ओ षु स्वसार कारवे श्रृणोते ययां वौ दूरादनसा रथेन | Other language |
| 14 | आणखी वाचा सविस्तर बेकायदेशीर खाणकाम वरदान की शाप आर्थिक विकासाच्या आंधळ्या स्पर्धेत भारताने आपल्या विशाल नैसर्गिक साधनसंपत्तीला दुर्लक्षित करणे सुरू केले आहे । | Other language |

| | | |
|----|--|----------------|
| 15 | केवळ आठ वर्षांच्या काळात हिरवीगार शेते असलेल्या जमिनी उघड्याबोडक्या झाल्या । | Other language |
| 16 | मैने बहुत सी डायलाक कहानी लिख कर रखी है | Hindi |
| 17 | रक्तभूमि जिसके मुख्य कलाकार सुपर स्टार रवि किशन है बन कर तैयार है और यह कुछ ही दिनों में रुपहले पर्दे पर दर्शकों का मनोरंजन करने | Hindi |
| 18 | मैने बहुत सी डायलाक कहानी लिख कर रखी है | Hindi |
| 19 | रक्तभूमि जिसके मुख्य कलाकार सुपर स्टार रवि किशन है बन कर तैयार है और यह कुछ ही दिनों में रुपहले पर्दे पर दर्शकों का मनोरंजन करने | Hindi |
| 20 | सुक्रीत फिल्मस एवं जन एकता फिल्मस के बैनर तले बनीइस भोजपुरी फिल्म के निर्माता कृष्णा यादव रवि किशन ला रहें हैं | Hindi |
| 21 | रक्तभूमि एक्शन रोमांस के साथ थ्रिलर का चस्का वाली फिल्म | Hindi |
| 22 | अदालत ने कहा कि अगर कोई पूरी जानकारी के साथ इस्लाम और कुरान में विश्वास करता है तो उसका धर्म बदलना समझा जा सकता है | Hindi |

In the table given above, the first eight sentences are examples of Sanskrit data in the corpus, the next three are English sentences written in Devanagari, sentences from 12 to 15 are from some language other than Hindi, Bhojpuri and Sanskrit and sentences from 16 to 22 are the examples of Hindi data as found in the corpus. A detail list of sentences from other languages found in the corpus has been added in the appendices at the end.

2.3.1.2 Header on the webpage

The Title, sub-title and headers are all clubbed together while crawling. The section headers given at the top of the web page are merged within the text and as the crawler move from one page to another within a given link; these headers are copied every time along with a new text

page. Some are separate strings whereas others are in continuation with the text which is difficult to find and remove. Other description like title of the text, publishing date and author are also copied with the text or sometimes, with the header itself. There is a list of crawled sections headers and titles given in the table 12 below:

Table 12 List of Section headers found in the data

| S No. | Example | Description |
|-------|---|-------------|
| 1 | उत्तर प्रदेश बिहार राजनीति देशदुनिया खेल व्यापार अपराध मनोरंजन शिक्षा फर्जी यूनिवर्सिटी खबर भोजपुरी में मंगल जनवरी | Headers |
| 2 | भोजपुरी का डॉट कॉम खबर पत्रिका किताब आ वेबसाइट्स भोजपुरी वेबसाइट्स ब्लॉग एक तरफा विचार चौपाल निजी नीकजबून मनोरंजन गीतगवनई टीवी फिल्म रंगमंच सरोकार कोर्स आ कैरियर देश आ समाज नारीजगत पर्व त्योहार भोजपुरिया लाल योग राजनीति सभा समारोह स्वास्थ्य साहित्य उपन्यास कविता कहानी निबन्ध पुस्तक चर्चा भाषा समीक्षा स्तम्भ अगड़म बगड़म कतरब्योत कोलकाता मेल ज्योतिष आ वास्तु बतकुञ्चन भउजी हो भोला बाबू रामझरोखा से लस्टम पस्टम कार्टून कोना सभकर राय । | Headers |
| 3 | टटका खबर उत्तर प्रदेश बिहार राजनीति देशदुनिया खेल व्यापार अपराध मनोरंजन शिक्षा फर्जी यूनिवर्सिटी खबर भोजपुरी में शुक जनवरी बियफे | Headers |
| 4 | देशदुनिया खबर भोजपुरी में गाजीपुर चंदौली जौनपुर बलिया रोहतास सारण खबर भोजपुरी में सोमार दिसंबर अतवार | Headers |
| 5 | साहित्य भाषा टी॰वी॰ सरोकार सिनेमा योग संगीत गीत गवनई भउजी हो चर्चा बा लस्टम पस्टम अगड़म बगड़म भोला बाबू मारीशस से । | Headers |

| | | |
|----|---|---------|
| 6 | देशदुनिया खबर भोजपुरी में खबर भोजपुरी में अतवार दिसंबर शनिचर दिसंबर के खबर | Headers |
| 7 | कात्यायनी पर्व त्योहार | Title |
| 8 | बतकुञ्चन बतकुञ्चन | Title |
| 9 | औरत आ जुआ | Title |
| 10 | भोजपुरी वेबसाइट्स | Title |

The first six sentences are the examples of the section headers crawled by the crawler in a linear fashion, which gives no sense if read as a sentence. The next four are example of titles of the columns from which the data has been extracted, the titles has no proper grammatical structure and hence, has been eliminated from the corpus. A list of varieties of headers found in the data has been included in the appendices.

2.3.1.3 Some other common Crawler errors

The crawled data is short of the periods in abbreviation and the comma between the serial nouns is missing. There is more than one space found in the corpus which might be the result of change in the format of data saved on web and in the corpus. Some examples of these types are listed in the table 13 below:

Table 13 List of other common crawler errors

| S No. | Error Type | Description | Examples | Gloss |
|-------|----------------|--|----------|---------------------|
| 1 | Multiple space | More than one space between two words | | |
| 2 | Punctuations | Comma and periods missing between all multiple entities and abbreviations respectively | यू पी | U P (Uttar Pradesh) |

| | | | | |
|---|----------|--|---------------------------------|------------------------------------|
| 3 | Numerals | All Arabic numerals like 1, 2, 3... and numeral as part of all ordinal 'th' (before 'वां' in the corpus) like '52वां', '47वां' etc. Are missing | ऋग्वेद के वां स्तोत्र के प्रमाण | proof of the the source of Rigveda |
| | | | मैं साल का हूँ | I am _ years old |

In the above table, three types of crawler errors are mentioned. This includes the problem of multiple spacing between two words in a string, the issue with the punctuation marker, as the crawler does not account for the periods in abbreviations and the commas are also omitted by it. The third is the omission of numerals, if written in roman numbers, at some places they are missing while the numerals are randomly present in the corpus.

2.3.2 Corpus Validation

Once the data has been cleaned by passing through the Sanitizer much of the problems are resolved like fragments, shuffling etc. But for authenticating the corpus the proper validation is necessary. The corpus validation is a manual task where each sentence is revisited for error detection and data authenticity. The ILCI tool used for the validation can be found on <http://sanskrit.jnu.as.in/ilciann/index/jsp>

2.3.2.1 Validation Challenges

After validation the corpus accounts for the following challenges:

2.3.2.1.1 Genre specific sentences

Corpus is collection of natural language data but there are some limitations to it. A lot of drops and ellipses are found in the natural language use which cannot be supported to include as it is in the corpus unless it is a dialogue. Similarly, different form of writing employ different techniques like reporting sentence in news and metaphorical in poetry. The line of poetry, though has complete meaning, are not always complete. The news headlines are also far from the language in use. Considering these, the text for corpus must be complete sentences. Therefore, genre specific sentences which might bring ambiguity must be avoided or edited if included. In

the present corpus, almost all the songs and poetries are separated out and the reporting sentences has been edited to the declarative ones. Some of the discrepancies are shown in table 14 below:

Table 14 Discrepancy with genre specific sentences

| Example | Gloss | Error type |
|--|--|--------------|
| 1. बोलेला देख के कजरा | Your eyes speaks | poetry lines |
| 2. बोलेला देख के कजरा बोले होठ लाली | Your eyes and the redness of your lips says | poetry lines |
| 3. रानी हो डोली तोहर हमरे घर आई | You will be my bride | poetry lines |

2.3.2.1.2 Repetitions

There are many instances of repetitive words or phrases. These are sometimes part of the sentence as misprinting or typo error and other times it is due to the linear crawling which brought the section header and title together in due process. Some of the examples from the same are given in the table 15 below:

Table 15 Examples of repeated phrases/words

| Example | Gloss | Error type |
|---|--|-------------------|
| कब का | long ago | repeated phrase |
| नरेन्द्र मोदी | Narendra Modi | repeated phrase |
| फिल्म की की शूटिंग | The shooting of the movie | repeated word |
| फिल्म भाभी गंगा पार का म्यूजिक रिलीज | Music release of the movie 'Bhabhi Ganga Par' | repeated sentence |

The first row in the above table exemplifies the phrase which has been replicated within a sentence; next two rows are examples of repeated words like *Narendra Modi* and *KI*. These are

clear case of typo based errors and the last one is due to the crawling fashion that the title of the article has been included into the text itself.

2.3.2.1.3 Unaccepted terminology

In this regard, Mishra (2003) says that the speech behaviour, decency in speech and correct language use are necessary elements for the standardization of a language. Though slangs they are important feature of a language which describes the social status of the language, should be uniformly rejected if makes language a ‘bad Bhojpuri’. The slangs like भँडुआ (bʰəɽua), हरामजादा (həraməɟaða) etc are not included in the corpus. Table six contains some examples for the same.

Table 16 List of unaccepted terminology/ slangs

| Example | Gloss | Error type |
|---|--|--------------------------------------|
| ऊ ससुराछाती ठोक ठोक के अपना के हरामजादा साबित करे में लागल रहुवे। | He is crazily proving himself a moron | slangs and unaccepted terminology |
| बापो | Father | slangs and unaccepted terminology |

2.3.2.1.4 Space missing between words or collocations

The crawled data has another discrepancy with the word spacing. There are several instances where the space between two or more words is missing. Such error pattern cannot be rectified by the tool and hence making the manual labour even intense. There are some examples from the corpus with similar case as shown in table 17.

Table 17 Table listing the words/ phrases without space in between

| Example | Gloss |
|---|-----------|
| तीजतेवहार रोपनीसोहनी सुनर सुनर गीत संगीतन (tjɽevəhar ropənisoɦni sunər ɡit saŋɡitən) | Festivals |
| ढेरढेर (dʰerdʰer) | a lot |

| | |
|---|---|
| घरवापसी (g ^h ərəvəpəsi) | coming back home |
| पढललिखल (pəṭ ^h əllik ^h əl) | Learnt |
| देखरेख (dek ^h rek ^h) | care taking |
| डब्ल्यू॰डब्ल्यू॰ई॰ (dəblju dəblju i) | D.L.W. |
| हंसीखुशी (həsik ^h ʊʃi) | Happily |
| हैंपर (hɛpər) | is but |
| किकागज़ (kikagəz) | that (subordinator) paper |
| र हल ना चलेलानायिका कुमकुमको इ (rə həl na çələlanajika kumkumko i) | |
| अइसनवइसन (əisənəvəisən) | Ordinary |
| प्रेमीप्रेमिका (premi ^h premi ^h ka) | Lovers |
| परहमनी (pərhəməni) | but I |
| चिट्ठीपत्री (ʃit ^h ṭ ^h i ^h pətri) | Letters |
| जूझतजागत ठोकतेठावत (dʒudʒ ^h əʃdʒagət ^h t ^h ökəṭ ^h t ^h əvət ^h) | Struggling |
| एक दृश्य हैगाँव (ekə d ^h r ^h ʃjə hɛgāvə) | The village is a scene |
| लिखलापढला (lik ^h əlapəṭ ^h əla) | Educated |
| रचनाप्रक्रिया एक डिसिप्लिन की मांग करती हैजिसमें कहानी (kəhani) | art of creativity; in which |
| गीतसंगीतअभिनयसम्पादनसिनेमेटोग्राफी (gīt ^h səŋgīt ^h əb ^h inəjəsəmpaḍənəsine ^h mətograp ^h i) | song, music, performance, publication, cinematography |
| साथेसाथे (saṭ ^h esaṭ ^h ə) | Along |
| आमनेसामने (amənesaməne) | in front |
| बरनाठझरनाठ (bərənəṭ ^h dʒ ^h ərənəṭ ^h) | Manner |
| कोशिशबहस (koʃiʃbəhəs) | effort and argument |
| बाउर (baorə) | auxiliary verb and coordinator |
| तिलकदहेज़ (ṭiləkəḍ ^h hedʒə) | custom and dowry |

On the contrary, there are also sentences which are found to have multiple spacing between two words/ tokens.

2.3.2.1.5 Typing errors

A large number of mistakes in the typing have been encountered throughout the data. The letters and diacritics are found both misplaced as well as reduplicated. A list of such words is presented in the table below:

Table 18 List of words with typing mistakes

| Example | Gloss |
|---|-------------------|
| देिने (ðine) | Day |
| फुुआ (p ^h əgʊɑ) | FaguA (a sason) |
| अमृृतेश(əmrɪtɛʃə) | Amritesh |
| वॉल्यूम (vɔljumə) | Volume |
| मेंं (mē) | in (postposition) |
| काँपी (kɔpi) | Copy |
| डाॅट काॅम (dɔt kɔm) | dot com |
| उँँचा (ũtʃɑ) | High |
| काॅलोनी (kɔloni) | Colony |
| अभिजीत दत्ता राॅय (ab ^h ɪdʒɪt ðəttɑ rɔjə) | Abhijit Rai Dutta |
| काॅंग्रेसच्या (kɔŋɡresʃjɑ) | Congress |

2.3.2.1.6 Word fragments

This has already discussed in brief in the crawler based issues that the web drawn data has many fragmented tokens which are present in the sentences at some or the other position like the *cI* (first example from the table below) was found in the middle of the sentence and *DxA* (fifth row) was present at the end. This might be names of the authors which are incorrectly separated as they stand for ‘Mister’ and ‘Doctor’ respectively. Others are names of places, authors and dates. Such tokens are erroneous and hence have been eliminated from the corpus.

Table 19 Fragmented words in the corpus

| Example | Gloss |
|-----------------------------|---------|
| चि (tʃi) | Mr. |
| आ (ɑ) | And |
| लखनऊ (lɔk ^h nəu) | Lucknow |

| | |
|---|-----------------------|
| रामचंद्र यादव (raməʃənðrə jaðəvə) | Ramchandra Yadav |
| डॉ (dɔ) | Dr. |
| गंगा देवी (gəŋga ðevi) | Ganga Devi |
| राउर ओम (raurə omə) | Yours Om |
| शशिकांत सिंह (ʃəʃikant̪ siŋgʰ) | Shashikant Singh |
| फाग गीत २ फाग गीत १ (pʰagə giṭ 2 pʰagə giṭ 1) | Fag song 1 Fag song 2 |
| नई दिल्ली (nəji ðilli) | New Delhi |
| प्रशांत निशांत (niʃant̪ prəʃant̪) | Prashant Nishant |
| फरवरी (fərəvəri) | February |
| अप्रैल (əprelə) | April |
| नीतीश सितंबर (niṭiʃə siṭəmbərə) | Nitish September |
| कॉम (kɔm) | Com |

2.3.2.1.7 Errors in Hindi Sentences

There are high chances of having Hindi sentences in the corpus as it is also written in Devanagari. According to the suitability of context writer easily switch from one language to another as a result of which there is an average of 30 Hindi sentences found in the corpus per file (1,000 sentences approx). The code mixing and code switching between Bhojpuri and Hindi is very common. Words like time, launch, release, blackmail, show and music etc are loan words in Bhojpuri and used very frequently in blogs and narratives.

2.3.2.1.8 Some other errors

Apart from these there are some other errors types found in the corpus as given in the table 20 below.

Table 20 List of miscellaneous errors

| S No. | Examples | Error type | Action taken |
|-------|---------------------------------|--------------------------------------|-------------------------|
| 1 | यू पी (ju pi) | Abbreviation | space removed |
| 2 | हम जब आवता था नू तो देखता था कि | Mixed construction (Bhojpuri> Hindi) | tagged according to the |

| | | |
|---|--|---------------------------------|
| <p>साहिब के मेमिन फटर फटर बतियावता था रंगरेजी में से हमहूँ थोड़ बहुत सीखने लगा आ कुछ सीखियो गया हूँ । (həmə dʒəb avəʈa t̪ʰa nu t̪o ðekH əʈa t̪ʰa k ɪ saɦibə ke meminə pʰəʈər pʰəʈər bəʈt̪jəvəʈa t̪ʰa rəŋɡredʒi mē se h əməhū t̪ʰo ɽ bəɦuʈə sikʰəne ləga a kuʈʰ sikʰiyo gəya hū)</p> | | <p>grammatical category</p> |
|---|--|---------------------------------|

In the first example, the abbreviation for a compound noun is space separated. For example, Uttar Pradesh is the name of a state in India also known as U.P. If there is a space added in between, the tagger will treat it as two separate tokens and does not recognize it as a single *proper noun*. The second one is the example of mixed dialect. The present sentence is neither Hindi nor pure Bhojpuri but a variety of Hindi spoken in Bihar. The lexicons are intelligible to both Hindi as well as Bhojpuri speaker and therefore, left unedited and tagger according to the grammatical category to which the words belong.

3 ANNOTATED CORPUS FOR BHOJPURI

3.1 POS TAGGING

The POS (Parts of Speech) tags are labels given to the tokens in a text. These labels are technically known as ‘tag’ which is a grammatical category label featured on either morphological or syntactic or both properties of the word. The process begins with creating a definite tagset for the language which can be adopted from among different tagging schemes present, so far. In the words of Hardie,

“POS tagging (or morpho-syntactic tagging) is the process of assigning a label to each word in a text which indicates the status of that word within a system of that language on the basis of their morphological and/or syntactic properties” (Hardie, 2003).

Since the struggle for the recognition of Bhojpuri began, scholars are making efforts in different directions to bring it to a national platform. The stream of literature, philosophy, entertainment, cinema and revisiting folks all are witness to its expansion. But the technological domain was left untouched. The abandonment from the technology might prove fatal for a language, no matter how popular the language has been or continues to be. The question of technical advancement for Bhojpuri is a major issue. The present work aims at this aspect of the problem by addressing some basic but important techniques for building language resources.

This chapter will study the second level of initiation, which is to create an annotated Bhojpuri corpus. The annotation of the corpus will take place over the raw monolingual corpus studied in the last chapter. The annotation work began soon after the corpus was ready. As part of this chapter we will see the annotation framework, parent schema, the process of annotation, guideline for annotation and the challenges met throughout, in detail.

3.2 POS ANNOTATION SCHEME

Annotation scheme is a mutually agreed set of rules, formed on the basis of pilot studies, workshops or some small scale experiments for tagging the corpus concerning the morphological properties of the words within. The first standard and most used POS scheme is the Penn Tree tagset for English which is still followed by a number of European and Asian languages. There

are different tagsets adopted or created for tagging Indian languages like IIIT tagset, ILMT tagset, tagset designed under TDIL program, IL-POST etc.

But the main point of concern, here, is the selection of the scheme for the present task. Which tagset to choose and how to choose it . All the tagging schemes are authenticated by some institute or the other, they all have their pros and cons, their wows and flaws. For example, the IL-POST tagset is a very elaborate one but with all the finer linguistic details which might not be of the primary concern at the level of Parts of Speech. The fine-grained-ness of a tagset might overburden the system and adversely affect the performance. Therefore, it would be better to keep the schema coarse and let the features like inflection and derivations be handled by the morphological analyser. But the plus point with IL-POST is its hierarchical arrangement. A hierarchical tagset is capable of accessing the attributes and values of the categories and its sub categories (Narayan, 2011). Where, IL-POST is fed with hierarchy, the IIIT tagset lacks the sub categorization of the tags but provide the practitioner with a very compact tagset with limited and only necessary tags.

D'ejean (2000) also opines that a tagset must be designed keeping two criteria in mind. Firstly, the external criterion was fulfilled by the maximum extent of retrieval of the grammatical distinction in the language. Secondly, the quality of successful disambiguation and accuracy result as internal criteria (as cited in Nainwani, 2011).

Hence, the struggle ended with the idea of choosing such a scheme which inherits both the necessary features required for tagging. The decision was made to follow the BIS (Bureau of Indian Standard) Indian national standard for the annotation of Bhojpuri for the present annotation task in hand.

3.2.1 The BIS Scheme

Bureau of Indian Standards (BIS) is a super ordinate Parts of Speech framework designed for annotating all Indian Languages. The different tagsets for particular language subordinated to it. It is designed such that all Indian languages fit into this scheme with some related changes according to their linguistic features and grammatical functions.

The BIS v.1.1 has been followed for the present annotation task, which has undergone revision during ILCI-ILMT POS workshop held in July 2012³². The total number of annotation tags in this scheme is 38 divided under 11 major categories. The sub-types were extended upto only two level at max. Therefore, the classification included the categories in the following order of the grammatical categories, sub type level 1 i.e. the sub-type of the major categories like verb will include main and auxiliary verb. The further sub-division was the subtypes of the level 1 category i.e. the subtype level 2. This level of tag is found attested in the verb category only. The major verb category includes two types, main verb and auxiliary verb as pictured in the table below and the main verb is further segregated into four- finite verb, infinite verb, infinitive verb and gerund.

The BIS guideline also states that there are three such categories which are themselves the tag level categories with no sub-divisions, they are, adjectives, adverbs and the postpositions and one category with 4 tags under the subtype level 2, i.e., the main verb. Rest all the categories have one or more subtypes up to level 1 only.

3.2.2 POS-Tagset for Bhojpuri

This tagset was first produced in the 2nd Workshop on Indian Language Data: Resource and Evaluation at LREC 2014, titled ‘*Annotating Bhojpuri Corpus Using BIS Scheme*’(Singh, 2014).The present tagset is a multi-categorial frame set of 34 tags³³ for Bhojpuri with 11 tag level category including sub-types. The tag level/major grammatical categories are same as seen in the original BIS frame above comprising nouns, pronouns, demonstratives, verbs, adjectives, adverbs, postpositions, conjunctions, qualifiers, particles and residuals. These have been further sub-categorized into its sub types like common, proper and spatio-temporal nouns under the noun head and main verb and auxiliary verb under verb heads, etc. The inclusion criteria of these categories are completely language specific. Let’s take the example of Hindi. The tags for Hindi tagset must be selected according to the linguistic characteristic features of Hindi and the intended granularity of the tagset. Verbal nouns or gerunds are found in Hindi. Despite having a

³² <http://sanskrit.jnu.ac.in/ilciann/index.jsp>

³³ The earlier tagset had 33 tag level categories (as listed in Singh and Banerjee, 2014), Echo before has been included in the revised version.

nominal function, the verbal nouns were suggested to be tagged as the main verb in the sentence. The compatibility factor also applies in case of adverbs, where only manner adverbs were advised to keep within this category and the rest depicting time, place and frequency etc are to be tagged as spatio-temporal (N_NST) noun. The inclusion- exclusion criteria for Hindi allows only the major sub-types of verb for Hindi (main and auxiliary) and the deciphered sub categories of the main verb (finite, non-finite, infinitive and gerund), though present in the language, are not included as the part of the tagset. For more on this please refer to the BIS guideline (<http://sanskrit.jnu.ac.in/ilciann/index.jsp>).

3.2.3 A Preliminary Comparison of Hindi and Bhojpuri tagsets

Though Bhojpuri and Hindi are closely related languages and share similar syntactic structure, they may be hypothesized to have more or less similar kind of POS schema. But the vast difference in the morphological processes of the language brought some considerable changes to the tagging scheme as well as technique. The inclusion of classifiers, the tagging pattern for the classifier inflected cardinals and particles inflected open ended classes are explained with examples under section for the tagging guideline.

There was one more feature, originally proposed for Malayalam which has reduplicated compounds in which the echo part of the word occurs before the word itself. It was agreed to mark such cases as echo_before (ECH_B). Some such constructions were found in the Bhojpuri corpus data which made its inclusion into the tagset during revision.

3.2.4 Revised Bhojpuri Tagset

The initial tagset has been improved and revised and the challenges met during the early annotation as discussed in Singh and Banerjee (2014) are addressed. This revised tagset with the addition of one tag level category; echo before, in the tagset serves as the basis for annotation. Rest of the categories are imported as they were. Revisiting the tagset helped in coming up with the first guideline for annotating Bhojpuri corpus. The revised tagset looks like the one presented in the table 21 below:

Table 21 Bhojpuri POS tagset (Revised)

| Sl. No. | Category | | | Label | Annotation Convention ** | Examples | Remarks |
|---------|-----------|-------------------|--------------------|-------|-----------------------------|--------------------------------|---------|
| | Top level | Subtype (level 1) | Sub type (level 2) | | | | |
| 1 | Noun | | | N | N | दरवाजा, बुढिया, समय, | |
| 1.1 | | Common | | NN | N_NN | दरवाजा, बुढिया | |
| 1.2 | | Proper | | NNP | N_NNP | भर्बित्तन, गनेस, | |
| 1.3 | | Nloc | | NST | N_NST | अगवें, पछवें, उपरां, निचवें | |
| 2 | Pronoun | | | PR | PR | रउरा, जउन, जे | |
| 2.1 | | Personal | | PRP | PR_PRP | राउर, रहुआ, हम | |
| 2.2 | | Reflexive | | PRF | PR_PRF | आपन, अपन | |
| 2.3 | | Relative | | PRL | PR_PRL | जउन, जेके, जउन, जहाँ | |
| 2.4 | | Reciprocal | | PRC | PR_PRC | एक-दूसर, आपस, | |
| 2.5 | | Wh-word | | PRQ | PR_PRQ | कउन/के, कव, | |

| | | | | | | | |
|-----|---------------|------------|--|------|--------|------------------------|--|
| | | | | | | केहर | |
| 2.6 | | Indefinite | | PRI | PR-PRI | कोई, किसी | |
| 3 | Demonstrative | | | DM | DM | ई, ऊ, जे, जउन | |
| 3.1 | | Deictic | | DMD | DM_DMD | ई, ऊ | |
| 3.2 | | Relative | | DMR | DM_DMR | जउन , जे | |
| 3.3 | | Wh-word | | DMQ | DM_DMQ | का, कउन कब | |
| 3.4 | | Indefinite | | DMI | DM_DMI | कोई, किसी | |
| 4 | Verb | | | V | V | कहलन, बइठें, रोये, | |
| 4.1 | | Main | | VM | V_VM | कहलन, बइठें, नहवाए | |
| 4.2 | | Auxiliary | | VAUX | V_VAUX | रहल, बा, लागल | |
| 5 | Adjective | | | JJ | JJ | छोट, खुले, काला | |
| 6 | Adverbs | | | RB | RB | चाहे जइसे भी, तबही | |
| 7 | Postpositions | | | PSP | PSP | से, में, का, के, ला | |
| 8 | Conjunctions | | | CC | CC | अउर , अगर, जबकि, कि | |

| | | | | | | | |
|------|-------------|--------------|--|------|---------|---------------------------|--|
| 8.1 | | Co-ordinator | | CCD | CC_CCD | अउर , पर, भा | |
| 8.2 | | Subordinator | | CCS | CC_CCS | मगर , तो , कि , | |
| 9 | Particles | | | RP | RP | तो, ही, भी ,जी | |
| 9.1 | | Classifier | | CL | RP_CL | गो, गु, ठो | |
| 9.2 | | Default | | RPD | RP_RPD | तो, ही, भी , ना , जी | |
| 9.3 | | Interjection | | INJ | RP_INJ | अरे, हे, ए , हो | |
| 9.4 | | Intensifier | | INTF | RP_INTF | सा, खूब, इनता, बहुत, मारे | |
| 9.5 | | Negation | | NEG | RP_NEG | नाही , मत | |
| 10 | Quantifiers | | | QT | QT | पूरा, सब, खूब, एक | |
| 10.1 | | General | | QTF | QT_QTF | पूरा, सब, खूब, सारा | |
| 10.1 | | Cardinals | | QTC | QT_QTC | एक, दू, तीन | |
| 10.3 | | Ordinals | | QTO | QT_QTO | पहिला , दूसर | |
| 11 | Residuals | | | RD | RD | | |
| 11.1 | | Foreign word | | FW | RD_FW | | 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 | | Echo words | | ECH | RD_ECH | (चुप-चाप), (सच-मुच) | |
| 11.6 | | Echo before | | ECH_ B | RD_ECH_B | (अदला-बदली) | |

3.3 DATA FOR ANNOTATION

Out of the whole validated corpus of approximately 169k words the annotated corpus of 90k words has been used in the present work. The annotation of the 90k token was done partially manually and partially through a semi automated annotation tool. The annotated corpus after tagging was found to include 5198 sentences and 89999 tokens from 6 sets of files with the size of around 1000 sentences each. Blogs and miscellaneous domains were selected for the annotation as they cover the vast variety of linguistic data. The length of each file is listed in the table 22 below:

Table 22 Composition of the Annotated corpus

| S. No. | Set of Files | No. of Sentences | No. of Tokens |
|--------|--------------|------------------|---------------|
| 1 | blogs set 1 | 917 | 15192 |
| 2 | blogs set 2 | 936 | 15200 |
| 3 | blogs set 3 | 933 | 16032 |
| 4 | blogs set 4 | 811 | 14390 |

| | | | |
|---|------------|------|-------|
| 5 | misc set 1 | 1066 | 19322 |
| 6 | misc set 2 | 535 | 9863 |
| | Total | 5198 | 89999 |

3.4 ANNOTATING BHOJPURI CORPUS

Creating resources for a resource poor language is a challenge. The POS tagging of Bhojpuri was done semi-automatically using ILCIANN³⁴ tool. Six files of 1,000 sentences were annotated at this stage. This annotated corpus of 90k tokens comprised of mixed data from different corpus domain sets labelled as miscellaneous and blogs. The average length of sentence is 16 words/tokens per sentence. Rest of the corpora is undergoing further annotation and has not been included here. The cross-validation of the annotated corpora was done manually as it will serve as the training file for the tagger.

Next sub-section is dedicated to the description of the tool used for annotation.

3.4.1 ILCIANN

ILCIANN is an annotation tool developed under the ILCI project in (phase 1) which being semi-automatic facilitates the annotator and reduces labour. The tool was first introduced in a conference paper entitled '*Issues in annotating less resources languages -the case of Hindi from Indian Language Corpora Initiative (ILCI)*' in 5th Language Technology Conference (2011), Poland.

3.4.1.1 Online interface of the ILCIANN tool

The online tool has two interfaces, first the main login page followed by the annotation page where the input files are uploaded and tagging takes place. The main page of the tool looks like as shown in fig. 6 below:

³⁴ ILCIANN (Indian Language Corpora Initiative Annotation tool) is a tool for semi-automated POS tagging developed as part of ILCI project.

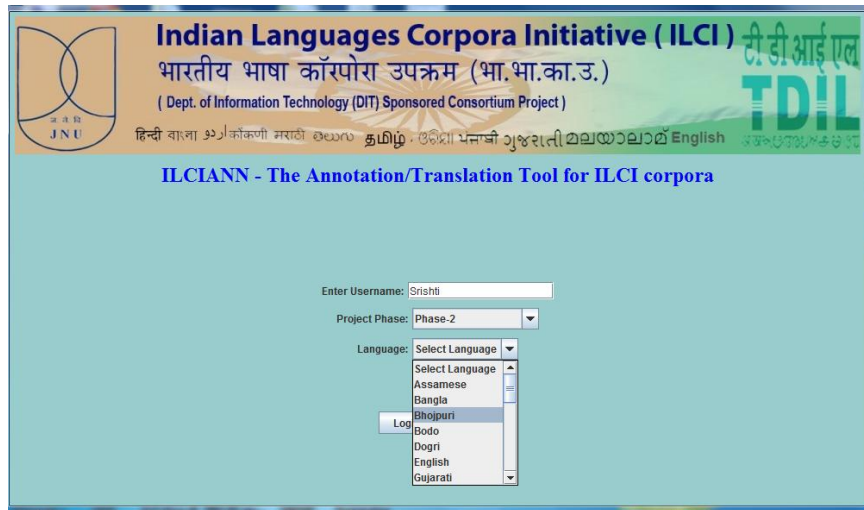


Figure 6 ILCIANN main page

This page is the initial page of the interface. After successful registration at the admin end the users can login with his/her ID, phase of the project currently running and the languages he/she is dealing with. Once the users have logged in, the annotation screen pops up as shown in fig 7.



Figure 7 ILCIANN annotation page (a)

The login details are mentioned at the top most row after the banner name followed by four options for editing the auto tags, uploading the file, logout and help.

The edit auto-tag contains the list of all the tags which are tagged by the tool automatically. This list can be updated by the users. The second icon is for uploading the file. This is helpful for a

completely annotated file, to connect it to the server and upload the new data on it. The third icon is for user to log off. According to the data format each file contains one thousand sentence which is not possible for a human annotator to go through with tagging in one go. Therefore, the user is facilitated with the option of signing off from the tool whenever required and complete the rest in the next session. Lastly, the help icon contains the tool handling instructions like any other software or program does. The annotation began with the selection of corpus file to be annotated and the sentence ID of the sentence. This also helps the user to pick up from where he left.

