A POS ANNOTATED CORPUS OF AWADHI LANGUAGE

Dissertation submitted to

Dr. Bhim Rao Ambedkar University, Agra

In partial fulfillment of the requirements for the award of the degree of

MASTER OF PHILOSOPHY

ABDUL BASIT

DEPARTMENT OF LINGUISTICS

KANHAIYALAL MANIKLAL MUNSHI INSTITUTE OF HINDI AND LINGUISTICS

DR. BHIM RAO AMBEDKAR UNIVERSITY

AGRA

2017

Date: 08-08-2017

DECLARATION BY THE CANDIDATE

This dissertation titled "*A POS ANNOTATED CORPUS OF AWADHI LANGUAGE*" 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.

ABDUL BASIT
M.Phil. Student
Department of Linguistics
Kanhaiyalal Maniklal Munshi Institute of Hindi and Linguistics
Dr. Bhim Rao Ambedkar University, Agra

Department of Linguistics

Kanhaiyalal Maniklal Munshi Institute of Hindi and Linguistics

Dr. Bhim Rao Ambedkar University, Agra

Dated: 08-08-2017

CERTIFICATE

This dissertation titled "A POS ANNOTATED CORPUS OF AWADHI LANGUAGE" submitted by Abdul Basit, Department of Linguistics, Kanhaiyalal Maniklal Munshi Institute of Hindi and Linguistics, Dr. Bhim Rao Ambedkar University, Agra, 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. Ritesh Kumar)
SUPERVISOR

(Prof. Pradeep Shridhar)
DIRECTOR

“*Dedicated to my loving **parents** and my **family** who are my greatest source of inspiration.*”

Acknowledgement

I would like to express my deep sense of reverence and gratitude for my supervisor ***Dr. Ritesh Kumar****, who motivated me in taking this research topic. He had faith in my ability to rise on the occasion and deliver the best work. And also sincerely thanks for patiently teaching me several things. I will always remember his passion, dedication, and thirst for perfection. He taught me the importance of self-evaluation. I will remain indebted to him all my life for guiding me in some of the toughest phases of my life. I am lucky to work with him.*

I owe everything to my family, for being with me at all the times and believing in what I wanted to pursue.

I am very thankful to ***Dr. Neelam Yadav, Dr. Ranjeet Bharati*** *and* ***Pallavi Arya*** *for their encouragement during my research.*

I would like to thank my friends and batchmates Mohammad Faizee, Anam Ahsan, Mohammad Anas Ansari, Mohammad Javed, Yogesh Dawer, Mohammad Usama and Ravina Toppo for supporting me in this journey. I thoroughly enjoyed our debates and discussions on a wide range of subjects and issues. They gave me the courage to pursue my real interests.

Abdul Basit

List of Tables

List of Figures

List of Abbreviations

List of Contents

List of Tables

List of Figures

List of Abbreviations

CL	Computational Linguistics
NLP	Natural Language Processing
PO	Part-of-Speech
TDIL	Technology Development for Indian Languages
ILCI	Indian Language Corpora Initiative
LCD	Language Consortium Database
IL	Indian Language
BNC	British National Corpus
CEU	Corpus of European Union
EMILLE	Enabling Minority Language Engineering
MRD	machine readable dictionary
OCR	Optical Character Recognition
BIS	Bureau of Indian Standard
IIIT	Indian Institute of Information Technology
ILMT	Indian Language
HMM	Hidden Markov Modal
SWL	Selected Word from Lexicon
CRF	Conditional Random Field Model
SVM	Support Vector Machine
CIIL	Central Institute of Indangered Language
MSRI	Micro Soft Research India
UTF	Unicode Transformation Format

List of Contents

Chapter – 1

INTRODUCTION

This dissertation contributes to the subject area of Part-of-Speech (POS) tagging which is an important part of Computational Linguistics (CL) or Natural Language Processing (NLP). Corpus Linguistics is a sub-branch of Computational Linguistics which seeks to study natural languages through empirical linguistic information. The invention of computers has added a new dimension to this area. POS tagging is a process of assigning part of speech to every token in a corpus. It plays an essential role in developing any serious applications for processing the natural languages. It is a first step towards the development of language processing systems like, information retrieval systems, machine translation, text to speech synthesis systems etc. The main focus of the present research is to prepare a POS-tagged corpus of Awadhi language.

1.1 What is Corpus ?

The word *Corpus,* derived from Latin word, which means '*body'***,** refers to any large collection of human language in written or spoken form. It generally represents samples of a particular variety or use of language(s) which are presented in machine readable form. Sinclair (1996) states that "a corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language. A computer-readable corpus is a corpus which is encoded in a standardized and homogenous way for open-ended retrieval tasks."

Crystal (1980) defines it as "*A collection of of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting point of linguistic description or as a means of verifying hypotheses about a language."*

Bird and Liberman, (2000): "*Linguistic annotation covers any descriptive or analytic notations applied to raw language data. The basic data may be in the form of time functions - audio, video and/or physiological recordings or it may be textual."* Leech (1997)

"*(corpus annotation) can be defined as the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data. 'Annotation' can also refer to the end-product of this process".*

1.2 Corpus Linguistics and Computational Linguistics

Corpus linguistics explores properties of a language by analyzing large collection of text or speech data. It has been used in a number of research area ranging from descriptive study of a language, to language education, lexicography, etc. "The basic philosophy behind corpus linguistics has two aspects: (a) We have a cognitive drive to know how people use language in their daily communication activities, and (b) if it is possible to build up intelligent systems that can efficiently interact with human beings". (Dash, 2010) With these motivations, a corpus can be used for designing intelligent systems like machine translation system, language processing system, speech understanding system, text analysis and understanding system, computer aided instruction system, etc. for the benefit of the language community at large.

Computational linguistics is an interdisciplinary field of study dealing with the rule based or the statistical modeling of natural language from a computational perspective. Computational linguistics has theoretical and applied components, where theoretical computational linguistics takes up issues in theoretical linguistics and cognitive science, and applied computational linguistics focuses on the practical technique for the realization of linguistic theory to facilitate real-world applications. *(wikipedia)*

1.3 Types of Text Corpus

A corpus contains language data collected from various written, printed, published, and electronic sources. These could be of various types depending on different factors. They are as follows.

1.3.1 *Comparable Corpus*

A type of corpus which is used for comparison of two or more languages. For example: *Corpus of European Union* (Dash, 2010)

1.3.2 *Annotated Corpus*

Any annotated corpus may be considered to be a repository of linguistic information, it contains tags and codes inserted form outside to record some additional information. For example: *ILCI Corpus*, *British National Corpus* (Dash, 2010)

1.3.3 *Monitor Corpus*

It is a growing , non finite collection of text. It is of primary use in lexicography. Monitor corpus reflects language changes over a large period of time.
For example: *Bank of English* (Dash, 2010)

1.3.4 *Multilingual Corpus*

Multilingual corpus contains representative collections from more than two languages. Basically, here as well as in bilingual corpus, similar text categories and identical sampling procedures are followed although the text itself belongs to different languages. For example: *Gyan Nidhi Parallel Corpus, Crater Corpus* (Dash, 2010)

1.3.5 *Parallel Corpus*

Parallel corpus is a kind of multilingual corpus, where text data of one language and their translation into other languages are aligned either sentence-by-sentence or preferably phrase-by-phrase. For example: *Joshua decoder Indian language parallel corpora*, *The EMILLE corpus*, *Chemnitz German-English Corpus*, *ILCI Corpus* (Dash, 2010)

1.3.6 *Reference Corpus*

It is made to supply comprehensive information about a language. It is large enough to represent all relevant varieties of language and characteristic vocabulary. So, it can be used for writing grammars, dictionaries, thesauruses and other materials. It includes spoken and written language representing various social and situational usage of the language. For example: *Bank of English* (Dash, 2010)

1.3.7 *Monolingual Corpus*

It is a type of corpus, which contains texts in a single language. For example: *EMILLE*

Corpus, *ISI Bengali Corpus* (Dash, 2010)

1.3.8 *Unannotated Corpus*

This type of corpus represents a plain text without additional linguistic or non-linguistic information. For example: *TDIL Corpus* (Dash, 2010)

1.4 Usage of Corpus

Corpus are very useful in the field of language description, study of syntax, phonetics and phonology, prosody, intonation, morphology, lexicology, semantics, lexicography,, discourse, pragmatics, language teaching, language planning, sociolinguistics, psycholinguistics, cognitive linguistics, computational linguistics etc. In fact, there is hardly any area of linguistics where corpus has not found its utility. This has been possible due to great possibilities offered by computer in collecting, storing and processing natural language databases. The availability of computer and machine readable corpus has made it possible to get data quickly and easily. (Dash, 2010)

1.4.1 *Corpus as knowledge resource*

Corpus is used for developing multilingual libraries, designing course books for language teaching, compiling monolingual and bilingual dictionaries (printed and electronic), various reference materials (printed and electronic version), developing machine readable dictionaries (MRDs), developing multilingual lexical resources etc.

1.4.2 *Corpus in language technology*

Corpus is used for designing tools and systems for word processing, spelling checker, text editing, morphological processing, sentence parsing, frequency counting, item search, test summarization, text annotation, information retrival, concordance, word sense disambiguation, WordNet, semantic web, semantic net, POS tagging, local word grouping, etc.

1.4.3 *Corpus for translation support systems*

Corpus is used for Language Resource Access Systems, Machine translation systems, multilingual information access systems and cross language information retrieval etc.

1.4.4 *Corpus for human-machine interface system*

Corpus is used for voice recognition, text speech, e-learning, online teaching systems, e-text preparation, question-answering, computer assisted language education, Computer-aided instruction, e-governance etc.

1.5 Limitations of Corpus

According to Dash (2010), there are some limitations of the corpus. These are discussed below.

1.5.1 *Lack of linguistic generativity*

Chomsky argued that “Any natural corpus will be skewed. Some sentences won’t occur because they are obvious; others because they are false, still others because they are impolite. The corpus, if natural will be so wildly skewed that the description (based upon it) would be no more than a mere list.”(Chomsky,1962:159) Generativists argue that corpus can not provide evidence for linguistic innateness. By virtue of its structure and content, it only can represent the linguistic performance, but does not reflect on the linguistic ‘competence’ and ‘generativity’ of the users. A corpus, which records only the examples of performance, can not be useful to linguists, who seek to understand the tacit, internalized knowledge of language rather than the external evidence of language use in various context.

