AUTOMATIC IDENTIFICATION AND ANALYSIS OF COMPLEX PREDICATES IN MAGAHI

Thesis submitted to

Jawaharlal Nehru University

in partial fulfilment of the requirements for the award of the degree of

DOCTOR OF PHILOSOPHY

Submitted by

SHIVEK KUMAR SICKY

Supervised by

PROF. PRADEEP KUMAR DAS
CENTRE FOR LINGUISTICS, JNU
SLL & CS, JNU, NEW DELHI

Co- Supervised by

PROF. GIRISH NATH JHA
SSIS, JNU, NEW DELHI
CHAIRMAN, CSTT
MINISTRY OF EDUCATION
GOVT. OF INDIA

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

2022

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

Dated 23/01/2023

CERTIFICATE

This is to certify that the thesis titled **"AUTOMATIC IDENTIFICATION AND ANALYSIS OF COMPLEX PREDICATES IN MAGAHI"** submitted by **Mr. SHIVEK KUMAR SICKY**, in partial fulfillment of the requirements for award of degree of Ph.D. of Centre for Linguistics, School of Language, Literature and Culture Studies-I, Jawaharlal Nehru University, New Delhi, has not been previously submitted in part or in full for any other degree of this university or any other university/institution.

We recommend this thesis be placed before the examiners for evaluation for the award of the degree of Ph.D.

PROF. PRADEEP KUMAR DAS

SUPERVISOR

Prof. PRADEEP K. DAS
Centre for Linguistics
School of Language, Literature & Culture Studies
Jawaharlal Nehru University, New Delhi-110067

PROF. GIRISH NATH JHA

Co- SUPERVISOR

PROF. PRADEEP KUMAR DAS

CHAIRPERSON

Chairperson
CL/SLL&CS
J.N.U., New Delhi-67

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

Dated 23/01/2023

Declaration by the Candidate

The Thesis titled "**Automatic Identification and Analysis of Complex Predicates in Magahi**" submitted by me for the award of degree of **Doctor 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.

Shivek Kr. Sicky

Mr. Shivek Kumar Sicky
Ph.D Student
Centre for Linguistics
SLL & CS
JNU, New Delhi-110067.

CONTENTS

List of Abbreviations Used

AI	: Artificial Intelligence
APP	: Application
AT &T	: American Telephone and Telegraph Company
AV	: Adjective Verb Constructions
BERT	: Bidimensional Encoder Representations Transformers
BIS	: Bureau of Indian Standardization
BNC	: British National Corpus
CBR	: Case Based Learning
CDAC	: Centre for Development of Advanced Computing
CGC	: Computational Grammar Coder
CIEFL	: Central Institute of English and Foreign Languages
CIIL	: Central Institute of Indian Languages
CLAWS	: Constituent Likelihood Automatic Word Tagging System
CNN	: Convolutional Neural Networks
CNV	: Conjunct Verbs
COCA	: Corpus of Contemporary American English
COLING	: Computational Linguistics
COLLAB	: Collaborator
CPR	: Complex Predicates Rules
CPs	: Complex Predicates
CRF++	: Conditional Random Fields
CRULP	: Centre for Research in Urdu Language Processing
CV	: Compound Verbs
CVC	: Complex Verb Construction
DeitY	: Department of Electronics and Information Technology
DFA	: Deterministic Finite Automatic
DNS	: Domain Named Server
ECV	: Explicator Compound Verb
EFLU	: The English and Foreign Languages University
GLM	: Global Linear Model

HMM : Hidden Markov Model

HTML : Hypertext Markup Languages

IA : Indian Languages

IBM : International Business Machines

IIIT : International Institute of Information Technology

IISC : Indian Institute of Science

IIT : Indian Institute of Technology

IL : Indian Language

ILCI : Indian Languages Corpora Initiative

ILCIANN : Indian Languages Corpora Initiative Annotation

ILMT : Indian Language Machine Translation

IP : Internet Protocol

IR : Information Retrieval

ISI : Indian Statistical Institute

ISO : International Organization for Standardization

JJP : Adjectival Phrase

JNU : Jawaharlal Nehru University

JSP : Java Server Pages

LDC-IL : Linguistic Data Consortium for Indian Languages

LRL : Left-Right-Left

LSTM : Long or Short- Term Memory

LTRC : Language Technologies Research Centre

MA : Morphological Analyser

MBT : Memory-Based Tagger

ME : Maximum Entropy

MECIT : Ministry of Electronics Communication and Information Technology

MFT : Most Frequent Tag

MIA : Middle Indo-Aryan

MSRI : Microsoft Research India Private Limited

MT : Machine Translation

MWE : Multi-word Expressions

NER : Named Entity Recognition

NIA : New Indo-Aryan

NLP : Natural Language Processing

NLU : Natural language Understanding

NP : Noun Phrase

NTNU : Norwegian University of Science and Technology

NUCES : National University of Computers and Emerging Sciences

NV : Noun Verb Constructions

OIA : Old Indo-Aryan

OOA : Object Oriented Approach

PNG : Person Number Gender

POS : Parts of Speech

RCVs : Reverse Compound Verbs

RNN : Recurrent Neural Network

SL : Source Language

SMT : Statistical Machine Translation

SOAP : Simple Object Access Protocol

SOV : Subject, Object, Verb

SP : Shallow Parser

SVC : Serial Verb Construction

SVM : Support Vector Machine

TAM : Tense Aspect Mood

TC : Typecraft

TDIL : Technology Development for Indian Languages

TICMTLAA : Teddington International Conference on Machine Translation of Languages and Applied Analysis

TL : Target Language

TMC : The Time Magazine Corpus

TTS : Text To Speech

UCRL : University Centre for Corpus Research

UDT : Universal Dependency Technique

ULMFIT : Universal Language Model Fine-Tuning

UP : Utasarga Apavada

URL : Uniform Resource Locator

USA : United States of America

UTF : Unicode Transformation Format

V1 : Verb One or Polar Verb

V2 : Verb Two or Vector Verb

VP : Verb Phrase

WSD : Word Sense Disambiguation

WSJ : Wall Street Journal

WWW : World Wide Web

XML : Extensible Markup Language

List of Abbreviations used for Complex Predicates Tagsets at Token Labels

COMP_EXPL : Explicator Compound

COMP_REVS : Reverse Compound

COMP_SV : Serial Verbs Compounds

COMP_SV1 : Serial Verbs Compounds Category 1

COMP_SV2 : Serial Verbs Compounds Category 2

COMP_SV3 : Serial Verbs Compounds Category 3

COMP_SV4 : Serial Verbs Compounds Category 4

COMP_V : Compound verb

CONJ_JJ : Adjective Conjunct

CONJ_NN : Noun Conjunct

List of Abbreviations used for Complex Predicates Tagsets for Tokens at Sentence Labels

B_CP : Beginning Complex Predicate Sentence

I_CP : Immediate Constituent Complex Predicate Sentence

I_CP_0 : Immediate Constituent Complex Predicate Sentence followed by no constituent

List of Figures, Screenshots, Charts and Tables, Snippets

Acknowledgements

Mere imagination of this long journey fills me with joy, compassion, enthusiasm and a lot of excitement. I still remember this afresh long journey of JNU, when it all started with clearing the entrance and viva and getting in B.A. 2nd year of German studies in Centre for German studies in the same School of languages in the year 2010. This long journey has not only been an educational and academic pursuit but a voyage of memories blended with lot of hardships, ups and downs, achievements and failures for all theses long years that I spent in JNU as a student, learner and then transforming myself as a PhD scholar. Getting into the India's top most Jawaharlal Nehru University for a student like me who comes from a normal social background with having no interest in education during my school days was itself a dream come true reality, which could never have been possible up to this large extent without the two prominent reasons one the blessings of Goddess of Wisdom and education Sarasvati and second the faith of my school principal Dr. C.M. Singh, who once said to my "never stop his education, just for the sake of some handful of money, no matter even if you sleep empty stomach someday.". I must say it here, big thank you Sir, for your firm belief and confidence in me. Your confidence and belief in me have made me to reach an attain this educational height today, which itself is a dream for many students in India and abroad toady. Along with this it was Ms. Sujata Mishra mam, who once said "be self-dependent and advised my elder brother to just become a torch bearer for life but not the one who could feed the education within." This PhD is a bestowed present from my end to these people of my entire school Career.