For tagging, the user is provided with another set of tags as select box options below the sentence. These are the actions taken on the corpus during tagging. Whether the translation for the present sentence must be saved, go to the next sentence, annotating a sentence, going back to the previous one, keeping notes and editing the sentence if required. These many options though look simple, but all becomes necessary when some automated method is used for reducing the burden of manual processing. There are list of tags provided after each token, tokens must take the tag from the respective category by selecting the tag name via either of the input devices i.e. keyboard and mouse. Figure 8 states that once the tagging is complete for a sentence, it must be saved before moving to the next one.

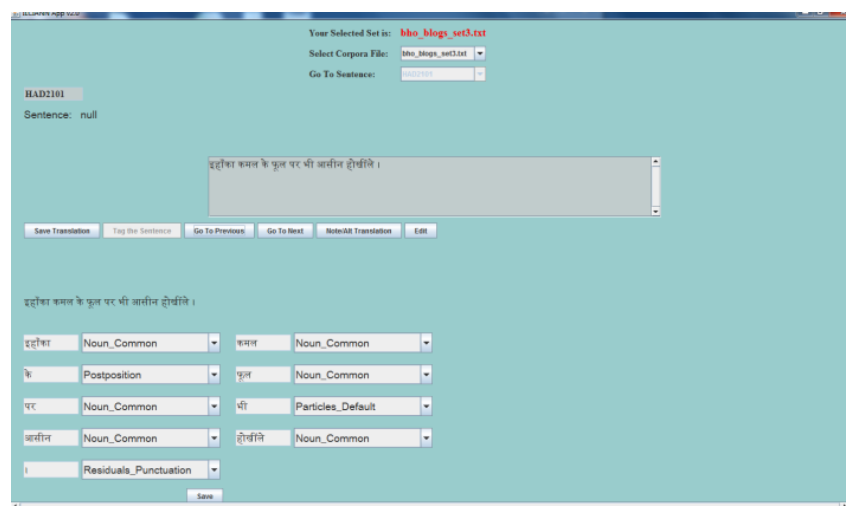


Figure 8 ILCIANN annotation page (b)

The tool is said to perform partial tagging better, for some categories like postpositions, conjunctions, qualifiers, auxiliary verbs, pronouns and other closed categories. Open ended

categories like nouns verbs and adverbs should be tagged manually because closed categories are less ambiguous than the open ended categories. Though there are exceptions to the same. Some postpositions and particles are also found ambiguous in like *par* (but), *ke bare meṅ* (about) and *to* (then) are problematic cases for the tool.

One major reason for keeping closed classes automated is that, these classes generally have a limited number of definite words in a given language and the chances of ambiguity in these are very low. Whereas, in open ended class, one single token can account for 2 to 3 different functions, sometimes even more. To save the machine from this confusion and overload, these must be handled manually. To know more about the tool design and application, please refer to Nainwani (2011).

3.4.2 Bhojpuri POS annotation guideline

Following the BIS standards and conventions, the present guideline for Bhojpuri has been produced which explains the annotated summary for the language. The limitation and challenges of the annotation are discussed in the next section.

3.4.2.1 NOUN (N)

Nouns, by the standard definition, are the names given to a person, place or thing. The nouns have three sub-divisions namely common noun, proper noun and temporal nouns (according to BIS scheme for Bhojpuri).

3.4.2.1.1 Common Noun (N_NN)

Common nouns are all abstract, concrete, mass and countable nouns. The animate and inanimate can also be a common noun unless it is a specific person or place. Common nouns in Bhojpuri are डॉक्टर(doctor), लोगन(people), साँप (snake), परीक्षा (exam) etc.

Abbreviations are given no separate tag but are to be marked under common and proper nouns according to their functions. The abbreviations for designation like डॉ. (Dr.), प्रो. (Prof.), एम.एल.ए. (M.L.A.), पं (abbreviation for saint) are to be marked as common noun whereas places like यू.पी. (U.P.) for ‘Uttar Pradesh’, is proper noun.

[1] कुछ भारी भरकम समस्यो/N_NN रहली सन

IPA: kuʈʰ bʰaɾɪ bʰərkəm səməsʃo rəhəli sən

Translation: Some big problems were also there

3.4.2.1.2 Proper Noun (N_NNP)

Proper nouns are the names of specific entities like person, place or thing. The names of product brands, name of a community or organizations, and their acronyms pronounced as complete names also fall under this category. For example, in addition to कोलकाता (Kolkata), रेशमा(Reshma), रायबरेली (Raibareli) which are names of people and place, brand names like Coca-Cola, organizations like भोजपुरी अकादमी (Bhojpuri Academy), नागरी प्रचारणी सभा (Nagari Pracharani Sabha) and acronyms like बीजेपी (B.J.P.)and सीबीआई (C.B.I.) will also be marked as Proper nouns.

[2] शाबाशी देत मणि/N_NNP शंकर/N_NNP अय्यर/N_NNP कहलन,

IPA: ʃabaʃi ðeʈ maɳi ʃənkər əjər kəhələn

Translation: Blessing the boy, Mani Shankar Iyer said,

3.4.2.1.3 Temporal Nouns (N_NST)

This contains a special set of verbs functioning both as a postposition and argument of a verb. These are invariably NST irrespective of their syntactic function. These words are as follows : आगे (front), पीछे (back), उपर (above), नीचे (below), अन्दर (in), बाहर (out), पहले (before), बाद (after). These words denote the time and place therefore are known as spatio-temporal nouns. The BIS scheme also suggests the inclusion of all adverbs except manner under this category, i.e., temporal nouns. More about this will be discussed under adverbs.

Some examples from them are:

[3] आगाआगा/N_NST तेज/RB चलत/V_VM युवक/N_NN हँस/V_VM के/V_VAUX

बोलल/V_VM

IPA: aga aga tɛdʒ ʈələt juvək h̄s ke boləl

Translation: The fast walking boy ahead said smiling

[4] भारतेन्दु/N_NNP युग/N_NN का/PSP पहिलहूँ/N_NST निबंध/N_NN लिखाते/V_VM
रहे/V_VAUX

IPA: b^hartənðʊ jʊgə kə pəhɪləhũ nibəndhə lɪkhatə rahe

Translation: Essays were being written even before the Bhartendu era.

3.4.2.2 PRONOUN (PR)

Pronouns are the words used in place of nouns. The category of pronoun consists of six self explanatory sub-categories. This includes personal, reflexive, relative, reciprocal, wh-word and indefinite pronouns which can be understood as follows:

3.4.2.2.1 Personal Pronoun (PR_PRP)

Personal pronouns are the pronouns denoting to any person, place or thing. These are the words replacing nouns. हम , हमका, हमरा, हमनी for ‘I’, तूं, तोहे, तहार, रउरा, राउ for ‘you’, ओकर, ओकरा, उनकर,(Possessive, him/her), ई, उ, ए, ऊहा (Deictic) etc are the personal pronouns in Bhojpuri.

[5] हम/PR_PRP अपना/PR_PRF कान्ह/N_NN पर/PSP बइठा/V_VM के/V_VAUX चल/V_VM
सकीला/V_VAUX

IPA: həm əpəna kə pər bəɪtʰə ke tʃəl səkɪlə

Translation: I can walk having him seated on my shoulders.

3.4.2.2.2 Reflexive Pronoun (PR_PRF)

The pronouns that denoted ownership to their antecedents are known as reflexive pronouns. The antecedents can either be a noun or a pronoun. अपन, आपन, अपनेआप and खुद (himself) are some limited examples of reflexives in Bhojpuri which are inflected to form their variants with the following meaning.

Table 23 Reflexive Pronouns

| Words | Gloss | Meanings |
|------------|-------------|----------------|
| अपना/ अपनी | ApanA/apanI | his or her own |

| | | |
|-------------|--------------|-----------------------------|
| आपने/ अपने, | Apane/ apane | himself/herself |
| अपनो | Apano | His/her + EMPH 'bhi' (also) |
| खुदे | Khude | Himself + EMPH 'hi' (only) |

[6] सुकुवार/JJ लउके/V_VM वाला/V_VAUX रूपवान/JJ भाई/N_NN आपन/PR_PRF राय/N_NN

दिहलस/V_VM

IPA: sukuwar ləuke wala rupawən bhai əpan raj ðihəles

Translation: The tender looking Sukumar presented his opinion

3.4.2.2.3 Relative Pronoun (PR_PRL)

Relative pronouns are pronouns which show the relation with their antecedent, either a noun or a pronoun. There is no change in the word form of a relative pronoun with the change in person, number and gender of the pronoun. जे, जवन, जेही, जेकरा, जिनकर, जिनका, are the relative pronouns in Bhojpuri.

Relative pronouns and relative demonstratives are similar in shape (form) but it is important to make out clear cut distinction between the two. Relative pronouns are used in place of noun and have its antecedent somewhere earlier in the sentence or discourse whereas relative demonstratives are the demonstratives indicating the noun (following noun in Indo Aryan languages), for which it stands. See the following examples:

[7] छपरा/N_NNP से/PSP कृष्ण/N_NNP कुमार/N_NNP वैष्णवी/N_NNP के/PSP फोन/N_NN

अउवे/V_VM जे/PR_PRL बहुते/RP_INTF उग्र/JJ रहलन/V_VM एह/DM_DMD फैसला/N_NN

के/PSP खिलाफ/N_NN

IPA: tʃʰəpəra ke kriʃɳə kumarə vɛʃɳəvi ke pʰon auve dʒe bəhuʈe ugrə rəhələn eh fəsəla ke kʰilapʰ

Translation: Krishna Kumar Vaishnavi called from Chapara who was very aggressive against the decision.

3.4.2.2.4 Recipocal Pronoun (PR_PRC)

Recipocal pronouns are indicative of reciprocity among two or more nouns/pronouns. एक दूसर, आपस में (each other) are the only reciprocal pronouns found in Bhojpuri. For example:

[8] धीरे धीरे आपस/PR_PRC मे बोलत बतियावत ई परिवार चलल जात रहे

IPA: ð^hire ð^hire apusə me boləʈ bəʈiyavəʈ i pərivarə ʃələl dʒəʈ rəhe

Translation: This family was moving gradually talking among themselves.

3.4.2.2.5 Wh-word Pronoun (PR_PRQ)

Wh-pronouns are the question words used typically to ask question. The following table enlists not all but some exemplar interrogative pronouns from Bhojpuri.

Table 24 List of Wh-Pronouns

| Word | Gloss | Meaning |
|--------------|---------------|-----------------|
| कइसे/कइसन | Kaise/kaisan | How |
| का/क | kA/ka | What |
| कतने | Katane | what time |
| कब | Kab | When |
| कहाँ/कहाँवाँ | kanA~/kahawA~ | Where |
| केने | Kene | which direction |
| के/कवन | Ke/kawan | Who |
| केकरा | kekarA | Whose |
| काहें | kAheM | Why |

This table contains the list of some of the wh- pronouns found in Bhojpuri.

3.4.2.2.6 Indefinite Pronoun (PR_PRI)

Indefinite pronoun refers to the objects or places which are not specified. This includes केहू, किसी, कोई, का, कवनो, कई, केकरो etc.

[9] बस केहू/PR_PRI तरह पइसा बनावे के बा

IPA: bəs kehu ʈərəh pəɪsə bənawe ke ba

Translation: He just wants to make money, by any means

The major difference between demonstrative indefinite and pronoun indefinite is that , the pronoun occur with some prior reference like केहू आदमी (someone).

3.4.2.3 DEMONSTRATIVE (DM)

Demonstrative has a major role in indicating about a noun without acting as a noun or anaphora. These make a distinct category from pronouns. It has four divisions- deictic, relative, wh-word and indefinite demonstratives.

3.4.2.3.1 Deictic Demonstrative (DM_DMD)

Deictic demonstratives demonstrate the noun it modifies. These are default demonstratives. In Bhojpuri deictic are typically ई, ऊ, ओ, एह, एहमे, ऐसन, इहो.

[10] ओह/DM_DMD घरी के निबनधकारन मे राम प्साद बिस्मिल हउएं

IPA: ohə gʰəri ke nibəndʰəkərən me ram prəsəd bisəmilə həuə̃

Translation: Ram Prased Bismil was one among the contemporary essayist of that time.

[11] एह/DM_DMD चलते गलन बढ गईल बा

IPA: ehə ʈələt̪e gələn bəd̪ʰə gəilə ba

Translation: Due to this the winter increased

3.4.2.3.2 Relative Demonstrative (DM_DMR)

Relative pronouns and relative demonstratives occur in similar fashion just the difference is that, a demonstrative always occur before a noun it modifies whereas a pronoun is used in place of noun.

[12] जे/DM_DMR लोग ए. सी. डब्बा मे जाला

IPA: dʒe logə e.si. dʌbba me jala

Translation: Those who travel in the A.C. class in the train.

[13] आजकाल्ह के लइका जवन/DM_DMR गावतारे सन ऊ गाइबि

IPA: ajəkəlhə ke ləika jəvən gavəʃare sən u gəibi

Translation: How will also sing the songs being sung by the today's generation kids.

3.4.2.3.3 Wh-word Demonstrative (DM_DMQ)

Wh-pronouns and wh-demonstratives share the words of interrogation but the demonstratives does not ask question, rather demonstrates using the wh-words. Wh-demonstratives in Bhojpuri are का, कब, कतना, कवने, काहे etc.

[14] लोग पूछल की भइल का/DM_DMQ रहे

IPA: logə puʃhələ ki bʰailə ka rəhe

Translation: People asked about what has happened.

[15] ए के पढेवाला केतना/DM_DMQ लोग बा भोजपुरी मे

IPA: e ke pəðhə wala keʃəna log ba bʰojpuri me

Translation: There are only a few Bhojpuri speakers who read this

3.4.2.3.4 Indefinite Demonstrative (DM_DMI)

Indefinite demonstratives also specify places, things and objects like indefinite pronouns. The words fall under this category are कोई, कवनो, कई,

[16] दाम मे कवनो/DM_DMI कमी नइखे भइल

IPA: ðamə me kəvəno kəmi nəikʰe bʰailə

Translation: There is no reduction in the price.

[17] सब केहू/DM_DMI जाना सवधान मुद्रा मे आ गइल

IPA: səb kehu dʒana səvəðʰanə muðra me a gail

Translation: Everybody became attentive.

3.4.2.4 VERBS (V)

The verbs have been divided into two sub-categories- main verb and auxiliary verb. Like Hindi Bhojpuri also has a closed set of helping verbs or auxiliaries and the main verb can be formed by inflecting the root verbs taking different verb forms. According to BIS each sentence or clause must have one main verb, and one or more auxiliaries can also be found. The second level of categorization has not been included in the existing Hindi guideline, based on which it has been excluded in the present Bhojpuri guideline to reduce the load on the machine. This discrimination can later be derived at the chunking level as the chunk tagset do have categories for different types of verb phrases.

Some of the constructions of verbs have been exemplified below:

3.4.2.4.1 Simple verbal occurrences

Verbs in simple sentence include one main verb which may or may not be followed by one more auxiliary verbs.

[18] डाढि/N_NN से/PSP लपटल/V_VM सरप/N_NN के/PSP फूत्कार/N_NN

IPA: dɑdʰɪ se ləpətəl səɾəpə ke phʊtkɑrə

Translation: The hissing of the snake wrapped around the branches of a tree.

[19] पहिला/QT_QTO बेर/N_NN माई/N_NN के/PSP माई/N_NN कहल/V_VM

सिखावल/V_VM जाला/V_VAUX

IPA: pəɦɪlə berə məi ke məi kəɦələ sikʰavələ dʒələ

Translation: The kid is learnt to speak mother to his/her mother for the first time.

[20] अँजोरिया/N_NN फिल्मी/JJ पत्रिका/N_NN जइसन/DM_DMR होखत/V_VM

जात/V_VUAX बियाV_VAUX

IPA: ʌ̃dʒorɪjɑ fɪlmi pət̪ɾɪkɑ dʒəɪsənə həkʰətə dʒatə bɪjɑ

Translation: Anjoria is becoming more like a movie magazine.

3.4.2.4.2 Compound verbs

Compound verbs in Bhojpuri are very similar to Hindi word formation where there are two root verbs. The first one act as the main verb with functional content and the other verb, takes a new meaning which adds to the completion of the meaning expected out of the main verb.

[21] सब कोई खाए/V_VM लागल/V_VAUX

IPA: səbə koi khae lagələ

Translation: Everybody started eating.

[22] नाचे वाली के इशारा कर/V_VM दिहल/V_VAUX

IPA: nɪʃe wali ɪʃara kərə ðihələ

Translation: He signalled the person standing underneath.

[23] लोग सुबहित नाम ना खोज/V_VM सके/V_VAUX आपन फिल्मन के

IPA: logə subəhiʈə namə na kʰodʒə səke

Translation: People cannot find a correct name for their movie.

3.4.2.4.3 Conjunct Verbs

Conjunct verbs are made with the combination of a noun and a verb. In this the verb is preceded by either a noun or an adjective. While tagging, the noun/ adjective counterpart are tagged with their own category and the next coming verb (also known as light or vector verb) are tagged as the main verb. For example, हासिल करना.

[24] थाना मे हाजिर/N_NN होखे/V_VM कि जरूरत ना पड़ी

IPA: ʈʰana me haʒɪrə hokʰe ki dʒərurətə na pəɽi

Translation: He does not require to present in the police station.

[25] ट्रेन सफ़र के मजा खतम/N_NN हो/V_VM जाई/V_VAUX

IPA: ʈrenə səpʰərə me mədʒa kʰəʈəmə ho dʒəi

Translation: The enjoyment will be lost in the journey by train.

[26] धर्म बदलवा कर दुसरे धर्म मे शामिल/N_NN कराने/V_VM की आजादी दी जाये

IPA: ðərəmə bəðələwa kərə ðUsərə ðərəmə me ʃamilə kərane ki adʒaði ði jae

Translation: To give the freedom for changing and adopting any other religion.

[27] फिलिम मे काम/N_NN कइला/V_VM से खुश मनोज

IPA: p^hɪlɪmə me kamə kəɪlə se k^huʃ mənɔdʒə

Translation: Manoj is happy working in the movies.

3.4.2.4.4 Explicator Compound Verbs

Explicator verbs are one of the major areal features of South Asian Languages. It is the sequence of two verbs V1 and V2. The first verb in the sequence is called the ‘main’ or ‘polar verb’ and the V2 is called ‘operator’, ‘explicator’, or ‘vector verbs’ (Abbi, 2001). Kachru (2006) lists some definite verbs forming the sequence in explicator as- *aa* ‘come’, *jaa* ‘go’, *le* ‘take’, *de* ‘give’, *uth* ‘rise’, *baithh* ‘sit’, *paDx* ‘fall’, *Dal* ‘drop, pour’, *rakh* ‘keep, put’, *choDx* ‘give up’, *mAr* ‘hit’, *nikal* ‘emerge’, *dhamak* ‘thump’, and *pahuMc* ‘arrive’. Bhojpuri explicator verbs are listed in the examples below:

[28] पढ़ के मन खुश हो/V_VM गइल/V_VAUX

IPA: pəɽə ke mənə k^huʃ ho gəɪlə

Translation: You will be happy to read it.

[29] सभे पीएम शुरू/N_NN कर/V_VM दीहल/V_VAUX

IPA: səb^he piʃe ʃurɔ kəɾə ðihələ

Translation: All started drinking.

[30] कहानी के शीर्षक बाद में कहाउत बन के रहि/V_VM गइल/V_VAUX

IPA: kəhəni ke ʃiɽʃəkə bəðə me kəhəuɽə bənə ke rəhi gəɪlə

Translation: The title of the story later became a proverb.

3.4.2.4.5 Verbal Nouns/Gerunds

The BIS guideline does not accounts for the category of verbal noun or gerund as an independent group but it includes this as part of main verb. Because, although, the verbal noun in Hindi or Bhojpuri functions as a noun but they can take their own argument which is an unique feature of verbs.

[31] सभे पिए/V_VM शुरु कर दीहल

IPA: səb^he piɛ furo kər ðihələ

Translation: Everybody started drinking.

[32] जब कबबो कही जए/V_VM के होत रहे

IPA: dʒəbə kəbbə kəhi dʒæ ke hoʈə rəhe

Translation: whenever he had to go somewhere.

[33] सभे लोग के नहाए/N_NN फीचे/N_NN में आसानी होखे

IPA: səb^he logə ke nəghæ p^hiʃe mē asani hoʃe

Translation: Everybody will find it easy to take bath and all.

3.4.2.4.6 Nouns Derived from Verbs

Nouns derived from verbs in Bhojpuri are similar to the derived nouns in Hindi, and they are marked as a noun due to inability to take a new argument on their own. Some of the derived nouns are listed in the table 25 below:

Table 25 Derived nouns from verbs (V<N)

| S. No. | Verbs | Gloss | V<N | Gloss |
|--------|-----------------|-----------|------------------|-----------|
| 1 | पढना (paDhanA) | to read | पढाई (PaDhAI) | Study |
| 2 | कमाना (kamAnA) | to earn | कमाई (kamAI) | Earning |
| 3 | मना (manA) | to reject | मनाई (manAI) | Abolished |
| 4 | सुनाना (sunAnA) | to hear | सुनवाई (sunawAI) | Hearing |

[34] सगरी मामिला क एके साथ सुनवाई/N_NN करे के होइ

IPA: səgəri mamila kə eke sat^he sunəwai kəre ke hoi

Translation: the whole matter has to be heard altogether.

3.4.2.4.7 Participle Construction of Verb Acting as Modifier

The participle verb form in Bhojpuri is quite different from that of Hindi. In Hindi the participles are formed with the occurrence of *huA* after the main verb which makes it look like *daudtA huA* (running), *AtA huA* (coming) and *haMstA huA* (smiling). Unlike Hindi, Bhojpuri participle are marked by ‘-ət’ marker inflected after the main verb and the verb form appears as *dauDat*, *At* and *hasat*, respectively. These are invariably marked as the main verbs.

[35] आगा आगा तेज/RB चलत/V_VM युवक हस के बोलल

IPA: aga aga tɛdʒə tʃələt̪ juvəkə həsə ke bolələ

Translation: The fast walking boy in front said smiling

[36] सहायता करे क घोषणा करत/V_VM जिलाधिकारी कहलन

IPA: səhajət̪ə kərə kə ɡʰoʃəɳə kərət̪ dʒilad̪hikari kəhələnə

Translation: Announcing about the support, the District Magistrate said.

3.4.2.4.8 Conjunctive participle

According to Yamuna Kachru (2006) “Conjunctive participles are used as temporal, manner, causal, concessive and esthetical verbs.” Conjunctive participle is formed by adding auxiliary *kar* in Hindi which changes to *ke* of Bhojpuri. The action verb in this series is marked as the main verb and the *ke* part is tagged as an auxiliary. For example:

[37] जुलाई अंक मिलल त देख/V_VM पढ़/V_VM के/V_VAUX मन खुश हो गइल

IPA: dʒulai aŋkə mɪlələ t̪ə ðekʰə pəɖʰə ke mənə kʰuʃə ho gəɪlə

Translation: I was happy to read the July edition when I found it.

[38] अपना कान्ह पर बइठा/V_VM के/V_VAUX चल सकीला

IPA: apənə kanhə pərə bəɪt̪ʰa ke tʃələ səkɪlə

Translation: I can walk having him seated on my shoulders.

3.4.2.5 ADJECTIVES (JJ)

Adjectives are qualifiers of the noun. This category does not include any sub-division and it is self explanatory as the attributives of a noun. And the quantifiers make a separate category in this scheme which is distinct from adjectives. Some common adjectives are अच्छा(good), छोटा(small), मुख्य (main), बड़ (big), बढ़िया (nice),etc.

[39] मन खुश/JJ हो गइल

IPA: mənə kʰuʃə ho gəɪlə

Translation: I became happy.

[40] सगरी/JJ प्रान्ते ना पूरा देश में नीतीश के जयकारा

IPA: səgəri prant̪e na pura ɖeʃə me~ niːʃə ke dʒəjəkara

Translation: Not only the state but the whole country is praising Nitish.

Some other adjectives are विवादास्पद (controversial), सामूहिक (in a group), विरोधी (rival), धार्मिक (religious) which are found in constructions as shown in the following example:

[41] दुनू देश सकारात्मक/JJ राह/N_NN देखवले बानी

IPA: ɖunu ɖeʃə səkarətməkə rahə ɖekʰəvələ bani

Translation: Both the countries are showing the positive pathways.

In some cases an adjective functions more like a verb as in *bahatA jharanA*. Such instances are accompanied with the *wAlA* particle in Bhojpuri and become *bahe wAlA jharanA*. This is tagged according to the guideline convention for *wAlA*. BIS guideline states that *wAlA* when occur as part of the noun like *sabji wAlA*, must be tagged as a postposition (PSP) and when come along with a verb, should be marked as an auxiliary to the verb. See examples blow:

[42] ऊ/DM_DMD वाला/PSP गाना/N_NN गावे के प्रेसर त नाहिं दी

IPA: u vala gana gəve ke presərə t̪ə nahĩɛ ɖi

Translation: Please do not pressurize me to sing that song, at least.

[43] पाती का मंच से जुड़े/V_VM वाला/V_VAUX

नया लोगन के स्वागत करत बा

IPA: paŋi ka mǝŋʈə se dʒuɽe ʋala nəya logəɳə ke sʋagəɽə kəɽəɽə ba

Translation: Pati welcomes all the new joinees to the group

3.4.2.6 ADVERB (RB)

Like adjectives, adverbs are also the single tag category. Adverbs are the qualifier of verbs, mainly. The BIS standard suggests only manner adverbs under this category. Words like धीरे (slow), जल्दी (hurry), and तेज (fast) falls under this in Hindi as well as in Bhojpuri. Kachru has categorised adverbs based on their nature and function, most of them do occur in the corpus and to reduce ambiguity, they are tagged as follows:

Table 26 Annotation for other types of adverbs

| S.NO. | Types of Adverbs | Features/forms | Examples | Gloss | Tag Category |
|-------|---|--|------------------------|----------------------------------|------------------|
| (A) | Basic Adverb | all place, manner and time adverbs | ab, phir, dhIre, jaldI | now, again, slowly, fast | |
| 1 | All manner adverbs | | dhIre, jaldI | slowly, fast | RB |
| 2 | N/PR<RB (adverbs from nouns and pronoun) | derived from noun and pronoun | Aj, kal, andar, bAhar | today, tomorrow, inside, outside | N_NST |
| 3 | DM, PRL, <RB (adverbs from deictic and relative pronoun) | derived from demonstrative, relative and interrogative pronoun | ab, tab, jab, | now, then, when, | DM_DMR and N_NST |
| 4 | JJ<RB (adverbs | derived from adjectives | aise, vaise, jaise, | this way, that way, like this, | DM_DMR |

| | | | | | |
|-----|---------------------------------|-----------------------|---------------------------------------|-------------------------------------|--|
| | from adjectives) | | | | |
| 5 | | | itnA and utnA | this much, that much | QT_QTF |
| 6 | V<RB (adverbs from verbs) | derived from verbs | | | will go under the verb (as already discussed under the verbs) |
| 7 | Conjunctive Participle | V+ke | | | will go under the verb (as already discussed under the verbs) |
| (B) | Complex Adverb | | | | |
| 1 | Temporal | | abhi, pehle | now, earlier | N_NST |
| 2 | Locational/ Directional | | kamre me, mez par, sateshan tak | in room, on table, at station | N_NN + PSP (mark according to the word category) |
| 3 | Manner | | dhyan se and shantipurvak | carefully, quietly | N_NN + PSP and JJ (mark accordin to the word category) |
| 4 | Instrumental | | kaicI se | with scissors | N_NN + PSP (mark according to the word category) |
| 5 | cause, reason | | kehne se, wajah se, dard se | by saying, due to, with pain | V_VM + PSP, N_NN + PSP, N_NN + PSP (mark according |

| | | | | | |
|---|--------------|--|----------------------|---------------------|--|
| | | | | | to the word category) |
| 6 | committative | | ke sAth, ke sahit | with, along with | PSP + N_NN, PSP + N_NN (mark according to the word category) |

As shown in the table above, all manner adverbs are tagged as RB, are placed under the spatio-temporal noun, adverbs acting as relative demonstrative are placed under the same. Adverbs derived for adjectives fall under relative demonstrative as well as under the general quantifier. Conjunctive participle adverbs derived from verbs are to be placed within the verb itself. And the complex adverbs are marked as nouns, postpositions and verbs depending upon the nature of its use as given in the table above. The adverb derived from noun and pronoun like शायद (perhaps), हमेशा (always), बिल्कुल (definitely), एकदम (surely) etc. has been included as part of RB (manner adverb, due to the lack of separate division for this category).

3.4.2.7 POSTPOSITION (PSP)

Postpositions are those parts of speech which marks cases. According to Kachru (2006), postpositions signal relations between the two words. का, के, मे पर, ला, ले, and बदे(के लिए), are some commonly found postpositions in Bhojpuri. Apart from this there are several examples where more than one word/ token forms a postposition, they are called complex postposition. के बाद (later), के ऊपरा (over), के पाछा (after) are parts of complex postpositions in Hindi and in Bhojpuri, both.

[44] फिलिम/N_NN में/PSP काम कइला/V_VM से/PSP खुश मनोज

IPA: pʰɪlɪm mɛ̃ kamkəɪla se kʰuʃə mənɔdʒə

Translation: Manoj is happy to act in the movie

[45] अखबार/N_NN से/PSP लेके टीवी चैनल/N_NN ले/PSP सभे चिचियात रहे

IPA: əkʰəbərə se leke tɪvi tʃənəl le səbʰe tʃitʃijaʈə rəhe

Translation: Everyone was gossipping from news paper to T.V.channel.

3.4.2.8 CONJUNCTION (CC)

Conjunctions act the conjoiners of two phrases or clauses. It has two sub-categories namely co-ordinator and subordinator.

3.4.2.8.1 Co-ordinators (CC_CCD)

Co-ordinating conjunctions are those which joins two phrases or clauses of the same category, like अउर (and), आ(and), लेकिन(but), बाकिर(and), बल्कि(although), भा(or). For example:

[46] एगो सूई लगवलें आउर/CC_CCD ईलाज तुरंते शुरू हो गइल

IPA: ego sUi ləgəvələ əuə iləjə tʊrəntə fʊrʊ ho gəilə

Translation: He gave an injection and the treatment began

[47] भोजपुरी/N_NNP में/PSP लिखल/JJ आ/CC_CCD कृति/N_NN फान्ट/N_NN

भा/CC_CCD यूनिकोड/N_NN में/PSP टाइप/N_NN कइल/V_VM

IPA: bʰodʒəpuri mē likʰələ a kɾitɪ pʰantə bʰa jʊnikodə me tʌipə kəilə

Translation: It was written in Bhojpuri and typed in Kruti font.

3.4.2.8.2 Subordinators (CC_CCS)

Subordinators, on the other hand, conjoin two clauses or sentences in which the second clause is subordinated by the first clause. Subordinating conjunctions are typically कि (that), क्योंकि (because), हालांकि (although), ताकि (so that), चाहे (whether) etc in Bhojpuri.

[48] काहें/DM_DMQ कि/CC_CCS उ गराज अब उनुका रहे

IPA: kahē kɪ ʊ gəradʒə əbə unuka rəhe

Translation: Because that baggage belongs to him now.

3.4.2.9 PARTICLES (RP)

Particles are those words which do not fall under any other category. Five types of particles has been included under the Bhojpuri tasget, namely- default, classifier, interjection, intensifier and negation. These are described below with examples.

3.4.2.9.1 Default (RP_PRD)

Default particles do not have any independent meaning but it appears with the head words for the emphasis or focus. ही(only), तो (still), भी (also), ना etc are the default words in Bhojpuri.

Although in Bhojpuri they appear in more than one orthographic form like *to*, *tA*, *ta* for *to*, *na* and *nA* for *nA* etc. Another unique feature of Bhojpuri is that most of the particles in Bhojpuri unlike Hindi are inflected with the head word. These head word may belong to any of the following categories forming a new word form like अबहींए (abahiyeM) for ‘right now’, चढईबो (caDhaibo) for ‘also offer’ and जिभिए(jibhiye) for ‘tongue only’ etc. This will be discussed in detail under the POS tagging issues for inflected categories.

3.4.2.9.2 Classifier (RP_CL)

Classifier is present only in some languages; these are known as the referent of a countable noun. In languages like Bangla, Bhojpuri, Odia and Maithili classifiers are majorly present (it appears with a numeral or demonstratives). Due to their presence with a noun when counted, these are also called measure word.

Bhojpuri classifiers differ on the variety of language spoken in a particular region. गो ‘go’ and ठो ‘Tho’ are the most popular varieties spoken in the region of Bihar and U.P, respectively. Moreover, another form खो ‘kho’ is also found but only as the speech of the lower communities/ local variety. The so called standard variety of Bhojpuri i.e., the one spoken in the Bhojpur region makes use of गो ‘go’ and is commonly followed in the written form of language also. These classifiers also appear sometimes along with the numeral and other times as an inflected part of the numeral, as shown in the example below:

[49] एगो/QT_QTC बात हमहू पुछीं

IPA: ego baʈə həməhə pʊʈʰi

Translation: May I also ask something.

[50] दू/QT_QTC गो/RP_CL आरोपियन/JJ के/PSP थाना/N_NN में/PSP हाजिरी/N_NN

IPA: ðu go aropiyənə ke tʰana mẽ haʤiri

Translation: Two culprits have their hearing in at the police station.