1.5.2 *Technical difficulties*

Corpus building is a large scale, multi-directional, enterprising work. It is a complex, time consuming, error-prone and expensive task. The whole enterprise requires an efficient data processing system, which may not be available to all, particularly in our country.

1.5.3 *Lack of texts from dialogues*

Present day corpus fails to consider the impromptu, non-prepared dialogues taking place spontaneously in daily linguistics exercises. Absence of texts from dialogues interactions make a corpus cripple lacking in the aspect of spontaneity, a valuable trait of human language. Corpus, either in spoken or written form, is actually a database

detachment from the contexts makes a corpus a dead database. Which is devoid of many properties of living dialogue interactions, discourse and pragmatics. It fails to reveal the real purpose underlying a linguistic negotiation, identify the language-in-use, determine the verbal action involved within the dialogues, describe the background where from the interlocutors derive cognitive and perceptual mean of communication.

1.6 Part of Speech(POS)-Tagging

The raw text corpus can be provided with additional linguistic information, known as annotation. These linguistic information can be of different kinds like part-of-speech (POS), prosodic, semantic, anaphoric, discoursal annotation etc. The POS tagged corpus is the most basic form of annotated corpus where the words are assigned part-of-speech information. There are many types of corpus like, “Brown Corpus, LOB Corpus and BNC which are grammatically annotated. The LLC is prosodically annotated while the Susarme Corpus is syntactically annotated.”(Dash, 2010) In the present research, I have produced a pos-annotated corpus of Awadhi.

“Part-of-Speech (POS) tagging is a process of assigning part-of-speech tags to each and every word used in a piece of text after the word is passed through the stages of morphological analysis and grammatical interpretation.”(Garside. 1995)

1.7 Brief description of Awadhi Language

Awadhi is an Indo-Aryan language, spoken in the eastern region of Uttar Pradesh *viz. Lucknow, Raebareli, Sitapur, Unnao, Allahabad, Faizabad, Sultanpur, Behraich and Pratapgarh etc.* There are 38 millions native speakers of Awadhi language (Census, 2001). It is the official language of Nepal and Fiji. The writing system of Awadhi is Devanagri, Kaithi and Perso-Arabic. (wikipedia)

1.7.1 *Grammatical information about the Awadhi*

The word order of Awadhi is Subject Object Verb (SOV). The use of postposition like मा, से, का etc. indicate possession in Awadhi. The verbal affixation marks person, number and gender of subject and object. There is no ergative case (ने) in Awadhi.

There are 30 consonants and 8 vowels in Awadhi. The morphological typology of Awadhi language is fusional. (Awadhi/Ethnolgue)

1.8 Present Research

In present scenario of India, there are several attempts to collect the corpus of Indian languages and few corpora are available in some of the major languages of India. However, there is no corpus available for Awadhi till now. In the present research, the data of Awadhi language is collected from Eastern region of Uttar Pradesh. In this research, I have developed a corpus with approximately 70,000 tokens. Approximately 20,000 tokens of the corpus data has been annotated with the POS information. It is the first POS-Tagged corpus of Awadhi language. I have also developed the first POS tagset of Awadhi based on the general BIS tagset for Indian language.

1.9 Outline of dissertation

This dissertation is divided into four chapters.

Chapter 1 briefly explains corpus linguistics, different kinds and usages of corpus and also its limitations. It also give a brief description of Awadhi language and a summary of the present research. Chapter 2 presents a brief survey of part-of-speech taggers in Indian languages including different approaches. It also discusses various tagsets that are used in POS tagging of Indian languages. Chapter 3 presents the method of data collection, sources of data, format of data and metadata for current research. It also discussed the challenges and issues faced in the development of the corpus, Bureau of Indian Standard (BIS) tagset for Awadhi and POS Annotation tool used for the manual annotation of the data Chapter 4 gives a summary of the dissertation and shows directions towards future research.

Chapter – 2

Part-of-Speech taggers in Indian Languages

This chapter presents a brief survey of part-of-speech taggers in Indian languages. It also discusses various tagsets that are used in POS tagging of Indian languages.

2.1 Different Approaches to Parts-of-Speech (POS) tagging

Different approaches have been used by different researchers for POS tagging. These approaches to POS tagging can be divided into three broad categories- rule-based, statistical/machine learning based and hybrid systems. In a rule-based system, a set of hand written rules are applied and also contextual information is used in order to assign POS tags to words. The statistical and machine-learning based POS tagging works on the basis of the most frequently used tag for a specific word in the annotated training data, As the name suggests, hybrid approach combines both of the above-mentioned approaches and it may perform better than statistical or rule based POS tagging. The taggers developed using these three methods are discussed in the following sections.

2.1.1 Rule Based Approach

The rule based POS tagging approach uses a set of hand written rules. It gives the contextual meaning to assign POS tags to any word. The tagger is divided into two stages. First, it finds the word in dictionary and then, if ambiguous, it assigns a tag by resolving the ambiguity of tag using linguistic information of the word.

Mazhar, & Memon (2012) discusses rule based part of speech tagging of Sindhi language. In the first stage, the system takes input text and creates token. Once token is created, it is searched in the lexicon (SWL). If the word is found one or more times, then the associated tag(s) are stored. If the word is not found then that word is added into lexicon by creating linguistic rule for new word. The tagset contains sixty seven tags. A lexicon of 26,366 tokens, named SWL, is developed for this. The lexicon also

contained the frequency of each tag. A set of 186 disambiguation rules are used for the tagging system. The contextual information is used for this rule-based approach and it manually assigns a part of speech tag to a word. The system achieved an accuracy of 96.28% . As more words will be tagged and rules will be added, the accuracy of the system is expected to increase.

In a rule-based system, the lexical rules act at the word level while the context sensitive rules act at the sentence level . However, the main disadvantage of rule-based approach is that it fails when the text is not present in lexicon.

2.1.2 Machine Learning based Parts-of-Speech (POS) tagging

Supervised Parts of speech taggers are based on pre-tagged corpora, which are used for training to learn information about the word-tag frequencies, rule and tagset, etc. The performance of the models generally increases with the increase in size of the annotated corpora.

2.1.2.1 *Supervised Machine Learning Taggers*

The development of POS tagger for Marathi using statistical approaches. They developed the statistical tagger using Unigram, Bigram, Trigram and Hiddden Markov Modal (HMM) Methods. In order to achieve higher accuracy they use a set of hand coded rules, as well as frequency and probability. They train and test their model by calculating frequency and probability of words of given corpus. In unigram technique, they find out how many times each word occur in corpus and then assign each word to the most common tag. Bigram tagger makes tag suggestion based on preceding tag i.e. it take two tags - previous and the current tag. Trigram provides the transition between the tags and helps to capture the context of the sentence. The probability of a sequence is just the product of conditional probabilities of its trigrams. Hidden Markov Models (HMM) assigns the best tag to a word by calculating the forward and backward probabilities of tags along with the sequence provided as an input. One of the powerful features of HMM is context description. The POS tagger described here is very simple and efficient for automatic tagging, but it is difficult for Marathi as it is morphologically rich language.

2.1.2.2 *Unsupervised Parts of Speech (POS) tagging*

The unsupervised POS tagging approaches, unlike supervised POS tagging approach, do not require any pre-tagged corpora. Evaluation of the probabilistic information or the contextual rules needed by rule based systems or transformation based systems is performed for Stochastic Taggers in these approaches. In these approaches an untagged text is run through a tagging model to generate initial output. This is one approach for automatic rule induction after the output error correction is done. This way the taggers learn the correction rules by comparing the two sets of data. For obtaining the better performance, this process is repeated a number of times.

2.1.3 The hybrid Approach

The hybrid approach is a combination of machine-learning based and rule based approach. In this approach most probable information/tag is assigned to the text using machine-learning method. After that, if ambiguity is found then disambiguation is carried out by applying grammar rules. The construction of such taggers contain trained machine learning model which includes approximated rules as well as exact disambiguation rule.

A survey on developments of different POS tagger systems as well as POS tagsets for Indian languages. (Antony & Soman, 2011) the existing approaches that have been used to develop POS tagger tools. They concluded that almost all existing Indian language POS tagging systems are based on statistical and hybrid approach.

A rule based part-of-speech tagger for Hindi developed by Garg, Goyal & Preet (2012). The system is evaluated over a corpus of 26,149 words with 30 different standard part of speech tags for Hindi. The evaluation of the system is done on the different domains of Hindi Corpus. These domains include news, essay, and short stories and system achieved an accuracy of 87.55%.

2.2 POS-Tagger for Indian languages

The journey of language technologies for Indian language began in a big way in 1991, when the Department of Electronics, Govt. of India, initiated a project called

Technology Development for Indian Languages (TDIL) to develop machine readable corpus of texts in all the major Indian languages. The emphasis was laid on development of software for language processing like POS tagging, text encoding, frequency counting, spelling checking, morphological processing, etc. as well as for designing of higher-level applications like machine translation systems, especially from English to Indian languages.

There are a lot of corpus based POS annotation research that has been carried out in Indian languages like- Bengali, kannada, Hindi, Sinhala, Sanskrit, Urdu, Panjabi, Oriya, Tamil, Telugu, Malayalam, Kokani, Gujarati, Marathi etc. I shall discuss some of these in the sections below.

2.2.1 Hindi

Garg, Goyal, & Preet (2012) discusses the development of a rule based POS tagger for Hindi. Their system is tested over dataset of 26,149 words with 30 different POS tags for Hindi. They have evaluated their system on data from different domains including news, essay, and short stories. The system achieved an accuracy of 87.55%. The system mainly works in two steps- first, the input words are searched in the database; if it is present then it is tagged. If it is not present then various rules are applied. If a sentence consists of 12 words out of which 8 words are unknown, then the system fails to tag them. It is hard to decide which rules should be handled first because word tagging resolution is based on neighboring word and hence it fails.

Shrivastava & Bhattacharya (2006) discusses a simple hidden markov model based POS tagger. As a pre-processor they have employed longest suffix matching stemmer and claim to achieve an accuracy of 93.12%. The core idea of their approach is to “explode” the input in order to increase the length of the input and to reduce the number of unique types they encountered at the time of learning. This increases the probability score of the correct choice while simultaneously decreasing the ambiguity of the choices at each stage.