Along with this, the constant belief and firm dedication of my parents especially my father, Shri Jitendra Prasad, who is also my role model in my extra ordinary capabilities has continuously given me a lot more strengths. It is my father, from whom I have learnt not to run from situations and face them bravely. I must say, thank you Papa ji. I have achieved this great height of education because of your hardships today. Thanks a lot. This inspiration of acquiring a doctorate degree from India's prestigious university is one among the millions of unfulfilled dreams of my father. I am also indebted and blessed to have mother, Smt. N.K. Baranwal, who stood besides me ahead of all my odds. It is she who tend to crave the passion of becoming a good person by heart, and its her divine education who put me always forward from all those situations whenever I felt exhausted, during this journey. It is her eternal love, affection and care which worked like a nectar for me. My elder brother Vivek kr. Vicky, who always filled me in with extreme confidence, and have stayed constantly as a guardian throughout my PhD journey. It was he, who stood beside me, when I was in my toughest times of academic career. Thank you, Bhaiya, for being always there like a support system. I am also thankful to his better half and my sister-in-law Ms. Madhu Baranwal, who will be witnessing this high rise of my career with bared eyes and will pray to God that her next generation also my cutie-pie niece Ms. Dhruvi Kashyap may achieve greater heights in future, than I am today. A lot of love and affection from chachu, my lovely. This is a small beautified present from my side to you.

I am indebted from the core of my heart, to my loving Sister Ms. Shalini Kumari, who supported and cared for me since childhood. I still remember those days, when she acted as a

life support system and pulled me all along her back for giving me a sight seen of my native cities, whenever, there existed any fairs like Dusshera. I am really indebted Didi. This is an auspicious present from babu, to you and your hardships that you bestowed on me during my childhood days. I am also indebted and thankful to my brother-in-law Shri Manish Kumar, who supported and stood with me all the time whenever, I felt low, during my entire academic career, and most importantly during all my hardships. A big Thanks to you Jija ji, for standing behind me and tolerating all my shits and nuances. It is also an inspirational moment for both my nephews, Tanishq Goyal and Tejas Goyal, that today with all your love and support I am standing confidently at the high rise of my career. Apart from all these, I am also dejected and regretful today, that two of my strongest family members, my uncle Late, Shri Upendra Kumar and my aunty Late, Smt. Pushpa Rani, who are not here with us on this planet earth to witness this height of my career. These two were the most awaited people of the family who might have got filled up with joy and courage when they have seen me today in reaching such academic altitudes. It is also a love bestowed upon my departed pious soul of my youngest cousin Late Aditya, who left us bereaved at a very early age. May all your souls rest in peace both of you as this an homage from my end to each one of you personally. Please bestow your love and affection from wherever you are in the heaven today. I am also thankful to my uncle Shri Surendra Kumar, who induced all his long-indebted efforts in me and help in in achieving to what I am today, right from my childhood, when studies were millions of miles away for me. Trust me, I still remember those long unslept nights, wherein you have taught me passionately taking out time from your busy schedules. It is also a gift of love and eternity to my aunties who are also like mothers to me namely, Smt. Poonam and Smt. Doli. I still remember those days when my youngest aunty Ms. Doli, hold my hands saying “aap jivan me bohot kuch kar skate hai is situation se bahar aayie”. Trust me, these acted as golden words for me during my toughest days, when I felt the lowest during my entire career till date. This is also a gift of love for my cousin brother Soham Raj, and cousin sister, Ananya Raj, from my end and I believe that he will also be able to reach such existential heights of his career whenever required. All the best to both of you for your shining futures. I cannot be more thankful to the almighty for bestowing me such a lovable and affectionate family, who all supported me collectively in someway or the other.

The real architect of my academic manor is non other than my M.Phil. supervisor and PhD co-supervisor Prof. Girish Nath Jha, Prof. SSIS, JNU and Chairman CSTT, Ministry of Education, Govt. of India, under whose ample guidance I have learnt the tricks of technologies and software. It is he, who always has firm belief in me, that I would be doing the best, no matter whatever, the situation will be. My basics of Computational researches is totally sane to imagine without him. Thank you, Sir, for all your friendly support and able guidance. I am among one of those who got a great chance to shape my PhD basics i.e., M.Phil. under your able guidance. My PhD work could not have seen the day of light without the able guidance of my supervisor Prof. Pradeep Kumar Das, who is also currently the chairperson of the centre. His fatherly attitude, ample and deep linguistics knowledge on each and every sphere, friendly nature, kind support, affection and necessary required instruction have made me today able in completing this great work. Thank you, sir, for this true love, support and accurate guidance. It was you who was always there, whenever I needed any suggestions for correcting my knowledge concerning the basics of linguistics in the centre. You are truly a man of honour sir. The entire linguistic centre is fortunate to have you as a faculty. Both these Professors are not just my supervisors rather, the torch bearer for

my entire research career. It is because of these two people that I am able to see a different world of academics today and would be able to shine much in future. I am also thankful to all the esteemed faculties in the centre for linguistics namely, Prof. Ayesha Kidwai, Prof. Hari Madhab Ray, Prof. Vadthya Kishore, Prof. Gopal Ram, Prof. Franson Manjali, Prof. Anvitta Abbi, Prof. Franson Manjali, Prof. P.K.S. Pandey, Prof. Vaishna Narang, Prof. Pauthang Haokip and all the distinct members who taught me during my M.A and M.Phil. days in centre for Linguistics. I owe special thanks to both the kind hearted office staff members Naveen Bhaiya and Gopal Sir. The pages of my PhD would always reminisce me of the endless good times spent here as a student at first and then as a researcher. Above all, a million thanks to the Computational Linguistics Research and Development Lab in the Special Centre for Sanskrit Studies for providing me access to use computers during the designing of the tool.

Working in this area was a bit challenging task as the topic needed deep ground work, involving a lot more interactions for collecting the primary data and validating the same. The primary work of data collection for my PhD would not have been possible without the kind guidance of Mr. Pitambar Behera, who always stood with me side by side in every odd situation. It was him, who made me learn all the basic necessary steps of Computational methods and NLP related tasks. It was him who made me to learn not only these but have supported like a friend, philosopher and guide in every sphere of my life. Thank you, Pitambar Sir, for all your love and support. This work is just a mere imagination without your support. It is a total injustice if I forgot to mention some great names like Prof. Umesh Kumar Khute Sir, JNU, with whom I started my JNU journey as a hostel mate and who have always made us learn from his struggling life. I believe that you will always guide us in getting for what you have achieved today in your life. Pankaj Srivastav Sir, a currently working Assistant Commissioner Income tax, and a JNU alumnus is another name, who stood out for me when I was at the lowest point of my career. I still remember the long talkative nights, where he encouraged me in all the way he could. A big thanks to you as well Sir. I am lucky and fortunate to have seniors and pathfinders like you.

Friends deserve special acknowledgments in this tedious task, without whose support and love, the goal of completing my PhD journey is always half-way. I take the opportunity to thank Abhay my first ever classmate cum friend in JNU in Centre for Germanic studies, my roommate who will always be my lovable younger brother Shubham Poddar, my current roommates Amit Gupta and Pushkar Chaudhary along with the very kind hearted Sushil ji who also is now a JNU alumnus from Spanish centre. I also thank Mr. Ajay Verma, Gulshan ji, Raees and everyone who always forced me for an often-late night chai and Maggi party. I also thank Mr. Harshit Jha, who is already an MPhil from Centre for Indian languages, who continuously praised me to finish off my work on time. it is his kind words that have made me to finish this task despite the long odd period of COVID-19 pandemic. I am also thankful to all the hostel and mess staff members including mess workers, managers and staffs, who have made me feel in JNU a truly home environment and made me feel this a home away from home in all these twelve years of my long journey.

-Shivek Kumar Sicky

ABSTRACT

The current undertaken research sights the interest of designing a computational model and tool based on the Support vector machine (SVM) statistical algorithm. It is a computational approach and amalgamation of two different fields of linguistics consisting of theoretical linguistics as one while the technical aspects of language also known as computational linguistics as the other. This is the very first research of its kind concerning any language like Magahi for the identification and classification of complex predicates at two labels wherein tokens or word labels are the one while sentence labels are the other. The research is broadly divided into two major parts wherein the first part covers the theoretical aspects of Complex Predicates or CPs in Magahi along with its characteristics, existence and formulating process whereas, the technical aspects and identification techniques comprise the other. For this research task, two different approaches have been followed wherein a total of about more than 5lac data tokens of Magahi has been collected and then annotated and verified both first by the human annotator who are mostly the local Magahi speakers and then by the machine itself. This was first crawled by using an Indian language crawler and then sanitized and cleaned by using an Indian language sanitizer. These are also termed as IL-Crawler and IL-Sanitizer. The annotation process also followed a two-way process wherein it was first annotated manually and then by the machine. This later process was done by creating a set of around 6k gold tokens, which were primarily not raw but the actual and correctly verified annotated data. These 6k gold tokens were then made to train a total of around 45k Magahi word tokens, which were collected mostly from the blogging websites comprising word classes of each category and domains such as entertainment, literature, health, short stories etc. the training and testing of the machine and the tool developed was performed and completed at a three-fold long process. In this three-way long process, every phase was trained with a total of around 2k gold tokens supplied with a total of around 15k raw tokens, which were annotated manually after the designed tagsets of their respective complex predicate labels also known as CP labels. The tool finally gave an output of 64.57% accuracy, which is a bit lower because most of the corpus contained much of the Hindi constructions. This is so because, Hindi and Magahi both belong to the same language family which is the most common Indo -Aryan language family. Due to this reason, most of the CP constructions of Magahi got overlapped with Hindi hence, affecting the final output of the tool.