[51] तकरीबन/RB दस/QT_QTC गो/RP_CL हथियारबन्द/JJ घुड़सवार/N_NN रहलन/V_VM

IPA: təkəribənə ðəsə go hətʰijarəbəndə ɡʱuṛəsəvərə rəhələnə

Translation: There were around 10 armed horseriders.

3.4.2.9.3 Interjection (RP_INJ)

Interjection particles are those particles which exclaim the emotions in the sentence. These are हे

(hay!), ओह (oh!), अरे(are!), हाए(hAe!), रे, री, भाई, हैं, जय and राम (oh god!) are some common

interjections in Bhojpuri. For example:

[52] बाजलि बैरनि रे/RP_INJ बाँसुरिया/N_NN

IPA: baʤəli bəɾəni re bāsurija

Translation: How rude is this pity flute!

[53] अरे/RP_INJ भया/RP_INJ ,/RD_PUNC मजूर/N_NN कहां/DM_DMQ से/PSP

मिली/V_VM

IPA: əre bhəija! məʤʊərə kəha se mili

Translation: Oh! How would I find a labour.

[54] हैं/RP_INJ हो/RP_INJ ढेर/QT_QTF दिन/N_NN त/RP_RPD होइए/V_VM

गइल/V_VAUX बा/V_VAUX

IPA: h̄ə ho ! dʰerə ðimə t̄ə hoije ɡəɪlə ba

Translation: Yes! It has already been a long time.

3.4.2.9.4 Intensifier (RP_INTF)

The words which intensify the quality of an adjective or an adverb are known as intensifiers. In Bhojpuri, intensifiers are namely बहुते (too much), सबले(all), खूबे(very much), इतना (this much), मार (very), बड़ा (big) etc.

[55] कमजोरी बतावे में सबले/RP_INTF आगा/N_NST रहेला

IPA: kəmədʒori bəʒəve me səbəle aga rəhəla

Translation: Everyone is ready at pointing out others.

[56] ई/DM_DMD काम/N_NN अतना/RP_INTF बड़हन/JJ बा/V_VM कि/CC_CCS

अकेले/JJ ना/RP_NEG सँपरी/V_VM

IPA: i kamə ətəna bəʒhənə ba ki əkeke na səpari

Translation: This is a bigger task which cannot be handled alone.

[57] मास्टर/N_NN चलि/V_VM जालें/V_VAUX त/RP_RPD ओनकर/PR_PRP

खूब/RP_INTF आदर/JJ होला/V_VM

IPA: mastərə ʃəlɪ dʒalɛ̃ t̪ə onəkərəkʰobə aɖər hola

Translation: A teacher is very much respected on his visit.

3.4.2.9.5 Negation (RP_NEG)

The words indicating negation are categorized under the negation particles. ना (no), नाँहि(no), नइखे (not), मत(don't), बिना(without), बगैर (without) are the negations in Bhojpuri.

[58] बिसवास नइखे/RP_NEG होत बाकिर ई साँच बा

IPA: bɪsəwasə nəɪkʰe hoʒə bakɪrə i səɪʃ ba

Translation: I cannot believe this but it is the truth .

[59] ना/RP_NEG त/RP_RPD तहार/PR_PRP बियाहो/N_NN ना/RP_NEG होखत/V_VM

रहे/V_VAUX

IPA: na t̪ə təharə bɪyaho na hokaahəʈə rəhe

Translation: Otherwise you would not have got married.

3.4.2.10 QUANTIFIERS (QT)

Quantifiers are the modifiers of noun or adjective which indicates quantity. Three sub-categories have been created to cover the quantifiers in the tagset - general, cardinal and ordinal.

3.4.2.10.1 General (QT_QTF)

The precision in the quantity is not mentioned under the general quantifiers. They simply make an estimation of the quantity, such as सभे(all), कम(less), कुछ(some), थोड़ा (little), ढेर(a lot), कुल (all) etc.

[60] लड़की कुल/QT_QTF पढिहें/ त इ नौबत काहेआई

IPA: ləʈəkɪ kulə pəʈh̃h̃ɛ̃ t̪ə ɪ nɔbət̪ə kahe ai

Translation: I girl education would begun; this situation will never rise again.

[61] फगुआ/N_NN आउर/CC_CCD होली/N_NN के/PSP ढेर/QT_QTF ढेर/QT_QTF

शुभकामना/N_NN

IPA: p̪h̃əɡua aurə holi ke d̪herə ʃubhəkaməna

Translation: Wishing you Fagua and holi.

[62] उनकर/PR_PRP बड़/JJ बेटा/N_NN बेसी/QT_QTF चालाक/JJ रहल/V_VM

IPA: unəkəɾə bəʈə beʈa besi ʃʌlakə rəhələ

Translation: His elder son was quiet clever.

3.4.2.10.2 Cardinals (QT_QTC)

Cardinals are the absolute numbers in digits or in words e.g. one, two, 3, 4, 5 etc.

[63] बाकिर एगो/QT_QTC बात हमहू पुछीं

IPA: bakɪrə ego baʈə həməhu pɔʃh̃ĩ

Translation: And may I also ask something

[64] हैलो/RP_INJ भोजपुरी/N_NNP के/PSP दिसम्बर/N_NNP अंक/N_NN पाती/N_NNP

के/PSP अंक/N_NN संख्या/N_NN ६९७०/QT_QTC

IPA: hɛlo bʰodʒəpuri ke ðisəmbərə əŋkə pati ke əŋkə səŋkʰja 6970

Translation: December edition of Hello Bhojpuri, Paati edition no. 6970.

3.4.2.10.3 Ordinals (QT_QTO)

The ordinal denotes the order of the digit like पहिला, दूसर, तीसर etc. In Bhojpuri the ordinals make gender distinction also, very much like Hindi but no number distinction. For example- पहिला, दूसरकी, तीसरके etc.

[65] कांग्रेस अपना उम्मीदवारन के पहिलका/QT_QTO सूची/N_NN जारी क दिहलसि

IPA: kəŋɡres əpəna ummiðəwarənə ke pəhɪləkə suʃi dʒəri kə ðihələsi

Translation: Congress has released the first list of its candidates.

[66] दूसरका/QT_QTO दिन/N_NN के/PSP खेल/N_NN खतम/JJ होखे/V_VM

बेरा/V_VAUX

IPA: ðusərəkə ðɪnə ke kʰelə kʰətəmə hokʰe bera

Translation: The game of the second day is about to end.

[67] भारत/N_NNP एक/QT_QTC विकेट/N_NN गवां/V_VM के/V_VAUX रन/N_NN

बनवले/V_VM रहवे/V_VAUX

IPA: bHarətə ekə vikeʃə gəvā ke rənə bənəvələ rəhuve

Translation: India has made runs at the loss of one wicket.

3.4.2.11 RESIDUALS (RD)

From the standard guidelines, residuals are the words which are not the intrinsic part of the language. This category has been divided into six further categories with the inclusion of one new 'Echo before' as part of the guideline. These categories are namely foreign words, symbols, punctuations, unknown, echo-words and at last, echo before.

3.4.2.11.1 Foreign Word (RD_RDF)

Foreign word includes all those words which are written other than the script of the language i.e., Devanagari script in case of Bhojpuri.

3.4.2.11.2 Symbols (RD_SYM)

Symbols are the characters which are neither part of alphanumeric chart nor related to the script. These includes \$%^@* etc.

3.4.2.11.3 Punctuations (RD_PUNC)

Punctuations are the regular punctuation markers of the language. For Bhojpuri the punctuation includes the sentence boundary marker (।), comma(,), colon(:) , exclamation mark(!), semi colon(;) etc.

3.4.2.11.4 Unknown (RD_UNK)

Unknown words are the words for which the annotator is not able to decide a particular tag category. This majorly includes the words from the other languages (some other language, related or unrelated) written in the script of the language.

3.4.2.11.5 Echo-Words (RD_ECH)

Echo formation is a common feature of Indo Aryan languages. This can be seen as part of reduplication as described in Abbi (2001). The word forms generated after the process of echo formation are called echo words. Following are the examples of echo words from the present corpus, as tabulated in below:

Table 27 Echo words from the Corpus

| S.NO. | Echo words | Transliteration | Tag | IPA |
|-------|------------|-----------------|--------|----------|
| 1 | नोंक | noMk | N_NN | Argument |
| | झोंक | khoMk | RD_ECH | |
| 2 | सीधा | sIdhA | JJ | Simple |
| | साधा | sAdhA | RD_ECH | |

| | | | | |
|----|---------|-----------|--------|----------------|
| 3 | अगडम | agaDham | N_NN | non-sense |
| | बगडम | bagaDham | RD_ECH | |
| 4 | रिपोर्ट | Riport | N_NN | report and all |
| | फिपोर्ट | Phiport | RD_ECH | |
| 5 | जूझत | Jujhat | V_VM | Struggling |
| | जागत | jAgat | RD_ECH | |
| 6 | ठोंकत | ThoMkat | V_VM | Hitting |
| | ठेठावत | TheThAwat | RD_ECH | |
| 7 | मिला | milA | V_VM | more or less |
| | जुला | julA | RD_ECH | |
| 8 | साफ | sAph | JJ | Clean |
| | सुथरा | sutharA | RD_ECH | |
| 9 | घूम | ghUm | V_VM | Roaming |
| | घाम | ghAm | RD_ECH | |
| 10 | होखे | Hokhe | V_VM | have done |
| | हवाखे | hawAkhe | RD_ECH | |
| 11 | भोली | bholI | JJ | Innocent |
| | भाली | bhAlI | RD_ECH | |
| 12 | गोताइल | Gotail | N_NN | |
| | बोथाइल | bothAil | RD_ECH | |

| | | | | |
|----|-------|-----|--------|--|
| 13 | बाप | bAp | N_NN | Father |
| | वाप | wAp | RD_ECH | |
| 14 | गिटिर | | N_NN | manner of speech (hard to understand) |
| | पिटिर | | RD_ECH | |

3.4.2.11.6 Echo before (RD_ECH_B)

Echo Before is the final category of the tagset. ‘Echo Before’ has been included for the first time, in the tagset for Indo Aryan Languages. This category explains such constructions in which the echo counterpart of the word appears before the content word. This feature was found to appear in Dravidian languages like Malayalam and was proposed for them (in the revised BIS guideline, 2012).³⁵

There are some occurrences of such echo formation found in the present Bhojpuri corpus, they are given in the table 28 below:

Table 28 Echo Before words from the corpus

| S. No. | Echo Before | Transliteration | Tag | Gloss |
|--------|-------------|-----------------|----------|--------------|
| 1 | अदला | aḍḍala | RD_ECH_B | Interchange |
| | बदली | Baḍḍali | N_NN | |
| 2 | आसे | ase | RD_ECH_B | Close by |
| | पास | pase | N_NST | |
| 3 | अगल | aḡaḡal | RD_ECH_B | Side by side |
| | बगल | baḡaḡal | N_NN | |

³⁵ <http://sanskrit.jnu.ac.in/ilciann/index.jsp>

The examples shown in the table above are almost similar in form for both Hindi and Bhojpuri and there can be more such examples which are not part of present corpus.

3.5 ISSUES IN POS TAGGING

Digitization of a less resource language is full of challenges. The very first in the queue, is the availability of the data in the desired domain and format. Due to this scarcity of machine readable data the corpus is created on the web drawn data. Though Bhojpuri has no prescribed standard variety to be followed uniformly, in written form all over the world, the so called standard dialect with *rauwa* and *bAnI* construction is accepted for this purpose. The advantage with the online text is that, most of columns, new and blogs maintain this so called standard which brings uniformity in the data. Bhojpuri is an ergative less and classifier rich language. Being a spoken tradition for centuries, there are variations in pronunciation which is also traced to the writing system of the language. The different realizations of the same lexicon and their correct categorization is as challenging as finding out the homophonous words in the data and differentiating their meanings contextually. Particles are mostly inflected with the head categories. Their floating nature and occurrence within a single phrase, though not so big a problem at this level of manual tagging. But these do generate high degree of ambiguity in automated tagging and disturbs the tagger's performance. It might also seek significant attention at other levels of annotations like chunking, parsing, etc.

3.5.1 Challenges in Manual Tagging

The issues found in manually tagging the corpus has been divided into different sub sections based on the nature and types ambiguity. These are namely- unidentified tokens and rare occurrences, dialectal variation, inflected categories, homophones and different realization (of one lexeme).

3.5.1.1 Unidentified Tokens and rare occurrences

3.5.1.1.1 Unidentified Tokens

While tagging the annotator came up with some words which were not encountered before and their categorization was difficult. In case of such words, the guideline suggests to keep them

under the default category. But from the context, and discussions the words were deciphered with the following meaning, as listed in table 29 below:

Table 29 List of unidentified words

| S. No. | Unidentified Words | Gloss | Tags |
|--------|--------------------|---------------------|-------|
| 1 | हिरऊ (hirəu) | Qualifier | JJ |
| 2 | एने (ene) | here and there | N_NST |
| | ओने (one) | | N_NST |
| 3 | रोगान (rogan) | name of some entity | N_NN |
| | शेषान (ʃeʃan) | | N_NN |
| 4 | टिकैत (tikɛtə) | name of a race | N_NN |
| 5 | गँवे (gəmvɛ) | to every village | N_NN |
| | गँवे (gəmvɛ) | | N_NN |

The list of items included in the table was sought for their meaning and the decision was finally made on the basis of context in which it appears in the corpus. The corpus data gives the item 1 in the table as a qualifier to the noun ‘well wishing’ (*mangalkAmanA*), *ene-one* means ‘here and there’ 2nd, 3rd *rogAn-seshAn* stands for some object of entity, 4th *Tikait* is deciphered to be the name of some race and the 5th *ga~Mve ga~Mve* is formed with the process of reduplication in Bhojpuri meaning ‘to every village’.

[68] लोकप्रियता/N_NN क/PSP साल/N_NN होखे/V_VM एकरा/PR_PRP खातिर/PSP

हमार/PR_PRP हिरऊ/JJ मंगलकामना/N_NN

IPA: lokəprɪjəʈa kə salə hokʰe ekəra kʰaʈɪrə həmarə hirəu məngələkaməna

Translation: The coming year brings you popularity, it is my heartiest wish.

[69] ढेर/QT_QTF एने/N_NST ओने/N_NST भागब/V_VM त/RP_RPD गति/N_NN

खराब/JJ हो/V_VM जाई/V_VAUX

IPA: d̪herə ene one bʰagəb̪ə gət̪i kʰərabə ho jai

Translation: If I take more cuts, the speed will go slow.

[70] हम PR_PR रटीला V_VM रोजे N_NST रोगान N_NN शेषान N_NN , RD_PUNC

रोगान N_NN शेषान N_NN

IPA: həmə rət̪ilə rod̪ʒe roɡənə ʃeʃənə, roɡənə ʃeʃənə

Translation: I am learning rogan-sheshan, rogan-sheshan daily, by heart.

[71] टिकैत N_NN का PSP बारे N_NN में PSP अधिका QT_QTF लोग N_NN इहे

DM_DMD जानत V_VM होखी V_VAUX कि CC_CCS ई DM_DMD कवनो DM_DMI खास

JJ जाति N_NN के PSP उपनाम N_NN ह V_VM

IPA: t̪ikət̪ə ka bare m̪ẽ aḏʰika logə d̪ʒənət̪ə hokʰi ki i kəvəno kʰasə d̪ʒət̪i ke upənəmə hə

Translation: Most people know only this much about tikait that it might have been a mickname of some special race.

3.5.1.1.2 Rare Occurrences

‘*katabyoMt*’ is one example of rarely found words falling under this category. In the present corpus the word ‘*katabyoMt*’ is found to function as a qualifier though the meaning is still not very clear. On the basis of the context and its function in the string, *katabyoMt* has been tagged as an adjective, invariably.

3.5.1.2 Dialectal Variation

Dialectal variations, here, refers to the words which are used in either variety of language, like the word *bhA* and *San*.

3.5.1.2.1 ‘bhA’ (as coordinator)-

bhA has been least heard in the spoken language and are present in the corpus repeatatively. It can also be put under the category of rare occurances as used only in some particular variety of Bhojpuri and not known too much of the speakers. From the phrases like ‘*rAt bhA dine*’

meaning ‘night or day’ this word seems to function as a coordinator. The examples below are justifying this fact:

- [72] कवनो/DM_DMI फिलिम/N_NN मुसलमान/N_NN भा/CC_CCD ईसाईयन/N_NN
के/PSP खिलाफ/N_NN बनल/V_VM रहीत/V_VAUX त/RP_RPD अबले/N_NST पूरा/JJ
देश/N_NN में/PSP बवाल/N_NN मच/V_VM गइल/V_VAUX रहीत/V_VAUX

IPA: kəvəno p^hɪlɪmə musələmanə b^ha isaiyənə ke k^hɪlap^hə bənələ rəhɪtə t̪ə abələ pura ðeʃə
mɛ̃ bəvalə məʃt̪ə gəɪlə rəhɪtə

Translation: If some movie would have been made after Muslims or Christians, would led to the riots throughout the country.

- [73] भोजपुरी/N_NNP से/PSP बस/RP_RPD अतने/DM_DMR नाता/N_NN कि/CC_CCS
जनम/N_NN वाला/PSP गाँव/N_NN भा/CC_CCD जिला/N_NN भोजपुरिया/JJ इलाका/N_NN
के/PSP रहुवे/V_VM

IPA: b^hodʒəpuri se bəsə aʈnə nata kɪ dʒənəmə vala gãvə b^ha dʒɪla b^hodʒəpurɪja ɪlaka ke
rəhuve

Translation: He has just this much bonding with Bhojpuri that he was born in the Bhojpuri speaking district and state.

3.5.1.2.2 ‘san’(as an auxiliary or particle) –

‘san’ is found only in the standard variety of Bhojpuri and is not present in the westerns dialects and Bhojpuri speaking regions. This appears as the final entity at the end of the verb phrase like V_VM+ V_VAUX +san. It was difficult to decide for ‘sa~M’ or ‘San’ as in ‘rahat rahalan san’ means ‘used to live’, whether to keep under auxiliary or tag as a discourse particle. Though it occurs particularly after the auxiliary verb but ingerited with minus (-) TAM feature and an auxiliary do have some TAM agreeing to the argument of the verb. Therefore, the dicision was made that it functions as particle with no concrete meaning, therefore, it was better to keep it under default particles only. For example:

[74] दरबार/N_NN पांच/QT_QTC घंटा/N_NN से/PSP बेसी/QT_QTF बतिया/V_VM
लिहें/V_VAUX सन /RP_RP

IPA: ðərəbarə pā̃ʃə ɡʰə̃ʈa se besi bətija lihē sənə

Translation: The darbar is running for more than five hours.

[75] सुरसा/N_NNP जइसन/DM_DMR मुँह/N_NN फइलवले/V_VAUX खड़ा/V_VM
कुछ/QT_QTF भारीभरकम/JJ समस्यो/N_NN रहली/V_VM सन/RP_RPD

IPA: surəsa dʒəisənə mūhə pʰəiləvələ kʰə̃ʈa kuʃʰ bʰaribʰərəkəmə səməsyo rəhəli sənə

Translation: there were also some bigger troubles like Sursaa.

3.5.1.3 Inflected categories

Bhojpuri morphology differs greatly from Hindi in certain constructions. Considering the case of default particles, in Hindi the emphatic particles like ‘hi’, ‘to’ and ‘bhi’ are independent lexemes which either follow or precedes the category that it emphasizes or focuses. Whereas, Bhojpuri have a different case. In Bhojpuri, emphatics are sometimes an independent or separate unit but most of the times the host category are inflected for these particles and forms a new kind of word. See the following examples (from Singh, 2014) for a clear picture:

| | | | |
|----|----------|------|-----------------|
| 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 |

Based on this some example from the corpus are enlisted in the following table along with the categories to which the newly coined word belongs.

Table 30 Word formations in Bhojpuri with inflected particles.

| S. NO. | Words | Tag | Token | Gloss | Default | Gloss |
|--------|---|-------|---------|--------------------------|---------|--------------------|
| 1 | युपीए (jupie) | N_NNP | यूपी | Uttar Pradesh (place) | ए | ही (only) |
| 2 | भोजपुरिए (b ^h oɖʒəpurije) | N_NNP | भोजपूरी | Bhojpuri (language) | ए | ही (only) |
| 3 | कसाबो (kəsabo) | N_NNP | कसाब | kasAb (proper name) | बो | भी (also) |
| 4 | जिभिए(ɖʒib ^h ie) | N_NN | जीभ | tongue | ए | ही (only) |
| 5 | अबहींए (abahIME) | N_NST | अबहीं | Now | ए | ही (just) |
| 6 | अभिए (abhie) | N_NST | अभि | Now | ए | ही (only) |
| 7 | कबो (kabo) | N_NST | कबहीं | Somemtime | ओ | भी (indefinite) |
| 8 | बतइबो (bataibo) | V_VM | बताइब | Tell | ओ | भी (also) |
| 9 | चढईबो (caDhalbo) | V_VM | चढाइब | Offer | ओ | भी (also) |
| 10 | होइए (hoie) | V_VM | होइ | will be | ए | ही (only) |
| 11 | शुरूए (shurue) | V_VM | शुरू | Start | ए | ही (only) |
| 12 | उतरिए (utarie) | V_VM | उतरि | climb down | ए | ही (only) |
| 13 | करीबिए (karIbie) | JJ | करीबी | Close | ए | ही (only) |
| 14 | जल्दिए (jaldie) | RB | जल्दी | Quite early | ए | ही (only) |
| 15 | एकदमें (ekadameM) | RB | एकदम | Completely | ए | ही (only) |

| | | | | | | |
|----|-----------------|--------|------|-----|----|------------|
| 16 | दूइए (duie) | QT_QTC | दुई | Two | ए | ही (only) |
| 17 | एकहू (ekahUM) | QT_QTC | एक | One | हू | ही (only) |
| 18 | दूगो (dUgo) | QT_QTC | दू | Two | गो | classifier |
| 19 | नाहिंए (nAhiMe) | RP_NEG | नहिं | No | ए | ही (only) |

Based on the classification made in the table above, there are six possible categories found which are capable of inflecting for a particle.

3.5.1.3.1 Inflected Noun

In the above table 30, serial number 1 to 7 is the examples of nouns inflected with particles. All the three categories of noun namely proper, common and temporal can inflect for particles in order to form a new word which is the emphasized form of the same original word from that category. For example, *yUple* (N_NNP) means not from any other place but from U.P.only, in the case of proper noun. Similarly, *jibhie* (N_NN) means by tongue only, in the case of common noun and *abahIME* which mean 'right now only', in the third case, N_NST (temporal nouns).

3.5.1.3.2 Inflected Verb

From among the verbs only main verbs are found to inflect for emphatics. There is not such concept of action verb, any action verb if appear in the construction other than the main verb, like as part of compound or serial verb, will not be able to take this form. S.No. 8 to 12 are examples of verb agglutinated with emphatics where *bataibo* mean will tell u also (due to presence of *bhi* particle) and *shurue* mean from the beginning (due to the presence of *hi* particle).

3.5.1.3.3 Inflected Adjectives

No. 13 shows the word formation with adjectives. *karIbie* literally mean a person who is close only, neither too close but not a distant relative. The particle *hi* is added to the word here, in this example.

3.5.1.3.4 Inflected Adverbs

No. 14 and 15 from the table are from the adverb class of word formation. The adverbs are *jaldie* and *ekadameM* meaning quite early and completely, respectively. Both the adverbs in the examples belong to manner adverb therefore the transformed words are tagged as RB which stands for adverbs, in the tagset.

3.5.1.3.5 Inflected Cardinals

The main property of a cardinal (1,2, 3..) is their absoluteness. But in Bhojpuri cardinals are often inflected for particles either with a default particle or with a classifier. Classifier being the referent does come along with cardinals but as a separate lexeme in Bangla and other languages. Bhojpuri cardinals are on the other hand are glued with these particles and take a new form as shown in the example no. 16, 17 and 18. First two are the examples of default and the last one is of the classifier getting attached to the cardinal leading to the construction like *duie* and *dUgo*. The former *duie* mean ‘only two’ and the later *dUgo* mean the ‘specific two’.

Classifiers are not always part of cardinals, but when they are, the first two cardinals in the number system changes its form from *ek*, *dUi* to *ego*, *dugo*.

3.5.1.3.6 Inflected Negations

The negatives when inflects for particles takes the shape of *nahiMe* from *nahiM* meaning no with the emphasis on it.

3.5.1.4 Homophones

Natural languages are inherited with ambiguities, Bhojpuri morphology also carries such ambiguities where one word form can be interpreted many meanings. Homophones are found very common in the basic Bhojpuri vocabulary. The presence of such words with more than one meaning will make the tagger burdened with the load of disambiguation. It seeks concentration even at the level of manual tagging; the words can be tagged considering the meaning and the context both at the same time. This includes *par*, *lA*, *le*, *ka* and *ke*, *mAre*, *nA*, *jI* etc.

- a) *lA* and *le*- Bhojpuri *lA* and *le* are homophonous as both are used as the postposition meaning ‘for’ in English or ‘ke liye’ in Hindi. And in other cases they also perform the verbal function where *lA* mean ‘to bring’ like ‘lAnA’ in Hindi and *le* mean ‘to take’ same as *lenA* in Hindi, derived from the root verbs *lAnA* and *lenA*, respectively.

- b) ‘*ka*’ and ‘*ke*’- Bhojpuri use of postposition are distinct from that of Hindi. Hindi noun are sometimes inflected with possessive markers and datives as in *usake*, *mujhko* or *mujhe* etc followed by the mainverb. Whereas similar constructions are treated differently in Bhojpuri. Bhojpuri uses *okarA* for *usake*, and *hamnI* or *hamarA* for *mujhe* though these words already include the possessives still the postposition follows them when used in constructing a sentence. For example:

[76] ओकरा के/PSP घरियाँ दी आवा

IPA: okəra ke gʰərɪjã ði avə

Translation: OkarA **ke** ghariyA~M dI AwA

[77] हमनी के/PSP चाह पसन्द बा

IPA: həməni ke tʃahə pəsəndə bə

Translation: hamnI **ke** chah pasand BA

Besides this, at some places, these words are often confused either as a postposition or an auxiliary. But it was found that *ke* and *ka* in Bhojpuri actually belongs to three possible grammatical categories - subordinator, postposition and auxiliary form of verb depending upon the context. For example:

[78] काहे के/CC_CCS सबले मन्जूर रहे

IPA: kahe ke səbəle məndʒʊrə rəhe

Translation: because all agreed to it

[79] ला के/V_VAUX दे दा

IPA: la ke ðe ða

Translation: bring it for him

[80] हम सूरज डूबे के/PSP बादे आइब

I will go only after the sun set

Where, ‘*ke*’ is a subordinating conjunction in example (67), functioning as an auxiliary verb in (68) and as a part of postposition in (69).

- c) *nA* – Bhojpuri *nA* is also a homophonous word which functions both as a negation particle and default particle, based on the context. It has been asked as negation where exhibits the meaning of complete negation is and where there is no definite meaning to the word then it was tagged as a negation particle.

[81] ईसाईयन/N_NN का/PSP बारे/N_NN में/PSP कवनो/DM_DMI गलत/JJ बात/N_NN
ना/RP_NEG कह/V_VM सके/V_VAUX

IPA: isaijənə ke bare mẽ kəvəno gələtə bəʈə na kəhə səke

Translation: Nobody can speak a word against Christians.

[82] बोले/V_VM के/PSP चाही/V_VAUX ना/RP_NEG

IPA; bole ke tʃahi na

Translation: You should have said it!

- d) *par*- Bhojpuri *par* shows similar function as the Hindi *par* meaning both a postposition ‘on’ and a coordinator ‘but’. For example-:

[83] ऑस्ट्रेलिया JJ दू QT_QTC विकेट N_NN का PSP नुकसान N_NN पर PSP रन N_NN
बना V_VM लिहले V_VAUX बावे V_VM

IPA: ɔstrelia dū vɪkətə ka nukəsənə pərə rənə bəna lihəle bəve

Translation: Australia has scored at the loss of two wicket

[84] जा/V_VM सकीला/V_VAUX पर/CC_CCD रूके/V_VM के/PSP बिचार/N_NN
बा/V_VM

IPA: dʒa səkɪlə pərə ruke ke bɪtʃərə bə

Translation: I can go but I am thinking of staying here.

Besides this, there can be another possible occurrence of *par* as noun, which means feather. Feather is called *paMkh* in Hindi, sometimes *par* but very rarely.

- e) *mAre*- Another ambiguous token is *mAre* which act both as verb and intensifier. For example:

[85] सब/N_NN मारे/V_VM जात/V_VAUX रहलन/V_VAUX

IPA: səbə mare dʒatə rəhələnə

Translation: All were about to beat him

[86] सेर/N_NN मारे/RP_INTF दहाड़/V_VM लगल/V_VAUX

IPA: serə mare ðəhaɽe ləgələ

Translation: the lion started roaring aloud.

The above example (85) explains the verbal function and example (86) is the intensifier use of the word *mAr* which is inflected form of the root *mAr*.

There are many other homophones like *A* or *aa* meaning ‘to come’ when functions as a verb and ‘and’ as a conjunction, *calate* which mean both ‘to walk’ as verb and ‘due to’ as reason and many more.

3.5.1.5 Different Realizations of one lexeme

Bhojpuri lexemes often show some variations in pronunciation. A longer oral tradition of Bhojpuri is the main reason of the emergence of this feature. A single word is realised in different forms in the spoken speech which is also observed in texts. These varied realisations can be found to occur with conjunctions, particles and postpositions. It is quiet tricky for the tool to decide the tag for the variations of a token which occurs for the first time in given data. *to* (conjunction), *aur* (conjunction), *san* (particle), *bAki*(conjunction) etc. are carriers of this feature. Some examplar realizations for these words are mentioned in the table 31 below:

Table 31 Different Realizations of the Bhojpuri words

| S.No. | Word | Gloss | Realization 1 | Realization 2 | Realization 3 | Realization 4 |
|-------|----------------|------------------------------|------------------|------------------|------------------|------------------|
| 1 | बाकि (bakɪ) | coordinating conjunction | बाकिर (bAkir) | | | |
| 2 | तो (tɔ) | subordinating conjunction | त (ta) | ता (tA) | | |
| 3 | सन(sən) | dafault particle | सँ (Sa~M) | स (sa) | | |
| 4 | अऊर | coordinating | अ (a) | आ (aa) | अउ (au) | |

| | | | | | | |
|---|------------------|------------------------|-------------|-------------|-------------------|------------------|
| | (और) | conjunction | | | | |
| 5 | नाहीं (nəhim) | nsegration particle | नहीं (nahI) | नाहि (nAhi) | नाँहि (nA~Mhi) | नइखे (naikhe) |

From the above table, we infer that the first example of *bAki* can be realised in two ways, *bAki* and *bAkir* referring to just one entity. Similarly, example two shows three realizations of *to* namely *to*, *ta* and *tA*, all functioning as the subordinator in a given sentence. Third point for *san* also has three realizations as *san*, *sa~M* and *sa* which is a default particle. Last two examples for coordinator and negations are found to be realized in four to five ways. *aUr* for ‘and’ has three more forms as *a*, *A* and *au* except *aUr* and the final negation has been realized in five different word forms as *nAhIM*, *nahl*, *nAhi*, *nA~Mhi* and *naikhe* which stands for the sense of complete negation, formed out of the stem word *nahi* ‘no’ in Bhojpuri.

3.5.1.6 Tagging inconsistency

Though much part of data was tagged using ILCIANN, some initial sections were tagged manually in first two files from blogs. The manual tagging sometimes encounter some typo errors. There were some similar errors found after the first validation of the data which were noticed during the testing session. The tagger report included such errors by generating classes for them. These were enlisted which lead to the second fold of validation of the annotated data. Some of the tagging inconsistency has been mentioned in the table 32 below:

Table 32 Inconsistent tags

| | Tagging Inconsistency | Incorrect Tag | Correct Tag |
|---|----------------------------------|---------------|-------------|
| 1 | CC_CCD_CC_CCS_CC_cCD_V_VAUX_V_VM | CC_cCD | CC_CCD |
| 2 | JJ_V_VM_V_vM | V_vM | V_VM |
| 3 | PSP_V_VAUX_V_vAUX | v_vAUX | V_VAUX |
| 4 | RD_RD_PUNC | RD | RB |
| 5 | PSP_RD_RPD_RP_RPD | RD_RPD | RP_RPD |

The present table shows the typo errors like cCD for CCD, vAUX for VAUX etc. these has been cross checked and the validated before the further process of the training of the tagger started.

4 SVM BASED BHOJPURI TAGGER

4.1 BHOJPURI AND SUPPORT VECTOR MACHINE

The present tagger is the first statistical tagger for Bhojpuri with a representative annotated corpus and tagging scheme following national standards for Indian languages. As already discussed in the introductory chapter that some work is in progress in universities like IIT BHU, Wardha and others but no significant result or module has yet appeared

4.1.1 Why SVM

Indian languages are morphologically rich. This property demand for such a tool which is flexible enough to handle the morphological complexities of the language. So far, SVM has been proved to be an efficient device with its variant, for resolving different kinds of issues like quadratic programming, optimization problem, problem of dual form, soft margin approach for the problem of mislabelled examples, and many more. The relative success history in most of these areas and achieving high accuracy for European as well as Indian languages like Malayalam and Bangali (ranging between 86 to 90%) draws the researcher's attention to take this as the training model for the Bhojpuri corpus.