Agrawal & Mani (2006) discusses part of speech tagging and chunking with conditional random fields (CRF) for Hindi. In this research, training is performed using a morph analyzer and CRF++ to provide extra information like root word and possible part of speech tags for training. While training on 21,000 tokens and best feature set, they claim to have achieved 82.67% accuracy.

Joshi, Darbari, & Mathur (2013) proposed a system on POS tagger using hidden markov model. They have used 15,200 sentences (3,58,288 words) from tourism domain to train the system. They have used two special tags <S> and </S> to denote starting and ending of the sentence which was added to all the sentences of the training corpus. The accuracy of 92.13% on test data was attained.

2.2.2 Marathi

Singh, Joshi and Mathur, (2013) have developed a POS tagger for Marathi. They have used trigram method using statistical approach. The concept mainly used here is to explore the most likely POS tag for a current word based on given knowledge of previous two tags by calculating probabilities to determine which is the best sequence of tag. For testing the performance of the system, they have developed a test corpus of 2000 sentences. They claim to have got an accuracy of 91.63% .

Patil, Pawar & Patil (2014) discusses another rule-based POS tagger developed for Marathi. In their approach, the hand–constructed rules are learnt from corpus and some manual additions after studying the grammar of Marathi language were added. They have tried to disambiguate tagging by analyzing the grammatical feature of the current word, its antecedent word, its succeeding word, etc. After testing their tagger, they claimed to have an accuracy of about 78.82% on three different types of data sets.

2.2.3 Bengali

Dandapat, Sarkar & Basu, (2004) presented a paper and described about a model that uses composition of supervised and unsupervised learning techniques using a Hidden Markov Model. They have made use of small tagged corpus and also large untagged corpus. They also make use of Morphological Analyzer that takes a word as input and

gives all possible POS tags for the word. They took 1003 words from CIIL corpus and tagged it manually. They have obtained an overall accuracy of 95% .

A tagger for Bengali using maximum entropy, which makes use of the different circumstantial information of the words with the variety of features that are helpful in predicting the various Part of speech classes. (Ekbal, Haque, & Bandopadhayay, 2008) Their tagger has been trained with a training corpus of 72,341 word forms and it uses a tagset of 26 different POS tags, defined for the Indian languages. The tagger has demonstrated an accuracy of 88.2% for a test set of 20K words.

A POS tagger using Conditional Random Field have developed. (Ekbal, Haque, & Bandopadhayay, 2007) They have used tag set of 26 POS tags, which are defined for the Indian languages. They have used tag set of 26 POS tags, which are defined for the Indian languages. The POS tagger has been trained and tested with the 72,341 words and 20k word forms, respectively. Their experimental results show that the CRF based tagger achieves an accuracy of 90.3%.

2.2.4 Manipuri

Singha, Purkayastha, & Singha (2013) have presented a paper discussing an attempt to develop a lexicon based POS Tagger for Manipuri. They have applied a set of hand written language specific rules of Manipuri language. In this paper they have designed a 3-tier tagset for Manipuri. This tagset consists of 97 tags including generic attributes and language specific attribute values.

Another tagger has been developed for Manipuri using stochastic approach namely Hidden Markov Model (Singha, Purkayastha, & Singha, 2012). Manipuri rule-based tagger gives tagged output that is used as corpus for training. In order to measure performance of tagger they have used manually annotated test set data that consist of 97 category of Manipuri language. They have claimed to have achieved the accuracy of 92%.

2.2.5 Kannada

A POS tagger for Kannada was developed by Shambhavi and Kumar (2012). In this POS tagging task of Kannada language they have chosen Second order Hidden Markov Model and Conditional Random Fields. Their training data consists of 51,269 tokens and test data set incorporate around 2,932 tokens. Both data set are taken from EMILLE corpus. Corpus was partitioned into 95% for training and 5% for testing. Their experimental result shows the accuracy of the HMM-based tagger to be 79.9% and that of CRF-based tagger to be 84.58%.

2.2.6 Malayalam

A stochastic Hidden Markov Model (HMM) based part of speech tagger has been proposed for Malayalam. To perform the Part-of-Speech tagging using stochastic approach, an annotated corpus is required. Due to the non-availability of annotated corpus, a morphological analyzer was also developed to generate a tagged corpus from the training set. Antony, Mohan, & Soman (2010) developed tagset and tagged corpora of 180,000 words for Malayalam language. This tagged corpus is used for training the system. The accuracy of the SVM based tagger is 94 % and shows an improved result over HMM based tagger.

2.2.7 Tamil

A rule based morphological analyzer and POS tagger for Tamil have been developed by Selvam, Natarajan, & Thangarajan (2009). They improved the above systems using Projection and Induction techniques. Rule based morphological analyzer and POS tagger can be built from well defined morphological rules of Tamil. Projection and induction techniques are used for POS tagging, base noun-phrase bracketing, named entity tagging and morphological analysis from a resource rich language to a resource deficient language. They applied alignment and projection techniques for projecting POS tags, and alignment, lemmatization and morphological induction techniques for inducing root words from English to Tamil. Categorical information and root words are obtained from POS projection and morphological induction respectively from English via alignment across sentence aligned corpora. They generated more than 600 POS tags for rule based morphological analysis and POS tagging.

2.2.8 Sinhala

A Sinhala language based Part of Speech (POS) Tagger using lexical semantics and HMM has been presented by Jayaweera, Antony, Dias, & Mohan (2014). They have used a statistical approach, in which the tag identification process is done by calculating the probability of tag sequence and the probability of word-likelihood from the given dataset, where the language specific knowledge is axiomatically extracted from the annotated dataset. In this research they used the Beta version of the UCSC Corpus which contains around 6,50, 000 words and from which definite words are 70,000. For known words they claim to have an accuracy of more than 90%.

2.2.9 Sanskrit

A treebank based deep Grammar acquisition and Part-Of-Speech tagging for Sanskrit sentences is discussed by Tapaswi, & Jain (2012). The input for the system is one sentence per line. The sentence is then split into words called lexeme. The system reads each word to find longest suffix and eliminates the suffix until the word length is 2. Then it applies the lexical rules and assigns the tag. It handles disambiguity using context sensitive rules. For experimental result, authors took a set of 100 words and manually evaluated those. The system gives 90% correct tags for each word. The evaluation was done in two stages. Firstly by applying the lexical rules and secondly, after applying the context sensitive rule. The POS tagger described here is very efficient for Sanskrit but it is difficult for Marathi as the root undergoes a change after affixation Marathi but not so in Sanskrit.

2.2.10 Gujarati

Patel & Gali (2008) introduced a machine learning algorithm introduced for Gujarati Part of Speech tagging. The machine learning part is performed using a CRF model. The algorithm has achieved an accuracy of 92% for Gujarati texts where the training corpus is of 10,000 words and the test corpus is of 5,000 words. From the experiments they observed that if the language specific rules can be formulated as features for CRF then the accuracy can be substantially enhanced.

2.2.11 Panjabi

In their paper Kaur, Aggerwal, & Sharma (2006) proposed an improvement for Punjabi Part of Speech tagger by using reduced tagset. They try to improve the accuracy of HMM based Punjabi POS tagger by reducing the tagset. The tagset has been reduced from more than 630 tags to 36 tags. We observed a significant improvement in the accuracy of tagging. Their proposed tagger shows an accuracy of 92-95% whereas the existing HMM based POS tagger was reported to give an accuracy of 85-87%.

2.3 POS Tagsets for Indian Languages

There are several POS Annotation tagsets have been developed by different research group working on Indian languages.

i. IIIT-H (ILMT) tagset
ii. MSRI tagset (IL-POSTS)
iii. LDC-IL tagset
iv. BIS tagset
v. AUKBC tagset
vi. JNU-Sanskrit tagset (JPOS)
vii. Sanskrit consortium tagset (CPOS)

In the following sections, I shall discuss some of the major POS annotation schemes for Indian languages.

2.3.1 IIIT (ILMT) Tagset

IIIT tagset uses a modified version of the basic Penn Treebank tagset for English. It modifies some of the basic tags and introduces some additional ones to address the unique features of Indian languages. It divides the tags into three sections.

I. Section 1

This section contains those tagset which are similar to Penn tagset. IIT tagset is a coarse tagset and hence it disregards notion of plurality and gender in its tagset(**See Table No. 2.1**).

S.No.	Categories	Annotation Convention
1.	Noun	NN
2.	Proper Noun	NNP
3.	Pronoun	PRP
4.	Verb Aux.	VAUX
5.	Adjective	JJ
6.	Adverb	RB
7.	Conjunction	CC
8.	Interjection	UH
9.	Foreign Symbols	SYM

Table No.2.1 (ILMT) Tagset Section 1

II. Section 2

The section contains tagsets that are modified form of the tags from Penn Treebank tagset. For example, the Penn tagset has 'Preposition' but in Indian languages there are postpositions. The Penn tagset had a category called Preposition with the tag 'PREP'. In the ILMT tagset, postpositions is called PREP. All Quantifiers in the language are tagged as QF. Any word denoting numbers are tagged as QFNUM. Main verb of a finite verb group of a sentence is considered as VFM. All non-finite verbs which are used as adjectives are tagged as VJJ. Nonfinite forms of verbs which are used as adverbs are tagged with VRB. Gerunds are marked as VNN. All wh words are marked as QW (**See Table No.2.2**).

S.No.	Categories	Annotation Convention
1.	**Postposition**	PREP
2.	**Quantifiers**	QF
3.	**Quantifiers** Number	QFNUM

4.	**Verb Finite Main**	VFM
5.	**Verb Non-Finite Adjectival**	VJJ
6.	**Verb Non-Finite Adverbial**	VRB
7.	**Verb Non-Finite Nominal**	VNN
8.	**Question Words**	QW

Table No.2.2 (ILMT) Tagset Section 2

III. **Section 3**

The tags in this group are completely new and address the peculiar features of Indian languages. Indian Languages like Hindi contains words like aage, upar, pahele, etc..These words can behave like adverbs, nouns, and postposition in different contexts. All these words are tagged as noun locative(NLOC). Words like 'bahuta', 'kama', etc. are tagged as intensifier(INTF). Negatives like 'nahi', 'na', etc. are marked as negation(NEG). The tag NNC is used for compound nouns. All words except the last one, of compound words are marked as NNC. The tag for compound proper nouns is NNPC and all compound proper nouns are tagged as NNPC excluding the last one. Conjunct are verbs formed by combining a noun or an adjective or an adverb with a (helping) verb. The conjuct formed by joining a noun are marked as NVB, those formed with an adjective are tagged as JVB and those formed by joining adverbs are tagged as RBVB.