The entire work has been divided into seven different chapters which covered all the major and necessary requirements of the same. In the last few chapters, a brief detailed

discussion of issues and challenges along with errors and ambiguities encountered in the tool has also been thoroughly discussed. Along with this a list of all possible and suggested solutions has also been given in the work. This research work in Magahi will undoubtedly serve as an inspiration for many future researchers who wishes to analyse and preserve any languages like Magahi, which are on the verge of extinction and is also very less recognised. It will help them in future to work on several other regional vernaculars and give them a significant standard recognition using computational and corpus linguistic methodologies.

Keywords: - Magahi Language, Complex Predicates (CPs), Corpus, Computational Linguistics (COLING), Natural Language Processing (NLP) Classification, Identification, Bihar, Jharkhand

Chapter-1 Introduction

1. Introduction

The research currently being conducted is an attempt to design a computational tool and model for automatically identifying complex predicates in Magahi. As the title of the work suggests, the main and sole focus of the entire work will be on the technical approach to Magahi complex predicates. The study will try to find out how a low-resourced language like Magahi will be able to cope with the new advances in technologies and software. It is an attempt to technically enrich all languages like Magahi in India, which are resource-poor and have not received much broad attention in society at a broader level, except outside of their own language community. The research currently being conducted is an attempt to bridge the link between computational software technologies to that of language engineers, which we call linguists. It also tries to establish a good link between the two different groups of language areas, namely general theoretical linguistics on the one hand and computational linguistics on the other by incorporating the latest modern techniques.

Complex predicates (hereafter CPs) are verb categories that use two different elements for their formation but together function as a single verbal unit, Rakesh and Kumar (2013). It consists of a few syntactically independent elements whose argument structures are brought together by an argument fusion mechanism, Butt (1995). This means that the conserved structure of a CP differs from that of all other types of general expressions. This is because all other expressions are not subjected to rigorous procedures to form a complete expression. Also known as special multi-word expressions, CPs follow a variety of structures that are combined into a single entity. Some of the common structures for forming CPs are compound verbs, i.e. (verb + verb), and conjunct verbs, i.e. (adjective + verb) or (noun + verb).

So, it is a collative form of speech that has two distinct groups of words, where the first element can be a noun, an adjective, a verb, etc., while the second will certainly be a verb, more precisely a light verb. For such constructions, a compound verb is a good example where the second element, V2, is grammaticalized. By grammaticalization, we mean that it is semantically bleached and is actually the bearer of the grammatical information for the entire predicate. As already mentioned, a CP can have many different combinations, such as Noun + Verb (NV), Adjective + Verb (AV) and Verb + Verb (VV); However, the resulting elements are always a VP. All of these combinations except (V+V) are called conjunctions, while the (V+V) or the verbal combinations are called explicators or reversed with V2 acting

as an explicator or light verb. This explicates the semantics of the entire noun-verb compound construction. The present work will attempt to consider all of such complex formations and structures in Magahi and will attempt to address the computational processes involved in their formations. It will also address the steps and methods by which all of these CPs can be identified and analysed computationally, involving several different processes of automatically identifying and analysing these complex linguistic expressions with the possible positive result of the tool at the end.

The entire research is divided into seven different chapters. The first chapter of the work deals with the introductory part of the work. It is further divided into nine distinct sections that will discuss the origin of Magahi as a language, the scope and goals of the research undertaken, the possible research questions, the hypothesis and the final process of explaining this work as to how it is different in comparison to any other existing research in Magahi and other Indian languages.

The second chapter of the study then looks at the existing work on the Magahi language and theoretical linguistics in general, which encompasses several Indic languages, most of which belong to the South Asian language family. These are then further expanded to include computer-based backgrounds, tools and technologies.

Chapter three of the work focuses on CPs in general and their specific and desired linguistic features. An attempt is made to give a clear picture of all types of CPs formed in Magahi. In the different sections and subsections, the different types of Magahi CPs such as Conjunct Verbs (CNV), Compound Verbs (CV), Reverse Compound Verbs (RCV), Noun-Verb Constructions (NV), Adjectives Verb Constructions (AV) are discussed in detail. Therefore, this chapter succinctly addresses only the Magahi CPs, their types and their respective formulation processes and properties, with possible suitable examples of each category.

The other three following chapters of the work, namely 4, 5 and 6, will thoroughly discuss the methodology used for the research carried out, the computational framework designed for the work and the evaluation and analysis of the developed system. All of these are discussed in these chapters with the help of the various sections and subsections, where not only the errors and disadvantages are discussed in detail, but also the specific measures which are suggested to minimize such errors of the tool.

The final chapter, seven, deals with the final aspects of the work. This includes the summary of the entire work, and its findings, along with the specific potential limitations and implications of the same. This chapter will also show how this research conducted will provide a clear insight into the technological advances towards all these lesser-known languages. Hence, this research is conducted with the idea and vision that in the near future it will pave a better path for all linguistical research undertaken of this sort.

1.1.Magahi: Language, People and Culture

India is a multilingual country of over 130 million people. It is a country where the language changes within every single mile and small distance. For a country that has such a large diverse language population, and where the language changes every single mile, it is quite unfortunate to note that the Constitution has recognized only a few languages by giving them high status and importance. Despite having different languages and cultures, the Indian Constitution only included twenty-eight languages in its eighth schedule. Apart from these, others have been overlooked and have not received much recognition and importance despite being rich in culture and traditions.

To proceed with the core idea of this work which is dealing with the identification of CPs in Magahi, at first it is very important to know Magahi as a language, culture and tradition. Unlike all other languages, Magahi is also a rule-driven language. This section of the chapter is about detailing Magahi as a language, its cultural developments over time along with the history and traditions that have existed in society since incredibly ancient times. Magahi, as a language, will follow a gradual process of exploration which is subjected to several detailed linguistic descriptions such as typology, morphology, syntax, etc. The coming sections and subsections of this chapter will gradually deal with the concepts of its writings also known as scripts, varieties of Magahi in society, the number of speakers, geographical distributions etc. So, one can say that this chapter is a detailed account of exploring Magahi not only as a language but also in terms of its vivid culture and traditions. Such elementary descriptions are very useful and crucial when conducting research on any language.

Later on, the author will move to the core idea of this work which is identifying the complex predicates (CPs), in 'Magahi', in the upcoming further chapters. Magahi as a language has its origin from the ancient word Magadhi and the word Magadhi has its origin from the term 'Magadha[1]', which means central. The ancient root of the word Magadha has

[1] Magadha- It was an ancient kingdom in southern Bihar. It was one of the great countries of ancient India.

its roots from the very ancient Vedic ages. During those ancient times, Magadha was immensely powerful among all the other sixteen different 'Mahajanpadas[2]'.

The Magadha Empire was immensely developed under the leadership of King Chandragupta Maurya and along with many of his successors such as Bindusara and Ashoka the Great. The empire extended its borders to the Himalayas in the north and also east to what we now call Assam. The province of Magadha in ancient times was not only rich in culture and traditions but also very well endowed with many profound centres of higher study. These include several major ancient universities in the regions of Nalanda[3], Takshila[4], Vikramshila[5] etc. Besides Magahi, there are also Maithili and Bhojpuri languages which are chiefly spoken mainly in Bihar. The group of all these three languages together form a cluster of Bihari languages. These languages are the subgroup of Indo-Aryan languages spoken in the eastern zones of Bihar (Grierson 1968). With the above facts, it is evident that it is a language belonging to the Eastern zone group of the Indo-Aryan language family.

If we analyse Magahi linguistically, it is a head-final language that follows sometimes a proper subject, object, verb (SOV) order and is sometimes rigid. The only difference between Magahi and other Indic languages is that it has features where the subject is sometimes both marked and unmarked. This peculiarity of Magahi makes it typologically very interesting. Magahi does not have too many cases like the Hindi language to denote its nominal features, instead, it has only three, namely direct, oblique and vocative.

As compared to other Indic languages, it also does not have any ergative case markers. Sentences in this language are formed by putting indirect objects in front of direct objects. Magahi often uses two verbs during sentence formation. While using these verbs, the main verb does not exhibit any form of changes while the auxiliary shows all possible inflections of the sentence such as person, number, gender and tense, aspect or mood etc. Hence, it is understood that in Magahi it is the auxiliary verb which carries all possible inflections of the sentence except gender. Magahi, allows a very fixed pattern of movement of its word components for sentence formation. It uses postpositions to identify case markers.