4.1.2 Efficiency of the tagger

All higher levels of application demands Parts-of-Speech or knowledge base as first steps to proceed with more complex data. At this level, the automatic Parts-of-Speech tagger used, must be very flexible and should be performing equally well on both qualitative as well as quantitative scales. The usefulness of a tagger is calculated in terms of its accuracy and efficiency. Support Vector Machine is capable of producing the accuracy result of 97.11 % and 97.24%, for Wall Street Journal Corpus for English, as reported in Giménez (2006). The LEXESP Spanish Corpus of 106 k words also found to have similar accuracy with the WSJ one. (as cited in Giménez and Màrquez, 2006).

Because of such convincing results, it has also had its spread over Indian languages where the Dravidian language Malayalam and an Indo Aryan Bengali were trained. The initial accuracy results for both the languages were ranging between 86-89% for the former language and reached 86.8% for the latter. Currently, along with 'Bhojpuri' some other languages like Hindi

and Odia are also being tested for the suitability of the model. These experiments are running by the NLP practitioners at JNU, there are no related publication yet as the experiments are in progress.

4.2 DESCRIPTION OF THE TOOL

The tagger presented in the study was introduced for English by Jesús Giménez and Lluís Màrquez in the 4th International Conference on Language Resources and Evaluation (LREC 2004) at Lisbon, Portugal. The work entitled *SVMTool: A general POS tagger generator based on Support Vector Machines*. Since then, the tool has witnessed several revisions within the span of not two years and the detailed manual for using the tagger can be seen in the volume 1.3, released in 2006.

Giménez and Màrquez (2006) states in their manual that the study has been developed under the partial sponsorship of Spanish Ministry of Science and Technology (MCyT's projects), European Commission, and by the Catalan Research Department (CIRIT's consolidated research group). And the C++ version of it is freely downloadable at "<http://www.lsi.upc.es/~nlp/SVMTool/>".

4.2.1 Property of the Tool

The tagger is said to have the following properties, as mentioned in Giménez and Màrquez (2006) :

- 1) The tool has very simple configuration and installation procedure. Training of the tagger is comparatively easy and has very few parameters to fix.
- 2) It is flexible at defining wider feature pattern and justifies with the shape and size of the context.
- 3) It is a robust machine which allows sentence level analysis and other strategies. The soft margin learning algorithm by tuning the parameter C is introduced. The soft margin classification utilizes the following equation

$$\{ z : \langle w, z \rangle + b = 0 \} \quad (ii)$$

Soft margin can be understood as a variant of optimizations problem where the parameter C is used to balance the margin maximization and errors occurred in training as shown in the following figure 9.

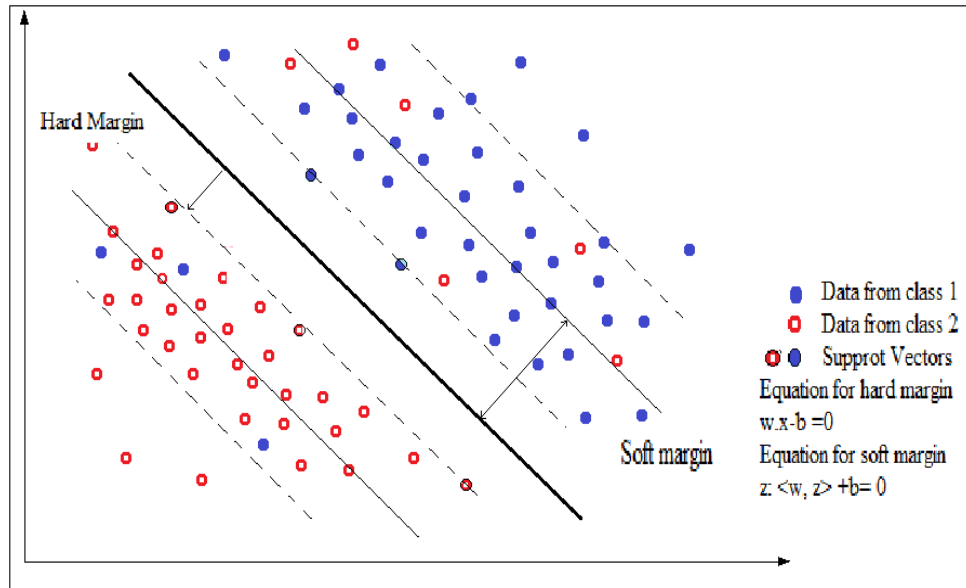


Figure 9 **Soft Margin vs. Hard Margin**

- 4) The ability to learn with very limited language data accounts for its portability. It is a semi-supervised machine and does not require full-fledged knowledge of language prior to the training.
- 5) SVM can work accurately in indefinite feature space with similar level of accuracy.
- 6) The use of linear kernel helps accelerating the tagging speed of the present Perl prototype tagger to 1500 words per second and the C++ version with the speed of 10,000 words per second.

4.2.2 Design of the tagger

The tagger has been designed to have three main components SVMTlearn, SVMTagger and SVMTeval. These components responsible for creating model file for the training data, tagging the test files and evaluation of the results obtained by the tagger, respectively.

4.2.3 Tagger Models

So far, the SVM has been tested for five different models with slight but noticeable differences. The models implemented, are named 0 to 5 which stands for the features contained in them (Giménez, 2006). These models are as follows:

1. *Model 0*- This is the default model trained on one pass scheme. This is a unidirectional tagger which tags either in left to right or right to left direction. Being a one pass, the tokens not disambiguated in advance remains ambiguous.

2. *Model 1*- This disambiguates the unseen context of the previous text. Hence, known as the second pass schema.
3. *Model 2*- The POS features are not the concern of this model. Working on the one pass it calls for the second pass to review the tagging results.
4. *Model 3*- This considers the unambiguous words from the annotated corpus for training the tagger.
5. *Model 4*- This handles the errors caused by the unknown words. This is done by creating different folders and generating dictionary. Before tagging the tagger looks into the dictionary and the words found in any folder but not in the rest of the corpus are marked unknown.

4.2.4 Configuration

There are many options like verbose, sliding windows and feature sets etc. For more on this please visit Giménez³⁶. The verbose throughout the experiment, was set as medium throughout the experiment. The tagger was set to perform a two pass, unidirectional tagging from right to left.

The tagger employs different tagging strategies at different tagging levels like strategy for running in one or two pass, choosing the direction for the tagger, filtering out thresholds for known and unknown tokens, prediction of POS to be tagged, backup lexicon, lemma lexicon etc as part of internal process done by the tagger, as the tagging continues.

4.2.5 Format of the data

The training data is formatted as one word per line/sentence. This column wise data setting further contains the tag in the second column and metadata or other relevant information (if necessary) in next following columns. The sentence boundary markers and symbols are also treated as unambiguous tokens. Both the columns of the content word and tag (including other columns, if present) must be separated with a single space in between. Taking example of a sentence from the data itself, we get the following setting:

³⁶ *ibid*

Exemplar sentence: लोग अपनी माईभाखा खातिर बहुत कुछ करेला

Gloss: logə əpəni maib^hak^ha k^hat̪irə bəhuṭə kuɟ^hə kərela

Free translation: Speakers of the language do a lot for preserving their mother tongue

Formatted form:

लोग N_NN

अपनी PR_PRF

माईभाखा N_NN

खातिर PSP

बहुत RP_INTF

कुछ QT_QTF

करेला V_VM

In the following example, the first column contains words tokens from the exemplary Bhojpuri sentence, the tag for each word makes the second column and the gloss of the words in English is written in the third column which may not necessarily present. The unannotated test file has to be formatted accordingly because the tagger processes the tokens sentence by sentence in a linear fashion.

4.2.6 Tagger Architecture

The tagger under development has the following system architecture.

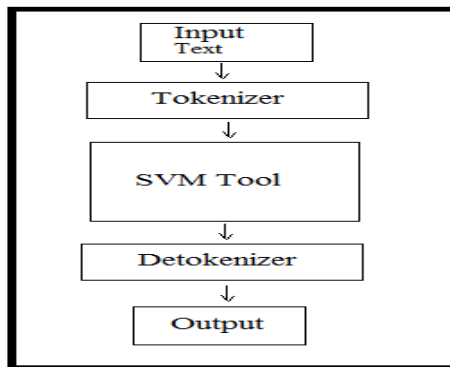


Figure 10 Tagger Architecture

The tagger under development has the following system architecture. The string of sentences from the corpus is the input to the tagger which processes the input at three levels, tokenization, tagging and detokenization. The input data is first sent to the tokenizer where each sentence is broken into its tokens. The tagger processes tokenized data and the tagged output is sent to the detokenizer. The detokenizer here conjoins all the tagged tokens, back to the string and the output is generated. The codes to tokenizer and detokenizer were written in java and the tagger has been hosted in the 'jsp' format on the cloud. The tagger trained in the present experiment can be found on the website of Sanskrit website under the given web link <http://sanskrit.jnu.ac.in/bhopos/index.jsp>

4.3 TRAINING THE TAGGER

The validated and annotated corpus serves as the training model for the tool. The 90k annotated corpus of Bhojpuri as mentioned in the previous chapter (3) is the input upon which the model file has been created. The assistance for the training section has been extracted from the SVM in C by Vapnik which was implemented by Thorsten Joachims for training the models (as cited in Giménez³⁷). The training process is part of SVMtlearn module, as in this the tool learning is going in form of creating a model file of further tagging/testing.

The experiment was carried out in two phases. The first phase was trained upon the model file of 30k with exactly 30,391 tokens. For the second phase, the training was extended to 90k with 89,997 tokens with the modified tagset described in the previous chapter. The training set includes miscellaneous data from all five domains. The gold set was created out of the training set after undergoing three further manual revisions for making it error free. The training set covers wide range of disciplines for the robustness of the tool including four sets of 1000 sentences from blogs and 2 sets from miscellaneous domains. The time taken by the tagger for training of 90k tokens was much higher than took for tagging a set of file, as the tagging was done in just a few seconds. The model file has been trained in both the directions for whole training data. After the completion of the training process the tagger displays the detailed mapping of the internal tagger process, which has been attested in figure 11 below:

³⁷ ibid

```

----- KNOWN WORDS... [MODE = 0 :: DIRECTON = RL] -----
FEATURE EXTRACTION from </home/sanskrit/svmtool/models/bho/BHO.REV > onto <
/home/sanskrit/svmtool/models/bho/BHO.M0.RL.SAMPLES >
.....10000.....20000.....30000.....40000.....10000.....89997 WORDS [DONE]
MAKING MAPPING from </home/sanskrit/svmtool/models/bho/BHO.M0.RL.SAMPLES>... [DONE]
FILTERING MAPPING (minfreq = 5 :: maxmapsize = 10000)
WRITING MAPPING </home/sanskrit/svmtool/models/bho/BHO.M0.RL.SAMPLES.MAP>... [FREQ = 0] :: [SIZE = 9655] [DONE]
.....10000.....20000.....30000.....40000.....MAPPING DATA FEATURES from
</home/sanskrit/svmtool/models/bho/BHO.M0.RL.SAMPLES> to
</home/sanskrit/svmtool/models/bho/BHO.M0.RL.SAMPLES.DSF>.....44796 [DONE]
LEARNING SVM MODELS... [C = 0]
..CC_CCD..CC_CCS..CC_cCD..DM_DMD..DM_DMI..DM_DMQ..DM_DMR..JJ..N..NN..N..NNP..N..NST..PR_PRC..PR_PRF..PR_P
RI..PR_PRL..PR_PRP..PR_PRQ..PSP..QT_QTC..QT_QTF..QT_QTO..RB..RD..RD_ECH..RD_ECH_B..RD_PUNC..RD_RPD..RD_
UNK..RP..CL..RP_INJ..RP_INTF..RP_NEG..RP_RPD..V_VAUX..V_VM..V_vAUX..V_vM [DONE]
BENCHMARK TIME: 60.0000
----- UNKNOWN WORDS... [MODE = 0 :: DIRECTON = RL] -----
FEATURE EXTRACTION from </home/sanskrit/svmtool/models/bho/BHO.REV > onto <
/home/sanskrit/svmtool/models/bho/BHO.UNK.M0.RL.SAMPLES >
.....10000.....20000.....30000.....40000.....50000.....60000.....70000.....80000.....89997 WORDS
[DONE]
MAKING MAPPING from </home/sanskrit/svmtool/models/bho/BHO.UNK.M0.RL.SAMPLES>... [DONE]
FILTERING MAPPING (minfreq = 5 :: maxmapsize = 10000)
WRITING MAPPING </home/sanskrit/svmtool/models/bho/BHO.UNK.M0.RL.SAMPLES.MAP>... [FREQ = 5] :: [SIZE = 1292]
[DONE]
..MAPPING DATA FEATURES from </home/sanskrit/svmtool/models/bho/BHO.UNK.M0.RL.SAMPLES> to
</home/sanskrit/svmtool/models/bho/BHO.UNK.M0.RL.SAMPLES.DSF>.....2492 [DONE]
LEARNING SVM MODELS... [C = 0]
..CC_CCD..CC_CCS..DM_DMD..DM_DMI..DM_DMQ..DM_DMR..JJ..N..NN..N..NNP..N..NST..PR_PRC..PR_PRF..PR_PRI..PR_P
RL..PR_PRP..PR_PRQ..PSP..QT_QTC..QT_QTF..QT_QTO..RB..RD..RD_ECH..RD_RPD..RD_SYM..RD_UNK..RP_INJ..RP_INTF..RP
_NEG..RP_RPD..V_VAUX..V_VM..V_vAUX..V_vM [DONE]
BENCHMARK TIME: 4.0000
READING MAPPING FOR KNOWN WORDS...</home/sanskrit/svmtool/models/bho/BHO.M0.RL.SAMPLES.MAP>
READING SVM-MODELS FOR KNOWN WORDS...
CC_CCD..CC_CCS..CC_cCD..DM_DMD..DM_DMI..DM_DMQ..DM_DMR..JJ..N..NN..N..NNP..N..NST..PR_PRC..PR_PRF..PR_PR
I..PR_PRL..PR_PRP..PR_PRQ..PSP..QT_QTC..QT_QTF..QT_QTO..RB..RD..RD_ECH..RD_ECH_B..RD_PUNC..RD_RPD..RD_U
NK..RP..CL..RP_INJ..RP_INTF..RP_NEG..RP_RPD..V_VAUX..V_VM..V_vAUX..V_vM [DONE]
MERGING MAPPING and MODELS FOR KNOWN WORDS...
WRITING MERGED MODELS FOR KNOWN WORDS...</home/sanskrit/svmtool/models/bho/BHO.M0.RL.MRG>
READING MAPPING FOR UNKNOWN WORDS...</home/sanskrit/svmtool/models/bho/BHO.UNK.M0.RL.SAMPLES.MAP>
READING SVM-MODELS FOR UNKNOWN WORDS...
CC_CCD..CC_CCS..DM_DMD..DM_DMI..DM_DMQ..DM_DMR..JJ..N..NN..N..NNP..N..NST..PR_PRC..PR_PRF..PR_PRI..PR_PR
L..PR_PRP..PR_PRQ..PSP..QT_QTC..QT_QTF..QT_QTO..RB..RD..RD_ECH..RD_RPD..RD_SYM..RD_UNK..RP_INJ..RP_INTF..RP_
NEG..RP_RPD..V_VAUX..V_VM..V_vAUX..V_vM [DONE]
MERGING MAPPING and MODELS FOR UNKNOWN WORDS...
WRITING MERGED MODELS FOR UNKNOWN WORDS...</home/sanskrit/svmtool/models/bho/BHO.UNK.M0.RL.MRG>
BENCHMARK TIME: 246.0000

TERMINATION... [DONE]

```

Figure 11 Training of the model file

4.3.1 Cautions for the training file

Several kinds of errors might remain in the training file after validation and formatting. One has to be cautious about the proper formatting of the data. Some common errors in the file might

prove to be a nuisance during the training process. Some of them encountered in due course are shared below:

- No numbering must be put before the word tokens.
- There should be no blank line or extra space between the two line or words/tokens.
- There should be a proper one to one, parallel alignment of words and tags. If the tagger finds either of the elements missing in a single line, might cause system failure.
- The extra information and columns except the two (for word and token) must be avoided during training.

4.4 TESTING THE TAGGER

SVMTagger module is synonymous to testing. After the successful training of the model file, the tagger was set at the task of tagging the test files. The testing is done in two phases- initial testing and currently running. Testing was done on seen and unseen data for both the test phrase.

The initial testing was done in January on the model file trained with 30391 tokens from blogs domain. Whole contemporary training set of 30391 was tested as part of seen data (part of training set). The unseen data (other than the training set) was taken from miscellaneous domain with 9,863 tokens. This was later on served as the seen data set for the second phase. Current phase of testing has been running on the model file trained with 90k tokens. The first test file (seen data) was chosen from miscellaneous domain with 9,863 tokens with 535 sentences and the second file from politics domain. This data set from politics was also selected for the comparative study of Hindi and Bhojpuri tagger, which is discussed later in this chapter.

The tagger generated display during testing has been captured in form of screenshot as shown in figure 12. The time taken by tagger has been noted in fractions of a second. From start up to tagging to extraction and processing, all done within 1.53 seconds for one file (which is more or less similar for all test files), as visible in the given figure.

```

READING DICTIONARY </home/sanskrit/svntool/models/bho/BHO.DICT>...
[DONE]
READING MODELS < DIRECTION = right-to-left :: MODEL = ambiguous context >
(1) READING MODELS (weights and biases) FOR KNOWN WORDS </home/sanskrit/svntool/
models/bho/BHO.M0.RL.MRG>...
(2) READING MODELS (weights and biases) FOR UNKNOWN WORDS </home/sanskrit/svntoo
l/models/bho/BHO.UNK.M0.RL.MRG>...
TAGGING < DIRECTION = right-to-left >
.....64 sentences [DONE]
=====
START-UP: 0.3600 secs
=====
TAGGING: 1.1700 secs
=====
F.EXTRACTION: 0.4800 secs
SVM: 0.2500 secs
PROCESS: 0.4400 secs
=====
OVERALL = START-UP + TAGGING = 0.3600 secs + 1.1700 secs = 1.5300 secs
=====

```

Figure 12 Screen shot of the tagger generated tagging display

Tagger tagged files were later on validated manually for the inspection of the type and pattern of errors generated by the tagger. Though, similar description is also updated by the tagger during evaluation process, they only allow access to the nature of ambiguity and its classes (the class tags). Manual validation brings the researcher, directly to the issues and the nature of issue. The tagging issues are discussed in the following section. These discrepancies are helpful in the analysis and further development of the tagger.

4.5 EVALUATION OF THE TAGGER

The SVM Tagger's evaluation of the performance is calculated in terms of the accuracy of the comparison between tagger generated tagged output and corresponding gold annotated corpus (Gimenez and Marquez, 2006).

4.5.1 Gold corpus

First of all, the gold corpus is created out of the training set. The gold corpus is a thoroughly revised training data which account for 100% accuracy (ideal state). The gold corpus created in this experiment, went through two fold validation of the training sets and evaluated in the following stages:

1. Test result evaluated for seen data (part of training set) in the initial phase (eval A, hereafter).
2. Test result evaluated for unseen data (other than training set) in the initial phase (eval B, hereafter).
3. Test result evaluated for seen data (part of training set) in the current phase (eval C, hereafter).
4. Test result evaluated for unseen data (other than training set) in the current phase (eval D, hereafter).

4.5.2 SVMTEval

The next module ‘SVMTEval’ stands for the evaluation of the tagger generated output. SVM is capable of producing a detail analysis of the result obtained. The evaluation result has been divided in various sub sections like tokens and their accuracies, accuracy as per level and class of ambiguity, overall accuracy etc. These are as follows:

- a) The first section is a simple listing of different kinds of tokens including known, unknown, and ambiguous and their accuracies followed by the total no. of hits and trials for each known and unknown class of tokens.
- b) The second section is the accuracy calculated ‘per class of ambiguity’. The tagger, here, constructs classes for levels of ambiguity depending upon the hits and trials attempted by the tagger. These witnessed a shift in the ambiguity ratio as moving from one level to another. The ambiguity as per the level, postulated 6 classes for both the sets tested in the current phase.
- c) Next section is about ‘accuracy per class of ambiguity’. This entails a detailed classification of the ambiguous tags. There are 222 and 230 classes postulated by the tool for unseen and seen test set, respectively. More about it will be discussed in the next section 4.5.2.1.
- d) This second last section is labelled as ‘accuracy per Parts-of Speech’. This gives a description of average accuracy acquired by tagger for each and every tag.
- e) The last section is the announcement of the overall accuracy achieved by the tool for the present test sets.

Tagger generated evaluation report has been enlisted in the appendices, at the end of the dissertation.

4.5.2.1 Accuracy as per class of ambiguity

For SVM tool, the ambiguity reported by the tagger is inversely proportional to the complexity in morphology. Indian languages are rich in morphology as compared to English. Giménez and Màrquez (2006) also confirms to the hypothesis that morphologically rich languages are found to have lesser degree of ambiguity as the SVM Tool trained as part of their study claimed 93.91% ambiguity for Wall Street Journal, corpus of English and 95.04% for LEXESP, corpus of Spanish language.

The ambiguity classes are formed considering the possible tags given to each word token. A list of all unique ambiguity classes can be found in appendices on the basis of which these classes are further classified into the following categories:

- (i) *Major classes*: The major class in this categorization includes all unique classes of ambiguity found in all the evaluation reports generated by the tagger. They are 263 in numbers, at the present stage. A list of all ambiguity classes can be found from appendices.
- (ii) *Classes with erroneous tags*: The annotated corpus is semi-automatically tagged corpus with manual validation. Therefore, it is quite possible to have some typo errors like misspelled tags, as N_NM for N_NN i.e. tag for a common noun. There are five such inconsistencies found with a total of only nine occurrences in both the tested data sets. The inconsistent tag list has already shown in section 3.5.1.6 in chapter three.
- (iii) *Classes with single tag*: This class includes all tags as presented in the POS tagset.
- (iv) *Classes with two tags*: This includes those words which have at least two possible tags. Let say, any word which functions both as a noun and a verb will fall in this category. For example, CC_CCD and V_VM will refer to the words which occur both as a coordinator and a verb in Bhojpuri. 'aa' in Bhojpuri means 'and' when used as a conjunction and in verbal sense it means 'to come'.

(v) *Classes with three tags*: Similarly, words with three possible tags are listed under this category. For example, PSP, PR_PRQ and N_NNP. Bhojpuri ‘*ke*’ can function as a genitive ‘*oke*’, interrogative (+ gen) ‘*ke (+ke)*’ where the first *ke* means ‘who’ and the second *ke* gives a genitive meaning. And third *ke* is a proper noun. It can be initial of a name like ‘*ke srivAstava*’ etc.

(vi) *Classes with more than three tags*: This category ranges up to maximum six possible function tags for a given word/token as listed by the tagger. Though no such instances noticed during manual annotation. Up to four usage can be accounted from the corpus.

4.5.3 Evaluation of the Tests

The experiment was done in two phases, as mentioned in the beginning of this chapter. Currently, the tagger has been trained on the model file of 90k tokens. The accuracy of the tagger was calculated in four stages for both the testing phases namely, eval A and eval B for the 30k model file; and eval C and eval D for 90k model file (as already mentioned above). On the basis of tests, following results have been found:

4.5.3.1 Evaluation of the first phase (model file with 30k tokens):

The initial phase of the experiment was conducted in January of 2015 with 30k trained tokens. The tagger was tested on the seen data (part of training set) with 30,391 tokens and unseen data (other than training set) set with 9,863 tokens. Tested files were evaluated on the gold corpus created out of the tested data after validation. This leads to the following accuracy results,

4.5.3.1.1 Test result evaluated for seen data (part of training set) in the initial phase (eval A, hereafter).

The tagger tested on set of 30k tokens was found to have all known and no unknown token. There were 13619 ambiguous tokens which makes 41.8% for the test sample. Based on token results, the tagger showed the MFT results of 91.8%. The overall accuracy reached 87.3910% with the average ambiguity of 1.8 tags per token. As can be inferred from the table below:

Table 33 Test result for seen data of 30k tokens (initial phase)

| | Known | Unknown | Ambiguity | MFT | Tagger Output |
|--|-------|---------|-----------|-----|---------------|
| | | | | | |

| | | | | | |
|---------------------------------|-------|---|-------|-------|-------|
| Total Tokens (30391) | 30391 | 0 | 13619 | 27912 | |
| Accuracy (%) | 100 | 0 | 41.81 | 91.84 | 87.39 |

The mentioned accuracy result is supported by the screenshot for the same, taken at the time of testing (see fig. 13).

```
* testset (predicted) = [/home/sanskrit/svmtool/bho.eval]
* =====
EVALUATING </home/sanskrit/svmtool/bho.eval> vs. </home/sanskrit/svmtool/bhogold
.txt> on model </home/sanskrit/svmtool/models/bho/BHO>...
.....10000.....20000.....30000...30391 tokens [DONE]
* ===== TAGGING SUMMARY =====
=====
#TOKENS          = 30391
AVERAGE_AMBIGUITY = 1.8144 tags per token
* -----
-----
#KNOWN           = 100.0000% -->          30391 / 30391
#UNKNOWN         =  0.0000% -->           0 / 30391
#AMBIGUOUS       = 44.8126% -->        13619 / 30391
#MFT baseline    = 91.8430% -->        27912 / 30391
* ===== OVERALL ACCURACY =====
=====
              HITS          TRIALS          ACCURACY          MFT
* -----
              26559          30391          87.3910%          91.8430%
* -----
=====
sanskrit@sanskrit-ThinkCentre-M57e:~/svmtool/bin$
```

Figure 13 Screenshot of test on 30k seen data

4.5.3.1.2 Test result evaluated for unseen data (other than training set) in the initial phase (eval B, hereafter).

The tagger tested on unseen data of 9,863 tokens, which was found to have 7724 known and 2139 unknown token that makes 78.3% and 21.6 %, respectively. The ambiguous tokens in the set were 41.62% with 4105 ambiguous tokens. The MFT, based on this the tagging description, was calculated to be 68.2% with the average ambiguity of 7.8 tags per token. The overall accuracy for unknown data was found to reach 74.24%, as shown in the table below:

Table 34 Test result for unseen data of 10k token (initial phase)

| | Known | Unknown | Ambiguous | MFT | Tagger Output |
|---------------------------------|-------|---------|-----------|------|---------------|
| Total Tokens (30391) | 7724 | 2139 | 4105 | 6734 | |

| | | | | | |
|---------------------|-------|-------|-------|-------|-------|
| Accuracy (%) | 78.31 | 21.68 | 41.62 | 68.27 | 74.24 |
|---------------------|-------|-------|-------|-------|-------|

The mentioned report is supported by the screenshot for the same, taken at the time of testing (see fig. 14)

```
* testset (predicted) = [/home/sanskrit/svntool/bhotest.out]
* =====
EVALUATING </home/sanskrit/svntool/bhotest.out> vs. </home/sanskrit/svntool/bhog
old.txt> on model </home/sanskrit/svntool/models/bho/BHO>...
.....9863 tokens [DONE]
* ===== TAGGING SUMMARY =====
=====
#TOKENS          = 9863
AVERAGE_AMBIGUITY = 7.8146 tags per token
* -----
#KNOWN           = 78.3129% -->          7724 / 9863
#UNKNOWN         = 21.6871% -->          2139 / 9863
#AMBIGUOUS       = 41.6202% -->          4105 / 9863
#MFT baseline    = 68.2754% -->          6734 / 9863
* ===== OVERALL ACCURACY =====
=====
                HITS          TRIALS          ACCURACY          MFT
* -----
                7323          9863          74.2472%          68.2754%
* =====
sanskrit@sanskrit-ThinkCentre-M57e:~/svntool/bin$
```

Figure 14 Screenshot of test on 10k unseen data

From the above section we infer that the overall accuracy of the tagger during initial testing phase was 87.3 % for the known set of data which was reduced to 74.24 % for unfamiliar or unknown data set. This fall in accuracy is due to the increased number of unknown token in the file. Where the first set had no unknown token the later set encountered 2000 approx. unknown tokens. A decrement of 0.2% was noticed in the ambiguity. The ambiguity for the seen data was reported to be 41.8 % which was reduces to 41.6 % for unseen data.

The limited training data at this level, results in higher degree of unknown tokens for the test set which has been overcome, to some extent, in the next phase. Therefore, from the above calculations we get that the accuracy for the SVM based Bhojpuri tagger was, initially, 87.39%.

4.5.3.2 Evaluation of the current phase (model file with 90k tokens):

4.5.3.2.1 Test result evaluated for seen data (part of training set) in the current phase (eval C, hereafter).

The tagger in the current phase was, firstly, tested on sample of 9,863 tokens selected from among the training set. As a result, there were all known and no unknown tokens found on testing. There were 5057 ambiguous tokens which makes 51.2 % of the test sample. Based on token results, the tagger showed the MFT results of 94.24%. The overall accuracy was calculated to be 88.6242% with the average ambiguity of 1.7 tags per token. The same has also been mentioned in the table below:

Table 35 Test result for seen data of 10k tokens (current phase)

| | Known | Unknown | Ambiguity | MFT | Tagger Output |
|-----------------------------|-------|---------|-----------|-------|---------------|
| Total Tokens (30391) | 9863 | 0 | 5057 | 9295 | |
| Accuracy (%) | 100 | 0 | 51.27 | 94.24 | 88.62 |

The above mentioned test result is supported by the screenshot for the same, taken at the time of testing (see fig. 15)

```

=====
EVALUATING </home/sanskrit/svmtool/bho.eval> vs. </home/sanskrit/svmtool/bhogold
.txt> on model </home/sanskrit/svmtool/models/bho/BHO>...
.....9863 tokens [DONE]
* ===== TAGGING SUMMARY =====
=====
#TOKENS          = 9863
AVERAGE_AMBIGUITY = 1.7249 tags per token
* -----
#KNOWN           = 100.0000% -->          9863 / 9863
#UNKNOWN         =  0.0000% -->           0 / 9863
#AMBIGUOUS       = 51.2724% -->        5057 / 9863
#MFT baseline    = 94.2411% -->        9295 / 9863
* ===== OVERALL ACCURACY =====
=====
                HITS          TRIALS          ACCURACY          MFT
* -----
                8741          9863          88.6242%          94.2411%
* =====

```

Figure 15 Screenshot of test on 10k seen data

4.5.3.2.2 Test result evaluated for unseen data (other than training set) in the current phase (eval D, hereafter).

The sample data for this section was chosen from the politics domain. A set of 105 sentences with 1967 tokens have been tested upon. There were 1698 known and 269 unknown tokens evaluated to make 86.3% and 13.6% in the sample, respectively. 43.1 % of ambiguity was noticed with 849 tokens and the MFT result was 87.3920%. The overall accuracy reached 87.3920% with the average ambiguity of 6 tags per token. Same can be inferred from the table below:

Table 36 Test result for seen data of 2k tokens (current phase)

| | Known | Unknown | Ambiguity | MFT | Tagger Output |
|-----------------------------|-------|---------|-----------|-------|---------------|
| Total Tokens (30391) | 1698 | 269 | 849 | 1574 | |
| Accuracy (%) | 86.32 | 13.67 | 43.16 | 80.02 | 87.39 |

The mentioned accuracy result is supported by the screenshot for the same, taken at the time of testing (see fig.16)

```

EVALUATING </home/sanskrit/svntool/bho.eval> vs. </home/sanskrit/svntool
.txt> on model </home/sanskrit/svntool/models/bho/BHO>...
...1967 tokens [DONE]
* ===== TAGGING SUMMARY =====
=====
#TOKENS          = 1967
AVERAGE_AMBIGUITY = 6.0005 tags per token
-----
#KNOWN           = 86.3244% -->          1698 / 1967
#UNKNOWN         = 13.6756% -->          269 / 1967
#AMBIGUOUS       = 43.1622% -->          849 / 1967
#MFT baseline    = 80.0203% -->          1574 / 1967
* ===== OVERALL ACCURACY =====
=====
          HITS          TRIALS          ACCURACY          MFT
-----
          1719          1967          87.3920%          80.0203%
=====

```

Figure 16 Screenshot of test on 2k unseen data

4.5.3.3 Results Obtained

The above evaluation reports that the tagger when tested for the seen/familiar data in the initial phase, showed the accuracy result of 87.3% which was improved to 88.62 % in the later stage.

The average ambiguity of 1.8 tags per token at the initial phase was found to have reduced which currently is 1.7 tags per token.

When tested with unseen data which is new to the tagger in terms familiarity, the result showed considerable improvement in the accuracy. The accuracy was 74.2 % in the beginning which later on rose to 80.02 %. With this increase in the accuracy the ambiguity ratio also decreased from 7.8 to 6.0 tags per token. This fall in the ambiguity, is may be because of the nature of data as the data was selected from different language domains.

Along with the increasing number of tokens in the training file, the ambiguity ratio also increased. The ambiguity for seen data in the earlier stage was calculated at 44% which gradually increased to 51% in the later experimental phase and the ambiguity for unseen data was 41% which was found rising to 43%. This is so because of the comparatively small data sample. With the increase in the no. of sentences in the data the language complexity also increased and as a result of which the ambiguity per token was found getting high.