S.No.	Categories	Annotation Convention
1.	Noun Locative	NLOC
2.	Inensifier	INTF
3.	Negation	NEG

4.	Compound Noun	NNC
5.	Compound Proper Noun	NNPC
6.	Noun Verb	NVB
7.	Adjective Verb	JVB
8.	Adverb Verb	RBVB

Table No.2.3 (IMLT) tagset Section 3

2.3.1 MSRI Tagset

Microsoft Research India Pvt. Ltd developed this tagset in 2008. It aims to provide a comprehensive tagset that captures as much information as possible from tagging. The guidelines of this tagset contains 9 categories - Nouns, Pronouns, Verbs, Nominal Modifier, demonstrative, Adverb, Particle, Punctuation and Residual. These categories branch out in types such as common, proper, verbal and spatio-temporal .This tagset has been subdivided into further categories.(Baskaran et al., 2008) (See **Table No.2.4**)

S.No.	Category	Types	Attributes
1.	**Noun**		
1.1		Common(NC)	gender, number, case, nominal, declension
1.2		Proper(NP)	gender, number, case, nominal declension
2.	**Pronoun**		
2.1		Pronominal(PPR)	gender, number, person, case, nominal declension, emphatic, honorificity, distance
2.2		Reflexive(PRF)	gender, number, case, nominal declension
2.3		Reciprocal(PRC)	gender, number, case, nominal declension
2.4		Relative(PRC)	gender, number, person,

			case, nominal declension
2.5		Wh-word(PWH)	gender, number, person, case, nominal declension
3	Verb	Verb(V)	pada, number, person, tense\mood, honorificity
4	**Demonstrative**		
4.1		Absolutive (DAB)	gender, number, person, case, nominal declension, distance, honorificity
4.2		Relative(DRL)	gender, number, person, case, nominal declension, distance, honorificity
4.3		Wh-(DWH)	gender, number, person, case, nominal declension, distance, honorificity
5.	**Nominal Modifier**		
5.1		Adjective(JJ)	gender, number, case, nominal declension, emphatic, negative, honorificity
5.2		Quantifier(JQ)	gender, number, case, nominal declension, numeral, emphatic, negative
6.	Adverb	Adverb(RB)	
7.	**Particle**		
7.1		Negative(CNG)	
7.2		Empatic(CEM)	
8.	Punctuation	Punctuation(PC)	
9.	Residual		
9.1		Foreign word(RDF)	
9.2		Symbol(RDS)	
9.3		Others(RDX)	

Table NO. 2.4 MSRI Tagset

2.3.3 LDC – IL Tagset

Language Development Consortium (LDC) developed a hierarchical tagset while the previously discussed IIIT and MSRI tagsets is also a flat tagset. Flat tagsets are a list of mutually exclusive categories and though they are easier to process they cannot capture higher level of granularity. Also if we want to develop such tagsets which are reusable across different linguistic boundaries we have no option but to constantly lengthen the tagset that facilitates addition of new tags. Hierarchical tagsets, on the other hand, have parameters which can take values for a certain language. This enables it to address the requirements of a group of related languages. Structurally hierarchical tagsets contain top-level tags which are further split into other bottom level more specific tags. The morpho-syntactic details are encoded in separate layer of hierarchy beginning from the major categories of the top and gradually progressing down to cover more granular features. Decomposability is another feature of a hierarchical tagset design as it allows different features to be encoded in a tag by separate substrings. A tag is considered decomposable if the string representing the tag contains one or more shorter sub-strings that are meaningful out of the context of the original tag. Decomposable tags help in better corpus analysis by allowing to search with an underspecified search string. LDC-IL has 13 top-level categories (Chandra, Kumawat, & Srivastava, 2014) (see **Table No. 2.5**).

S. No.	Categories	Annotation Convention
1.	Noun	N
2.	Pronoun	P
3.	Demonstrative	D
4.	Nominal Modifier	J
5.	Verb	V
6.	Adverb	A
7.	Postposition	PP
8.	Particle	C

9.	Numeral	NUM
10.	Reduplication	RDP
11.	Residual	RD
12.	Unknown	UNK
13.	Punctuation	PU

Table NO. 2.5 LDC IL Tagset

2.3.4 BIS Tagset

The Bureau of Indian Standards (BIS) Tagset has recommended the use of a common tagset for the part of speech annotation of Indian languages. The tagset, incorporating the advice of the experts and the stakeholders in the area of natural language processing and language technology of Indian languages, has to be followed in the annotation tasks taking place in Indian languages. It is also a hierarchical tagset and contains 11 top-level tags. As of now, the annotations taking place under the Indian Languages Corpora Initiative (ILCI) program is following the BIS tagset (Chaudhary, 2010).(See Table No. 2.5)

S.NO.	Categories	Subtypes Level 1	Level 2	Annotation Convention
1	**Noun**	**N**		**N**
1.1		Common		N_NN
1.2		Proper		N_NNP
1.3		Verbal		N_NNV
1.4		Nloc		N_NST
2	**Pronoun**	**PR**		**PR**
2.1		Personal		PR_PRP
2.2		Reflexive		PR_PRF
2.3		Relative		PR_PRL
2.4		Reciprocal		PR_PRC
2.5		Wh-word		PR_PRQ

2.6		Indefinite		PR_PRI
3	**Demonstrative**			**DM**
3.1		Deictic		DM_DMD
3.2		Relative		DM_DMR
3.3		Wh-word		DM_DMQ
3.4		Indefinite		DM_DMI
4	**Verb**			**V**
4.1		Main		V_VM
4.1.1			Finite	V_VM_VF
4.1.2			Non-Finite	V_VM_VNF
4.1.3			Infinitive	V_VM_VINF
4.1.4			Gerund	V_VM_VNG
4.2		Verbal		V_VN
4.2		Auxiliary		V_VAUX
4.2.1			Finite	V_VAUX_VF
4.2.2			Non-finite	V_VAUX_VNF
4.2.3			Infinitive	V_VAUX_VINF
4.2.4			Gerund	V_VAUX_VNG
4.2.5			PARTICIP-LE NOUN	V_VAUX_VNP
5	**Adjective**			**JJ**
6	**Adverb**			**RB**
7	**Postposition**			**PSP**
8	**Conjunction**			**CC**
8.1		Co-ordinator		CC_CCD
8.2		Subordinator		CC_CCS
8.2.1			Quotative	CC_CCS_UT
9	**Particles**			**RP**
9.1		Default		RP_RPD
9.2		Classifier		RP_CL
9.3		Interjection		RP_INJ

9.4		Intensifier		RP_INTF
9.5		Negation		RP_NEG
10	**Quantifier**			**QT**
10.1		General		QTF
10.2		Cardinals		QT_QTC
10.3		Ordinals		QT_QTO
11	**Residuals**			**RD**
11.1		Foreign word		RD_RDF
11.2		Symbol		RD_SYM
11.3		Punctuation		RD_PINC
11.4		Unknown		RD_UNK
11.5		Echo-words		RD_ECH

Table No. BIS Tagset for Indian Languages

As one would notice, BIS tagset bears close resemblance to the LDC-IL tagset. In addition to one type of a category. It also introduces another subtype. BIS tagset groups together unknown, punctuation and residual in one top-level tag – Residual while LDC_IL tagset had 3 different tags for these. Nouns and, Pronouns in the two tagsets are almost identical in the two tagsets. Verb (V), too, has the same subtypes - main verb (VM) and auxiliary verb (VAUX). Adjective and adverb has no subtype whereas we have two new categories in BIS tagset - one is conjunction (CC) which has two subtypes namely coordinator (CCD) and subordinator (CCS). These subtypes were grouped under particle (RP) in LDC-IL tagset. As a result, particles (RP) in BIS contains default(RPD), classifier(CL), interjection(INJ), Intensifier(INF) and Negation(NEG) as its subtype. The other category not in BIS tagset is numerals (NUM) – it is replaced by by Quantifier(QT) with General(QTF), cardinal(QTC) and ordinal(QTO) as its subtypes. Except for the three categories of adjective, adverb and postposition, all the categories have two or more sub-categories. Moreover, the category of residual, although not part of the language, it is part of the text which is to be annotated and so included in the tagset.

2.4 Summary

This chapter deal with different approaches for POS taggers of Indian languages. It also discussed various tagsets including IIIT-H (ILMT), MSRI, LDC-IL and BIS tagset for tagging part-of-speech. This chapter provides a brief survey of many POS-tagged Indian languages like- Hindi, Marathi, Bengali, Manipuri, Gujarati, Punjabi, Sanskrit, Tamil, Malayalam, Kannada, and Sinhala.

Chapter - 3

Development of POS-Tagged Corpus of Awadhi

This chapter presents the the process of the development of the post-tagged corpus of Awadhi. It includes the methods of data collection, sources of data, format of data and metadata for current research. It also discusses the challenges and issues in Optical Character Recognition (OCR) of Awadhi texts using a Hindi OCR system and how I worked around the problems. I also discuss the part-of-speech (POS) tagset of Indian languages approved by the Bureau of Indian Standards (BIS) and the POS Annotation tool that I have used for annotating the data.

3.1 Corpus Collection

Corpus provides an empirical base for various linguistic observations, hence, it is a primary source of data for the purpose of linguistic studies and for developing various tools for Computational Linguistics and Natural Language Processing. The ideal aim of data collection is to include as much diversity of a language as possible. As such it is carried out to include millions of words collected from different domains. The current research, however, aims to collect at least fifty thousand tokens of Awadhi language for the corpus formation.

3.1.1 *Source of data*

The data for the current research has been collected from Uttar Pradesh Hindi Sansthan's Library and various publication house in Lucknow. The corpus data has been collected from primary sources i.e., textbooks, short stories and novels.

First of all, the data was converted into computer-readable form. In order to do this, the written data was scanned using an OCR system that was originally developed for Devanagari. As such, this system was not accurate 100%. Therefore, there was a need to perform some postprocessing and careful proofreading.

3.1.2 *Format*

The format of current data is plain text in UTF-8 format with no additional linguistic information.

3.1.3 Metadata

Metadata plays a role in systematizing the ways in which a language corpus can be processed. It stores the interpretive framework within which the building block of a corpus were selected and are to be understood (Burnard, 1998) It is required for efficient use of any corpus. Metadata includes all the information about the corpus.