[2]Mahajanapadas - sixteen kingdoms or oligarchic republics that existed in Northern ancient India from the sixth to fourth centuries BCE during the second urbanisation period.

[3] Nalanda- An ancient Mahavihara, a revered Buddhist monastery which also served as a renowned centre of learning, in the ancient kingdom of Magadha.

[4]Takshila- It is located at the pivotal junction of the Indian subcontinent and Central Asia, on the eastern coast of the Indus River.

[5]Vikramshila- It was one of the two most important centres of learning in India during the Pala Empire, along with Nalanda.

For example, it uses /-ke/ for cases like dative, genitive, and accusative, /se/ for instrumental and ablative, whereas /-lɑ/ for benefactive, and /meⁿ/, /pərə/ for locative.

Before proceeding with the work further, this section has given a small account of Magahi including small basic details including the case marking properties, word order patterns, language family affiliations and a little bit about its ancient historical origins and backgrounds etc.

In addition, this work also portrays few major researches of South-Asian languages carried out by linguists like Aryani (1965) and Grierson (1927). This not only includes languages such as Magahi but also some other major languages of Bihar like Maithili and Bhojpuri along with a few major Indic languages. This dissertation will also try to give a little insight to show how closely Magahi is related to some of the ancient Indic languages like Pali and Prakrit. This work is an attempt to study Magahi in terms of its culture, traditions, historical and geographical backgrounds, along with a few detailed theoretical and computational linguistic aspects.

1.1.1. Geographical Distribution

We all know that Bihar is the 13th largest state in India and is located in the eastern region of the country. On its eastern side lies the state of West Bengal while on the west is the state of Uttar Pradesh. To the north are the regions of Nepal while to the south are the regions of Jharkhand. With this short geographical demography, one can see how closely the various geographic borders surround the borders of Bihar. It is due to this reason that Magahi also has its effect and influence over these geographical regions as well.

Apart from the above facts, Bihar also has two parts with unequal distribution namely North and South. The famous holy river Ganges divides these two halves equally. This sacred river flows between these two halves of Bihar from west to east. There are also various rivers in Bihar, on the banks of which lie several different cities of Bihar. These are known as the tributaries of the famous river Ganges. Few of them are Phalgu, Gandak, Damodar, Son etc.

Bihar has a latitude distance in a range of about 24-20-10 N, 27-31-15 N to the north while it has a longitude distance of about 82-19-50 E-88-17-40 to the south. Due to the wide coverage of longitude and latitude, the languages of Bihar have spread their wings to a wider area, allowing one to see its impact on the other states as well.

Fig.1. Political map of Bihar

1.1.2. Speakers of Magahi

Magahi has never gained much prominence in terms of language and research except in some parts and regions of Bihar. Due to this, it is impossible to provide the correct and accurate numbers for its speakers. Another important reason behind this is until now, no suitable statistics are available from any of the existing sources that can identify the correct figures. This great lack of data on Magahi speakers is due to two main reasons firstly, most Magahi speakers have shifted their mother tongue to Hindi because they consider it as a matter of their own prestige or else to make their work easier, faster and understandable to the masses and society. The other reason behind this is that many of the Magahi speakers, who lived primarily in and around the Bihar and Jharkhand regions, believed it to be a variety of Hindi and not a separate language. These factors have primarily hampered the further growth of Magahi, despite being much older. Both of the above reasons have hampered not only the further growth of Magahi but also its further expansion outside Bihar. This was not the case with other Bihari languages such as Maithili and Bhojpuri. Because of this, Magahi

as a language in Bihar has lost its identity over time, while the latter two languages have stood out strongly and successfully gained wide popularity. It has also always been believed that Magahi is nothing more than a dialect or local variety of Hindi. But the fact of the matter is that Magahi itself is an independent language having its own morphology and scripts. But adhering to the core concern of this research which is the automatic identification of CPs in Magahi, areas such as scripts, the number of speakers and other sections will not be of great importance. However, these are tackled a bit through possible graphic representations wherever required. Since determining the exact number of Magahi speakers is a trivial task, hence roughly approximate data have been used herein just for the purposes of illustration.

Magahi is spoken as a native language in an area of approximately fourteen thousand square miles in Bihar. It is the area that formed the core areas of the Magadha region in ancient times. These include the major areas and districts like Gaya, Aurangabad, Jehanabad, Patna, Palamu, Munger etc. These are the areas or regions that use Magahi as their first language. Based on all the above facts and data, one can deduce that the total number of Magahi speakers in Bihar would be around 2.5 crores. According to the 1881 census report, the total number of Magahi speakers was around 6,504,817. Later, in 1961, the same census recorded a huge drop in this number to around 3,792,447, Grierson (1927). This huge decline in the number of Magahi speakers has suffered another major decline over time due to several reasons such as language prestige, the influence of Hindi on Magahi and several other sociolinguistic phenomena. There were also speakers who preferred Hindi instead of Magahi as their standard and decorated form of communication in many Magahi-speaking regions of Bihar. So, sticking to the above facts, it can be said that the number of Magahi speakers has not actually declined but has been overtaken by the standard societies that favoured Hindi as a mother tongue or otherwise the first language of communication in Bihar. These are the same Magahi-speaking people and regions who report themselves as Hindi speakers in the given census data. Hence, language change is another major reason that has led to this incredibly dwindling number of Magahi speakers in Bihar. However, even after this sharp decline in the number of Magahi speakers, there are around thirteen million native Magahi speakers present today in Bihar and its neighbouring regions, according to the 2002 census data. The regions of Bihar that speak Magahi as their local language have also been shown in detail in the given figure.

Fig.2. Magahi speaking regions of Bihar shown in green colour.

The accurate figures for Magahi speakers are still not known till date however, as per the report of census data 1951 it is as follows: -

Year	figures	place of returns
1901	18,147	Santhal Paragana - 12,393
1911	501	Manbhum - 5,373
1951	3,728	Santhal Pargana -2847 Palamou -835
1921		accurate figures not available
1931		figures not available as speakers got merged with Hindi figures in 1921 and Hindusthani figures in 1931.

Table 1 Magahi speakers as per census data 1951

O Malley, (1913; p. 382) was not very satisfied with the given census data as according to him there were several speakers who spoke many Bihari languages besides Magahi, such as Maithili and Bhojpuri. These speakers later switched towards Hindi for their own ease of access preferably at work. It was at this time that all of the regional languages of

Bihar including Magahi lost their importance and got mixed with the standard Hindi, establishing mass contacts with Hindi-speaking people. This was another important reason for Magahi to lose its original Kaithi script over time and got it mixed with Hindi's Devanagari script.

1.1.3. Dialects of Magahi

Like any other language of the country, Magahi has its own dialects and different forms which are not only spoken differently but also spoken in different regions of India and Bihar. These dialects are geographically distributed in different language regions. For example, in the language of the educated community in Bihar, the influence of standard Hindi can be easily seen if they speak Magahi. This influence can also be seen or monitored while using the vocabulary. This is because most of them didn't know much about Magahi for their daily use. Another important reason is that the schools and educational institutions of these Magahi-speaking regions have chosen to use Standard Hindi as the source language for delivering education. This is the reason why people today consider Magahi to be quite inferior compared to Hindi. This has led people to view Magahi as their local variant rather than a standalone or primary language of communication.

In popularity, there are only three main Magahi dialects. These are standard Magahi spoken in the Gaya, Patna and Hazaribagh regions. Eastern Magahi is spoken in the regions of Jharkhand such as Ranchi and Hazaribagh and near some border areas of Odisha such as Kharsawan, Mayurbhanj and Bamra. The third dialect type is influenced a little by Maithili. These are spoken in the eastern parts of Bihar such as Munger, Begusarai and Bhagalpur. With this influential evidence, we were able to see how Magahi changed shape and led to a new shape in the different regions of Bihar. This is because the main spoken language of this area influenced the original form of the Magahi and gave it a new shape and form.

Despite the fact as mentioned above, currently there are only three major varieties of Magahi that are being very popular and spoken in and around the major regions of Bihar. These are mainly the central Magahi which are spoken and can be found in and around the central regions of Bihar like Gaya and Patna. Apart from these the other two forms that can be found are south-eastern form of Magahi and eastern Magahi. One is being spoken in south-eastern regions of Jharkhand such as Ranchi and Orissa, while the other is being spoken in the regions of Begusarai and Munger.