Therefore, on the basis of the calculations made above, the overall accuracy of the tagger is 88.62% with the ambiguity ranging between 7.8 to 6.0 tags per token which is quite convincing at this stage of evolution of the tool.

4.6 TAGGER BASED TAGGING ISSUES

Unlike manually tagged data, the tagger generated output reveals some types of errors occurring throughout the data. Some of the errors shows a definite pattern, some occur due to the ambiguous nature of the tokens and others are random error. All the errors has been listed down and classified into three different error types:

4.6.1 Pattern based errors

There are some errors which follows a certain pattern. Verb series and conjunct verbs are the two major components of this error type. The patterns for both the constructions are:

4.6.1.1 Error Pattern for Verb series

Tagger very often tags the verbs incorrectly. One of the most repeated error is with the auxiliary '*ba*'. *ba* appears either as the only auxiliary or as the last auxiliary in a Bhojpuri sentence, depending upon the

sentence construction, but it has been tagged as a main verb in almost all the cases. Some of them are given below:

Table 37 Error pattern for serial verbs

| S.No. | Token | Gloss | Erroneous Tag | Correct Tag |
|-------|------------------|-----------|---------------|-------------|
| 1 | गया (gayA) | go | N_NN | V_VM |
| | है (hai) | be | V_VM | V_VAUX |
| 2 | बन (ban) | make | N_NN | V_VM |
| | कर (kar) | do | V_VM | V_VAUX |
| 3 | दाखिला (dAkhilA) | admission | V_VM | N_NN |
| | ले (le) | take | PSP | V_VM |
| | सकेले (sakeleM) | can | V_VAUX | V_VAUX |
| 4 | लिखले (likhle) | write | V_VM | V_VM |
| | बानी (bAnI) | be | V_VM | V_VAUX |
| 5 | दिहल (dihal) | give | V_VAUX | V_VM |
| | गइल (gail) | go | V_VAUX | V_VAUX |
| | बा (bA) | be | V_VM | V_VAUX |
| 6 | ले (le) | take | V_VAUX | V_VM |
| | चली (call) | walk | V_VM | V_VAUX |
| 7 | लागत (lAgat) | feel | V_VAUX | V_VM |
| | बा (bA) | be | V_VM | V_VAUX |
| 8 | करत (karat) | do | V_VAUX | V_VM |
| | रहलन (rahalan) | live | V_VM | V_VAUX |

Table 37 above contains the examples of verb phrases with the token in the first column, the gloss of the token in the second column, the incorrect tag in the third column and the potential correct tags in written the final column. In first two examples, the main verbs of the phrase are tagged incorrectly as nouns as a result of which the next following verb (an auxiliary) was tagged by the tool, as the main verb. 3 stands for both the pattern for serial verb and the converbs. In this example the postposition tag for main verb is due to the ambiguity of the token, *le* is both a postposition and verb in Bhojpuri. The final auxiliaries in 4, 5, 6, 7 and 8 are consistently marked as main verb and vice a versa the main verbs in 5, 6, 7 and 8 are provided with the auxiliary tag.

4.6.1.2 Error pattern for conjunct verbs

The next error pattern is found for conjunct verbs. Conjunct verb is formed with the combination of two categories, noun and the verb. The first counterpart should be given a noun and the second must be a verb (main verb) tag. But the following has been incorrectly tagged by the tool. The

tool identifies the first counterpart as the main verb and following this, it tags the second counterpart of a converb as an auxiliary verb, examples for this has been tabulated below:

Table 38 Error pattern for conjunct verbs

| S.No. | Token | Gloss | Erroneous Tag | Correct Tag |
|-------|------------------|-----------|---------------|-------------|
| 1 | पीड़े (plDxe) | pain | V_VM | N_NN |
| | देबे (debe) | give | V_VAUX | V_VM |
| 2 | दाखिला (dAkhilA) | admission | V_VM | N_NN |
| | ले (le) | take | PSP | V_VM |
| | सकेले (sakeleM) | can | V_VAUX | V_VAUX |

From table 38, the correct label for first example should be given as a noun followed by the main verb which has been incorrectly tagged as main and auxiliary verbs. These pattern based errors are quite big in numbers for Bhojpuri (as these are inherent features of Indian languages) but not so trivial. These types of errors and the errors caused due to the ambiguity of the token (see next sub-section) can be overcome to a great extent by proposing linguistic rules for the same. Emphasis has been laid on this issue in the concluding chapter.

4.6.2 Errors due to ambiguous tokens

A number of errors have occurred due to the ambiguous word tokens. Ambiguous tokens are the tokens which function in more than one way. This appears to be simple process for a human analyst but for a machine to understand the correct usage of the token in particular context is quite difficult. Different levels of ambiguities has already discussed under evaluation of the tagger. Some examples of such tokens which causes the system failure are explained with the help of the table, given below:

Table 39 Errors due to ambiguous tokens

| S. No. | Ambiguous Tokens | Meaning 1 | Expected tag 1 | Meaning 2 | Expected Tag 2 |
|--------|-----------------------------|---------------|----------------|---------------------|----------------|
| 1 | हाल (hAl) | state | N_NN | recently (temporal) | N_NST |
| 2 | योगी (yogI) | saint (noun) | N_NN | saint (adjective) | JJ |
| 3 | हिन्दू (hindu) | Hindu (noun) | N_NNP | hindu (adjective) | JJ |
| 4 | कवना (kawanA) | who | PR_PRI | which | DM_DMI |
| 5 | की (kI) | genitive | PSP | to do (verb) | V_VM |
| 6 | लेवल (leval) | take (verb) | V_VM | level (noun) | N_NN |
| | जाई (jAl)/हो (ho)_जाई (jAl) | will be taken | V_VAUX | will be levelled | V_VAUX_V_VAUX |
| 7 | ला (lA) | for | PSP | bring | V_VM |
| 8 | आ (A) | and | CC_CCD | to come | V_VM |
| 9 | उहाँ (uhA~M) | there | N_NST | him/her | PR_PRP |
| 10 | ओहिजा (ohijA) | there | N_NST | him/her | PR_PRP |

The above table explains the ambiguous tokens which often cause disturbance to the tool's performance. 1, in the table, functions as both a common noun and spatio-temporal noun depending upon the meaning and the context. 2 and 3 appears as a noun in some context and as adjective in other contexts. 4 is the example of interrogative which can either be an interrogative pronoun or an interrogative demonstrative, depending upon the environment it receives. If occurs as a pronoun without any supporting noun phrase after it, it is advisable to tag it as an interrogative pronoun where as a supporting noun following the interrogative makes it a demonstrative. Similarly, point 5 explains Bhojpuri *ki*, functions both as a genitive and a verb, though verbal use of *ki* is not common to Bhojpuri but in some cases with a mixed contraction of Hindi and Bhojpuri, these may be found. In example 1 and 2 the temporal and nominal use of *haal* has been exemplified.

[87] अबही/N_NST हाल/N_NST ही/RP_RPD में/PSP सर्वर/N_NN मेंटेनेन्स/N_NN
का/PSP चलते/V_VM अँजोरिया/N_NNP परिवार/N_NN के/PSP साइट/N_NN बन्द/JJ रहली/V_VM

Gloss: əbəhī halə hi mē sərɖər mēʈenənsə ka ʃələtə ʔndʒorɪə pəriɖərə ke saɪt bəndə rəhəli

Translation: The website of anjoria was shut due to server maintenance, some times back.

[88] कांग्रेसो/N_NNP के/PR_PRQ इहे/DM_DMD हाल/N_NN हो/V_VM गइल/V_VAUX
बा/V_VAUX

Gloss: kãŋɡresə ke ihe halə ho gələ bə

Translation: Congress is also facing similar situation

Further, two usage of *leval*, can be observed from 6. First is the noun, meaning ‘the level or stage’ and second, the verb meaning ‘to take’ as explained with the help of following verbs in different context. The first *jAI* means ‘go’ when attached with *leval* gives the meaning ‘will be taken’ and on the contrary the same *jAI* ‘go’ when comes along with another auxiliary *ho* ‘to be’ change the phrase meaning as ‘will be levelled’. Next two points, 7 and 8 chiefly functions as a preposition ‘*IA*’ and a co-ordinator ‘*A*’ but contextual use of the words changes its meaning by forming verbs out of it, meaning ‘to bring’ and ‘to come’ respectively.

[89] मल्टी/JJ लेवल/N_NN मार्केटिंग/N_NN वाली/PSP कंपनियन/N_NN से/PSP अलगे/JJ

रहे/V_VM

Gloss: məlti levələ market[ɪŋ] vali kəmpənijən se ələge rəhe

Translation: Stay away from the companies offering multi level marketing

[90] किराना/N_NN बाजार/N_NN से/PSP लेवल/V_VM जाई/V_VAUX

Gloss: kirana bazarə se levələ dʒai

Translation: The household items will be bought from the grocery store.

Last two examples were not pretty sure till very late. The native Bhojpuri speakers are also barely able to define this occurrence unless were provided with the example from the corpus. ‘*ahA~M*’ is a reference term for him in Maithili and ‘*uhA~M*’ is used for there in many Bhojpuri varieties but the corpus examples did not fit to the regular usage of the words and meaning was not clear. Similarly, *ohijA* is mostly used as a there (place adverb) but some examples disagree this use. It was then decided after consulting Bhojpuri speaking friends of the researcher and colleagues that it also accounts for the pronominal, continually. Therefore, primarily it should be tagged as spatio-temporal nouns (following BIS for adverbs other than manner) and secondly as a pronoun.

[91] दिले/N_NN के/PSP दौरा/N_NN पड़ला/V_VM का/PSP बाद/N_NST उहाँ/PR_PRP

के/PSP निधन/N_NN हो/V_VM गइल/V_VAUX

Gloss: ðile ke ðəure pəɾlə ka baðə uhā ke niðʱənə ho gəɪlə

Translation: He died after he having heart attack

[92] जब/N_NST तब/N_NST उहाँ/N_NST ठहरल/V_VM अथाह/JJ सन्नाटा/N_NN

जीव/N_NN जन्तु/N_NN भा/CC_CCD पक्षियन/N_NN का/PSP फड़फड़ाहट/N_NN से/PSP

टूट/V_VM जात/V_VAUX रहे/V_VAUX

Gloss: Jəbə t̪əbə uhā ʈʰəhəɾələ uʈʰahə sənnat̪ə dʒiɪə jənt̪u bʰa p əkʃijənə ka pʰəɾəpʰəɾahət se t̪ʊt̪ dʒat̪ə rəhe

Translation: The deep silence of the plae was broken due to the flying sounds of the birds and animals, every now and then.

In the present examples 5 and 6 above, the pronominal and temporal use of *uhA~M* has been shown which makes the last two points the table. All the aforementioned cases are incorrectly tagged as their second counterparts by the tagger is a given context. Hence, increasing the number of errors and lowering the taggers performance.

4.6.3 Random errors

There are some more types of error which does not allow any definite pattern and are randomly occurring errors. Table below contains a list of such errors. Some of them, though being part of the training corpus, has been incorrectly tagged by the tagger like ‘Delhi’. See more in table 40.

Table 40 List of random errors

| S.No. | Token | Gloss | Erroneous Tag | Correct Tag |
|-------|---------------------|---|---------------|-------------|
| 1 | लुटावेली (lutAveII) | to over spend | JJ | V_VM |
| | अँजूरी (anjurI) | both palm are joined for holding something (fluid or uncountable nouns) | JJ | N_NN |
| 2 | बरिसावे (barisAwe) | to rain | JJ | V_VM |
| | अमिरित (amirit) | holy water | JJ | N_NN |
| 3 | दिल्ली (diIII) | Delhi (place name) | N_NN | N_NNP |
| 4 | अकसरहाँ (aksarhAM) | Often | V_VAUX | N_NST |
| 5 | मार (mAr) | fight (collocations) | V_VM | N_NN |
| | धाड़ (dhADx) | | N_NNP | RD_ECH |
| 6 | कहिया (kahiyA) | when | N_NNP | N_NST |
| 7 | फर्स्ट (farst) | first look (collocation) | N_NNP | N_NN |
| | लुक (luk) | | N_NNP | N_NN |

The mentioned 1, 2, 5 and 7 in the table are extracted fragments from the corpus sentence where 1 says *lutAveII anjurI se* mean ‘to spend with both hands’, 2 says *BarisAwe amirit* means ‘abundance of holy water’, 5 and 6 are collocations meaning ‘fight’ and ‘the first look’. Both tokens in example 1 and 2 are tagged as an adjective which is not correct. But actually, these tokens must make two phrases, in both the examples. First token should contribute to a verb phrase with main verb and second, a noun phrase with a common noun tag.

It is impossible to feed the tagger with the complete vocabulary and Proper nouns, in count are much higher in number as new names, brands, locations etc are being coined every day It is true

for all the open ended grammatical categories. Point 3 carries one such case, though here, *Dilli* being a part of training corpus, tagged incorrectly by the tagger but other pronouns might not be identified as a proper noun by the tagger unless it is known to it.

Fourth one *aksarhA~M* means ‘often’ and 6 *kahiyA* meaning ‘when’ annotated as an auxiliary and a proper noun , respectively, which should have been a part of spatio-temporal noun, in correct sense.

4.7 COMPARISON AND PERFORMANCE OF HINDI AND BHOJPURI TAGGER

This section is dedicated to the comparison of the evaluation results of the present Bhojpuri tagger with the under-development Hindi tagger. Hindi and Bhojpuri shares similar syntactic structure, though varied morphology. Therefore, the hypothesis made in the beginning of this work that Hindi and Bhojpuri are related languages and the tagger trained on the same model must show the similar kind of results for both the languages, has been tested here. The test is beneficial in the way that it will explain the strengths of the tagger for both the language; differentiate between the nature and pattern of errors and help analyzing the scope of development.

4.7.1 Data for Comparison

Data for the present comparison test has been borrowed from the Politics domain. The data consists of about 100 sentences from Bhojpuri corpus which makes 1,967 tokens. A parallel test set for Hindi has been prepared on this data by translating the Bhojpuri data into Hindi. Keeping the sentences constant, the translated Hindi data came up with 2,016 tokens. The idea behind creating parallel chunk of corpus was to test the results for both the languages keeping the variables like model of the tagger, length of the data, genre and the test data itself, all constant. This was done I order to get the difference in the accuracies and feature differences based on the nature of error, in both the languages.

4.7.2 Hindi Tagger

The Hindi tagger mentioned in this section is under development as part of ILCI project, JNU. To my knowledge, this tagger has been trained upon the corpus from Health and Tourism

domain, created earlier as part of ILCI corpus.³⁸ The tagger has reached the accuracy of 94.2% for the data from the same domain. There might be some fluctuation in the accuracy claimed by the tagger as the data for the comparison has been taken from another domain namely politics. This comparative analysis will also be helpful in a way to determine how far, would it be accurate for the data from the other domains, at this stage of its evolution.

The input Hindi data has been formatted not in tokens but in form of string/sentences. The data has been tagged by the online Hindi SVM tagger available on the ILCI website with link ‘sanskrit.jnu.ac.in/pos/index.jsp’.

4.7.3 SVM performance on Hindi Tagger

On testing the Hindi tagger with the data from politics domain, translated from Bhojpuri, following output was obtained:

Table 41 Accuracy result of Hindi Tagger

| | Known | Unknown | Ambiguous | MFT | Tagger Output |
|-----------------------------|-------|---------|-----------|-------|---------------|
| Total Tokens (30391) | 1922 | 94 | 1274 | 1777 | |
| Accuracy (%) | 95.33 | 4.66 | 63.19 | 83.18 | 93.20 |

Above table 41 depicts that for the data of 2016 tokens, the Hindi tagger was able to identify 95% known, 4.6 % unknown and 63% ambiguous tokens which made 1941, 95 and 1286 tokens out of the total 2016 token, respectively. The average ambiguity was calculated to have 3.8 tags per token with the tagging based MFT result of 83%. The overall accuracy of the Hindi tagger for an unknown data set from politics domain was 93.2%

The screenshot of the evaluation result for the Hindi tagger has been provided in the figure 17 below:

³⁸ <http://sanskrit.jnu.ac.in/lciann/index.jsp>

```

EVALUATING </home/atul-thinkcentre/Desktop/Hindi gold token.txt> vs. </home/atul
-thinkcentre/Desktop/Hindi Tag token.txt> on model </home/atul-thinkcentre/pos_t
agger/svntool_v1.3.2/models/HI/HIN>...
.....2016 tokens [DONE]
* ===== TAGGING SUMMARY =====
=====
#TOKENS          = 2016
AVERAGE_AMBIGUITY = 3.8229 tags per token
* -----
-----
#KNOWN           = 95.3373% -->          1922 / 2016
#UNKNOWN         =  4.6627% -->           94 / 2016
#AMBIGUOUS       = 63.1944% -->          1274 / 2016
#MFT baseline    = 83.1845% -->          1677 / 2016
* ===== OVERALL ACCURACY =====
=====
-----
*              HITS          TRIALS          ACCURACY          MFT
-----
*              1879          2016          93.2044%          83.1845%
-----

```

Figure 17 Screenshot of the evaluation for Hindi tagger

Present test also elaborates that the overall accuracy of Hindi tagger for a particular domain was 94.2 which was reduced to 93.2 % when tested with the data from another domain. This difference in accuracy is not so significant as it reports the fall of one percent. This fall was expected as the tagger is domain specific and the test data was out of that field. This can be improved upon, in due course of time by training the tagger with more versatile data. The tagger is underdevelopment and if it is capable of achieving 93% at this stage, the further development will bring it to more precise results. The nature of errors found in the data tagged by the Hindi tagger is discussed in the section 4.7.5 under Comparison of results.

4.7.4 SVM performance on Bhojpuri Tagger

The data from politics domain under unseen data tested by Bhojpuri tagger for 90k tokens, also makes the input for the present comparison. This data sample of 1967 tokens with about 100 sentences has been translated into Hindi. The tagger based accuracies has already been mentioned in table 44 under section evaluation of the tagger.

For this part of testing, the Bhojpuri tagger was mentioned to have 87.3% accuracy with 86% known, 13 % unknown and 43% of ambiguity. The average accuracy was listed to be 6.

4.7.5 Comparison of results

Both Hindi and Bhojpuri taggers are running in their development stage. The Bhojpuri tagger, being a general domain tagger, requires increasing the data for each domain and making it more robust one. The Hindi tagger was trained for specific domains with relatively huge number of

data which is shown by the result itself. This tagger also requires expanding its domains for its better performance over the unseen data.

When tested for a new domain, both the taggers showed a fall of 1 percent in the overall tagger accuracy. The new accuracy of Hindi and Bhojpuri tagger for this domain was 93.2 and 87.3 %, respectively. When compared, the Hindi tagger was found to have 95 % known tokens which was 86% for Bhojpuri tagger and the ambiguity of Hindi tagger was 63% which was higher than the Bhojpuri tagger found with 43% of ambiguity. The Hindi tagger having identified larger number of known tokens is due to the larger corpus data.

Table 42 Comparative results of Hindi and Bhojpuri taggers

| | Bhojpuri tagger | | Hindi Tagger | |
|-------------------------------|-----------------|----------|--------------|----------|
| Total no. of sentences | 103 | | 103 | |
| Total no. of tokens | 1967 | | 2016 | |
| | tokens | % | token | % |
| Known | 1698 | 86.32 | 1922 | 95.33 |
| Unknown | 269 | 13.67 | 94 | 4.66 |
| Ambiguity | 849 | 46.1622 | 1274 | 63.19 |
| MFT | 1574 | 80.02 | 1777 | 83.18 |
| tagger Output | | 87.392 | | 93.2044 |
| | | | | |

4.7.5.1 Error analysis of the Hindi tagger

The types of errors made by the Hindi tagger can be divided into the following categories:

4.7.5.1.1 Simple Verbs

Table 43 Issues in Hindi tagger for tagging simple verb phrases

| S. No. | Verb phrases | Transliteration | Gloss | Correct tagging |
|--------|---------------------------------|------------------------|-----------------------------------|-----------------------------------|
| 1 | समझने\N_NN लगे\V_VM हैं\V_VM | samajhane lage hain | have started understandi ng | समझने\V_VM लगे\V_VM हैं\V_VAUX |
| 2 | बढ़ा\N_NN दी\V_VM | baDhxA dI gayI hai | has been raised | बढ़ा\V_VM |

| | | | | |
|---|---|-------------------|------------------|--|
| | गई\V_VAUX है\V_VM | | | दी\V_VAUX गई\V_VAUX है\V_VAUX |
| 3 | बोली\N_NN जाती\V_VAUX है\V_VM \RD_PUNC | bolI jAtI hai | being spoken | बोली\V_VM जाती\V_VAUX है\V_VAUX \RD_PUNC |
| 4 | कहा\V_VM गया\V_VAUX है\V_VM | kahA gayA hai | has been said | कहा\V_VM गया\V_VAUX है\V_VAUX |
| 5 | हो\V_VM गया\V_VAUX है\V_VM | ho gayA hai | has happened | हो\V_VM गया\V_VAUX है\V_VAUX |
| 6 | कर\V_VM दी\V_VAUX गई\V_VAUX है\V_VM | ja dI gayI hai | has been done | कर\V_VM दी\V_VAUX गई\V_VAUX है\V_VAUX |
| 7 | लौटे\N_NN हैं\V_VAUX | lauTe hain | came baack | लौटे\V_VM हैं\V_VAUX |
| 8 | छोड़\N_NN कर\V_VM | choDx kar | leaving | छोड़\V_VM कर\V_VAUX |
| 9 | समा\N_NN गए\V_VAUX | samA gaye hain | has sunken | समा\V_VM |

| | | | |
|----------|--|--|------------|
| हैं\V_VM | | | गए\V_VAUX |
| | | | हैं\V_VAUX |

The verbs have been incorrectly marked for their order of appearance. For example, the verb *samajhane*, *baDhxA*, *bolI* and *samA* in 1, 2, 3 and 9 in the table, has been tagged as a noun in place of main verb. The final auxiliary in all the above phrases has been invariably tagged as the main verb, instead of an auxiliary verb. Similar inconsistency is found with the Bhojpuri tagger for tagging verbs and the rules has been tried to propound for resolving such errors.

4.7.5.1.2 Nouns

Table 44 Issues in Hindi tagger for tagging noun phrases

| S. NO. | Noun Phrases | Transliteration | Gloss | Correct tagging |
|--------|------------------------------------|-------------------------|-------------------------------|----------------------------------|
| 1 | आक्रामक\N_NN रुख\N_NN | AkrAmak rukh | agressive mood | आक्रामक\JJ रुख\N_NN |
| 2 | गैरमुस्लिम\N_NN स्टूडेंट्स\N_NN | gairmuslim stuDent | non- Muslin student | गैरमुस्लिम\JJ स्टूडेंट्स\N_NN |
| 3 | लंबा\N_NN रिहायश\N_NN | laMbA rihAyash | long chain | लंबा\JJ रिहायश\N_NN |
| 4 | राजनैतिक\N_NN बयार\N_NN | rAjnaitik bayAr | political agenda | राजनैतिक\JJ बयार\N_NN |
| 5 | पूर्वाचली\N_NNP संगठन\N_NN | purvAMcall saMgathan | Organisation of Purvanchal | पूर्वाचली\JJ संगठन\N_NN |
| 6 | सफल\N_NN प्रयास\N_NN | saphal prayAs | successful effort | सफल\JJ प्रयास\N_NN |

| | | | | |
|---|------------|------------|-----------|--------------------|
| | बीमार\N_NN | | | |
| 7 | इकाई\N_NN | bImAr ikAI | weak unit | बीमार\JJ इकाई\N_NN |

The adjective or modifier of a noun phrase has been incorrectly tagged as nouns in the table above. This error type is handled well by the Bhojpuri tagger and the modifiers coming before a noun in a noun phrase or before a verb in a conjunct verb are tagged correctly, in most cases.

Table 45 Issues in detecting noun types in noun phrases

| S. No. | Noun Phrases | Transliteration | Gloss | Correct tagging |
|--------|---|--------------------------|---------------------------------|---|
| 1 | डोमेसाइल\N_NN की\PSP आग\N_NNP | Domesail kI Ag | wave of domecile | डोमेसाइल\N_NN की\PSP आग\N_NN |
| 2 | सीएनटी\N_NN एक्ट\N_NNP | sIenTI aikt | CNT act | सीएनटी\N_NN एक्ट\N_NN |
| 3 | इम्तेहान\N_NNP में\PSP | imtehAn meM | in examination | इम्तेहान\N_NN में\PSP |
| 4 | चीनी\N_NN मिलें\V_VM खुलीं\N_NN | cInI mileM khulIM | sugar mills opened | चीनी\N_NN मिलें\N_NN खुलीं\V_VM |
| 5 | खरीद\N_NN बिक्री\N_NN से\PSP जुड़े\N_NN | kharId bikrI se juDxe | related to sale and purchase | खरीद\N_NN बिक्री\N_NN से\PSP जुड़े\V_VM |

From the above table, we infer that the tagger is tagging most of the nouns, which are unknown to it, as proper noun and other unknown tokens are tagged simply as nouns. The first rows in the table are examples of common nouns incorrectly tagged as proper and other two are examples of unknown words like *khulIM* and *juDxe* which are clear verbs and if would have been part of the trained corpus, was expected to be tagged as a variant of verb, but not a noun.

The Bhojpuri tagger also makes such mistakes of tagging a common noun in place of proper and vice versa, due to the limited training data and the later issue of tagging a unknown token as noun, is universal for all SVM models. This can be checked by exceeding the length of corpus to match the like elements in it. One major difference in tagging of the taggers for noun phrases is that, the proper nouns are tagged correctly by the Hindi tagger which is not the case with Bhojpuri tagger. The names like *sharat candrayAn* and *dayAnand gauDx* are merkek as proper noun in Hindi and common noun in Bhojpuri.

Verbs like *samA* (to get mingled), *badhxA* (to extend), *khuliM* (to open), *juDxeM* (to add) etc are all main verbs in both Hindi and Bhojpuri but the tagger for Hindi identifies it as noun whereas Bhojpuri tagger perhaps due to the word form *samA gail*, *badhxa gail*, *khul gail* and *juDxa gail* has been tagged correctly as the main verb.

4.7.5.1.3 Ambiguous tokens

There are a lot of ambiguities found in Indo Aryan languages, as a result of their rich morphological property. The tokens like *kI* and *par* are ambiguous in both Hindi and Bhojpuri and the tagging accuracy for such tokens, depends largely upon the occurrence of tokens in particular context. *kI* has been tagged as a postposition (PSP) by the Hindi tagger and *par* has been marked as but by the Bhojpuri tagger because the postposition use of *kI* in Hindi is much higher than when used as a verb. Similarly, in Bhojpuri, the coordinator use or *par* is higher than the locative/ postpositional *par* leading to the inconsistent tagging.

4.7.5.1.4 Other cases

Words like *zarur*, *mukhayatah*, *bAkI* and are tagged as common noun for the first two and verb for the third token. Where *zarur* and *mukhyatah* should be part of adverb as they mean ‘necessarily’ and ‘mainly’ and the third *bAkI* can have two meanings one as a conjunction

meaning and /or and secondly, as ‘rest’. Here in the context of conjunction, it has been marked as verb, which is not a correct tag.

4.7.6 Comparison Result

The Hindi tagger is found to have 94.2 % accuracy for the Hindi data from the trained domains and the Bhojpuri tagger has been reported to have achieved 88.6 % for the general domain corpus. On comparing both the taggers keeping dependent variables constant, like SVM model for tagging both languages, the size of input test data, the domain of the test data, number of input sentences. Under controlled environment the tagging results of both taggers are found to have 93.2 and 87.3 % for Hindi and Bhojpuri, respectively. The ambiguity ratio of the respective taggers is calculated as 63% and 43%. The number of known and unknown tokens is 1922 and 1698 (known); and 94 and 269 unknown, for Hindi and Bhojpuri.

The number of ambiguous tokens and known tokens are higher in Hindi tagger, this might be the result of bigger training data which facilitates the tool in detecting known tokens. There are 269 unknown tokens encountered by the Bhojpuri tagger which is 2.5 times higher than the Hindi tagger. This can be reduced by increasing the length of the training corpus and by expanding the domain of the tagger.

Both Hindi and Bhojpuri taggers are running in their development stage. The Bhojpuri tagger, being a general domain tagger, requires increasing the data for each domain and making it more robust one. The Hindi tagger was trained for specific domains with relatively huge number of data which is shown by the result itself. This tagger also requires expanding its domains for its better performance over the unseen data.

Both the taggers are encountered with different types of errors like the Hindi tagger is providing erroneous results for the conjunct verbs, simple noun phrases and adverbs whereas the problematic cases in Bhojpuri tagger are the cases of proper nouns, compound verbs and demonstratives and pronominal use of verb. This confusion of demonstratives and pronouns are handles very well by the Hindi tagger. There is one common error made by both the taggers is for the simple verb constructions. Both the taggers fail to mark the consecutive verbs (main +aux+ aux). They often mark the first verb in the series as an auxiliary and the final auxiliary is

mostly marked as the main verb. Rule regarding the verb and adjectives has been formulated, which can be found in the next sub section.

4.8 DEVELOPMENT OF THE BHOJPURI TAGGER

The Bhojpuri tagger has been trained upon 90k tokens from the corpus; rest 80K tokens are under development. Based on the results of the tagger's accuracy and discrepancy, the further input training file will be selected. The further training is expected to give a better accuracy result with the lesser ambiguity classes. Some rules for the major discrepancies of verb and modifiers are proposed below.

4.8.1 Rules proposed for improvement of pattern errors for Bhojpuri tagger

The Bhojpuri tagger is a statistical tagger, therefore the rules proposed here can be applied during the post processing stage to avoid the conflict as well as keeping a check on the alterations made after the implementation of the rules.

4.8.1.1 Verbs

In languages like Hindi and Bhojpuri, there should be at least one main verb in a clause. Other verbs in same clause could be 'main' as well as an 'auxiliary', depending upon the context and meaning. The standard tagging for verbs in series constructions should be - a main verb followed by auxiliary (ies) i.e.,

MAIN +AUX +AUX+ and so on

But the tagger generated output in cases of more than one verb in a string, even in simple constructions; were found to be tagged as

AUX + AUX + MAIN or

MAIN + MAIN or

MAIN + AUX +MAIN (tagger generated output)

Rule -1

Here, a rule can be proposed for the tool that wherever there is more than one verb in a sentence/clause and the second verb is not a main verb (or is among the list of auxiliary verbs) the tagger must tag the first verb in the series as main verb and rest of the verbs as auxiliary verbs, like:

MAIN +AUX +AUX+ and so on

It is possible to identify the list of auxiliaries present in the corpus and tag accordingly as the training data is available. Although this rule cannot be generalized for all cases, because there might be cases of compound verbs where main verb is followed by another main verb. This rule might not work for such instances.

Except of the linguistic rules, the tagger's performance can also be improved by feeding it with more data from the exemplar constructions which improves the performance gradually by decreasing the number of error for that particular case. This can be applied for adjectives and other tagging issues also.

4.8.1.2 Adjectives

Both Hindi and Bhojpuri tagger were found to have difficulty in identifying the modifiers of a noun. The cases of modifiers concerned here is for the simple noun phrase.

The proposed rules for adjective are as follows

Rule- 2

If a word preceding the noun in a noun phrase does not belong to any other categories like demonstratives, quantifier, classifier and adverb, must be tagged as a modifier. This rule must be applicable as a noun phrase can have only these preceding elements and other like intensifier, negation and pronoun, does not occur in this environment.

Rule -3

Another rule for adjectives, can be formed, based on the *-ik* & *-it* suffixes, that it takes. Most of the adjectives in Hindi and Bhojpuri have *-ik* or *-it* suffix like sAmAjika, AdhArita, vyavasAyika, kAlpanika, AdhUnika, nAitika etc. Therefore, it can be generalised that a word

ending with *-ik* and *-it* suffix, if not identified as part of any other word category in the corpus, must be put under the adjectives.

These rules cannot be true for all cases and exceptions to it can be found.

5 DISCUSSION AND CONCLUSION

5.1 DISCUSSION

The present piece of work is an NLP oriented assignment. The objective of the study was to develop the first statistical tagger for Bhojpuri. For the development of the tagger, it was required to have a corpus of Bhojpuri, which was not available till date, in full form. Therefore, the study has been divided into three main units. The first unit deals with the corpus creation for Bhojpuri at a space where no such annotated corpus for the language is available. The process of corpus building right from the data collection till the validation and the challenges involved have been discussed in the concerning chapter. The second unit basically includes the tagging description for Bhojpuri. The process of tagging, the tagging scheme followed and the detailed guideline for POS annotation of the Bhojpuri corpus has been presented in this chapter. The last section of this chapter deals with the challenges met in manual tagging of data and the issues concerning the decisions undertaken to resolve them. The third unit is the development of the SVM based statistical tagger for the POS annotation of Bhojpuri. The system architecture, the development of the tagger by repeated training, testing and evaluation are the major contents of this chapter. The annotation challenges discussed here are based on observations made on the output result of the tagger. The final section of the chapter is the comparative analysis of the evaluation result of the SVM tagger on the parallel data for Bhojpuri and Hindi.