Metadata can be generated manually or by automated information processing. Manual generation tends to be more accurate, allowing the user to input any information they feel is correct or needed to describe the corpus. Automated metadata generation can be much more elementary, usually only displaying information such as file size, file extension, when the file was generated and who generated the file.

In the current research, metadata includes those information about the corpus data which makes it easier to find out the data. For example, it includes information like who is the writer, when it published, what is the title, writing script etc. It helps the users to find out the particular data that has been stored. A sample of the metadata is given in **Appendix A.**

3.2 Optical Character Recognition (OCR)

Optical Character Recognition (OCR) is a technology through which we covert various types of documents, scanned written images and PDFs into editable texts.

3.2.1 *Challenges in Optical Character Recognition (OCR) of Awadhi Texts*

The data collection process for current research was quite challenging from several perspectives. The very first and basic challenge was the absolute lack of corpus of Awadhi. And so it naturally follows that OCR system is not being developed for Awadhi. As a result, a lot of spelling errors were found in the OCRed Awadhi corpus

data when Awadhi texts were scanned using Devanagri OCR system. Some of the most common spelling errors in OCR are mentioned in Table 3.1.

Spelling Error	Correct words/Matras
◌ो (तो)	◌ौ (तौ)
◌े (ले)	◌ै (लै)
द।द 7	दादा
दें	दै

Table 3.1 Spelling Errors

As we could see, the errors seem to be largely because of the absence of such words in Hindi and the presence of a very closely-related but quite different word in Awadhi it could be hypothesised that these errors might be because of auto-correction by the 'Hindi' OCR system. Such errors necessitated a manual proofreading of the corpus. The proofreading was carried out a Java/JSP-based in-house editing tool 'editit'.

3.3 The Corpus Editing tool: Editit

As I mentioned above the corpus editing tool, Editit, has been developed using Java/JSP at the backed and runs on Apache Tomcat 8.5 web server. This tool helped in proofreading and correcting the errors that creeped into the corpus data after OCR.

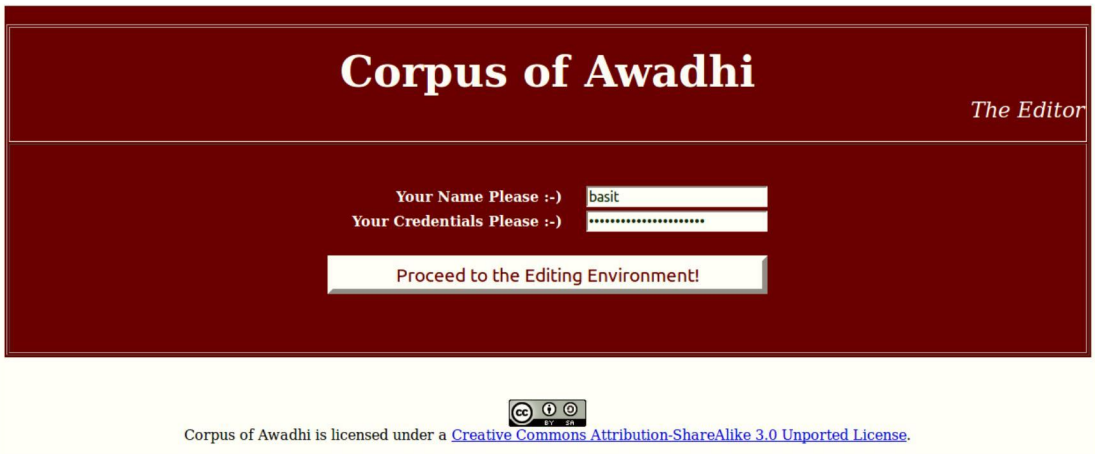

Figure 3.1: Apache Tomcat 8.5 Corpus Tool

At first, the whole OCRed as well as scanned Awadhi data was uploaded in this tool. The scanned data were in .jpeg format. The tool presents both the original JPEG files and the OCRed text together in a window. Using both of these, the some spelling errors as well as other kinds of errors in the data were corrected in this tool. The tool saved the edited files into .txt format (see Figure 3.2). The complete data was corrected using this tool.

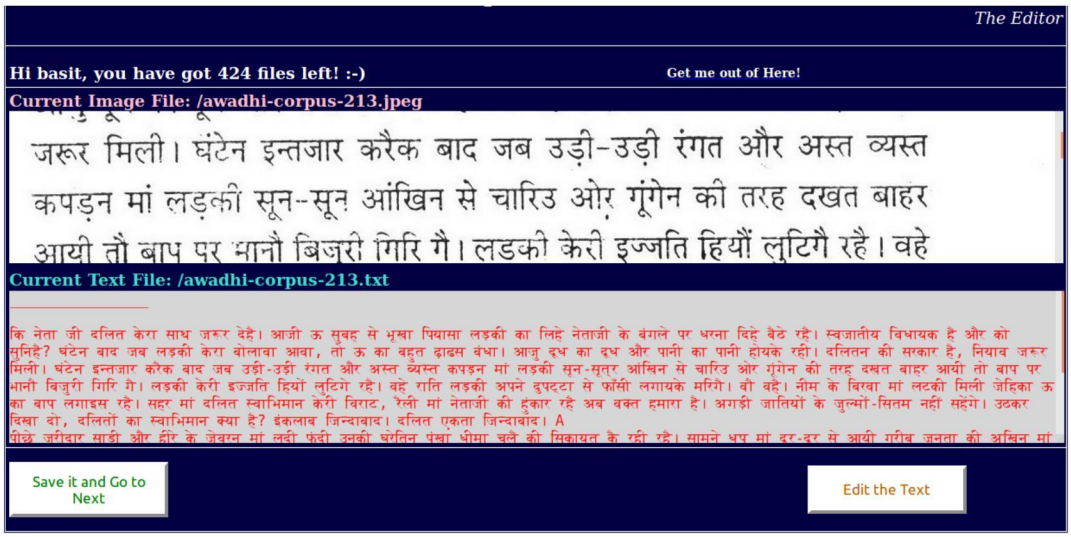

Figure 3.2: Corpus Text Edit Box

3.4 Annotation of the data: BIS Tagset

The Bureau of Indian Standards (BIS) Tagset has recommended the use of a common tagset for the part of speech annotation of Indian languages. The tagset, incorporating the advice of the experts and the stakeholders in the area of natural language processing and language technology of Indian languages, has to be followed in the annotation tasks taking place in Indian languages (Chaudhary, 2010).

Since there is no earlier tagset available for Awadhi, a POS tagset for the language was developed as part of this research. The tagset is a subset of the general BIS tagset. It is used for the POS tagging of Awadhi corpus of approximately 20 thousand tokens. The tagset has 32 different categories including punctuation, residual and unknown category. The complete tagset is given in Table 3.2

S.NO.	Categories	Subtypes Level 1	Annotation Convention	Examples
1	**Noun**	**N**	**N**	**मेहरारु,किताब दारोगा,मनसेदु**
1.1		Common	N_NN	चश्मा,गिलास, बासन,डाक्टर
1.2		Proper	N_NNP	अब्दुल, योगेश, रीना, अनम
1.3		Nloc	N_NST	ऊपरै,नीचै,आगै, पीछै
2	**Pronoun**	**PR**	**PR**	**वुइ, तुमे, यहे**
2.1		Personal	PR_PRP	वुइ, तुमरे,
2.2		Reflexive	PR_PRF	अपन, हमरे, खुदै
2.3		Relative	PR_PRL	जौ, जिस, जबै
2.4		Reciprocal	PR_PRC	दुनौ, आपसै
2.5		Wh-word	PR_PRQ	कबहूँ, काहे, का
2.6		Indefinite	PR_PRI	केउ, किस
3	**Demonstrative**		**DM**	**हिंया, हुंआ, जौ**
3.1		Deictic	DM_DMD	हिंया, हुंआ
3.2		Relative	DM_DMR	जे, जौन
3.3		Wh-word	DM_DMQ	के, काहे
3.4		Indefinite	DM_DMI	काउनौ, किस
4	**Verb**		**V**	**गवा, रहन**
4.1		Main	V_VM	कीन, कै, गवा
4.2		Auxiliary	V_VAUX	रहन, रहै, होय, है
5	**Adjective**		**JJ**	**बड़ा, अच्छै**
6	**Adverb**		**RB**	**तेजी,**
7	**Postposition**		**PSP**	**मा, से, का**
8	**Conjunction**		**CC**	**औ, अउर, बल्कि**
8.1		Co-ordinator	CC_CCD	औ, बल्कि

8.2		Subordinator	CC_CCS	तौ, कि
9	**Particles**		**RP**	**बहुत, हे, ना, भी**
9.1		Default	RP_RPD	भी, ही
9.3		Interjection	RP_INJ	अरै, हे, वाह
9.4		Intensifier	RP_INTF	बहुत
9.5		Negation	RP_NEG	नाही, ना, बिना
10	**Quantifier**		**QT**	**तनिक, एक, पहिला**
10.1		General	QTF	तनिक, बहुतै, कुछै
10.2		Cardinals	QT_QTC	एक, दुई, छे
10.3		Ordinals	QT_QTO	पहिला, दुसरका, तीसरका
11	**Residuals**		**RD**	
11.1		Foreign word	RD_RDF	**Other than script of the original text**
11.2		Symbol	RD_SYM	**$,&,*,(,)**
11.3		Punctuation	RD_PINC	**., :, ;, “,’, । ,?,!,**
11.4		Unknown	RD_UNK	
11.5		Echo-words	RD_ECH	**खाना-वाना, कुर्सी-उर्सी**

Table: 3.2 Awadhi BIS Tagset

3.5. Categories of tagset

The BIS tagset is one of the most essential parts of natural language processing. Here, I have discussed briefly about the various features of Awadhi.

3.5.1 Noun (N)

The first major category in the Tagset is ‘nouns’. A noun is a word that functions as the name of some specific thing or set of things, such as living creatures, objects,

places, actions, qualities, states of existence, or ideas. In the BIS scheme, noun has four subcategories viz common, proper, verbal and noun location. But we have studies here subcategories of Noun viz. common, proper, noun location which is best suited for Awadhi language.