1.1.4. Historical Development of Magahi

According to Tripathi (1993), there are five different varieties of Magahi. These are standard Magahi, spoken in and around the regions of Gaya. Inferior or impure Magahi is easily found in the regions of North Bihar ranging from Bihar Sharif to Patna. Talaha Magahi is spoken near the regions of Mokama, Badh, Barahiya and in some regions of Lakhisarai and Gidhaur. Apart from these two, there are also several types of Magahi like Sontatiya Magahi spoken in and around the Son River regions of Bihar like Aurangabad. The form of Magahi spoken near this region was heavily influenced by Bhojpuri thereby influencing its true form. This is because the Aurangabad and Son River areas are mainly a Bhojpuri language belt. The other varieties like Jangali Magahi or the forest-like form of this language are spoken in the forest areas of Bihar like Rajgriha and Gaya.

As mentioned, Magahi is related to the ancient province of Magadha, meaning the central region. Due to the continuous development or change of sounds, ancient Magadhi has become Magahi today. The source of the emergence of this language can be traced back to the earlier ancient periods, which are around 1200 AD to 1400 AD. It is so because during this given period of dissemination there were said to exist some Siddhas who were the Saints in the Magadha region who followed the Vrjayana school. This school is said to have practised the impure form of Buddhism. Because of this, people find popular Siddha literature in some important parts of Bihar like Nalanda, Vikramshila, Udantapuri etc. This is because these were the important centres of the Siddhas of that time. All of this literature was written in the form of Kaithi, the oldest script of Magahi. Because of this, Siddhas and their pre-existing literature are considered to be one of the oldest forms of Magahi. However, in the absence of suitable records and documents, it is quite impossible to trace when and how the current form of Magahi evolved into common practice.

All modern Indo-Aryan languages such as Gujrati, Marathi, Bengali, Maithili etc. are believed to have their distinct origin and existence somewhere around the 14th century. (Chatterjee; 1926). Magahi is said to have originated only during this time. As per the above fact, languages like Bengali, Oriya, Maithili, Bhojpuri and Magahi are the sister languages. Of these, Magahi and Maithili belong to the same community and therefore have sister relationships, but Bhojpuri and the Eastern Magadhan languages of Bengal, Assam and Orissa are their cousins. (ibid.).

However, from the above facts, one can conclude that Magahi is one of the languages which is much closer to Maithili than Bhojpuri. Despite this, it has been overlooked for a long time. The development of Magahi has also suffered somewhat, as it has a direct genetic influence and affiliation with Hindi. This is why today Magahi is still considered a dialect by most of its speakers. Therefore, all these factors and parameters have not resulted in Magahi developing as a distinct form, leaving it on the brink of extinction.

1.1.5. The Script

Of many other ancient Indic languages, there are many other representatives of the same modern Indo-Aryan language families which got its origin from the popular Magadhan subfamily, Comrie (2001). A few modern representatives are languages like Maithili, Bhojpuri and Magahi etc.

Magahi has one of the oldest and most traditional scripts of its own, known as Kaithi. In ancient times Magahi can be written in four major different scripts such as Devanagari, Kaithi, Odia and of course Bangla, Verma (2007). Bangla was used as the script of Magahi to write the eastern forms of Magahi (ibid.). However, over time, Magahi has not gained much prominence and therefore has not received its own script. Therefore, Kaithi as a Magahi script has lost its importance in modern times which was used earlier primarily for writing purposes.

Kaithi is the script written in italics or cursive. As the name suggests, the characters in this script have their existence from Kayath or the Kayastha caste. It was the community that came with the task of writing. The Kaithi script was gradually developed from the ancient Brahmi script. According to Grierson (1926), the writing of Kaithi was particularly important and prominent for all Bihari languages. It was very commonly used in conjunction with the popular Nagri script, which was the main script of the region of Upper India at the time. Therefore, since the writings of Nagri in the upper parts of India and Kaithi in the main regions of Bihar developed simultaneously, Kaithi can be seen as the distorted form of Nagri which has not gained much importance since.

Kaithi, as a script of Magahi, can still be seen in some parts of Bihar and eastern Uttar Pradesh. In ancient times, Kaithi was not only the script of Magahi but also of many other languages of Bihar and Uttar Pradesh such as Awadhi and Maithili, Pandey (2007). Although Kaithi is one of the most important and ancient scriptures of languages belonging to Bihar, Bengal and Orrisa, it does not deny the fact that it has been overlooked for ages and has lost

its importance among the common masses. Due to this, Devanagari arose to prominence in the early 20th century and gained much social and political prominence. This usage of Devanagari not only neglected Kaithi as the original script of Magahi, but also created certain major problems for the proper appearance of Magahi as a mainstream language. It is also due to the frequent use of the Devanagari script in Magahi that has made it impossible to recognize certain conventions of sounds or letters of Magahi. This is the reason why today Magahi sometimes cannot be understood or written without the use of Devanagari.

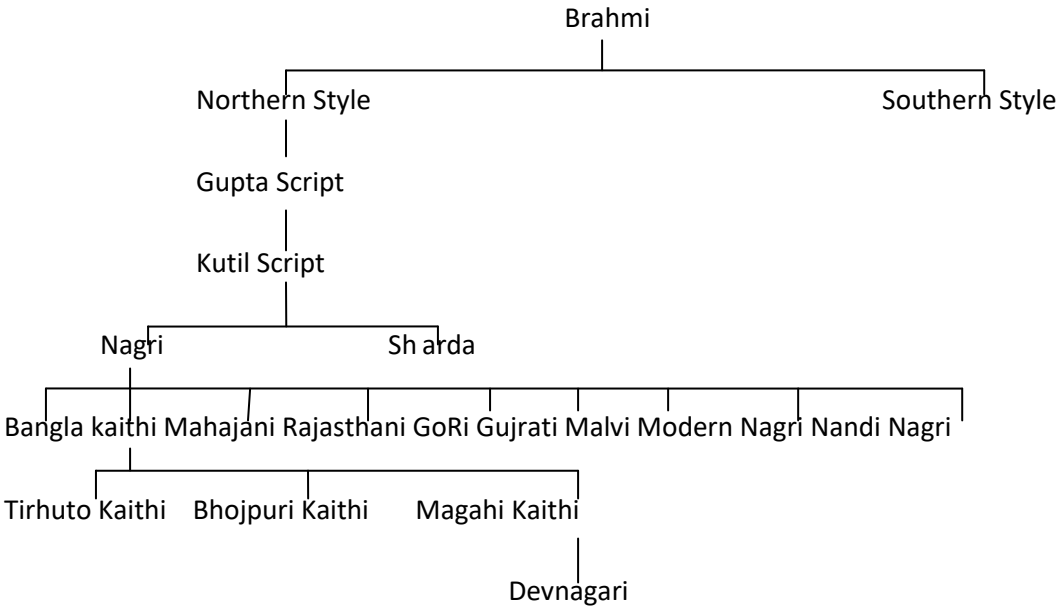

Fig.3. Origin of Magahi script Kaithi.

1.2. Status and Socio-Political Situation of Magahi in India

The status of Magahi in India can be easily represented and decided by its native speakers, who consider it an inferior language themselves. Because of this, more than half of the Magahi-speaking population now adopted Hindi as their local language before Magahi. A wide range of Magahi-speaking populations are now shifting to using Hindi for their area of operation as they see it as a matter of prestige. Not only that, but politically it has long been overlooked despite being one of the oldest and most ancient forms of language in India. For this reason, even today, after decades, it is an unscheduled language. Due to the lack of political strength, many Magahi speakers from Bihar regions are switching to Hindi and adopting the same as their dominant language for their daily needs and educational purposes.

People living in India speak different languages and all these languages belong to a specific language family. Of all the different language families according to general linguistics, the languages spoken in India broadly belong to four different language families such as Indo-Aryan, Tibeto-Burman, Austro-Asiatic and Dravidian. There are also some languages that belong to both Indo-Aryan and Indo-European language families. These classifications of languages into their respective language families are based on their respective historical development, Grierson (1927). Because of this classification process, Magahi was part of the eastern group of languages to what people also call Prachaya, Chatterji (1926). Chatterji and Grierson's detailed classification can be seen in the respective figures such as fig 4 and fig 5 mentioned below.