5.2 MOTIVATION OF THE STUDY

The first chapter of the study is self explanatory of the background and motivation of the study. It deals with the existing works in the field of NLP for Indian Languages, the description of the POS tagger already trained. The review of Bhojpuri from literary and technological aspect was presented. One main technological advent was noticed in form of Opus parallel corpus for Bhojpuri. This corpus was created by Open Linguistics Working Group (OLWG) group of NLP practitioners, who developed this corpus in a fully automated way. From collection of data to the corpus cleaning, all was done by the machine. Therefore, the reliability of the quality of its content is doubtful. There is also a mention of the first POS Tagset already devised for Bhojpuri, as part of the same project. This section explains the need for developing language resource for

Bhojpuri as there is no significant work done in this field and also discussed the SVM (Support Vector machine) model on which the tagger was trained.

5.3 STAGES OF DEVELOPMENT OF RESOURCE FOR BHOJPURI

The second chapter of this study titled ‘Corpus Creation for Bhojpuri’ began with the description of the process involved in the corpus creation. The present corpus is a general domain monolingual corpus with 1.69 lakh tokens and 9,019 sentences which includes six major domains for the data selection, namely blogs, entertainment, literature, politics, sports and miscellaneous. The corpus has been formatted according to the ILCI data format and validated using the ILCIANN tool for corpus validation, developed as part of ILCI project. The corpus data for each domain in the corpus has been drawn from the web using ILCrawler, developed by Computational Linguistics group at SCSS, JNU. The crawled data is first cleaned for noises by using ILSanitizer, where the data compilation, the spelling errors and others were resolved. After the data cleaning, the formatting and management of data began. This included the process of manual selection of the corpus data, the compilation of the text file and its formatting according to the requirement of the tool was done.

‘Annotated Corpus for Bhojpuri’ which makes the third chapter in this dissertation, is the description of the annotation process. This includes the details about the BIS (Bureau of Indian Standards) tagging scheme followed and the tagset for the language. The revised version of the tagset has been produced here with the addition of one tag level category called ‘echo- before’. The annotation is done both manually and semi automatically using the ILCIANN tool to POS tagging. The tool and its functioning have been explained inside the chapter, in detail. This section is followed by the extensive guideline for tagging the Bhojpuri corpus with the do(s) and don’t(s) for the language, in particular. The inflecting categories like classifiers and particles have been advised to be marked according to the category of the head word, as this inflected form of word does not have a clear meaning of its own.

The fourth chapter titled ‘SVM Based Bhojpuri Tagger’ discusses the utility of the SVM model for resolving different types of classification and optimization problems. This chapter explains the description of the tagger based on SVM, its property and architecture. This is followed by the format of the input data, training and the testing of the tagger. The tagger has been trained upon

the data set of 90k tokens with 5,201 sentences from blogs and miscellaneous domains. The testing has been done in two phrases. The initial experiment was done with the training file of 30,391 tokens with the tagger accuracy of 87 % which needed improvement due to limited number of training data. When the training data was increased three times the length of the initial , presently with 90k tokens, the accuracy of tagger is calculated with 88.6 %. Although there is improvement of only one percent in the accuracy of the tagger, the tagger tagged output was found with good result when validated for errors, manually. This ranged between 11 to 16 errors per 100 tokens with an average of 14.

5.4 CHALLENGES MET IN CORPUS CREATION AND ANNOTATION

The corpus related challenges were presented under the second chapter on corpus building. Some of the most often found errors in the data included the occurrence of section headers and title of the report/column as the part of the text. The repetition of headers and word fragments were also there. Data from other languages written in Devanagari were found in the corpus, from among which the Hindi data was present in a big number. Some other common errors like multiple spacing, missing of space between two words and punctuation error were checked during the validation process. There was an crawler generated issue that is used to skip the numerals written in Roman. This seemed to be a major fault because the numbers are the carrier of fact or information like , the number of people, the amount of money, year, quantity, which is necessary part of a factual report. This has been fixed for the later part other data collection. The genre specific sentences like lines of poetry, headlines of a news/report are eliminated from the corpus during the manual selection process along with the other language data. Similar action has been taken for the unaccepted terminology of the language like slangs and bad terms of addressing. Although slangs and discourse particles like ‘hmmm..’, ‘mmm..’, ‘aãã..’ are part of natural speech but taking care of the adaptability of the tool, such features are swept out. There are some common typing mistakes also found in the web data, there are listed among the challenges under the category of repetition, typo errors and Fragmented words in the corpus. These are verified and edited throughout the data in order to make it error free.

The Along with the issues while building corpus, some features of Bhojpuri have also been described. The absence of ergativity, the use of determiners, the morphological inflections like of classifiers, and particles were addressed.

The annotation challenges have been classified into two:

- Issues in manual tagging
- Tagger based annotation challenges

Issues in manual tagging

The annotation challenges in manual tagging of Bhojpuri corpus has been dealt as part of the chapter 3. This includes, first of all, discusses the problem of unidentified tokens like *hirəʋ*, *roganə- sefanə*, *ʈikɛʈə* and rare occurrences like *kəʈəbjōʈə* etc. Next is the problem with the dialectal differences, found in languages. Though it was not so big a problem in this case because the corpus data has been adopted from the authentic web sources which themselves select and follow a standard of writing for the maintaining consistency on their website. The word like *b^ha* and *sən* was the main concerns of this issue as these are found only in some variety of Bhojpuri, particularly, one spoken in Bhojpur region. There are two crucial issues found which are inherited feature of language. These are the homophonous words and the realizations of one lemma differently in different context. The homophonous property of a language is always a point of concern while creating any resource as this results n the increased ambiguity ratio in the output. The other problem of varied realizations is not found in Hindi. This feature might be present because of the two reasons, first, language being an oral medium of communication for a very long time and secondly, the lack of one common standard for Bhojpuri, to follow in the literary writing and formal engagements.

Tagger based annotation challenges

Apart from the tagging issues faced during the manual annotated, there are some more issues met in due course of the automatic tagging. These are categorised into three- 1) pattern based errors, 2) errors due to ambiguous token and 3) random errors. The pattern based errors contains list of those errors which follow a certain pattern. For example, in the present experiment, the error pattern was found with the verb constructions. The sentences with two or more verbs and

examples with conjunct verb showed such patterns. In simple verb constructions, the tagger often tagged the first verb of the series as an auxiliary instead of the main verb and the last verb in the series was mostly tagged as the main verb instead of an auxiliary. Similarly, in converbs, the tagger was very often tagging the first (noun/adjective) counterpart of the construction was incorrectly tagged as the main verb. The next type of errors was the result of ambiguity. The ambiguity is defined in terms of the number of tags a token can avail or take. Based on this list of ambiguous tokens with erroneous tags were presented. The last category was the list of random errors. These are errors which do not follow any pattern and does not occur because of some definite discrepancy. Such errors have been listed and will be taken care of during the development process.

5.5 RESULT OF COMPARATIVE ANALYSIS OF HINDI AND BHOJPURI

The final section of the experiment deals mainly with the comparative analysis of the SVM based taggers trained for Hindi and Bhojpuri. Our hypothesis was to find and explain the difference of accuracy results of Hindi and Bhojpuri tagger, their nature of errors and the fluctuation in the accuracies of the tagger when tested upon a the data set from a new domain. Both the taggers were found to reduce to 1 percent in their accuracy results. The accuracy of Hindi tagger for the politics domain was 93.2% and for Bhojpuri it was 87.2%.

On manual validation of the output generated by the tagger, the Hindi tagger was found with lesser errors than the Bhojpuri tagger. Mainly, verb and adjectives are the concerns for error in the Hindi tagger and nothing can be said about nouns for certain because for all unknown tokens, the SVM has invariably tagged it as noun. With Bhojpuri tagger, during validation, the issue of ambiguity among the pronoun and demonstratives were found. Apart from this the Bhojpuri tagger is working better for adjectives than on verbs but both the cases are most erroneous in Bhojpuri too. Therefore, towards the end of the chapter, some rules have been proposed for handling issues related to verb and adjectives which are to be applied at the post processing stage of annotation.

5.6 FINAL RESULT

In a nut shell, the present study is resulted in developing an ‘SVM based statistical tagger’ for Bhojpuri with the overall accuracy reaching 88.6%. In due course of developing the tagger, the stages of creating the corpus and annotation of the corpus were dealt. As part of the study, was developed, a general domain corpus of Bhojpuri with 1.69 lakh words out of which the first 90 thousand words have been annotated for the training of tagger. Therefore, an annotated corpus of Bhojpuri with 90 thousand words have been created along with the un annotated corpus. The tagger trained in the study was found to have 88.2% accuracy result on the data from the domain of the training set and 87.2% for data set from other domains. Therefore the overall accuracy of the tagger was found to range between 87-88%, at present.

Finally, the comparison of the present Bhojpuri tagger result with the Hindi tagger result shows that the accuracy of both the taggers were reduced to one percent when tested on a new domain and the ambiguity was increased. In comparison with the Hindi tagger, more error types were noticed for the output of Bhojpuri tagger. Apart from all other issues, the issue with the tagging of verb constructions and the adjectives were most common for both the taggers and rules have been proposed for the same. The performance of the tagger can be improved with application of these rules, to some extent and another means to improvise the tagger is by providing it with more data with the similar construction.

5.7 SCOPE FOR DEVELOPMENT

The technology development for Bhojpuri is a relatively untouched discipline. The corpus and the tagger introduced are probably among some initial experiments undertaken for the language. There is lot to discover in this area like the expansion of the corpus, developing domain specific corpus for Bhojpuri. The higher level linguistic analysis like chunker, parser, etc would be welcomed as this is a language of a large population. At a later stage, the inclusion of Bhojpuri MT systems can also be worked upon as there are speaker constantly in touch with the web sources (by visiting or posting) and who are inclined to get such applications in Bhojpuri which can be used by them as any other language.

REFERENCES

1. Abbi, A. (1994). *Semantic Universals in Indian Languages*. IAS, Shimla (pp: 138).
2. Abbi, A. (2001). *A Manual of Linguistics Field Work and structures of Indian Languages*. Lincom Europa.
3. Abe, Shigeo.2005. *Support Vector Machines for Pattern Classification*. Springer London Dordrecht Heidelberg, New York.
4. Baskaran, S.,Bali, k., Bhattacharya, T.,Bhattacharya, P., Jha, G.N., Rajendran,S., Saravanan, K., Shobha, L., Subbarao, K.V. 2008. Designing a Common POS-Tagset Framework for Indian Languages. In: Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Daniel Tapias (Eds.) Proceedings of the workshop on Asian Language Resources. In Proceedings of *Sixth Conference on International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.
5. Bennett, K. P. and Campbell, C. (2000). *Support vector machines: Hype or hallelujah?* SIGKDD Explorations 2 (pp: 1-13).
6. Bharati, A., Misra, D., Sharma, L. B., Sangal, R. 2006. *Anncorra: Annotating Corpora*. Language Technologies Research Centre, IIIT, Hyderabad
7. Bhattacharya, P. 2010. *Lecture 11, Parts-of-speech tagging continued*. National Programme on Technology Enhanced Learning, NPTEL.
8. Choudhary N. 2011. Web-Drawn Corpus for Indian Language: A case of Hindi. CCIS. ICISIL 2011 (pp:139).
9. Choudhary N. And Jha, G. N. 2011. Creating Multilingual Parallel Corpora in Indian Languages. In Proceedings of 5th Language Technology Conference, Poznan.
10. Choudhary, N.(2011). *Automatic Identification and Analysis of Verb Groups in Hindi*. Unpublished doctoral thesis submitted to Centre for linguistics, Jawaharlala Nehru University. <http://nkchoudhary.com/Research/lrc-107-Narayan.pdf>
11. D'ejean H. (2000). How to evaluate and compare tagsets? A proposal. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (Lrec 2007)*, Athens.
12. Dalal, A., Nagaraj, K., Sawant, U., and Shelke, S.(2006). Hindi Part-of-Speech Tagging and Chunking: A Maximum Entropy Approach. In *Proceedings of NLP AI Machine Learning Competition*.

13. Dash, N. S.(2010). *Corpus Linguistics: A General Introduction*. CIIL, Mysore.
14. Dash, N.S. (2005). *Corpus Linguistics and Language Technology: With Reference to Indian Languages*. Mittal Publication, new Delhi.
15. Davis, R., Buchanan, B. G., & Shortliffe, E. (1977). *Production rules as a representation for a knowledge-based consultation program*. *Artificial Intelligence*, 8(1), 15–45. doi:10.1016/0004-3702(77)90003-0
16. Giménez, J. and Màrquez, L. 2004. SVMTool: A general POS tagger generator based on Support Vector Machines. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.
17. Giménez, J. and Màrquez, L.(2006). *Technical Manual v1.3*. Universitat Politecnica de Catalunya, Barcelona.
18. Gopal, M. (2012). Annotating Bundeli Corpus Using the BIS POS Tagset. In *Proceeding of workshop on Indian Language Data: Resources and Evaluation (WILDRE). Conference on Language Resources and Evaluation (LREC 2012)*, pp. 50-56.
19. Guilder, L. V. (1995). *Automated Part of Speech Tagging: A Brief Overview, Handout for LING361*. Fall 1995, Georgetown University.
20. Hardie A. And McEnery, T. (2012). *Corpus Linguistics: Method, theory and practise*. Cambridge University Press.
21. Hasan, F. M. (2006). *Comparison Of Different POS Tagging Techniques for Some South Asian Languages*, thesis submitted in Bachelor of Science in Computer Science and Engineering from <http://dspace.bracu.ac.bd/bitstream/handle/10361/83/Comparison%20of%20different%20pos%20tagging%20techniques.pdf?sequence=1>
22. Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York. MR 1851606
23. Jha, G. N. (2009). *Indian Language Corpora Initiative (ILCI)*. Invited talk, 4th International Language and Technology Conference (4th LTC), Poland.
24. Jha, G. N. (2010). The TDIL program and the Indian language corpora initiative (ILCI). In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010)*. European Language Resources Association (ELRA), pp. 982-985.
25. Joachims, T.(1999). *Making large-Scale SVM Learning Practical*. Schölkopf B. and Burges C. and Smola A. (ed.). *Advances in KernelMethods - Support Vector Learning*. MIT-Press.

26. Jurafsky, D. & Martin, J. H. (2006). *Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing*. Prentice Hall Englewood Cliffs, New Jersey 07632.
27. Kachru, B. (2008). *Languages in South Asia*. Cambridge University Press.
28. Kachru, Y. (2006). *Hindi*. John Benjamins Publishing Company, Amsterdam/ Philadelphia.
29. Karthik, K. G, Sudheer, K, Avinesh, Pvs. (2006). Comparative Study of various Machine Learning Methods for Telugu Part of Speech Tagging. In *Proceedings of the NLP AI Machine Learning 2006 Competition*.
30. Kecman, Vojislav. (2006). *Support Vector Machines for Pattern Classification by Shigeo Abe*. SIAM Review, Vol. 48, No. 2 pp. Society for Industrial and Applied Mathematics (pp. 418-421)
31. Kumar, D., Josan & Gurpreet S. (2010). *Parts of speech tagger for morphologically rich Indian Languages: A Survey*. *International Journal of Computer Applications* (0975 – 8887. Volume 6– No.5.
32. Kumar, R., Kaushik, S., Nainwani, P., Banerjee, E., Hadke, S., & Jha, Girish Nath. (2012). Using the ILCI Annotation Tool for POS Annotation: A Case of Hindi. *IJCLA Vol. 3, NO. 2, Jul-Dec 2012*, pp. 93–104.
33. Litisseliti, L. (2010). *Research Methodology in Linguistics*.(ed) Litosseliti, L..Continuum International Publishing Group.
34. MacKinlay, A. (2005). *The Effects of Part-of-Speech Tagsets on Tagger Performance*. Undergraduate Thesis submitted to University of Melbourne, 2005.
35. Moguerza, J. M. (2006). Support Vector Machines with application. *Statistical Science Vol. 21*,(pp 322-336). Institute of Mathematical Statistics
36. Mukherjee, S. and Vapnik, V. (1999). *Multivariate density estimation: A support vector machine approach*. Technical Report, AI Memo 1653, MIT AI Lab.
37. Nainwani, P., Banerjee, E., Kaushik, S., & Jha, Girish Nath. (2011). Issues in annotating less resourced languages—the case of Hindi from Indian Languages Corpora Initiative (ILCI). In *Proceedings of 5th Language Technology Conference (LTC 2011)*.
38. Patnaik, B. N. 2001. *Nominative and non-nominative constructions in Oriya*. Retrieved on February 7, 2010 from <http://home.iitk.ac.in/~patnaik/documents/nnom.pdf>
39. Ristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press.

40. Samuelsson, C. and Voutilainen, A. (1997). Comparing a Linguistic and a Stochastic Tagger. In *Proceedings of Joint 35th Annual Meeting of the ACL and 8th Conference of the European Chapter of the ACL 1997*.
41. Shrivastava, M., Agrawal, N., Singh, S., Bhattacharya, P. (2005). Harnessing Morphological Analysis in POS Tagging Task. In *Proceedings of the ICON 2005*.
42. Shukla, S. 1984. *Adhunik Avadh aur Bhojpuri: Etahas aur Kavya*. Anubhav Prakashan, Kanpur.
43. Singh, S. (2009). *Bhojpuri aur Hindi*. Vishwavidyalaya Prakashan, Varanasi.
44. Singh, S. And Banerjee, E. (2014). Annotating Bhojpuri Corpus Using BIS Scheme. In *Proceedings of 2nd Workshop on Indian Language Data: Resource and Evaluation* at (LREC 2014).
45. Sober, M. M., & Benedito, J. R. M. (Eds.). (2010). *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. Information Science Reference.
46. Upadhyay, H. S. (1988). *Bhojpuri folksongs from Ballia*. India Enterprises Incorporated.
47. Upadhyay, K. D. (2008). *Bhojpuri Liksahitya*.(ed II). Vishwavidyalaya Prakashan, Varanasi.
48. Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer, New York. MR 1367965
49. Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, New York. MR 1641250

Internet links

1. <http://anjoria.wordpress.com/> accessed 15/04/2015
2. <http://datahub.io/dataset/opus>
3. <http://dspace.bracu.ac.bd/bitstream/handle/10361/83/Comparison%20of%20different%20pos%20tagging%20techniques.pdf?sequence=1>
4. <http://en.wikipedia.org/wiki/Malayalam>
5. http://folk.uio.no/jannebj/PUBLIKASJONER_TIL_NEDLASTING/J&M.pdf
6. <http://home.iitk.ac.in/~patnaik/documents/nnom.pdf>
7. <http://ilk.uvt.nl/mbt/>
8. <http://ilk.uvt.nl/mbt/>
9. <http://linguistics.okfn.org>
10. <http://norivers.org/>
11. <http://opus.lingfil.uu.se/>

12. <http://regicon2015.sanchay.co.in/files/schedule.pdf>
13. <http://sampark.iiit.ac.in/sampark/web/index.php/registration/validuser>
14. <http://sanskrit.jnu.ac.in/pos/bhu.jsp>
15. <http://sanskrit.jnu.ac.in/rstudents/Chandra/thesis.pdf>
16. <http://tatkakhabar.com/>
17. <http://trendsarrived.com/>
18. <http://www.bhojpurika.com/>
19. <http://www.ethnologue.com/language/bho>
20. http://www.indsenz.com/int/index.php?content=sanskrit_tagger
21. <http://www.jstor.org/stable/20453817> accessed: 15/01/2015 01:00
22. <http://www.jstore.org/stable/27645765> accessed on 15/01/2015 00:57
23. <http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-WILDRE%20Proceedings.pdf>
24. <http://www.thesundayindian.com/>
25. <http://www.thesundayindian.com/bh/category/4/> accessed 05/01/2015
26. <http://anjoria.wordpress.com>
27. <http://anjoria.wordpress.com/tag/कइसन-पियवा-के-चरित्तर-बा/> accessed 15/04/2015
28. <http://bhojpurika.com/wp-content/uploads/2014/06/HelloBhojpuri-June14.pdf> accessed 07/01/2015
29. <http://norivers.org/amazing/box/0SbjEn0vZ3k/.html> accessed 15/04/2015
30. <http://tatkakhabar.com/?cat=5> accessed 15/04/2015
31. <http://tatkakhabar.com/?s=अप्रैल%2C+2015> accessed 15/04/2015
32. <http://tatkakhabar.com/?s=जनवरी%2C+2015>
33. <http://tatkakhabar.com/?s=दिसम्बर%2C+2014> accessed 15/04/2015
34. <http://ethnologue.com/language/bho> accessed 05/06/2015 time 01:25:08 am
35. <http://ilk.uvt.nl/mbt/>
36. <http://linguistics.okfn.org>
37. <http://datahub.io/dataset/opus>

38. <http://opus.lingfil.uu.se/>
39. <http://regicon2015.sanchay.co.in/files/schedule.pdf>
40. <http://www.lancaster.ac.uk/fass/projects/corpus/emille/>
41. ¹<http://www.ciilcorpora.net/>
42. <http://www.ciil-spokencorpus.net/>
43. <http://www.ldcil.org/default.aspx>
44. <http://www.ldcil.org/workInProgress.aspx>
45. <http://sanskrit.jnu.ac.in/ilci/index.jsp>
46. <http://aclweb.org/anthology/W/W12/W12-3152.pdf>
47. <http://sampark.iiit.ac.in/sampark/web/index.php/registration/validuser>
48. <http://sanskrit.jnu.ac.in/pos/index.jsp>
49. <http://en.wikipedia.org/wiki/Malayalam>
50. http://www.indsenz.com/int/index.php?content=sanskrit_tagger
51. <http://tatkakhabar.com/?s=मार्च%2C+2015> accessed 15/04/2015
52. <http://trendsarrived.com/files/box/BPZwrf4M7Bk/तर-मई-क-द.html> accessed
15/04/2015
53. <http://www.bhojpurika.com/category/blogs/> accessed 07/01/2015
54. <http://www.bhojpurika.com/category/columns/> accessed 07/01/2015
55. <http://www.bhojpurika.com/category/concerns/>
56. <http://www.bhojpurika.com/category/concerns/राजनीति/> accessed 07/01/2015
57. <http://www.bhojpurika.com/category/entertainment/> accessed 07/01/2015
58. <http://www.bhojpurika.com/category/literature/> accessed 07/02015
59. <http://www.facebook.com/ApniBhojpuri/posts/438608122886529> accessed 15/04/2015
60. <http://www.facebook.com/JayBhojpuri/posts/682974015065210> accessed 15/04/2015

61. http://www.facebook.com/permalink.php?id=331015513661458&story_fbid=641342285962111 accessed 15/04/2015
62. <http://www.thesundayindian.com/bh/blogs/> accessed 05/01/2015
63. <http://www.thesundayindian.com/bh/category/2/> accessed 07/01/2015
64. <http://www.thesundayindian.com/bh/category/खेल/4/1> accessed 07/01/2015
65. <http://www.thesundayindian.com/bh/category/> accessed 07/01/2015

Appendices

Appendix A. Tagger generated evaluation report for misc test set (90 k training file)

| | | | | |
|---|------------------------|---|--|--|
| EVALUATION OF TRAINING SET ON GOLD CORPUS OF SAME TRAINING SET (EVAL B) | | | | |
| * ===== SVMTeval report ===== | | | | |
| * model | = | [/home/sanskrit/svmtool/models/bho/BHO] | | |
| * testset (gold) | = | [/home/sanskrit/svmtool/bhogold.txt] | | |
| * testset (predicted) = [/home/sanskrit/svmtool/bho.eval] | | | | |
| * ===== | | | | |
| EVALUATING </home/sanskrit/svmtool/bho.eval> vs. </home/sanskrit/svmtool/bhogold.txt> on model </home/sanskrit/svmtool/models/bho/BHO>... | | | | |
|9863 tokens [DONE] | | | | |
| * ===== TAGGING SUMMARY ===== | | | | |
| #TOKENS | = 9863 | | | |
| AVERAGE_AMBIGUITY = 1.7249 tags per token | | | | |
| * ----- | | | | |
| #KNOWN | = 100.0000 % --> | 9863 / 9863 | | |
| #UNKNOWN | = 0.0000% --> | 0 / 9863 | | |
| #AMBIGUOUS | = 51.2724% --> | 5057 / 9863 | | |

| | | | | |
|--|----------------------|------------------|-------------------|-----|
| #MFT baseline | = 94.2411% --> | 9295 / 9863 | | |
| * ===== KNOWN vs UNKNOWN TOKENS ===== | | | | |
| | HITS | TRIALS | ACCURACY | |
| * ----- | | | | |
| * ===== known ===== | | | | |
| | 8741 | 9863 88.6242% | | |
| ----- known unambiguous tokens ----- | | | | |
| | 4805 | 4806 99.9792% | | |
| ----- known ambiguous tokens ----- ----- | | | | |
| | 3936 | 5057 77.8327% | | |
| * ===== unknown ===== | | | | |
| | 0 | 0 | 0.00% | |
| * ===== | | | | |
| * ===== ACCURACY PER LEVEL OF AMBIGUITY ===== | | | | |
| #CLASSES = 6 | | | | |
| * ===== | | | | |
| | LEVEL | HITS | TRIALS ACCURAC | MFT |

| | | | | |
|--|-------|------|------------------------|-----------|
| | | | Y | |
| * ----- | | | | |
| | 1 | 4805 | 4806 99.9792% | |
| | | | 99.9792% | |
| | 2 | 2513 | 3317 75.7612% | |
| | | | 88.7850% | |
| | 3 | 1190 | 1448 82.1823% | |
| | | | 88.9503% | |
| | 4 | 188 | 238 78.9916% | |
| | | | 86.9748% | |
| | 5 | 45 | 47 95.7447% | |
| | | | 97.8723% | |
| | 6 | 0 | 7 | 0.0000% |
| | | | | 57.1429% |
| * ===== ACCURACY PER CLASS OF AMBIGUITY ===== | | | | |
| #CLASSES = 230 | | | | |
| * ===== | | | | |
| ===== | | | | |
| | CLASS | HITS | TRIALS ACCURAC Y | MFT |
| * ----- | | | | |
| CC_CCD | 11 | 11 | 100.0000% | |
| | | | 100.0000% | |
| CC_CCD_CC_CCS | 119 | 119 | 100.0000% | 100.0000% |
| CC_CCD_CC_CCS_CC_cCD_V_VAUX_V_V M | 6 | 7 | 85.7143% | |
| | | | 100.0000% | |
| CC_CCD_CC_CCS_DM_DMR_RP_RPD | 9 | 10 | 90.0000% | |
| | | | 90.0000% | |

| | | | |
|---|-----|-----|------------------------|
| CC_CCD_CC_CCS_N_NN_QT_QTF_V_VAU X_V_VM | 0 | 5 | 0.0000% 60.0000% |
| CC_CCD_CC_CCS_RB_V_VAUX_V_VM | 1 | 1 | 100.0000% 100.0000% |
| CC_CCD_DM_DMD_N_NNP_PR_PRP | 3 | 3 | 100.0000% 100.0000% |
| CC_CCD_N_NN | 6 | 7 | 85.7143% 100.0000% |
| CC_CCD_PSP | 82 | 83 | 98.7952% 98.7952% |
| CC_CCD_QT_QTF | 1 | 1 | 100.0000% 100.0000% |
| CC_CCD_QT_QTF_RP_INTF | 44 | 44 | 100.0000% 100.0000% |
| CC_CCD_RP_INJ | 40 | 40 | 100.0000% 100.0000% |
| CC_CCD_RP_INTF | 12 | 13 | 92.3077% 92.3077% |
| CC_CCD_RP_RPD | 10 | 10 | 100.0000% 100.0000% |
| CC_CCD_V_VAUX_V_VM | 11 | 21 | 52.3810% 71.4286% |
| CC_CCD_V_VM | 133 | 147 | 90.4762% 94.5578% |
| CC_CCS | 7 | 7 | 100.0000% 100.0000% |
| CC_CCS_DM_DMQ_PR_PRQ | 3 | 3 | 100.0000% 100.0000% |
| CC_CCS_DM_DMQ_PR_PRQ_V_VM | 7 | 8 | 87.5000% 87.5000% |
| CC_CCS_DM_DMR_N_NN_N_NST | 20 | 22 | 90.9091% 90.9091% |

| | | | | |
|---------------------------------------|----|--------------|-----------|-------|
| CC_CCS_DM_DMR_N_NST | 1 | 1 100.0000% | 100.0000% | |
| CC_CCS_DM_DMR_PR_PRP | 2 | 2 100.0000% | 100.0000% | |
| CC_CCS_DM_DMR_RB | 4 | 4 100.0000% | 100.0000% | |
| CC_CCS_JJ_N_NN_RB | 2 | 3 66.6667% | 66.6667% | |
| CC_CCS_JJ_RB | 0 | 1 | 0.00% | 0.00% |
| CC_CCS_N_NN_PSP_V_VM | 33 | 44 75.0000% | 79.5455% | |
| CC_CCS_N_NN_RP_RPD | 63 | 63 100.0000% | 100.0000% | |
| CC_CCS_PSP_V_VAUX_V_VM | 9 | 9 100.0000% | 100.0000% | |
| CC_CCS_RB | 1 | 1 100.0000% | 0.00% | |
| CC_CCS_RP_INJ_V_VAUX_V_VM | 2 | 3 66.6667% | 66.6667% | |
| CC_CCS_V_VAUX | 4 | 6 66.6667% | 100.0000% | |
| DM_DMD | 12 | 12 100.0000% | 100.0000% | |
| DM_DMD_DM_DM_Q_PR_PR_P_V_VAUX | 10 | 10 100.0000% | 100.0000% | |
| DM_DMD_DM_DMR_N_NST_PR_PRL_PR_P RP | 1 | 1 100.0000% | 100.0000% | |
| DM_DMD_DM_DMR_PR_PRL_PR_PR_P | 10 | 10 100.0000% | 100.0000% | |
| DM_DMD_DM_DMR_PR_PR_P | 1 | 2 50.0000% | 50.0000% | |

| | | | | |
|--|-----|--------------|-----------|----------|
| DM_DMD_DM_DMR_RB | 11 | 11 100.0000% | 100.0000% | |
| DM_DMD_N_NN_N_NNP_PR_PRP_QT_QTC _RP_INJ | 0 | 2 | 0.0000% | 50.0000% |
| DM_DMD_N_NN_PR_PRP | 3 | 3 100.0000% | 100.0000% | |
| DM_DMD_N_NN_PR_PRP_PSP | 3 | 9 33.3333% | 55.5556% | |
| DM_DMD_N_NST_PR_PRP | 1 | 1 100.0000% | 100.0000% | |
| DM_DMD_PR_PRP | 175 | 178 98.3146% | 97.7528% | |
| DM_DMD_PR_PRP_QT_QTC | 2 | 2 100.0000% | 100.0000% | |
| DM_DMD_PR_PRP_V_VM | 1 | 2 50.0000% | 100.0000% | |
| DM_DMI_DM_DMQ | 2 | 3 66.6667% | 33.3333% | |
| DM_DMI_DM_DMQ_DM_DMR | 2 | 2 100.0000% | 100.0000% | |
| DM_DMI_DM_DMQ_N_NST_PR_PRI_PR_PR Q | 9 | 9 100.0000% | 100.0000% | |
| DM_DMI_DM_DMQ_PR_PRI | 12 | 12 100.0000% | 100.0000% | |
| DM_DMI_DM_DMQ_PR_PRQ | 1 | 1 100.0000% | 100.0000% | |
| DM_DMI_DM_DMR | 3 | 4 75.0000% | 75.0000% | |
| DM_DMI_JJ_PR_PRI_QT_QTC_QT_QTF | 5 | 5 100.0000% | 100.0000% | |
| DM_DMI_PR_PRI | 7 | 7 100.0000% | 100.0000% | |

| | | | | |
|---------------------------------|----|--------------|-----------|----------|
| DM_DMI_PR_PRI_QT_QTF | 6 | 13 46.1538% | 46.1538% | |
| DM_DMQ | 2 | 2 100.0000% | 100.0000% | |
| DM_DMQ_DM_DMR | 2 | 3 66.6667% | 66.6667% | |
| DM_DMQ_N_NST_PR_PRQ | 2 | 2 100.0000% | 100.0000% | |
| DM_DMQ_N_NST_V_VM | 1 | 2 50.0000% | 50.0000% | |
| DM_DMQ_PR_PRQ | 9 | 12 75.0000% | 100.0000% | |
| DM_DMQ_PR_PRQ_PSP | 94 | 99 94.9495% | 94.9495% | |
| DM_DMR | 24 | 24 100.0000% | 100.0000% | |
| DM_DMR_N_NN | 9 | 10 90.0000% | 90.0000% | |
| DM_DMR_N_NN_N_NNP_PR_PRI_PR_PRL | 6 | 6 100.0000% | 100.0000% | |
| DM_DMR_N_NN_PR_PRL | 0 | 2 | 0.0000% | 50.0000% |
| DM_DMR_N_NN_RB | 1 | 1 100.0000% | 100.0000% | |
| DM_DMR_N_NN_V_VM | 2 | 2 100.0000% | 100.0000% | |
| DM_DMR_N_NST | 6 | 6 100.0000% | 100.0000% | |
| DM_DMR_N_NST_PR_PRP_PSP | 5 | 7 71.4286% | 71.4286% | |
| DM_DMR_N_NST_PSP_V_VM | 1 | 1 100.0000% | 0.00% | |