I. Common Noun (N_NN)

The nouns that simply function as noun and are content words. These are used to name general items. For example, लालटेन/Lamp, कुर्सी/chair, टी.वी/TV, खिड़की/window, पेंटिंग/painting, तकिया/pillow and मोमबत्ती/candle

For example:

Words	**झण्डा**	कइयो	**घरन**	के	ऊपर	लहराति	है
POS	N_NN	QT_QTF	**N_NN**	PSP	N_NST	V_VM	V_VAUX
Translation	Many flag waving above the house.						

II. Proper Noun (N_NP)

Proper nouns are generally names that stands for two distinct features, person or place. For example, अब्दुल, इलाहाबाद, सिवप्रासाद, etc.

For Example:

चंदावती	ईद	पे	हियाँ	आवति	रहै
N_NNP	N_NN	PSP	DM_DMD	V_VM	V_VAUX
Chandawati was came here on Eid.					

III. Noun Locative (N_NST)

Locative is a grammatical case which indicate a location. For examples, पीछै, ऊपरे, नीचै, आगै, etc.

For example:

Words	सीता	बस	के	**ऊपरे**	बैठी	है
POS	N_NNP	N_NN	PSP	**N_NST**	V_VM	V_VAUX
Translation	Sita sit on top of the bus.					

3.5.2 Pronoun (PR)

The second major category is pronoun. It is divided into six sub-categories. These are:

personal, reflexive, relative, reciprocal, wh-word and indefinite.

I. Personal Pronoun (PR_PRP)

Personal pronouns cover all the pronouns that denotes a person, place or thing. This includes all their cases as well for example: वा, वुइ, तुमरे, तुमै, etc.

For Example:

Words	**तुम**	चिंता	न	करौ	।
POS	**PR_PRP**	V_VM	RP_NEG	V_VAUX	RD_PUNC
Translation	You don’t worry.				

II. Reflexive Pronoun (PR_PRF)

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

For example:

Words	**हम**	दवाई	दै	द्याबै	तुमका
POS	**PR_PRF**	N_NN	V_VM	V_VAUX	PR_PRP
Translation	I will give you medicine.				

III. Relative Pronoun (PR_PRL)

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

For Example:

Words	**जौन**	बात	करैक	होय	हमरे	सामने	करौ
POS	**PR_PRL**	N_NN	V_VM	V_VAUX	PR_PRF	N_NST	V_VM
Translation	Whatever the matter, talk in front of me.						

IV. Reciprocal Pronoun (PR_PRC)

Reciprocal pronouns denote some reciprocity. This is commonly denoted by दुनहूँ, आपसै, etc.

For example:

Words	पुलिसवाले	**आपसै**	मा	भीड़	गयेन
POS	N_NN	**PR_PRC**	PSP	V_VM	V_VAUX
Translation	The Police fought each-other.				

V. Wh-word Pronoun (PR_PRQ)

An interrogative word or question word is a function word used to ask a questions. These words are काहे, के, का, etc.

For Example:

Words	वुइ	**का**	बात	कहेन
POS	PR_PRP	**PR_PRQ**	N_NN	V_VM
Translation	What he said.			

VI. Indefinite Pronoun (PR_PRI)

The indefinite pronouns refer to unspecified objects, places or things. These are, केउ, कि, कोई etc.

For example:

Words	**केउ**	बात	नही	रजाना
POS	**PR_PRI**	N_NN	RP_NEG	N_NNP
Translation	Don’t mind Rajana.			

3.5.3 Demonstrative (DM)

The next top level category is of demonstrative. Demonstratives have the same form of the pronouns, but distributionally they are different from the pronouns as they are always followed by a noun, adjective or another pronoun. The demonstratives have been sub-categorized into four divisions- deictic, relative, wh-words and indefinite.

I. Deictic (DM_DMD)

The deictic demonstratives are default demonstratives that demonstrate the noun it modifies. Such as, हिंया, हुंआ, ई etc.

For example:

Words	चंदावती	ईद	पे	**हियाँ**	आवति	रहै
POS	N_NNP	N_NN	PSP	**DM_DMD**	V_VM	V_VAUX
Translation	Chandawati was came here on Eid.					

II. Relative (DM_DMR)

Relative demonstrative are non-distinguishable from relative pronouns, except for that a demonstrative is followed by a noun, pronoun or adjective.

For example:

Words	**जौनी**	साक्षर	न	रहै	वहि	चिट्ठी	देखिन
POS	**DM_DMR**	N_NN	RP_NEG	V_VAUX	PR_PRP	N_NN	V_VM
Translation	Saw that letter which was illiterate.						

III. Wh-word (DM_DMQ)

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

For examples:

Words	के	हमरे	साथ	जाई
POS	DM_DMQ	PR_PRF	CC_CCD	V_VM
Translation	Who will go with me			

IV. Indefinite (DM_DMI): The indefinite demonstratives refer to unspecified objects, places or things. These words are कउनिव, किस, etc.

For example:

Words	मारपीट	**केउ**	समस्या	क	हल	नहीं	है
POS	N_NN	**DM_DMQ**	N_NN	PSP	N_NN	RP_NEG	V_VAUX
Translation	Fight is not the solution of any problem.						

3.5.4 Verb (V)

The category of verb is somewhat complicated in this framework. It has main and auxiliary.

I. Main (V_VM)

The main verb tag is given for all the forms that express the main predication of the sentence. For example:

Words	चंदावती	सीता	के	रोल	करत	रहै
POS	N_NNP	N_NNP	PSP	N_NN	**V_VM**	V_VAUX
Translation	Chandawati was playing the role of sita.					

II. Auxiliary (V_VAUX)

The auxiliary verb is a closed set of verb. A sentence can have one more auxiliary verbs.

For Example:

Words	हनुमान	चंदावती	के	दादा	थे
POS	N_NNP	N_NNP	PSP	N_NN	**V_VAUX**
Translation	Hanuman was grandfather of chandawati.				

3.5.5 Adjective (JJ)

Adjective modifies a noun.

For example:

Words	वुई	**कारिया**	घोड़ा	ते	सवार	रहन
POS	PR_PRP	**JJ**	N_NN	PSP	V_VM	V_VAUX
Translation	He was riding with black horse.					

3.5.6 Adverb (RB)

Adverb also is mono-category part-of-speech. The standards document says that the category of adverb (RB) is only for manner adverbs.

For example:

Words	सीता	का	किरदार	बहुतै	नीक	रहन	।
POS	N_NNP	PSP	N_NN	**RB**	JJ	V_VAUX	RD_PUNC
Translation	Sita’s role was very good.						

3.5.7 Postposition (PSP)

Postpositions are all the parts-of-speech that work as case marker. Words like, की, मा, से, ते, etc.

For example:

Words	वुई	कारिया	घोड़ा	**ते**	सवार	रहन
POS	PR_PRP	JJ	N_NN	**PSP**	V_VM	V_VAUX
Translation	He was riding with black horse.					

3.5.8 Conjunction (CC)

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

I. Co-ordinator (CC_CCD)

Coordinators are typically the words that join two phrases(noun or verb), of the same category or a clause. Some common conjunctions are औ, या, बल्कि, etc.

For Example:

Words	अमरूद	तोड़ेस	**औ**	भाई	का	दिहेसि
POS	N_NP	**V_VM**	**CC_CCD**	N_NN	PSP	V_VM
Translation	Plucked guava and gave to the brother.					

II. Subordinator (CC_CCS)

Subordinator typically conjoins two clauses and the second clause is subordinated. Some of the subordinate conjunctions are मगर, कि, तौ, etc.

For Example:

Words	रामबरन	आय	जाय	तब	हम	बात	करबै
POS	N_NNP	V_VM	V_VAUX	CC_CCS	PR_PRP	N_NN	V_VM
Translation	Rambaran will comes then we will talk.						

3.5.9 Particles (RP)

Particles play many roles in the language. In the tagset, default, classifier, interjection, intensifier and negation are subtypes of the particles category.

I. Default (RP_RPD)

Default Particle is a category that includes all those element of the language which

though do not have any lexical important but are salient functionally. Such as, ही, भी, जी etc.

For example:

Words	राजेस	जी	घरै	पे	आये	रहेन
POS	N_NNP	**RP_RPD**	N_NN	PSP	V_VM	V_VAUX
Translation	Rajes ji came at home.					

II. Interjection (RP_INJ)

Interjections are particles which denote exclamation utterances. The common exclamatory marks in Awadhi are अरे, हाय, हे etc.

For example:

Words	हे	राम	!	का	हूई	गवा	।
POS	**RP_INJ**	N_NNP	RD_PUNC	PR_PRQ	V_VM	V_VAUX	RD_PUNC
Translation	Hey Ram! What happened.						

III. Intensifier (PR_INTF)

Intensifiers are words that intensify the adjectives or adverbs. For example:

Words	तै	वहिके	साथ	जाय	केरि	**बहुत**	छोट	है
POS	PR_PRP	PR_PRL	CC_CCS	V_VM	PSP	**PR_INTF**	JJ	V_VAUX
Translation	You are too young to go with them.							

IV. Negation (PR_NEG)

The negation particles are the words that indicate negation. These include नाही, ना, मत, बिना, बिगैर etc.

For example:

Words	हम	जादा	नौटंकी	**नहीं**	कीन	है
POS	PR_PRF	JJ	N_NN	**RP_NEG**	V_VM	V_VAUX
Translation	I have not done much nautanki.					

3.5.10 Quantifiers (QT)

A quantifier is a word which qualities the noun, i.e., it expresses the noun's definite or

indefinite number or amount. The Quantifier category includes general, cardinal, and ordinal subtypes.

I. General (QT_QTF)

The general quantifiers do not indicate any precise quantity. For example, बहुत, कुछै, etc.

For example:

Words	शादी	मा	**बहुत**	लोगन	रहेन	।
POS	N_NN	PSP	QT_QTF	N_NN	V_VAUX	RD_PUNC
Translation	There were so many people in marriage.					

II. Cardinals (QT_QTC)

The numbers which quantify objects are cardinal quantifier. In Awadhi, एक, दुई, छे, etc.

For example:

Words	उ	**दुई**	चाय	बनाई	बा
POS	PR_PRP	**QT_QTC**	N_NN	V_VM	V_VAUX
Translation	He has made two tea.				

III. Ordinals (QT_QTO)

Quantifiers that specify the order in which a particular object is placed in a given world are ordinal quantifiers. Such as, in Awadhi, पहिला, दुसरका, तीसरका etc.

For example:

Words	चंदावती	**तिसरे**	दिन	हियाँ	आवति	है	
POS	N_NNP	**QT_QTO**	N_NN	DM_DMD	V_VM	V_VAUX	
Transaltion	Chandawati comes here on the third day.						