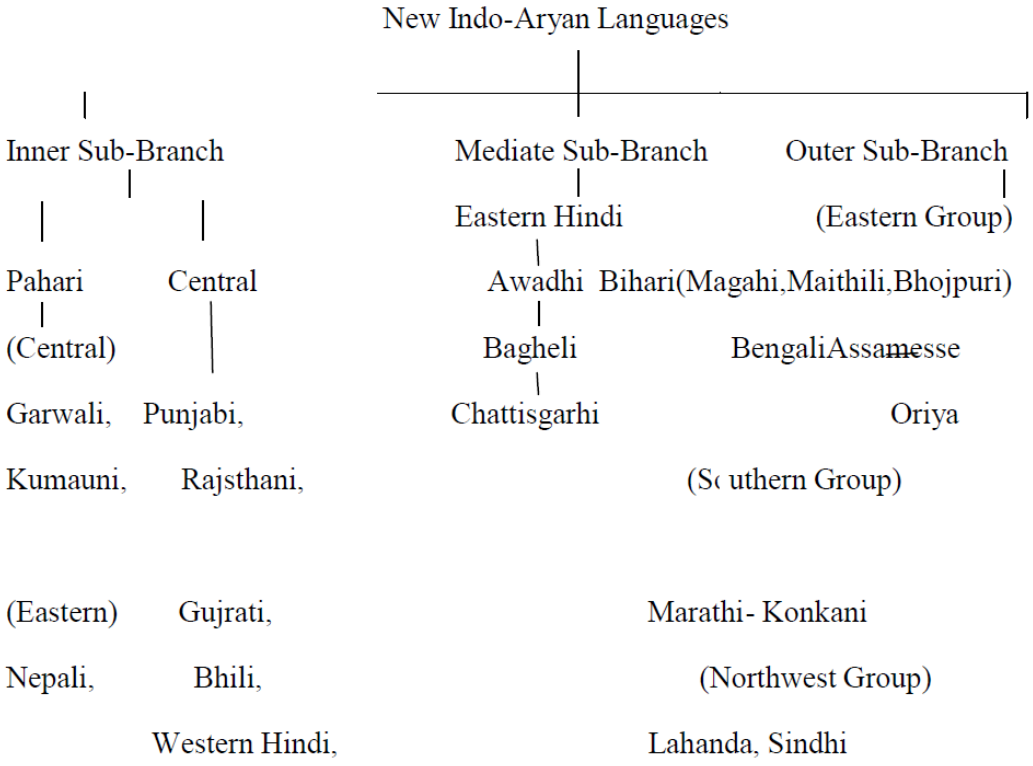

Fig 4. Classification of New Indo-Aryan languages by Grierson

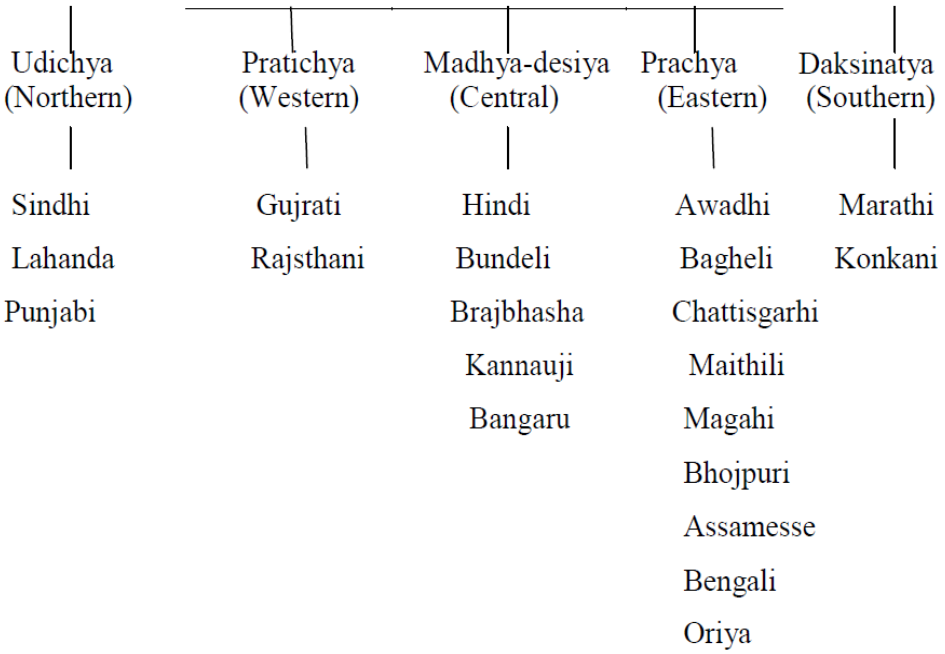

Fig 5. Classification of New Indo-Aryan languages by Chatterji

As for the status of Magahi in India, not only is it considered inferior in terms of usage, but also considered unacceptable by the people as it detracts from their prestige (ibid.). Because of this, other Bihari languages such as Maithili gained popularity at the same time, believed to be the languages of educated peoples and higher-class sects such as Brahmins. From the fact mentioned here, due to all these facts, Magahi is said not only to have been racially inferior in comparison to other languages of the Magadha region but also to have been considered the language of lower-class people with minimal or no education during the Vedic Age.

This prejudice about Magahi was so harsh that it still lingers today in modern times, and therefore the people who speak Magahi do not receive much attention as compared to those who speak a language other than Magahi, which is easily reflected in our everyday normal and social life.

1.3.Need for a CP Identifier Tool

It is a known truth that in today's scenario almost half of the population in India prefers only two prominent languages as a source of communication, namely Hindi and English. Of these two, English is primary while Hindi is secondary. Both of these are widely spoken by the masses. Because of this, many other languages like Magahi, Maithili, Bhojpuri, etc. have

been overlooked. In order to draw the sensitive attention of today's population to all these overlooked languages, we have taken on this task as one of our main research concerns.

We believe that with modern technological tools and facilities, this task of saving such languages and promoting them as an important platform for their recognition can be easily accomplished. Not only that, but it may also pave the way for future researchers who can help save all languages as endangered as Magahi, which are on the verge of extinction. The development of the CP identifier tool will not only help in the grammatical learning of the Magahi language but also give full throttle to the task of saving it from extinction. This is another important reason why developing such tools or corpora is important.

1.4. Scope of the Research

Magahi is a lesser-known language for Bihar and the whole country. Because of this, very little research has been done on the language. As a non-official language, the government has received almost no attention to any kind of language development concerning Magahi. The present work is the first of its kind to develop NLP applications in Magahi using computational methods. The proposed research primarily examines the description of the form and function of the complex predicates in Magahi. The work also helps in improving the language situation by providing a recognizable platform through NLP by creating tools and online corpora. This work is a small attempt to develop a suitable Magahi CP identification tool that analyses the structures of CPs and also identifies their challenges along with possible errors while computing the data. We are sure that once this work is available online along with the extensive corpora, many other works will follow in Magahi after it.

This dissertation presents the entire research with a single aim, which is to show that the structures of complex predicates are not only unique but also how widespread they are, especially in South Asian languages including Magahi. All the research has been done keeping only this aspect in mind. To study this language in detail with the help of technology, one can follow the upcoming chapters of this dissertation that will explain everything in detail.

1.5.Objectives of the Research

The objectives of this research conducted are to create a bridge between the two prestigious fields of modern times, one being modern software and technology, while the other being the language engineers to what we call linguists. It is also intended to examine and study the nature and structure of the CPs within the given framework or established

principles and parameters. It also seeks to show how these CPs of Magahi are not only polysemic in nature but how they play different semantic roles in the language in question. This work was hence attempted most importantly, to develop an appropriate identification and classification module for Magahi's CPs, which will allow Magahi to access new technological advances as required by today's time and development.

Apart from that, it will also address the issues and challenges that the author faces during the process of developing this tool. It is important to discuss these questions as this will help to see if the tool developed works well in the process of identifying the CPs, which is of course one of the main aims of this research. It is important to carry out such tasks for resource-poor languages like Magahi because in this era of digital India it is important to bring all these resource-poor languages into the main domain of society with the help of technology.

1.6.Hypothesis and Research Questions

The entire dissertation is based on one hypothesis, namely: Complex predicates are one of the most crucial and important syntactical features of the South Asian language family, of which Magahi is one. Through this research, the author has also tried to construct an idea that the Magahi complex predicate agreement system works in the same way as in the case of simple verb forms of this language. This means that the verbal arguments match the subject in Magahi, but at the same time, if the subject has no argument characteristics in that language, then it targets the next higher argument to show its sustainability. This will be clearly seen in the later coming chapters of the work when the author will discuss the noun-verb (N+V) constructions where the noun of the sentence agrees with the light verb. This is mainly seen in cases where the arguments of the sentences are case-sensitive, except for the CPs. At the same time, it also notes the fact that there are many similarities between Hindi CP constructions with those of Magahi as both are similar and identical. Because of this, most of Hindi CP's constructions were mixed with those of Magahi, which can be clearly seen in the collected corpus, which of course affected the final output of the developed tool in terms of its accuracy and ambiguity.