| | | | |
|-----------------------------|-----|---------------------------|--|
| DM_DMR_PR_PRI_PR_PRL | 14 | 14 100.0000% 100.0000% | |
| DM_DMR_PR_PRI_PR_PRL_PR_PRP | 7 | 7 100.0000% 100.0000% | |
| DM_DMR_PR_PRL | 6 | 8 75.0000% 62.5000% | |
| DM_DMR_PR_PRL_PR_PRP | 1 | 3 33.3333% 33.3333% | |
| DM_DMR_PR_PRP | 2 | 2 100.0000% 100.0000% | |
| DM_DMR_RB | 11 | 11 100.0000% 100.0000% | |
| DM_DMR_RP_INTF | 6 | 6 100.0000% 100.0000% | |
| JJ | 259 | 259 100.0000% 100.0000% | |
| JJ_N_NN | 96 | 147 65.3061% 78.2313% | |
| JJ_N_NNP | 81 | 89 91.0112% 95.5056% | |
| JJ_N_NN_N_NNP | 21 | 27 77.7778% 85.1852% | |
| JJ_N_NN_N_NNP_PR_PRQ | 11 | 12 91.6667% 91.6667% | |
| JJ_N_NN_N_NNP_RP_RPD | 2 | 2 100.0000% 100.0000% | |
| JJ_N_NN_N_NST | 4 | 5 80.0000% 80.0000% | |
| JJ_N_NN_N_NST_PR_PRP | 5 | 6 83.3333% 83.3333% | |
| JJ_N_NN_N_NST_RB | 2 | 2 100.0000% 100.0000% | |
| JJ_N_NN_PR_PRP_V_VM | 16 | 16 100.0000% | |

| | | | | |
|--------------------------|----|--------------|-----------|-------|
| | | 100.0000% | | |
| JJ_N_NN_PSP | 3 | 4 75.0000% | 75.0000% | |
| JJ_N_NN_PSP_V_VM | 2 | 4 50.0000% | 50.0000% | |
| JJ_N_NN_QT_QTF | 3 | 4 75.0000% | 75.0000% | |
| JJ_N_NN_V_VAUX_V_VM | 9 | 14 64.2857% | 92.8571% | |
| JJ_N_NN_V_VM | 26 | 34 76.4706% | 85.2941% | |
| JJ_N_NST | 11 | 11 100.0000% | 90.9091% | |
| JJ_N_NST_RB_V_VM | 1 | 1 100.0000% | 100.0000% | |
| JJ_PSP | 6 | 6 100.0000% | 100.0000% | |
| JJ_PSP_V_VM | 2 | 3 66.6667% | 66.6667% | |
| JJ_QT_QTF | 17 | 18 94.4444% | 83.3333% | |
| JJ_QT_QTF_RB | 0 | 1 | 0.00% | 0.00% |
| JJ_QT_QTF_RB_RP_INTF | 1 | 1 100.0000% | 100.0000% | |
| JJ_QT_QTF_RP_INJ_RP_INTF | 8 | 8 100.0000% | 100.0000% | |
| JJ_QT_QTF_RP_INTF | 3 | 3 100.0000% | 100.0000% | |
| JJ_QT_QTO | 1 | 2 50.0000% | 50.0000% | |
| JJ_RB | 14 | 15 93.3333% | 93.3333% | |

| | | | | |
|-------------------|------|----------------|-----------|-----------|
| JJ_RB_RP_RPD | 2 | 2 100.0000% | 100.0000% | |
| JJ_RP_INJ | 1 | 2 50.0000% | 100.0000% | |
| JJ_RP_INTF | 6 | 6 100.0000% | 83.3333% | |
| JJ_V_VAUX | 1 | 2 50.0000% | 100.0000% | |
| JJ_V_VAUX_V_VM | 3 | 14 21.4286% | 85.7143% | |
| JJ_V_VM | 14 | 25 56.0000% | 88.0000% | |
| JJ_V_vM | 0 | 1 | 0.0000% | 100.0000% |
| N_NN | 1837 | 1837 100.0000% | 100.0000% | |
| N_NNP | 598 | 598 100.0000% | 100.0000% | |
| N_NNP_N_NST | 1 | 1 100.0000% | 100.0000% | |
| N_NNP_PR_PRP | 12 | 12 100.0000% | 100.0000% | |
| N_NNP_PR_PRP_V_VM | 0 | 1 | 0.00% | 0.00% |
| N_NNP_QT_QTF | 35 | 35 100.0000% | 100.0000% | |
| N_NNP_RB | 1 | 1 100.0000% | 0.00% | |
| N_NNP_RD_ECH | 1 | 1 100.0000% | 0.00% | |
| N_NNP_RP_INJ | 27 | 30 90.0000% | 90.0000% | |
| N_NNP_RP_INTF | 3 | 3 100.0000% | 100.0000% | |

| | | | | |
|------------------------------|-----|-----|-----------|-----------|
| N_NNP_V_VAUX | 0 | 2 | 0.0000% | 100.0000% |
| N_NNP_V_VM | 1 | 4 | 25.0000% | 75.0000% |
| N_NN_N_NNP | 293 | 371 | 78.9757% | 87.6011% |
| N_NN_N_NNP_N_NST | 9 | 11 | 81.8182% | 72.7273% |
| N_NN_N_NNP_N_NST_V_VAUX_V_VM | 3 | 3 | 100.0000% | 100.0000% |
| N_NN_N_NNP_PR_PRP | 1 | 1 | 100.0000% | 0.00% |
| N_NN_N_NNP_RD_UNK | 2 | 2 | 100.0000% | 100.0000% |
| N_NN_N_NNP_RP_CL_RP_RPD_V_VM | 13 | 14 | 92.8571% | 92.8571% |
| N_NN_N_NNP_RP_INJ | 6 | 7 | 85.7143% | 85.7143% |
| N_NN_N_NNP_V_VAUX | 3 | 4 | 75.0000% | 100.0000% |
| N_NN_N_NNP_V_VAUX_V_VM | 3 | 9 | 33.3333% | 55.5556% |
| N_NN_N_NNP_V_VM | 21 | 27 | 77.7778% | 85.1852% |
| N_NN_N_NST | 29 | 34 | 85.2941% | 79.4118% |
| N_NN_N_NST_PR_PRP | 4 | 4 | 100.0000% | 100.0000% |
| N_NN_N_NST_PSP | 4 | 4 | 100.0000% | 100.0000% |
| N_NN_N_NST_QT_QTO | 4 | 4 | 100.0000% | 100.0000% |

| | | | | |
|----------------------------|----|---------------------------|---------|-----------|
| N_NN_N_NST_V_VM | 1 | 2 50.0000% | 0.00% | |
| N_NN_PR_PRP | 19 | 22 86.3636% 95.4545% | | |
| N_NN_PR_PRP_QT_QTF | 7 | 7 100.0000% 100.0000% | | |
| N_NN_PSP | 51 | 57 89.4737% 94.7368% | | |
| N_NN_PSP_RB | 15 | 15 100.0000% 100.0000% | | |
| N_NN_QT_QTC | 2 | 4 50.0000% 100.0000% | | |
| N_NN_QT_QTC_QT_QTO | 3 | 3 100.0000% 100.0000% | | |
| N_NN_QT_QTF | 14 | 14 100.0000% 92.8571% | | |
| N_NN_QT_QTF_RB_RP_INTF | 5 | 5 100.0000% 100.0000% | | |
| N_NN_QT_QTF_RB_RP_RPD_V_VM | 1 | 1 100.0000% 100.0000% | | |
| N_NN_QT_QTF_RD_RPD_V_VM | 1 | 1 100.0000% 100.0000% | | |
| N_NN_QT_QTF_RD_UNK | 0 | 1 | 0.00% | 0.00% |
| N_NN_QT_QTF_RP_INTF_V_VM | 0 | 3 | 0.0000% | 100.0000% |
| N_NN_QT_QTF_V_VM | 0 | 1 | 0.00% | 0.00% |
| N_NN_QT_QTO | 7 | 8 87.5000% 75.0000% | | |
| N_NN_RB | 19 | 25 76.0000% 88.0000% | | |
| N_NN_RB_RD_ECH_B | 1 | 1 100.0000% 100.0000% | | |

| | | | |
|---------------------|----|---------------------------|--|
| N_NN_RB_V_VAUX_V_VM | 1 | 8 12.5000% 100.0000% | |
| N_NN_RP_INJ | 3 | 3 100.0000% 100.0000% | |
| N_NN_RP_INTF | 3 | 6 50.0000% 83.3333% | |
| N_NN_RP_NEG | 8 | 8 100.0000% 100.0000% | |
| N_NN_RP_NEG_V_VM | 3 | 3 100.0000% 100.0000% | |
| N_NN_RP_RPD_V_VM | 8 | 8 100.0000% 100.0000% | |
| N_NN_V_VAUX | 10 | 12 83.3333% 75.0000% | |
| N_NN_V_VAUX_V_VM | 40 | 112 35.7143% 75.8929% | |
| N_NN_V_VM | 98 | 149 65.7718% 83.2215% | |
| N_NN_V_VM_V_vM | 4 | 4 100.0000% 100.0000% | |
| N_NST | 65 | 65 100.0000% 100.0000% | |
| N_NST_PR_PRP | 1 | 1 100.0000% 100.0000% | |
| N_NST_PR_PRQ | 9 | 9 100.0000% 100.0000% | |
| N_NST_PSP | 13 | 15 86.6667% 93.3333% | |
| N_NST_PSP_QT_QTO | 6 | 6 100.0000% 100.0000% | |
| N_NST_RB | 25 | 25 100.0000% 100.0000% | |

| | | | | |
|----------------------|-----|---------------|-----------|-----------|
| N_NST_V_VM | 2 | 2 100.0000% | 100.0000% | |
| PR_PRC_QT_QTC | 34 | 35 97.1429% | 97.1429% | |
| PR_PRC_QT_QTO | 1 | 1 100.0000% | 100.0000% | |
| PR_PRF | 51 | 51 100.0000% | 100.0000% | |
| PR_PRF_PR_PRL | 14 | 14 100.0000% | 100.0000% | |
| PR_PRF_PR_PRL_PR_PRP | 10 | 10 100.0000% | 100.0000% | |
| PR_PRF_PR_PRP | 13 | 16 81.2500% | 81.2500% | |
| PR_PRL | 8 | 8 100.0000% | 100.0000% | |
| PR_PRL_PR_PRP | 21 | 21 100.0000% | 100.0000% | |
| PR_PRL_V_VM | 0 | 1 | 0.0000% | 100.0000% |
| PR_PRP | 166 | 166 100.0000% | 100.0000% | |
| PR_PRP_QT_QTC | 5 | 5 100.0000% | 100.0000% | |
| PR_PRP_RP_INJ | 3 | 3 100.0000% | 100.0000% | |
| PR_PRP_V_VAUX | 1 | 1 100.0000% | 100.0000% | |
| PR_PRP_V_VAUX_V_VM | 1 | 8 12.5000% | 87.5000% | |
| PR_PRP_V_VM | 6 | 11 54.5455% | 72.7273% | |
| PR_PRQ | 3 | 3 100.0000% | 100.0000% | |

| | | | | |
|---------------------|-----|--------------|-------------------|--|
| PR_PRQ_PSP_V_VAUX | 579 | 607 95.3871% | 95.5519% | |
| PR_PRQ_RB | 2 | 2 100.0000% | 100.0000% | |
| PSP | 453 | 454 99.7797% | 99.7797% | |
| PSP_RD_RPD_RP_RPD | 1 | 1 100.0000% | 0.00% | |
| PSP_RP_INTF_V_VM | 0 | 2 | 0.0000% 100.0000% | |
| PSP_RP_RPD | 19 | 19 100.0000% | 100.0000% | |
| PSP_V_VAUX | 14 | 23 60.8696% | 56.5217% | |
| PSP_V_VAUX_V_VM | 51 | 87 58.6207% | 50.5747% | |
| PSP_V_VAUX_V_vAUX | 20 | 26 76.9231% | 92.3077% | |
| PSP_V_VM | 26 | 34 76.4706% | 94.1176% | |
| QT_QTC | 23 | 23 100.0000% | 100.0000% | |
| QT_QTC_QT_QTF_RP_CL | 29 | 29 100.0000% | 100.0000% | |
| QT_QTC_QT_QTO | 1 | 1 100.0000% | 100.0000% | |
| QT_QTC_RP_CL | 10 | 10 100.0000% | 100.0000% | |
| QT_QTF | 37 | 37 100.0000% | 100.0000% | |
| QT_QTF_QT_QTO | 1 | 1 100.0000% | 100.0000% | |

| | | | | |
|--------------------|-----|-----|-----------|-----------|
| QT_QTF_RD_UNK_V_VM | 0 | 1 | 0.0000% | 100.0000% |
| QT_QTF_RP_RPD | 0 | 2 | 0.0000% | 100.0000% |
| QT_QTF_V_VM | 4 | 4 | 100.0000% | 100.0000% |
| QT_QTO | 20 | 20 | 100.0000% | 100.0000% |
| RB | 36 | 36 | 100.0000% | 100.0000% |
| RD_ECH | 3 | 3 | 100.0000% | 100.0000% |
| RD_ECH_RD_ECH_B | 0 | 1 | 0.0000% | 100.0000% |
| RD_ECH_V_VM | 1 | 1 | 100.0000% | 100.0000% |
| RD_PUNC | 535 | 535 | 100.0000% | 100.0000% |
| RD_RD_PUNC | 284 | 284 | 100.0000% | 100.0000% |
| RD_SYM | 8 | 8 | 100.0000% | 100.0000% |
| RD_UNK | 1 | 1 | 100.0000% | 100.0000% |
| RD_UNK_V_VAUX | 0 | 2 | 0.0000% | 100.0000% |
| RP_INJ | 7 | 7 | 100.0000% | 100.0000% |
| RP_INJ_RP_INTF | 4 | 4 | 100.0000% | 100.0000% |
| RP_INJ_V_VAUX_V_VM | 7 | 35 | 20.0000% | 85.7143% |
| RP_INTF | 3 | 3 | 100.0000% | 100.0000% |
| RP_INTF_V_VAUX | 18 | 20 | 90.0000% | 95.0000% |

| | | | | |
|--|-----|---------------|------------------------|-----------|
| RP_INTF_V_VM | 1 | 1 100.0000% | 100.0000% | |
| RP_NEG | 24 | 24 100.0000% | 100.0000% | |
| RP_NEG_RP_RPD | 76 | 77 98.7013% | 98.7013% | |
| RP_RPD | 26 | 26 100.0000% | 100.0000% | |
| RP_RPD_V_VAUX | 5 | 5 100.0000% | 100.0000% | |
| RP_RPD_V_VM | 1 | 2 50.0000% | 50.0000% | |
| V_VAUX | 72 | 72 100.0000% | 100.0000% | |
| V_VAUX_V_VM | 362 | 854 42.3888% | 78.8056% | |
| V_VAUX_V_VM_V_vAUX | 0 | 4 | 0.0000% | 100.0000% |
| V_VM | 513 | 513 100.0000% | 100.0000% | |
| V_VM_V_vM | 5 | 7 71.4286% | 85.7143% | |
| V_vM | 1 | 1 100.0000% | 100.0000% | |
| * ===== ACCURACY PER PART-OF-SPEECH ===== | | | | |
| | POS | HITS | TRIALS ACCURAC Y | MFT |
| * ----- | | | | |
| CC_CCD | 279 | 290 96.2069% | 99.6552% | |
| CC_CCS | 94 | 97 96.9072% | 96.9072% | |

| | | | |
|--------|-------|--------------|---------------------------|
| DM_DMD | 153 | 154 99.3506% | 96.1039% |
| DM_DMI | 19 | 28 67.8571% | 71.4286% |
| DM_DMQ | 9 | 15 60.0000% | 66.6667% |
| DM_DMR | 95 | 101 94.0594% | 92.0792% |
| | JJ | 349 | 413 84.5036% 93.7046% |
| | N_NN | 2312 | 2393 96.6151% 97.3255% |
| | N_NNP | 821 | 885 92.7684% 95.7062% |
| | N_NST | 208 | 209 99.5215% 94.2584% |
| PR_PRF | 79 | 79 100.0000% | 100.0000% |
| PR_PRI | 15 | 15 100.0000% | 100.0000% |
| PR_PRL | 14 | 20 70.0000% | 80.0000% |
| PR_PRP | 327 | 339 96.4602% | 96.7552% |
| PR_PRQ | 27 | 31 87.0968% | 83.8710% |
| | PSP | 1408 | 1427 98.6685% 96.4961% |
| QT_QTC | 92 | 92 100.0000% | 100.0000% |
| QT_QTF | 107 | 109 98.1651% | 99.0826% |

| | | | | |
|--------------------------|-------|--------|---------------|-----------|
| QT_QTO | 28 | 29 | 96.5517% | |
| | RB | 48 | 55 87.2727% | 93.1034% |
| RD_ECH | 5 | 5 | 100.0000% | 80.0000% |
| RD_ECH_B | 0 | 1 | 0.0000% | 100.0000% |
| RD_PUNC | 819 | 819 | 100.0000% | 100.0000% |
| RD_SYM | 8 | 8 | 100.0000% | 100.0000% |
| RD_UNK | 1 | 1 | 100.0000% | 100.0000% |
| | RP_CL | 10 | 10 100.0000% | 100.0000% |
| RP_INJ | 8 | 13 | 61.5385% | 69.2308% |
| RP_INTF | 43 | 45 | 95.5556% | 93.3333% |
| RP_NEG | 111 | 111 | 100.0000% | 100.0000% |
| RP_RPD | 150 | 151 | 99.3377% | 98.6755% |
| V_VAUX | 239 | 723 | 33.0567% | 83.2642% |
| | V_VM | 862 | 1192 72.3154% | 84.3121% |
| | V_vM | 1 | 3 33.3333% | 66.6667% |
| * ===== OVERALL ACCURACY | | | | |
| ===== | | | | |
| | HITS | TRIALS | ACCURAC Y | MFT |
| * ----- | | | | |

| | | | | |
|--|------|------|--------|--------|
| | 8741 | 9863 | 88.62% | 94.24% |
| * ===== | | | | |
| ===== | | | | |
| sanskrit@sanskrit-ThinkCentre-M57e:~/svmtool/bin\$ | | | | |

Appendix B List of all unique classes of ambiguity

| S. No. | Original tags | Inconsistent tags | two tags | three tags | more than three tags |
|--------|---------------|--------------------------------------|----------------|-----------------------|--|
| 1 | CC_CCD | CC_CCD_CC_CCS_CC_cCD _V_VAUX_V_VM | CC_CCD_CC_CCS | CC_CCD_CC_CCS_DM_DMR | CC_CCD_CC_CCS_DM_DMR_RP_RPD |
| 2 | CC_CCS | JJ_V_VM_V_vM | CC_CCD_N_NN | CC_CCD_QT_QTF_RP_INTF | V_VM |
| 3 | DM_DMD | N_NN_V_VM_V_vM | CC_CCD_PSP | CC_CCD_V_VAUX_V_VM | CC_CCD_CC_CCS_RB_V_VAUX_V_VM |
| 4 | DM_DMI | PSP_V_VAUX_V_vAUX | CC_CCD_QT_QTF | CC_CCS_DM_DMR_PR_PRP | CC_CCD_DM_DMD_N_NNP_PR_PRP |
| 5 | DM_DMQ | RB_V_VM_V_vM | CC_CCD_RP_INJ | CC_CCS_DM_DMR_RB | CC_CCS_DM_DMQ_PR_PRQ_V_VM |
| 6 | DM_DMR | RD_RD_PUNC | CC_CCD_RP_INTF | CC_CCS_JJ_RB | CC_CCS_DM_DMR_N_NN_N_NST |
| 7 | JJ | PSP_RD_RPD_RP_RPD | CC_CCD_RP_RPD | CC_CCS_N_NN_RP_RPD | CC_CCS_JJ_N_NN_RB |
| 8 | N_NN | V_VAUX_V_VM_V_vAUX | CC_CCD_V_VM | DM_DMD_DM_DMR_PR_PRP | CC_CCS_N_NN_PSP_V_VM |
| 9 | N_NNP | V_VM_V_vM | CC_CCS_DM_DMR | DM_DMD_DM_DMR_RB | CC_CCS_PSP_V_VAUX_V_VM |
| 10 | N_NST | | CC_CCS_N_NST | DM_DMD_N_NN_N_NNP | CC_CCS_RP_INJ_V_VAUX_V_VM |
| 11 | PR_PRF | | CC_CCS_V_VAUX | DM_DMD_N_NN_PR_PRP | DM_DMD_DM_DMQ_PR_PRP_V_VAUX |
| 12 | PR_PRL | | DM_DMD_PR_PRP | DM_DMD_N_NST_PR_PRP | DM_DMD_DM_DMR_PR_PRL_PR_PRP |
| 13 | PR_PRP | | DM_DMI_PR_PRI | DM_DMD_PR_PRP_QT_QTC | DM_DMD_N_NN_N_NNP_PR_PRP_QT_QTC_R P_INJ |
| 14 | PSP | | DM_DMQ_PR_PRQ | DM_DMI_DM_DMQ_DM_DMR | DM_DMD_N_NN_PR_PRP_PSP |
| 15 | QT_QTC | | DM_DMR_N_NN | DM_DMI_DM_DMQ_PR_PRI | DM_DMD_N_NST_PR_PRP_RB |
| 16 | QT_QTF | | DM_DMR_N_NST | DM_DMI_DM_DMQ_PR_PRQ | DM_DMI_DM_DMQ_N_NST_PR_PRI_PR_PRQ |
| 17 | QT_QTO | | DM_DMR_PR_PRL | DM_DMI_PR_PRI_QT_QTF | DM_DMI_JJ_PR_PRI_QT_QTC_QT_QTF |
| 18 | RB | | DM_DMR_RB | DM_DMQ_N_NST_PR_PRQ | DM_DMR_N_NN_N_NNP_PR_PRI_PR_PRL |
| 19 | RD_ECH | | DM_DMR_RP_INTF | DM_DMQ_N_NST_V_VM | DM_DMR_N_NST_PR_PRP_PSP |
| 20 | RD_PUNC | | JJ_N_NN | DM_DMQ_PR_PRQ_PSP | DM_DMR_PR_PRI_PR_PRL_PR_PRP |
| 21 | RD_UNK | | JJ_N_NNP | DM_DMR_N_NN_RB | JJ_N_NN_N_NNP_RP_RPD |
| 22 | RP_INJ | | JJ_N_NST | DM_DMR_PR_PRI_PR_PRL | JJ_N_NN_N_NST_RB |
| 23 | RP_INTF | | JJ_PSP | JJ_N_NN_N_NNP | JJ_N_NN_V_VAUX_V_VM |
| 24 | RP_NEG | | JJ_QT_QTC | JJ_N_NN_QT_QTC | JJ_N_NST_RB_V_VM |
| 25 | RP_RPD | | JJ_QT_QTF | JJ_N_NN_QT_QTF | JJ_QT_QTF_RP_INJ_RP_INTF |
| 26 | V_VAUX | | JJ_RB | JJ_N_NN_V_VM | N_NN_N_NNP_N_NST_V_VAUX_V_VM |
| 27 | V_VM | | JJ_RD_ECH | JJ_QT_QTC_QT_QTO | N_NN_N_NNP_RP_CL_RP_RPD_V_VM |
| 28 | PR_PRQ | | JJ_RP_INJ | JJ_QT_QTF_RB | N_NN_N_NNP_V_VAUX_V_VM |
| 29 | RD_SYM | | JJ_RP_INTF | JJ_QT_QTF_RP_INTF | N_NN_QT_QTF_RB_RP_INTF |
| 30 | PR_PRC | | JJ_RP_RPD | JJ_RB_RP_RPD | N_NN_QT_QTF_RB_RP_RPD_V_VM |
| 31 | PR_PRI | | JJ_V_VM | JJ_V_VAUX_V_VM | N_NN_QT_QTF_RD_UNK |
| 32 | RD_ECH_B | | N_NNP_PR_PRP | N_NNP_PR_PRP_V_VM | N_NN_QT_QTF_RP_INTF_V_VM |
| 33 | | | N_NNP_PSP | N_NN_N_NNP_N_NST | N_NN_RB_V_VAUX_V_VM |
| 34 | | | N_NNP_QT_QTF | N_NN_N_NNP_RD_UNK | DM_DMD_DM_DMR_N_NST_PR_PRL_PR_PRP |
| 35 | | | N_NNP_RD_UNK | N_NN_N_NNP_RP_INJ | DM_DMR_N_NST_PSP_V_VM |
| 36 | | | N_NNP_RP_INJ | N_NN_N_NNP_V_VAUX | JJ_N_NN_N_NNP_PR_PRQ |
| 37 | | | N_NNP_RP_INTF | N_NN_N_NNP_V_VM | JJ_N_NN_N_NST_PR_PRP |
| 38 | | | N_NN_N_NNP | N_NN_N_NST_PSP | JJ_N_NN_PR_PRP_V_VM |
| 39 | | | N_NN_N_NST | N_NN_PR_PRP_QT_QTF | JJ_N_NN_PSP_V_VM |
| 40 | | | N_NN_PR_PRP | N_NN_QT_QTC_QT_QTO | JJ_QT_QTF_RB_RP_INTF |
| 41 | | | N_NN_PSP | N_NN_RB_RD_ECH_B | N_NN_QT_QTF_RD_RPD_V_VM |
| 42 | | | N_NN_PSP_RB | N_NN_RD_UNK_V_VM | |
| 43 | | | N_NN_QT_QTC | N_NN_RP_RPD_V_VAUX | |
| 44 | | | N_NN_QT_QTF | N_NN_V_VAUX_V_VM | |
| 45 | | | N_NN_QT_QTO | N_NST_PSP_QT_QTO | |
| 46 | | | N_NN_RB | PR_PRF_PR_PRL_PR_PRP | |
| 47 | | | N_NN_RD_ECH | PR_PRP_V_VAUX_V_VM | |
| 48 | | | N_NN_RD_UNK | PR_PRQ_PSP_V_VAUX | |
| 49 | | | N_NN_RP_INJ | PSP_RP_INTF_V_VM | |
| 50 | | | N_NN_RP_NEG | PSP_V_VAUX_V_VM | |

| S. No. | Original tags | Inconsistent tags | two tags | three tags | more than three tags |
|--------|---------------|-------------------|-----------------|----------------------|----------------------|
| 51 | | | N_NN_V_VAUX | QT_QTC_QT_QTF_RP_CL | |
| 52 | | | N_NN_V_VM | QT_QTC_V_VAUX_V_VM | |
| 53 | | | N_NST_PR_PRP | QT_QTF_RD_UNK_V_VM | |
| 54 | | | N_NST_PR_PRQ | RP_INJ_V_VAUX_V_VM | |
| 55 | | | N_NST_PSP | CC_CCS_DM_DMQ_PR_PRQ | |
| 56 | | | N_NST_QT_QTO | CC_CCS_DM_DMR_N_NST | |
| 57 | | | N_NST_RB | DM_DMD_PR_PRP_V_VM | |
| 58 | | | N_NST_V_VM | DM_DMR_N_NN_PR_PRL | |
| 59 | | | PR_PRC_QT_QTC | DM_DMR_N_NN_V_VM | |
| 60 | | | PR_PRF_PR_PRL | DM_DMR_PR_PRL_PR_PRP | |
| 61 | | | PR_PRF_PR_PRP | JJ_N_NN_PSP | |
| 62 | | | PR_PRL_PR_PRP | JJ_PSP_V_VM | |
| 63 | | | PR_PRP_QT_QTC | JJ_QT_QTO | |
| 64 | | | PR_PRP_RP_INJ | N_NN_N_NNP_PR_PRP | |
| 65 | | | PR_PRQ_RB | N_NN_RP_NEG_V_VM | |
| 66 | | | PSP_RP_RPD | N_NN_N_NST_PR_PRP | |
| 67 | | | PSP_V_VAUX | N_NN_N_NST_QT_QTO | |
| 68 | | | PSP_V_VM | N_NN_N_NST_V_VM | |
| 69 | | | QT_QTC_QT_QTF | N_NN_QT_QTF_V_VM | |
| 70 | | | QT_QTC_QT_QTO | N_NN_RP_RPD_V_VM | |
| 71 | | | QT_QTC_RP_CL | DM_DMD_DM_DMR_N_NST | |
| 72 | | | QT_QTF_V_VM | DM_DMD_N_NST_RB | |
| 73 | | | RD_UNK_V_VAUX | DM_DMQ_PR_PRQ_V_VM | |
| 74 | | | RP_INJ_RP_INTF | N_NNP_RB_V_VAUX | |
| 75 | | | RP_INJ_V_VAUX | | |
| 76 | | | RP_INJ_V_VM | | |
| 77 | | | RP_NEG_RP_RPD | | |
| 78 | | | RP_RPD_V_VAUX | | |
| 79 | | | RP_RPD_V_VM | | |
| 80 | | | V_VAUX_V_VM | | |
| 81 | | | CC_CCS_RB | | |
| 82 | | | DM_DMI_DM_DMQ | | |
| 83 | | | DM_DMI_DM_DMR | | |
| 84 | | | DM_DMQ_DM_DMR | | |
| 85 | | | DM_DMR_PR_PRP | | |
| 86 | | | JJ_N_NN_N_NST | | |
| 87 | | | JJ_V_VAUX | | |
| 88 | | | N_NNP_N_NST | | |
| 89 | | | N_NNP_RB | | |
| 90 | | | N_NNP_RD_ECH | | |
| 91 | | | N_NNP_V_VAUX | | |
| 92 | | | N_NNP_V_VM | | |
| 93 | | | N_NN_RP_INTF | | |
| 94 | | | PR_PRC_QT_QTO | | |
| 95 | | | PR_PRL_V_VM | | |
| 96 | | | PR_PRP_V_VAUX | | |
| 97 | | | PR_PRP_V_VM | | |
| 98 | | | QT_QTF_RP_RPD | | |
| 99 | | | RD_ECH_V_VM | | |
| 100 | | | RP_INTF_V_VM | | |
| 101 | | | DM_DMD_N_NST | | |
| 102 | | | PR_PRL_RB | | |
| 103 | | | QT_QTF_RP_INTF | | |
| 104 | | | RD_ECH_RD_ECH_B | | |