3.5.11 Residual (RD)

Residual as a major category in this tagset has five subtypes; foreign word, symbol, punctuation, unknown and echo words as subtypes.

I. Foreign Words (RD_RDF): The foreign words are all the words that are not written in the Devanagari script.

II. Symbols (RD_SYM): The symbols are the characters that are not part of the regular Devanagari script such as *, @, #, $, % etc.

III. Punctuation (RD_PUNC): Only for punctuations like, -, ?, ; , “ , ।, etc.
For example:

Words	राजेस	कहेसि	-	बुआ
POS	N_NNP	V_VM	**RD_PUNC**	N_NN
Translation	Rajes said – Bua			

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

V. Echo-words (RD_ECH)

Echo words are two words that occur together and the second one has no meaning on its own and it cannot occur on its own. It enhances the meaning of the word with which it occurs. Such as, खाना-वाना, कुर्सी-वुर्सी etc.

3.6 Tool for Part of speech Annotation: WebAnno

For the current research, WebAnno web-based tool was used for POS tagging of the data. WebAnno is a general purpose web-based annotation tool for a wide range of linguistic annotations including various layers of morphological, syntactical, and semantic annotations. (Biemann, et al. 2013) Additionaly, custom annotation layers can be defined, allowing WebAnno to be used also for non-linguistic annotation tasks. “WebAnno is a multi-user tool supporting different roles such as annotator, curator, and project manager.”(Yimam, Gurevych, Eckart de Castilho, and Biemann. 2013) The progress and quality of annotation projects can be monitored and measured in terms of inter-annotator agreement. Multiple annotation projects can also be conducted in parallel.

The process of POS annotation of the data has been done in four steps.

i. Importing documents.

ii. Creating tagset.

iii. POS Annotation

iv. Exporting the Project

Importing documents: In the first step, the whole corpus of Awadhi was uploaded on the page of import document. These uploaded data of Awadhi were in plain text format (see Figure 3.3)

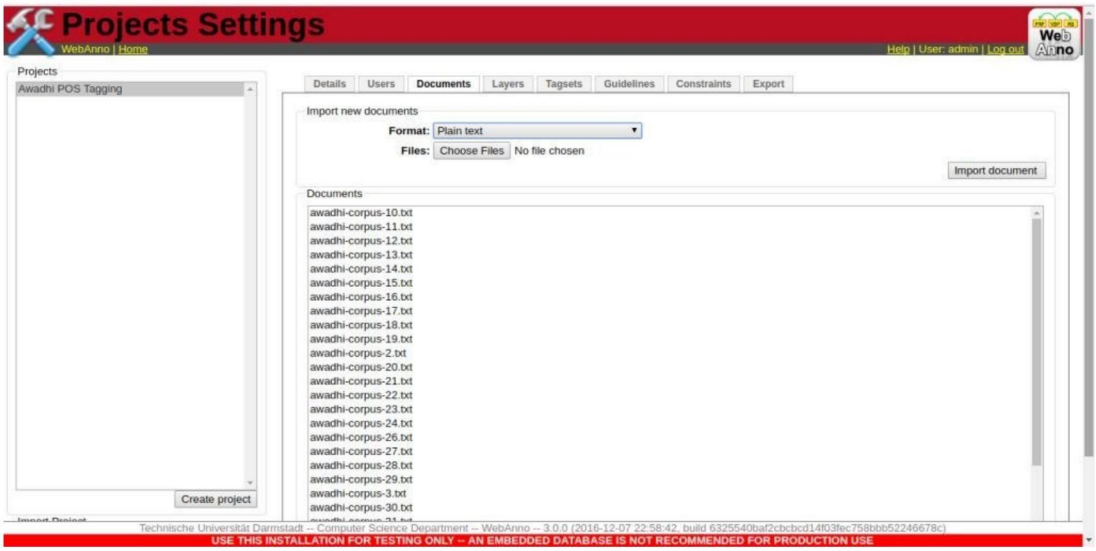

Figure 3.3: Import page

Creating tagset: In the second step,I entered the tagset of Awadhi in the tool with details like Name of the tagset, Name of the language and Description about user and project. After this, I created the POS tags - Noun, Pronoun, Demonstrative, Verb, Adjective, Adverb, Postposition, Conjunction, Quantifiers, Particles, Residuals – along with the description of each tag, as per the tagset that I had developed for the language. (see Figure 3.4).

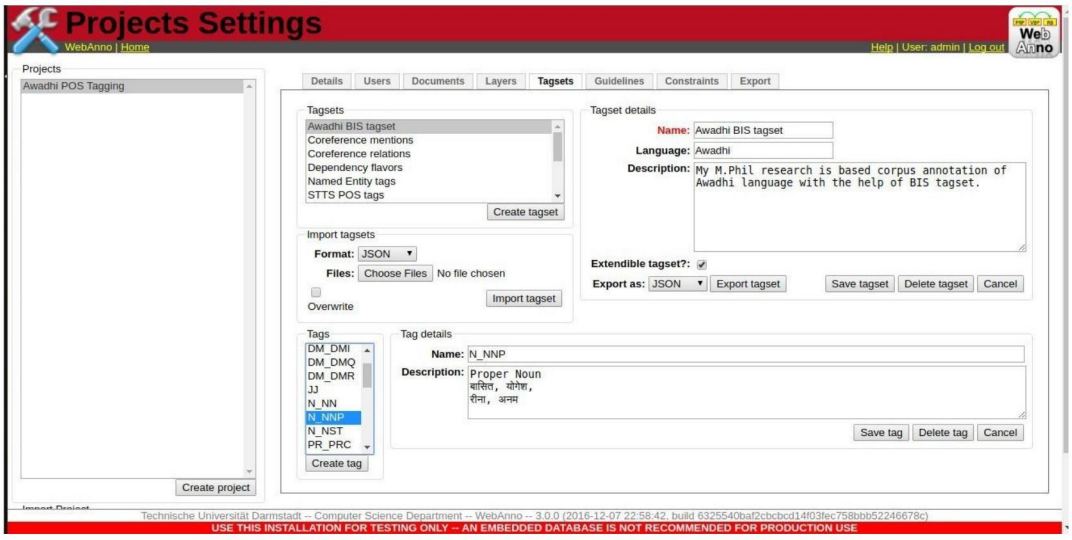

Figure 3.4: Tagset Page

POS Annotation: In the third step, the uploaded document of Awadhi language was opened in Annotation page. The page opens 5 sentences at one go . First, I selected the first token in the sentence and assigned the POS value to that token. After this, the tool automatically forwards the control to the next token and the POS tag is assigned to that token – thus it allows for quick annotation of the data. All the corpus data has been annotated through this process. (see Figure 3.5)

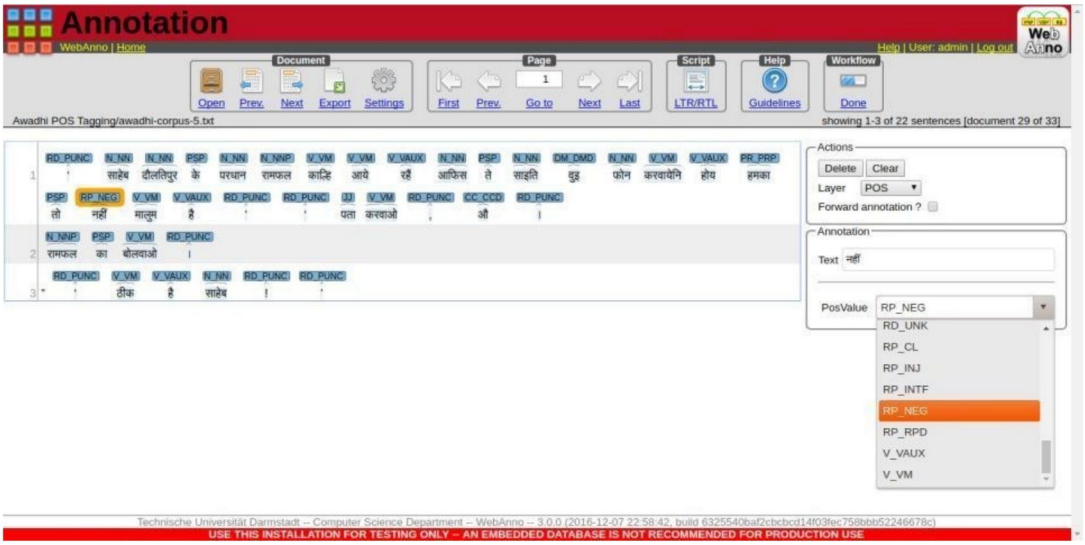

Figure 3.5: POS Annotation Page

Exporting Documents: In the last step, after the POS annotation of the corpus data is completed then all the data is exported in CoNLL format (see Figure 3.6).

The final output resembles the following sample.

e.g.,

हनुमान N_NN O

कहेनि V_VM O

- RD_PUNC O

‘ RD_PUNC O

द्याखौ V_VM O

ई DM_DMD O

सार N_NN O

पाखंडी N_NN O

बाबा N_NN O

कइस PR_PRQ O

दुनिया N_NN O

भरेक RB O

स्वाँग N_NN O

बनावति V_VM O

हैं V_VAUX O

? RD_PUNC O

' RD_PUNC O

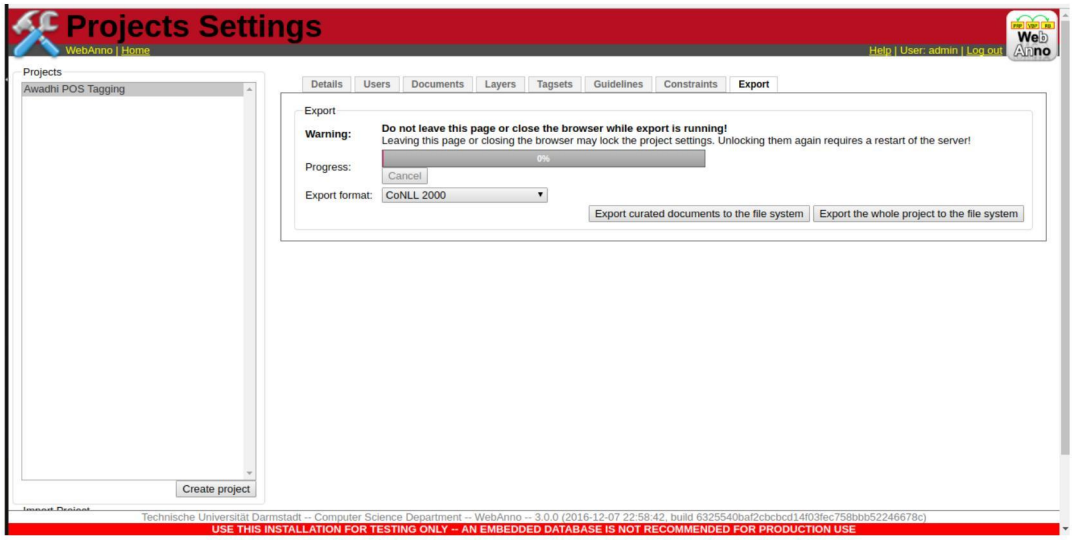

Figure 3.6: Export Document Page

3.7 Summary

This chapter discussed the methods of data collection, sources of data, format of data and metadata for current research. It also discussed various challenges and issues in OCR of Awadhi texts and how I worked around these problems. I also discussed the BIS tagset for POS Annotation and tool that I have used for annotating the data.