Through this research, we have also tried to construct an idea that the Magahi complex predicate agreement system works in the same way as in the case of simple verb forms of this language. This means that the verbal arguments match the subject in Magahi, but at the same time, if the subject has no argument characteristics, then it targets the next higher argument to

show its sustainability. This will be clearly seen in the later coming chapters of the work when the author will discuss the noun-verb (N+V) constructions where the noun of the sentence agrees with the light verb. This is mainly seen in cases where the arguments of the sentences are case-sensitive, except for the CPs. At the same time, it also notes the fact that there are many similarities between Hindi CP constructions with those of Magahi as both are similar and identical. Because of this, most the Hindi CP constructions were mixed with those of Magahi, which can be clearly seen in the collected corpus, which of course affected the final output percentage of the developed tool in terms of its accuracy and ambiguity.

1.7.Rationale of the Study

As mentioned earlier, Magahi has been a culturally rich language since ancient times, but despite this well-known fact, it has not attained great prominence and has been neglected since the pre-independence period. Although for Magahi, there is an enormous body of vast folk literature and literary traditions, however, no popular research employing modern technology has been reported. As a member of the Indo-Aryan family, Magahi follows the SOV order of sentence patterns, which is like all other languages falling in the same language group. However, sometimes it also follows the SVO or OVS order of sentence formation, which makes it slightly different. This implies that it does not have the property of rigidity and can be framed into any of the word order forms or sentence structures without changing the meaning of the context, but attaining the same is also possible wherever necessary and required.

In other words, as a member of the Indo-Aryan language family, Magahi also sometimes allows for the process of scrambling. The sole reason for conducting this research is to enrich this language with fewer resources to a broader recognizable platform in terms of technology. The research conducted aims not only to provide a tool to identify CPs in Magahi but also to help future researchers to better advance their investigations in this regard. Another important aspect of this research is to make the Magahi language and its resources available online using the technologies to help the masses not only to cater for their research needs in order to quench their thirst for general linguistic identifications rather, to give extremely wide access to this very resource-poor language in the area of NLP and COLING.

1.8. How is the present research different from existing ones?

Regarding question of research in Magahi, there is some work in the field of sociolinguistics, morphology and computational. But the proposed idea of the research area mentioned is the

first of its kind in the Magahi field. As already said and mentioned, in this task the collected corpus is crawled with the help of IL Crawler, which is later sanitized with the help of a tool called IL Sanitizer. Once the data is crawled and cleaned, it is annotated using the semi-automated online tool called ILCIANN, which is an Annotation Tool given by ILCI, Kumar et al. (2012)) and then tested for annotation accuracy to select a suitable model and tool to be developed for the proposed area. The collected and further refined data is then trained and evaluated differently on two different bases such as manual and machine training. Once the data is trained, it is scored to indicate the level or percentage of accuracy. The same step is repeated twice or thrice so that the data collected can be refined a little further and the tool created could improve its accuracy with further possible improved training and corrections made at different stages. The errors found throughout this process are improved and later injected into the system tool for better results or accuracy. Finally, the error that may have occurred is discussed together with the problems of ambiguity.

From more than 5lac Magahi corpora taken from various fields such as literature, entertainment, health, etc., a total of 45,000-word signs and sentences were taken for this research. Each token and sentence were first manually tagged and verified with the native speakers to eliminate any discrepancies. In order to estimate the accuracy of the data, both visible and invisible data are sent to the system, based on which the system evaluates and provides the best possible result. However, accuracy is judged on the output data collected from the same training data and passed to the system. Hence, one can see here that there are many more technical processes applied to a less recognized language, which of course distinguishes this research from any previous work by Magahi.

Chapter- 2 Literature Review

The concerned chapter will discuss the existing research work in the concerned area. It will work as the precursor, which helps in developing a sound piece of research work for Magahi. The chapter gives a brief account of work done in the field of complex predicates and Magahi as a whole. Through this one can have a broader overview of the mentioned area and the specific title to what the author has chosen to work on. It will also help in gaining a good piece of knowledge about the CPs and its different works cross-linguistically. This means that it will examine the different works of CPs in Magahi for sure but also will not limit to the mentioned language only. Therefore, it will give a detail insight upon the CPs of different languages also.

It will throw an insight upon the pre-existing researches under taken in the concerned field. Along with this, it will also discuss the previous and the current trends of various researches in Indian languages as well as all Bihari languages including Magahi, Bhojpuri and Maithili as well. It will discuss different natural language processing (NLP) works, that has been taken forward so far in terms of what so ever Indian languages like Hindi, Urdu, Marathi, Bengali, Maithili, Telegu, Tamil etc. not only this it will also discuss the theoretical linguistics works done in all these languages. All these will be discussed in detail under concerned sections and sub-sections, of this chapter into which it has been further divided. This is the only reason due to which this chapter of the presented work becomes relevant as it plays a vital role in bringing on to the different works of different researchers under one roof. One can say that it is the heart and soul of the entire research.

2.1. Complex Predicates – An Overview

Languages is that special human ability which enables us different in the entire universe as compared to all other species. Linguistics then later emerged as a field of science to investigate as to how humans became capable of understanding and speaking a language. Not only this but the understanding of this even gets better with time and ages of humans as the more they grow, the more they become intelligent in this regard. It is still a matter of research as to how at a very young age any human being became capable of not only speaking a complex sentence, indeed framing them quite beautifully too.

Knowing a language consists of several different things and constructing complex sentences with the help of predicates is one of them. It is a general understanding that any normal person could do this by following a set of rules that are required to frame such

sentences. But at the time its advancements and spread to several other local dialects is still a matter of research. It is a general understanding that many Eastern Indo-Aryan languages do share the same rules for its formation but apart from this there are many that exhibits certain differences. Goldberg (2003) is of the opinion that the construction of a CP begins in a human mind at a very early stage. Complex predicates are not only complex but also heavily loaded constructions as it incorporates several different linguistic items into one unit. It is due to this reason that any CP construction across languages will comprise of several linguistic properties such as lexical, structural and functional.

Complex Predicates constructions are abundantly found in many South-Asian languages. In many languages it is known with several different names such as Compositional Predicates, Clause Constructions, Idiomatic and Phrasal Constructions, Multi-Word Expressions (MWEs), Participles, Co-Verbs etc. for all such constructions linguists found a common umbrella term which is Complex Predicates. All such constructions in linguistics are now been referred with this name. In case of foreign languages like German and Dutch the process of clustering of main verbs with modal or auxiliary verbs are known as Complex Predicates, Bouma and Noord (1999; pp; 43-72).

2.2. Brief Descriptive Sketch of Important Linguistic works in Indian and Bihari Languages

In many Indian languages like Hindi, Urdu and Bangla the constructions comprising verb-verb, noun-verb and adjective-verb is termed as Complex Predicates. Complex predicates in general are morphological as well as syntactico-semantic phenomenon. The procedure of its formation is quite identical and similar to almost all South-Asian languages. The idea of formation of CP sets back to ancient times and marks its presence to Vedic Sanskrit period. These formations were also indigenous to almost all South-Asian languages which are found in some way or the other at different stages, Gambhir (1993). At that time, it was only limited to Vedic Sanskrit but now with the onset of time it has spread its wings to other languages as well. Such formations gained its momentum at much later stages as now the process of their formations is quite simplified and easier to understand. It is due to this reason that now such constructions have now gained momentum to several Modern Indo-Aryan languages such as Bhojpuri, Magahi, Tamil, Telegu, Marathi etc. apart from this there are also languages which does not have any CP types of their own then, such languages borrow it from their parent languages and has fulfilled its requirement.

In many languages CPs are termed as expressive as it is the core element that is very much required to express the meaning of an entire word or sentence, Abbi (1980). The excessive use of CPs across languages especially towards Indo-Aryans can be termed as a shift of language structure from inflectional to analytical (ibid). The CP formations are found quite easily in languages like Hindi and Bangla. Later it was further spread to several other Modern Indo Aryan languages such as Magahi, Maithili, Bhojpuri etc. This not only reduced the culture of borrowing words from one language to another rather, have made the communications among languages and people easier.

It is no denying the fact that the idea of analysing a CP is quite challenging today. It is due to the reason that while analysing the parameters of both syntax and semantics are being thoroughly examined before its final acceptance. The idea of identifying case markers in CPs was first discovered in English language Cattell (1969), Jackendoff (1972), Higgins (1974) Oehrle (1976). The revised theory of CP analysis required for thorough comparison between English and Malayalam was given by Jayaseelan (1988). In order to make the CPs more understandable and easier a new set of special rules for comparison and better interpretation of were proposed which is popularly known today as Complex Predicate Rules or CPR Jackendoff (1974). It was a rule that merges the host or the main verb with direct object which is most possibly a noun thus resulting into CP formation. With this CPR rule it tends to construct a new category of CP which eliminates the position of direct object from the sentence and makes it a host verb. This means that it will combine the direct object sequence along with the main verb in order to establish proper semantical relation between the object and verb. This entire phenomenon is termed as overlapping of semantic function of the verb with the desired direct object Jayaseelan (1988). As per him any CP formation is only possible through repetition of lexical rules. These lexical rules come in action only when they are attached to any nouns, adjectives or else any other linguistic elements.