Appendix C List of other language sentences from the corpus

| S.No. | Data from other languages | |
|-------|--|----------|
| 1 | इमे भोजा अंगिरसो विरूपा दिवस्पुत्रासो असुरस्य वीरा | Sanskrit |
| 2 | श्रीकाशी विश्वनाथो विजयतेतराम् | Sanskrit |
| 3 | श्वेते वृषे समारूढा श्वेताम्बरधरा शुचिः महागौरी शुभं दद्यान्महादेवप्रमोददा | Sanskrit |
| 4 | चन्द्रहासोज्ज्वलकरा शार्दूलवरवाहना कात्यायनी शुभं दद्याद्देवी दानवघातिनी | Sanskrit |
| 5 | एकवेणी जपाकर्णपूरा नग्रा खरास्थिता लम्बोष्ठी कर्णिकाकर्णी तैलाभ्यक्तशरीरिणी वामपादोल्लसल्लोहलताकण्टकभूषणा | Sanskrit |
| 6 | वर्धनमूर्धध्वजा कृष्णा कालरात्रिर्भयङ्करी | Sanskrit |
| 7 | सिंहासनगता नित्यं पद्माश्रितकरद्वया शुभदास्तु | Sanskrit |
| 8 | त्वदीयं वस्तु गोविन्दम् तुभ्यमेव समर्पयामि | Sanskrit |
| 9 | लास्ट सेवन डेज | English |
| 10 | आल इज वेल | English |
| 11 | टेकेन फार ग्रान्टेड | English |
| 12 | मैने बहुत सी डायलाक कहानी लिख कर रखी है | Hindi |
| 13 | रक्तभूमि जिसके मुख्य कलाकार सुपर स्टार रवि किशन है बन कर तैयार है और यह कुछ ही दिनों में रुपहले पर्दे पर दर्शकों का मनोरंजन करने | Hindi |
| 14 | मैने बहुत सी डायलाक कहानी लिख कर रखी है | Hindi |
| 15 | रक्तभूमि जिसके मुख्य कलाकार सुपर स्टार रवि किशन है बन कर तैयार है और यह कुछ ही दिनों में रुपहले पर्दे पर दर्शकों का मनोरंजन करने | Hindi |
| 16 | सुक्रीत फिल्मस एवं जन एकता फिल्मस के बैनर तले बनीइस भोजपुरी फिल्म के निर्माता कृष्णा यादव रवि किशन ला रहे हैं | Hindi |
| 17 | रक्तभूमि एक्शन रोमांस के साथ थ्रिलर का चस्का वाली फिल्म | Hindi |
| 18 | अदालत ने कहा कि अगर कोई पूरी जानकारी के साथ इस्लाम और कुरान में विश्वास करता है तो उसका धर्म बदलना समझा जा सकता है | Hindi |
| 19 | परन्तु बिना इस्लाम को समझे या कुरान की जानकारी लिए सिर्फ किसी लड़के से शादी करने के लिए धर्म बदलने की अनुमति नहीं दी जा सकती। | Hindi |
| 20 | बीजेपी ने का ऐसा वीडियो जारी किया है जिसमें नीतीश नरेंद्र मोदी की तारीफ करते हुए नजर आ रहे हैं। | Hindi |
| 21 | यह फिल्म बन कर तैयार है और जल्द ही रिलीज की घोषणा की जायेगी। | Hindi |
| 22 | भोजपुरी स्टार खेसारी लाल यादव और एक्शन स्टार विराज भट्ट तथा हॉट गर्ल पुनम दुबे को लेकर बन रही फिल्म इंतकाम की शूटिंग मुंबई में शुरु हो गयी है। | Hindi |
| 23 | तो क्या अब रुपये किलो बिकेगा प्याज दिल्ली में प्याज के दाम से रुपए तक पहुंच गए हैं। | Hindi |

| | | |
|----|--|-------|
| 24 | हम सेवाओं को सुरक्षित और उपयोग में आसान बनाए रखने में मदद करने के लिए कुकी जैसे टूल का उपयोग करते हैं। | Hindi |
| 25 | सामाजिक मुद्दा पर बनल फिल्म गंगा देवी के बेसी से बेसी दर्शकन ले चहुँपावे खातिर एह फिल्म के हिंदी संस्करण श्रीमती नेताजी के नाम से रिलीज होखे वाला बा। | Hindi |
| 26 | हम किसी विज्ञापनदाता या साझेदार की साइट पर मौजूद पिक्सेल द्वारा भी कुकी रख सकते हैं। | Hindi |
| 27 | साइन अप करें विज्ञापन नियंत्रण गोपनीयता की मूलभूत बातें कुकी नीति शर्तें अधिक संसाधन इंटरैक्टिव टूल अवयस्क और सुरक्षा गोपनीयता पेज सुरक्षा पेज साइट नियंत्रण पेज डेटा नीति हम दुनिया को और खुला बनाने और जोड़े रखने के लिए अपने मिशन के भाग के रूप में आपको साझा करने का सामर्थ्य देते हैं। | Hindi |
| 28 | कुछ पोस्ट करते समय मैं यह कैसे चुनूँ कि उसे कौन देख पाए। | Hindi |
| 29 | यह इश्क नहीं आसां बस इतना समझ लीजे कि एक आग का दरिया है ओर डूब के जाना है। | Hindi |
| 30 | जब हमारे पास स्थान जानकारी होती है तो हम आपके और अन्य लोगों के लिए अपनी सेवाओं को तैयार करने के लिए उसका उपयोग करते हैं जैसे चेकइन करने और अपने क्षेत्र में स्थानीय ईवेंट या ऑफ़र के बारे में पता करने में आपकी मदद करना या आपके मित्रों को बताना कि आपपास हैं। | Hindi |
| 31 | हम उन विक्रेताओं सेवा प्रदाताओं और अन्य साझेदारों को जानकारी स्थानांतरित करते हैं जो दुनिया भर में हमारे व्यवसाय की सहायता करते हैं जैसे तकनीकी अवसंरचना सेवाएँ प्रदान करना यह विश्लेषण करना कि हमारी सेवाओं का कैसे उपयोग किया जाता है विज्ञापनों और सेवाओं की प्रभाविकता को मापना ग्राहक सेवा प्रदान करना भुगतानों की सुविधा देना या अकादमी शोध और सर्वेक्षण करना। | Hindi |
| 32 | हम डेटा को तब तक संग्रहीत करते हैं जब तक यह ऊपर वर्णित लोगों सहित आपको और अन्य लोगों को उत्पाद और सेवाएँ प्रदान करने के लिए आवश्यक हो। | Hindi |
| 33 | यहाँ तरीका बताया गया है सेवाएँ प्रदान करें बेहतर बनाएँ और विकसित करें। | Hindi |
| 34 | हम कुकी डिवाइस पहचानकर्ता स्थानीय संग्रहण या समान तकनीकों का कब उपयोग कर सकते हैं। | Hindi |
| 35 | अधिक जानें। | Hindi |
| 36 | कैसा रहा बतौर निर्माता कजरा मोहब्बतवाला का अनुभव। | Hindi |
| 37 | है जिससे आप ऑनलाइन संपर्क कर सकते हैं या इस पते पर डाक भेज सकते हैं। | Hindi |
| 38 | हमारी कंपनी हमेशा नया करती है। | Hindi |
| 39 | हम कभीकभी कुछ उत्पाद और सेवाएँ प्रदान करने में अपनी मदद के लिए सेवा प्रदाताओं का उपयोग करते हैं। | Hindi |

| | | |
|----|---|-------|
| 40 | आप हमारे सार्वजनिक रूप से उपलब्ध ऑडिट के कुकी अनुभाग पर भी एक नज़र डाल सकते हैं जो कि आयरलैंड के डेटा सुरक्षा आयुक्त के कार्यालय द्वारा किया जाता है जो हमारे द्वारा उपयोग की जाने वाली कुकीक के बारे में अधिक विवरण देता है। | Hindi |
| 41 | इसमें आपके द्वारा देखी जाने वाली वेबसाइटों और एप्लिकेशन उन वेबसाइटों और एप्लिकेशन पर आपके द्वारा हमारी सेवाओं के उपयोग से संबंधित जानकारी और साथ ही उस एप्लिकेशन या वेबसाइट के डेवलपर या प्रकाशक द्वारा आपको या हमें प्रदान की जाने वाली जानकारी शामिल है। | Hindi |
| 42 | लॉग इन करने और अपना पासवर्ड बदलने में मदद प्राप्त करें एक खाता बनाएँ अक्षम खातों के बारे में जानकारी पाएँ देखें कि पर क्या नया है मुख्य प्रश्न मैं अपना पासवर्ड कैसे बदलूँ। | Hindi |
| 43 | आपके खाते से संबद्ध जानकारी को तब तक रखा जाएगा जब तक आपका खाता हटा नहीं दिया जाता उसके बाद उत्पाद और सेवाएँ प्रदान करने के लिए हमें डेटा नहीं चाहिए होता। | Hindi |
| 44 | हम उन लोगों को एयरलाइन सेल का विज्ञापन दिखाने में उस विज्ञापनदाता की मदद करने के लिए एक विज्ञापन कंपनी के साथ काम करते हैं। | Hindi |
| 45 | मैं के लिए कैसे साइन अप करूँ। | Hindi |
| 46 | आपका ब्राउज़र या डिवाइस इन तकनीकों से संबंधित सेटिंग प्रदान कर सकता है। | Hindi |
| 47 | हम ए महान विभूति लोगन स पुछत बानी कि भोजपुरी के संवैधानिक दरजा मिले एकरा खातिर एह लोग के कवनो फरज नइखे। | Hindi |
| 48 | उदाहरण के लिए लॉग इन स्वीकृतियों के साथ अगर कोई व्यक्ति किसी ब्राउज़र से आपके किसी ऐसे खाते में लॉग इन करता है जिसका आपने पहले कभी भी उपयोग न किया हो तो हम उन्हें ब्लॉक करके अधिक जानकारी की माँग कर सकते हैं। | Hindi |
| 49 | एक गोत में। | Hindi |
| 50 | हम कुकी और समान तकनीकों का उपयोग क्यों करते हैं। | Hindi |
| 51 | साइन अप करें साइन अप करें लॉग इन करें मोबाइल मित्रों को ढूँढें बैज लोग पेज स्थान खेल स्थान के बारे में विज्ञापन बनाएँ पेज बनाएँ डेवलपर करियर गोपनीयता कुकी शर्तें मदद हिन्दी। | Hindi |
| 52 | हमारी सेवाओं पर या उनका उपयोग करके वाले एप्लिकेशन वेबसाइटें और तृतीयपक्ष एकीकरण। | Hindi |
| 53 | ईमेल या फ़ोन पासवर्ड मुझे लॉगइन रखें या के लिए साइन अप करें अपना पासवर्ड भूल गए। | Hindi |
| 54 | हम प्राथमिकताओं को स्टोर करने में अपनी मदद करने यह जानने कि आपने सेवाओं की सामग्री को कब देखा या उनसे इंटरैक्ट किया और आपको और अन्य लोगों को सोशल प्लगइन और अन्य कस्टमाइज़ की गई सामग्री और अनुभव प्रदान करने जैसे कि आपको और अन्य लोगों को सुझाव देने के लिए भी कुकी और समान तकनीकों का उपयोग कर सकते हैं। | Hindi |

| | | |
|----|---|-------|
| 55 | इन साझेदारों को सख्त गोपनीयता ज़िम्मेदारियों का इस तरीके से पालन करना होगा जो कि इस डेटा नीति और उन अनुबंधों के अनुसार होता है जो हमारे बीच हुए हैं। | Hindi |
| 56 | हम प्रासंगिक विज्ञापन दिखाने के लिए कुकी और समान तकनीकों का कैसे उपयोग करते हैं। | Hindi |
| 57 | आपके या अन्य लोगों द्वारा इन एप्लिकेशन और वेबसाइटों से साझा की जाने वाली आपसे संबंधित जानकारी को आप कैसे नियंत्रित कर सकते हैं इस बारे में और जानें। | Hindi |
| 58 | उदाहरण के लिए कुकी और समान तकनीकें हमें बताती हैं कि आपने कब में लॉग इन किया हुआ है ताकि आपके द्वारा हमारे सामाजिक प्लगइन का उपयोग करने वाली अन्य वेबसाइटों पर विज़िट किए जाने पर हम आपको प्रासंगिक और सोशल जानकारी दिखा सकें। | Hindi |
| 59 | उदाहरण के लिए जब आप हमारी साइट पर जाते हैं या हमारे एप्लिकेशन का उपयोग करते हैं तब हम कुकी रख या पढ़ सकते हैं या आपके डिवाइस की जानकारी प्राप्त कर सकते हैं। | Hindi |
| 60 | आप हमारे अपनी जानकारी डाउनलोड करें टूल का उपयोग करके भी अपने खाते से संबद्ध जानकारी डाउनलोड कर सकते हैं। | Hindi |
| 61 | चैनल चैबीसों घंटे मनोरंजन करी जवना में भोजन स्वास्थ्य संगीत फिल्मस सोप्स रियलिटी शोज़ कोमेडी यात्रा आ लाइफ़ स्टाइल से जुड़ल कार्यक्रम देखावल जाई। | Hindi |
| 62 | हम अपनी सेवाएँ प्रदान करने और उनका समर्थन करने में अपनी मदद के लिए अपने पास मौजूद सभी जानकारी का उपयोग करते हैं। | Hindi |
| 63 | अपना खाता ढूँढें अपना खाता ढूँढें ईमेल फ़ोन उपयोगकर्ता नाम या पूर्ण नाम रद्द करें मैं मेरे खाते की पहचान नहीं कर सकता साइन अप करें लॉग इन करें मोबाइल मित्रों को ढूँढें बैज लोग पेज स्थान खेल स्थान के बारे में विज्ञापन बनाएँ पेज बनाएँ डेवलपर करियर गोपनीयता कुकी शर्तें मदद हिन्दी। | Hindi |
| 64 | मराठी फिल्मों की तरह भोजपुरी फिल्में भी मल्टीप्लेक्स में लगनी चाहिए। | Hindi |
| 65 | इसका यह मतलब भी है कि किसी व्यक्ति द्वारा किसी ऐसी तृतीयपक्ष वेबसाइट या एप्लिकेशन को एक्सेस किए जाने पर को ये कुकी भेजी जाती हैं जिनमें हमारी सेवाएँ एकीकृत होती हैं या जो उनका उपयोग करते हैं जैसे हमारा कोई प्लगइन। | Hindi |
| 66 | हम जिन विज्ञापन कंपनियों के साथ काम करते हैं वे सामान्य रूप से अपनी सेवाओं के भाग के रूप में कुकी और समान तकनीकों का उपयोग करती हैं। | Hindi |
| 67 | हम किसी विज्ञापनदाता या साझेदार की साइट पर मौजूद पिक्सेल द्वारा भी कुकी रख सकते हैं। | Hindi |
| 68 | साइन अप करें विज्ञापन नियंत्रण गोपनीयता की मूलभूत बातें कुकी नीति शर्तें अधिक संसाधन इंटरैक्टिव टूल अवयस्क और सुरक्षा गोपनीयता पेज सुरक्षा पेज साइट नियंत्रण पेज डेटा नीति हम दुनिया को और खुला बनाने और जोड़े रखने के लिए अपने मिशन के भाग के रूप में आपको साझा करने का सामर्थ्य देते हैं। | Hindi |
| 69 | साइन अप करें बैज अन्तरजाल पर कहीं भी बाँटें। | Hindi |

| | | |
|----|--|-------|
| 70 | इन एप्लिकेशन वेबसाइटों या एकीकृत सेवाओं द्वारा एकत्रित की जाने वाली जानकारी इनकी अपनी शर्तों और नीतियों के अधीन होती है। | Hindi |
| 71 | सार्वजनिक जानकारी ऐसी कोई भी जानकारी जिसे आप सार्वजनिक ऑडियंस से साझा करते हैं और साथ ही आपकी सार्वजनिक प्रोफ़ाइल में मौजूद जानकारी या आपके द्वारा अपने पेज या किसी अन्य सार्वजनिक फ़ोरम पर साझा की जाने वाली सामग्री होती है। | Hindi |
| 72 | अगर आपके पास एप्लिकेशन नहीं है तो हम उसकी बजाय आपको में भेज देंगे ताकि आप एप्लिकेशन डाउनलोड कर सकें। | Hindi |
| 73 | अगर आप खरीदारियों या वित्तीय लेनदेनों के लिए हमारी सेवाओं का उपयोग करते हैं जैसे जब आप पर कुछ खरीदते हैं किसी गेम में कोई खरीदारी करते हैं या कोई दान करते हैं तो हम खरीदारी या लेनदेन से संबंधित जानकारी एकत्रित करते हैं। | Hindi |
| 74 | हम अपने विज्ञापन और मापन सिस्टम को बेहतर बनाने के लिए अपने पास मौजूद जानकारी का उपयोग करते हैं ताकि हम आपको अपनी सेवाओं पर और उसके बाहर प्रासंगिक विज्ञापन दिखा सकें और विज्ञापनों और सेवाओं की प्रभावशीलता और पहुँच माप सकें। | Hindi |
| 75 | साइन अप करें की शर्तें और नीतियाँ वह सभी चीज़ें जो आप जानना चाहते हैं एक स्थान पर है। | Hindi |
| 76 | आप विवाद में भी रहे हैं। HAD2209 | Hindi |
| 77 | और जानने के लिए हमारी कुकी नीति पढ़ें। | Hindi |
| 78 | आपके मोबाइल ऑपरेटर या का नाम ब्राउज़र प्रकार भाषा और समय क्षेत्र मोबाइल फ़ोन नंबर और पता जैसी कनेक्शन जानकारी। | Hindi |
| 79 | इसमें आपकी भुगतान जानकारी शामिल है जैसे आपका क्रेडिट या डेबिट कार्ड नंबर और अन्य कार्ड जानकारी और अन्य खाता और प्रमाणीकरण जानकारी और साथ ही बिलिंग शिपिंग और संपर्क विवरण। | Hindi |
| 80 | साइन अप करें। | Hindi |
| 81 | उदाहरण के लिए लोग आपकी फ़ोटो साझा कर सकते हैं किसी पोस्ट में किसी स्थान में आपका उल्लेख कर सकते या आपको टैग कर सकते हैं या आपसे संबंधित कोई ऐसी जानकारी साझा कर सकते हैं जो आपने उनसे साझा की हो। | Hindi |
| 82 | भुगतानों से संबंधित जानकारी। | Hindi |
| 83 | हो सकता है कि हम ट्रैक करते समय ब्राउज़र या डिवाइस सिग्नल को न पहचानें या उनका जवाब न दें और कुछ सेटिंग हमारे द्वारा प्रदान की जाने वाली सुविधाओं के आपके उपयोग में हस्तक्षेप कर सकती हैं। | Hindi |

| | | |
|----|---|-------|
| 84 | स्पेस क्रिएटिव मीडिया के रपट इण्डस्ट्री न्यूज कलाकार गंगा देवी गुलशन जनवरी से गंगादेवी की शूटिंग करेंगे बिगबी दिसम्बर गंगा और गंगोत्री जैसी फिल्मों से भोजपुरी दर्शकों से रुबरु हो चुके अमिताभ बच्चन एक बार फिर गंगादेवी के जरिये भोजपुरी दर्शकों के सामने होंगे निर्माता दीपक सावंत और निर्देशक अभिषेक चड्ढा की इस फिल्म में बिग बी के अलावा जया बच्चन और गुलशन ग्रोवर पहली बार भोजपुरिया दर्शकों से रुबरु होंगे । | Hindi |
| 85 | हम आपको शॉर्टकट और सुझाव देने के लिए भी अपने पास मौजूद जानकारी का उपयोग करते हैं । | Hindi |
| 86 | आपके द्वारा की जाने वाली चीज़ें और आपके द्वारा प्रदान की जाने वाली जानकारी । | Hindi |
| 87 | आप गतिविधि लॉग टूल द्वारा का उपयोग करते समय अपने द्वारा साझा की जाने वाली सामग्री और जानकारी प्रबंधित कर सकते हैं । | Hindi |
| 88 | जब आप अपने खाते को हटा देते हैं तो हम आपके द्वारा पोस्ट की गई चीज़ों को भी हटा देते हैं जैसे आपकी फ़ोटो और स्थिति अपडेट । | Hindi |
| 89 | उदाहरण के लिए हमारी सेवाओं पर कुछ जानकारी सार्वजनिक है और इसलिए उसे इंटरनेट पर कोई भी व्यक्ति एक्सेस कर सकता है । | Hindi |
| 90 | आप सेवाओं का उपयोग करते समय जिन तृतीय पक्षों से इंटरैक्ट करते हैं वे भी विभिन्न उद्देश्यों के लिए इन तकनीकों का उपयोग कर सकते हैं । | Hindi |
| 91 | जब आप हमसे संपर्क करते हैं तब हम आपको जवाब देने के लिए भी आपकी जानकारी का उपयोग करते हैं । | Hindi |
| 92 | हम आपको सेवाओं पर या उसके बाहर विज्ञापन दिखाने के लिए जैसे कि आपके द्वारा विज्ञापनदाता की साइट या एप्लिकेश पर विज़िट करने के बाद या आपके द्वारा विज़िट की जाने वाली वेबसाइटों या आपके द्वारा उपयोग किए जाने वाले एप्लिकेशन के आधार पर आपको एक विज्ञापन दिखाने के लिए किसी विज्ञापनदाता या उसके मार्केटिंग साझेदारों के साथ भी काम कर सकते हैं सबकुछ इंटरनेट और मोबाइल इकोसिस्टम में । | Hindi |
| 93 | डबल मीनिंग गानों को खत्म करना चाहिए । | Hindi |
| 94 | उनके द्वारा एकत्रित या प्राप्त की जाने वाली जानकारी के बारे में अधिक जानने के लिए उनकी गोपनीयता नीतियों की समीक्षा करें । | Hindi |
| 95 | हमें लोगों के लिए दिलचस्प और कस्टमाइज़ किए गए अनुभव बनाने का जुनून है । | Hindi |
| 96 | हम आपको विपणन संचार भेजने आपके साथ अपनी सेवाओं के बारे में संचार करने और आपको अपनी नीतियों और शर्तों के बारे में बताने के लिए आपकी जानकारी का उपयोग करते हैं । | Hindi |
| 97 | किसी ने कहा स्क्रीन टेस्ट के बिना हीरो नहीं बनते । | Hindi |
| 98 | मैं ऐसे लोगों से बचने की सलाह दूंगा । | Hindi |
| 99 | हम आपको सेवाओं पर और उनके बाहर विज्ञापन दिखने के लिए कुकी और समान तकनीकों का | Hindi |

| | | |
|-----|---|-------|
| | उपयोग करते हैं। | |
| 100 | लॉग इन स्वीकृतियों के साथ अगर कोई व्यक्ति किसी ब्राउज़र से आपके किसी ऐसे खाते में लॉग इन करता है जिसका आपने पहले कभी भी उपयोग न किया हो तो हम उन्हें ब्लॉक करके अधिक जानकारी की माँग कर सकते हैं। | Hindi |
| 101 | कभीकभी हम वेबसाइटों एप्लिकेशन और उनके साझेदारों के साथ काम करते हैं ताकि आपके द्वारा तृतीयपक्ष सेवाओं को विज़िट करने पर हम आपके ब्राउज़र या डिवाइस पर कुकी रख सकें या पढ़ सकें। | Hindi |
| 102 | बाकिर भोजपुरिहा लोग अइसे चाहे जतना बतकुञ्चन कर लेव बाकिर वेबसाइट पर लिखल सामग्री का बारे में कवनो तरह के टिप्पणी करे से बहुते सकुचालें। | Hindi |
| 103 | हमारी सेवाओं पर मौजूद विज्ञापनों और आपको दिखाई देने वाले विज्ञापनों को वैयक्तिकृत करने के लिए आपकी जानकारी का उपयोग किए जाने के तरीके को आप कैसे नियंत्रित कर सकते हैं इस बारे में और जानें। | Hindi |
| 104 | इससे हम अपने सेवाओं पर और उनके बाहर आपके द्वारा उपयोग किए जाने वाले एक से अधिक डिवाइस या ब्राउज़र से कुकी को पढ़ने और उनका संदर्भ लेने जैसी चीज़ें कर सकते हैं ताकि हम आपको आपके सभी डिवाइस पर सेवाएँ प्रदान कर सकें तथा हमारे द्वारा इंटरनेट पर आपको और अन्य लोगों को प्रदान किए जाने वाले उत्पादों विज्ञापनों और सेवाओं के बेहतर बनाकर उन्हें समझ सकें। | Hindi |
| 105 | हम यह जानने के लिए भी कुकी का उपयोग कर सकते हैं कि क्या सेवाओं पर कोई विज्ञापन देखने वाले किसी व्यक्ति ने बाद में विज्ञापनदाता की साइट पर खरीदारी की या वह एप्लिकेशन स्थापित किया जिसका विज्ञापन दिया गया था। | Hindi |
| 106 | हम आपके द्वारा दी गई अनुमतियों के आधार पर उन कंप्यूटर फ़ोन या अन्य डिवाइस की या उनसे संबंधित जानकारी एकत्रित करते हैं जहाँ आप हमारी सेवाओं को स्थापित या एक्सेस करते हैं। | Hindi |
| 107 | आपके द्वारा उपयोग की जाने वाली सेवाओं के आधार पर हम आपसे या आपके बारे में विभिन्न प्रकार की जानकारी एकत्रित करते हैं। | Hindi |
| 108 | अगर आप पर अपने विज्ञापन अनुभव को नियंत्रित और प्रबंधित करना चाहते हैं तो आप अपनी विज्ञापन प्राथमिकताओं को समायोजित कर सकते हैं। | Hindi |

Appendix D. List of Section headers from the Crawled data

| S. No. | Headers |
|--------|---|
| 1 | खोजें लोग पेज स्थान एप्लिकेशन पेज समुदाय उत्पादसेवा उत्पादसेवा वेबसाइट उत्पादसेवा उत्पादसेवा उत्पादसेवा उत्पादसेवा उत्पादसेवा उत्पादसेवा कंप्यूटरइंटरनेट वेबसाइट संगीतकारबैंड भोजनपेय उत्पादसेवा संगीतकारबैंड टीवी कार्यक्रम अभिनेतानिर्देशक एप्लिकेशन पेज मूवी साइन अप करें लॉग इन करें मोबाइल मित्रों को ढूँढें बैज लोग पेज स्थान खेल स्थान के बारे में विज्ञापन बनाएँ पेज बनाएँ डेवलपर करियर गोपनीयता कुकी शर्तें मदद हिन्दी । |
| 2 | खोजें लोग पेज स्थान शहर शहर नगर शहर शहर नगर नगर नगर नगर के पास स्थान नगर के पास स्थान शहर नगर नगर शहर शहर नगर नगर शहर शहर साइन अप करें लॉग इन करें मोबाइल मित्रों को ढूँढें बैज लोग पेज स्थान खेल स्थान के बारे में विज्ञापन बनाएँ पेज बनाएँ डेवलपर करियर गोपनीयता कुकी शर्तें मदद हिन्दी । |
| 3 | औरत आ जुआ |
| 4 | लस्टम पस्टम |
| 5 | भोजपुरी वेबसाइट्स |
| 6 | पुस्तक चर्चा भाषा सरोकार |
| 7 | स्वतंत्र कार्टून कोना कार्टून कोना नीमन कार्टून कोना |
| 8 | खबर भोजपुरी में शुक जनवरी खबर भोजपुरी में बियफे जनवरी खबर भोजपुरी में बुध जनवरी खबर भोजपुरी में मंगल जनवरी खबर भोजपुरी में सोमार जनवरी खबर भोजपुरी में अतवार जनवरी खबर भोजपुरी में शनिचर जनवरी खबर भोजपुरी में शुक जनवरी खबर भोजपुरी में जनवरी खबर भोजपुरी में बुध दिसंबर के ट्वीट्स के द्वारा किये गए ट्वीट्स अमित शाह अल कायदा अलकायदा आजम खां आरएसएस इस्लाम उच्चतम न्यायालय उत्तर प्रदेश कांग्रेस केजरीवाल गठबंधन गडकरी गरबा चारा घोटाला चार्जशीट चुनाव जम्मूकश्मीर दिल्ली धर्मांतरण नरेंद्र मोदी नरेन्द्र मोदी फेसबुक बिहार बीजेपी भाजपा भारत महाराष्ट्र मुख्यमंत्री मोदी सरकार यूपी योगी आदित्यनाथ राम मंदिर राहुल गांधी रेप लव जिहाद लव जेहाद शिवसेना शीला दीक्षित सरकार सर्वे सुप्रीम कोर्ट हत्या हरियाणा हिंदू हिन्दू । |
| 9 | देशदुनिया खबर भोजपुरी में खबर भोजपुरी में बुध जनवरी मंगल जनवरी के खबर |
| 10 | उत्तर प्रदेश बिहार राजनीति देशदुनिया खेल व्यापार अपराध मनोरंजन शिक्षा |
| 11 | उत्तर प्रदेश बिहार राजनीति देशदुनिया खेल व्यापार अपराध मनोरंजन शिक्षा फर्जी यूनिवर्सिटी खबर भोजपुरी में मंगल जनवरी |
| 12 | टटका खबर उत्तर प्रदेश बिहार राजनीति देशदुनिया खेल व्यापार अपराध मनोरंजन शिक्षा फर्जी यूनिवर्सिटी खबर भोजपुरी में शुक जनवरी बियफे |
| 13 | देशदुनिया खबर भोजपुरी में गाजीपुर चंदौली जौनपुर बलिया रोहतास सारण खबर भोजपुरी में सोमार दिसंबर अतवार |
| 14 | साहित्य भाषा टी॰वी॰ सरोकार सिनेमा योग संगीत गीत गवर्नई भउजी हो चर्चा वा लस्टम पस्टम अगडम |

| | |
|----|---|
| | बगड़म भोला बाबू मारीशस से । |
| 15 | देशदुनिया खबर भोजपुरी में खबर भोजपुरी में अतवार दिसंबर शनिचर दिसंबर के खबर |
| 16 | खबर भोजपुरी में शुक्र जनवरी खबर भोजपुरी में बियफे |
| 17 | खबर भोजपुरी में बुध जनवरी |
| 18 | खबर भोजपुरी में मंगल जनवरी खबर भोजपुरी में |
| 19 | सोमार जनवरी खबर भोजपुरी में अतवार जनवरी खबर भोजपुरी में शनिचर जनवरी खबर भोजपुरी में शुक्र जनवरी खबर भोजपुरी में जनवरी खबर भोजपुरी में बुध दिसंबर के ट्वीट्स के द्वारा किये गए ट्वीट्स अमित शाह अल कायदा अलकायदा आजम खां आरएसएस इस्लाम उच्चतम न्यायालय उत्तर प्रदेश कांग्रेस केजरीवाल गठबंधन गडकरी गरबा चारा घोटाला चार्जशीट चुनाव जम्मूकश्मीर दिल्ली धर्मांतरण नरेंद्र मोदी नरेन्द्र मोदी फेसबुक बिहार बीजेपी भाजपा भारत महाराष्ट्र मुख्यमंत्री मोदी सरकार यूपी योगी आदित्यनाथ राम मंदिर राहुल गांधी रेप लव जिहाद लव जेहाद शिवसेना शीला दीक्षित सरकार सर्वे सुप्रीम कोर्ट हत्या हरियाणा हिंदू हिन्दू । |
| 20 | भोजपुरी का डाॅट काॅम खबर पत्रिका किताब आ वेबसाइट्स भोजपुरी वेबसाइट्स ब्लॉग एक तरफा विचार चौपाल निजी नीकजबून मनोरंजन गीतगवनई टीवी फिल्म रंगमंच सरोकार कोर्स आ कैरियर देश आ समाज नारीजगत पर्व त्योहार भोजपुरिया लाल योग राजनीति सभा समारोह स्वास्थ्य साहित्य उपन्यास कविता कहानी निबन्ध पुस्तक चर्चा भाषा समीक्षा स्तम्भ अगड़म बगड़म कतरब्योत कोलकाता मेल ज्योतिष आ वास्तु बतकुच्चन भउजी हो भोला बाबू रामझरोखा से लस्टम पस्टम कार्टून कोना सभकर राय बनचरी दुसरकी कड़ी बनचरी दुसरकी कड़ी उपन्यास |
| 21 | भोजपुरिया डाॅट काॅम भोजपुरी पोर्टल भोजपुरिका डाॅट काॅम भोजपुरी में भोजपुरी के बात टटका खबर भोजपुरी समाचार अंजोरिया डाॅट काॅम भोजपुरी में सबले पहिला वेबसाइट भोजपुरी नामा डाॅट काॅम भोजपुरी विडियो आ गाना एमपीभोजपुरी डाॅट काॅम भोजपुरी गाना मस्त भोजपुरी डाॅट काॅम भोजपुरी गाना भोजपुरी डाॅट को भोजपुरी सिनेमा भोजपुरी मीडिया डाॅट काॅम भोजपुरी सिनेमा आ संगीत सनिमाहाॅल डाॅट काॅम भोजपुरी सिनेमा आ संगीत भोजपुरी गाना डाॅट इन भोजपुरी गाना भोजपुरी सम्राट डाॅट काॅम भोजपुरी गाने डाॅट काॅम जोगीरा डाॅट काॅम भोजपुरी सिनेमा भोजपुरिया सिनेमा डाॅट काॅम दि भोजपुरी डाॅट काॅम भोजपुरी वेबसाइट हेलो भोजपुरी डाॅट काॅम भोजपुरी पत्रिका भोजपुरी स्टोर डाॅट काॅम भोजपुरी गाना आपन भोजपुरी डाॅट इन भोजपुरी वेबसाइट भोजपुरी दंगल डाॅट इन भोजपुरी सिनेमा आ संगीत भोजपुरी जोन चौरीचौरा डाॅट काॅम भोजपुरी टुडे डाॅट काॅम भोजपुरी के रसगर गाना आ विडियो भोजपुरी नेट डाॅट काॅम भोजपुरी मस्ती डाॅट इन सारन वैप डाॅट इन भोजपुरी किंग डाॅट इन मनोज भावुक डाॅट काॅम भोजपुरी वेबसाइट भोजपुरी खोज डाॅट काॅम भोजपुरी समाचार भोजपुरी माटी डाॅट काॅम भोजपुरी वेबसाइट जय भोजपुरी डाॅट काॅम भोजपुरी सोशल साइट भोजपुरी डाॅट काॅम पूर्वांचल एक्सप्रेस भोजपुरी वेबसाइट माय भोजपुरी वैप डाॅट काॅम ग्लेमरस डाॅट काॅम । |
| 22 | भोजपुरी वेबसाइट्स भोजपुरी वेबसाइट्स |
| 23 | टटका खबर उत्तर प्रदेश बिहार राजनीति देशदुनिया खेल व्यापार अपराध मनोरंजन शिक्षा फर्जी यूनिवर्सिटी खबर भोजपुरी में जनवरी बुध दिसंबर के खबर |

| | |
|----|---|
| 24 | भोजपुरी का डाॅट काॅम खबर पत्रिका किताब आ वेबसाइट्स भोजपुरी वेबसाइट्स ब्लॉग एक तरफा विचार चौपाल निजी नीकजबून मनोरंजन गीतगवनई टीवी फिल्म रंगमंच सरोकार कोर्स आ कैरियर देश आ समाज नारीजगत पर्व ल्योहार भोजपुरिया लाल योग राजनीति सभा समारोह स्वास्थ्य साहित्य उपन्यास कविता कहानी निबन्ध पुस्तक चर्चा भाषा समीक्षा स्तम्भ अगडम बगडम कतरब्योत कोलकाता मेल ज्योतिष आ वास्तु बतकुञ्चन भउजी हो भोला बाबू रामझरोखा से लस्टम पस्टम कार्टून कोना सभकर राय |
| 25 | टटका खबर उत्तर प्रदेश बिहार राजनीति देशदुनिया खेल व्यापार अपराध मनोरंजन शिक्षा फर्जी यूनिवर्सिटी खबर भोजपुरी में बुध जनवरी मंगल जनवरी के खबर |
| 26 | खबर भोजपुरी में |
| 27 | शुक जनवरी खबर भोजपुरी में बियफे जनवरी खबर भोजपुरी में बुध जनवरी खबर भोजपुरी में मंगल जनवरी खबर भोजपुरी में सोमार जनवरी खबर भोजपुरी में अतवार जनवरी खबर भोजपुरी में शनिचर जनवरी खबर भोजपुरी में शुक जनवरी खबर भोजपुरी में जनवरी खबर भोजपुरी में बुध दिसंबर ट्वीट्स के द्वारा किये गए ट्वीट्स |