Chapter – 4

Conclusion

This dissertation presents the first POS-tagged corpus of Awadhi language. There are a total of approximately 70,000 tokens in the corpus and approximately 20,000 tokens are tagged with POS information. For the current research, the data was collected from the primary sources. For the current research, the task of POS annotation has been done through WebAnno tool using the BIS tagset. The tagset was developed as part of this research.

Further research can be done in several directions. More resources for Awadhi language, such as annotated and unannotated corpus, needs to be developed. Moreover, automatic tagging systems may also be developed for the language using these resources.

References:

Baskaran, S., Bali, K., Bhattacharya, T., Bhattacharyya, P., Jha, G. N., Sobha, L., Subbarao, K. V. 2008. *Designing a Common POS-tagset Framework for Indian Languages.* The 6th Workshop on Asian Language Resources,. Retrieved from https://www.microsoft.com/enus/research/wpcontent/uploads/2016/02/I08-7013.pdf (28.07.2017)

Bhattacharyya, P., Chakrabarti, D., & Sarma, V. M. 2007. *Complex predicates in Indian languages and wordnets.* Language Resource Evaluation, 40 (3-4), 332-355.

Burnard, L. 2004. Metadata for corpuswork. Retrieved from http://users.ox.ac.uk/~lou/wip/metadata.html. (20.07.2017)

Chapelle, C., Garside, R., Leech, G., & Sampson, G. 1988. *The Computational Analysis of English: A Corpus-Based Approach*. TESOL Quarterly.

Collins, P. 2014. Review: Bas Aarts, Joanne Close, Geoffrey Leech and Sean Wallis (eds.). *The verb phrase in English: Investigating recent language change with corpora*. ICAME Journal

Crystal, D. 1991. *A Dictionary of Linguistics and Phonetics,* (3rd ed.). Blackwell.

Dash, N.S. 2007. *Language Corpora and Applied Linguistics*. Kolkata: Sahitya Samsad.

Dash, N.S. 2008. *Corpus Linguistics: An Introduction.* New Delhi: Pearson Education. Longman.

Dash, N.S. 2009. *Corpus Linguistics: Past, Present and Future*. New Delhi: Mittal Publications.

Dash, N. S. 2010. *Corpus Linguistics : A General Introduction.* Retrieved from http://www.ldcil.org/download/Corpus%20Linguistics.pdf. (08.08.2017)

Govilkar, S., J. W, B., & Rathod, S., 2014. *Part of Speech Tagger for Marathi Language*. International Journal of Computer Applications.

Garside, R., Leech, G. and McEnery, A. 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

Gupta, R., Joshi, N., & Mathur, I. 2013. *Analysing Quality of English-Hindi Machine Translation Engine outputs using Baysian Classification.* International Journal of Artificial Intelligence & Applications, 4(4), 165-171. doi:10.5121/ijaia.2013.4415.

Garside, R., Leech, G. and Sampson, G.1987. *The Computational Analysis of English: A Corpus Based Approach. London*: Longman.

Gerbig, A. 1997. *Lexical and Grammatical Variation in a Corpus: A Computer Assisted Study of Discourse on the Environment.* London: Peter Lang Publishing.

Joshi, N., Darbari, H., & Mathur, I. 2013. *HMM Based POS Tagger for Hindi. Computer Science & Information Technology* (CS & IT). doi:10.5121/ csit.2013.3639

Leech, G., Myers, G., and Thomas, J. 1995. *Spoken English on Computer: Transcription, Markup and Applications*. Harlow: Longman.

MASON, O. 1999. *Corpus Annotation: Linguistic Information from Computer Text Corpora. Natural Language Engineering*, 5(3), 301-307. doi:10.1017/s135132 4900212266

McCarthy, M., & O'Keeffe, A. 2012. *Analyzing Spoken Corpora. The Encyclopedia of Applied Linguistics.* doi:10.1002/9781405198431.wbeal0028.

McCarthy, J. 1982. *Formal Problems in Semitic Phonology and Morphology*. New York: Garland.

McCarthy, M. 1998. *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University press.

McEnery, T., and Wilson, A. 1996. *Corpus Linguistics. Edinburgh*: Edinburgh University Press.

Nelson, G., G. Sampson. 1997. *English for the computer: the SUSANNE corpus and analytic scheme.* Oxford: Clarendon Press.

Petrovic, S., Snajder, J., Dalbelo. B., & Kolar, M. 2006. *Comparison of Collocation Extraction Measures for Document Indexing. Journal of Computing and Information Technology.*, 14(4), 321.doi:10.2498/cit.2006.04.08

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Singha, R. K., Purkayastha, B., & Singha, K. D. 2012. *Part of Speech Tagging in Manipuri: A Rule based Approach.* International Journal of Computer Applications, 51(14), 31-36. doi:10.5120/8111-1727

Singh, S., Gupta, K., Shrivastava, M., & Bhattacharyya, P. 2006. *Morphological richness offsets resource demand- experiences in constructing a POS tagger for Hindi. Proceedings of the* COLING/ACL on Main conference poster sessions .doi:10.3115/1273073.1273173

Smith, N., & McEnery, T. 1998. *Issues in Transcribing a Corpus of Children's Handwritten Projects.* Literary and Linguistic Computing, 13*(*4), 217-225. doi:10.1093/llc/(13.4.2017)

Tapaswi, N., & Jain, S. 2012. *Treebank based deep grammar acquisition and Part-Of-Speech Tagging for Sanskrit sentences*. CSI Sixth International Conference on Software Engineering 2012. (CONSEG).

Yadava, Y. P., Hardie, A., Lohani, R. R., Regmi, B. N., Gurung, S., Gurung, A., Hall, P. 2008. *Construction and annotation of a corpus of contemporary Nepali.* Corpora, 3(2), 213-225. doi:10.3366/e1749503208000166

Yule, George. 2005, *The study of Language* 3rd edition. Cambridge University Press.

http://www.wikipedia.com/awadhi_language 20.07.2017

Appendix-A
Metadata of Awadhi

Code	Title of the article	Text type	Name of the newspaper/book/magazine	Date of publication	Original publication page number	Name of the original author	Original OCRed file path ----->>
ac001	नारासिंहल की पैकरमा	Short Story	चंदावती	2012	37	भारतेन्दु मिस्र	ocr-text/awadhi-corpus-1.txt
ac002	फैसला	Short Story	चंदावती	2012	96	भारतेन्दु मिस्र	ocr-text/awadhi-corpus-2.txt
ac003	नारासिंहल की पैकरमा	Short Story	चंदावती	2012	44	भारतेन्दु मिस्र	ocr-text/awadhi-corpus-3.txt
ac004	चंदावती की निकासी	Short Story	चंदावती	2012	73	भारतेन्दु मिस्र	ocr-text/awadhi-corpus-4.txt
ac005	देबीदल केरि चौकी	Short Story	चंदावती	2012	85	भारतेन्दु मिस्र	ocr-text/awadhi-corpus-5.txt
ac006	चंदावती की निकासी	Short Story	चंदावती	2012	30	भारतेन्दु मिस्र	ocr-text/awadhi-corpus-6.txt
ac007	जारी है लड़ाई	Short Story	चंदावती	2012	127	भारतेन्दु मिस्र	ocr-text/awadhi-corpus-7.txt
ac008	देबीदल केरि चौकी	Short Story	चंदावती	2012	79	भारतेन्दु मिस्र	ocr-text/awadhi-corpus-8.txt

---->>First proofread file path	Source of the data	Scan and OCR by	First proofread by	Path to the original scan	POS-tagged File Path	POS Tagging by
proofread-text1/awadhi-corpus-1.txt	Abdul	Abdul	Abdul	original-images/awadhi-corpus-0042B.jpeg	tagged-corpus-1/awadhi-corpus-1.txt	Abdul
proofread-text1/awadhi-corpus-2.txt	Abdul	Abdul	Abdul	original-images/awadhi-corpus-0023A.jpeg	tagged-corpus-1/awadhi-corpus-2.txt	Abdul
proofread-text1/awadhi-corpus-3.txt	Abdul	Abdul	Abdul	original-images/awadhi-corpus-0010A.jpeg	tagged-corpus-1/awadhi-corpus-3.txt	Abdul
proofread-text1/awadhi-corpus-4.txt	Abdul	Abdul	Abdul	original-images/awadhi-corpus-0020B.jpeg	tagged-corpus-1/awadhi-corpus-4.txt	Abdul
proofread-text1/awadhi-corpus-5.txt	Abdul	Abdul	Abdul	original-images/awadhi-corpus-0008A.jpeg	tagged-corpus-1/awadhi-corpus-5.txt	Abdul
proofread-text1/awadhi-corpus-6.txt	Abdul	Abdul	Abdul	original-images/awadhi-corpus-0008.jpeg	tagged-corpus-1/awadhi-corpus-6.txt	Abdul
proofread-text1/awadhi-corpus-7.txt	Abdul	Abdul	Abdul	original-images/awadhi-corpus-0063B.jpeg	tagged-corpus-1/awadhi-corpus-7.txt	Abdul
proofread-text1/awadhi-corpus-8.txt	Abdul	Abdul	Abdul	original-images/awadhi-corpus-0051B.jpeg	tagged-corpus-1/awadhi-corpus-8.txt	Abdul