There also exists several other important works in CPs in many other Indian languages like in Hindi, Verma and Mohanan (1993). It was also observed in Telugu Krishnamurti (1993). Later, it was Masaica (1993), who tend to discuss and elaborate the aerial features of CPs by specially giving attention to conjunct verbs. Verma (1993) tends to focus on identifying problems of CPs specially in Magahi and Hindi during agreements and case markings. After thoroughly examining these issues in the languages mentioned above, he suggested to attempt a valuable solution of amalgamation of any lexical or syntactic entity

which will help in giving a successful nominal verb or NV, which is proper incorporation of a noun with verb.

Much later than this it was Mohanan (1994) who proposed the issue of argument system in CPs of Hindi whereas it was Davison (1991) who was of the opinion that any CP of language can easily percolate with any immediate linguistic entity which is most possibly a verb, adjective or noun in order to form a CP. As per Mohanan in a CP the existence of removal of agreement is directly dependent upon the orderly relation between transitivity feature of the verb and the concerned noun which act as host in the entire sentence. This implies that if the host of the sentence is transitive in nature, then the logical object will act as grammatical object but at the same time if the host is intransitive in nature, then it is not possible for the logical object to agree with the verb in the sentence and therefore, it needs an extra element for its agreement to what he calls as light verb construction.

So far as the question of doing linguistic researches concerning CP in Indo-Aryan language is concerned, there exists too many prominent works in this accord. There are prominent researches done by Hook (1991) in Sanskrit and Hindi. It was also done in Bangla by Dasgupta (1989). He focused on the experiential construction of CPs in Bangla wherein he opined that the experiencer or the subject in Bangla always takes the genitive marker. Some major researches were primarily conducted by Abbi (1990, 1991). In Marathi it was Pandharipande (1990), who laid her full throttle for CPs.

Out of all these Abbi's (1990) work was basically concerning typological approaches in Indian datives. Besides all these there also exists several other researches in Dravidian languages such as in Malayalam by Jayseelan (2004), in Telugu by Balusu (2012). He executed this work by applying lexical semantic approach in order to identify the meaning and the argument structures of CPs primarily in Telugu. Apart from all these there also exists several impressive works done in this accord concerning many Tibeto Burman like Manipuri, Sinhala or Munda languages. All these prime works have been jointly assembled together which later published in the form of a book in the year 1990 by Stanford University. The title of the book reads as 'Experiential Subjects in South-Asian Languages by Manindra Verma and K. P. Mohanan.

The formation of CPs construction in many South-Asian as well as European languages were also termed as Psyche Verbs during earlier times as it contains the deepest meaning for the entire sentence or phrase. For CPs in Italian, one may refer to the interesting

works of Belletti and Rizzi (1988). They primarily focused on the positioning and assigning of such verbs in Italian language in their work.

Apart from these there also exist several other important works in the field of CP in linguistics. These includes Standard Incorporation Theory (STINT henceforth,) by Mark Baker (1985, 1988). It is a theory that focused on changes of grammatical functions of the linguistic elements as to how any linguistic element changes its form and function after being grammaticalized during CP formation. The Determinant Phrase or DP hypothesis proposed by Abney (1987) and Functional Phrase or FP hypothesis proposed by Leffel (1988) states that during a CP formation a Noun Phrase or a NP is always essential.

This dissertation finds out a common hypothesis after commuting the task that study of CP and its process of formation is one of the leading syntactic features in almost all South-Asian languages including Magahi. Most importantly the work done in CPs of Dravidian and Indo-Aryan languages have gained much importance than any other as there are number of issues and challenges encountered during their formation. In most Indo-Aryan languages like Bengali, Oriya and Assamese the main verb (MV) exists in participle form or root form while forming a Compound. This is due to the reason as the light verb bears all the inflections and hence gets semantically bleached. Kumar (2006) proposed the idea of negativization in CP formation. He also proposed through his research that its formation of negatives in CP prohibits the formation of compound verbs in Hindi. Davison (2004) proposed that no languages in the world can form similar and identical CPs by using same set of nouns and verbs. It is natural to have some alterations or changes in their formulating structures. There always exists a constraint for the languages to select an appropriate predicate for their formation. Ghosh (2008) in his work shows that while forming a conjunct verb in CP the agreement system of an adjective in a conjunct verb depends directly on the feature and nature of the immediate verb. Once these properties are available then only one can match and adjoin a verb constituent with another linguistic particle in order to form a predicate or else an adjectival conjunct.

In order to study thoroughly the idea of CPs in many Indo-Aryan languages such as Bengali, Hindi, Marathi, Oriya etc. one needs to study various aspects of its formations as well such as synchronic and Diachronic study, syntactic and semantic rules and moving a bit forward to technology it must also follow and study the ideas of machine learning. So far as the concept of studying CPs in Magahi is concerned a little bit of historical work has been

done in this accord by some prominent linguists like Grierson (1968) and Aryani (1965). Tripathi (1993) provides a detailed account of relative study between Hindi and Magahi morphological aspect. Verma (1985) gives a detailed idea of Magahi verb structure and also deals briefly with the idea of CP formation in the language concerned. Indradev (2007) in his work provides a detailed account for origin and development of CPs, that starts right from the ancient period of using it as Apabhramsa.[6] However, it is evident to note that the notion of morpho-syntactic and syntactic properties such as forming negatives in CP, phrasal integrity and the agreement structures of Magahi CP has yet not been discussed till date. Also, the idea of Polysemy[7] of Magahi light verbs have also not been discussed yet by any linguists so far.

In Magahi there are no any literary works done till date broadly either theoretical or empirical in the field of CP that can focus on its relevant properties. Therefore, keeping that in mind, this research work will try to emerge as a torchbearer for all the mentioned aspects in near future. It is one of the important reasons as to why this topic has been chosen and hence, the objectives of this dissertation have been defined.

Lohar (2020), has written a book concerning a detailed account for the study of Bhojpuri grammar. Mundotiya et al. (2021) have proposed their work for creating possible linguistic resources in major languages of Bihar namely Bhojpuri, Magahi and Maithili. Verma (1991) tried to show the nature and properties of agreement of Magahi verbs by exploring possible parameters of the language concerned. Alok (2021) put forward his research concerning the morphosyntactic properties of addressee agreement in Magahi. Singh et al. (2014) have tried to identify the nature of politeness in major languages of Bihar with special reference to Bhojpuri, Magahi and Maithili. Atreya (2016) gave a descriptive account on four various linguistic aspects of syntactic properties in Magahi. Atreya et al. (2014) focused on the idea as to how the formation of passives are possible in Magahi. Sicky and Behera (2022) have focused on the key achievements of the descriptive research of explicators and reverse compound verbs in Magahi, by closely examining their formation processes and highlighting several significant differences between the two.

Atreya et al. (2013) proposed their work concerning the idea of formation of lexical anaphors in Magahi through the treatment of government binding (GB) theory of grammar.[8] Atreya et al. (2014) through their detailed work tried to concentrate as to how the relative

[6] It is a Sanskrit word which means corrupt or un-grammatic for of a language.
[7] One that has more than one meaning.
[8] theory of syntax for phrase structure grammars which follows the central idea of transformations.

clauses are being framed in Magahi. Atreya and Sinha (2020) have also proposed a work in the field of Magahi to show as to how the term ‘waa’ functions phonologically as diminutive marker. This means that the work is done to show that how a non-meaningful word ‘waa’ of Magahi is attached with several new words in order to frame a meaningful term. Rakesh and Kumar (2013) and Atreya et al. (2014) have worked on different aspects of Magahi that paid attention not only towards Magahi as a language, but also to its grammatical patterns and constructions. Butt (1995) has described the form and function of CPs in Urdu. Goyal and Sinha (2009) have also induced their best efforts in detecting CPs in a parallel corpus of Hindi and English by applying simple methods. Das (2009; 2017; 2016; 2015; 2013; 2012; 2011; 2019) has done an outstanding work on CPs by paying attention towards the computational aspects of the verbal string generation and have suggested for a better model of morphological concatenation process. This suggestion could be taken positive for near future as this can help a lot if it is applied carefully in the area of COLING researches.

It is due to the above-mentioned researchers that the classifications of CPs in different grammatical categories like verb-verb (VV), Adjective-Verb (AV), Adverb-Verb (Adv-V), Noun-Verb (NV) etc. are easily done today. The grammatical classifications done above are termed as the linguistic tagsets which will be used by the author for tagging the corpus. Grierson (1927) proposed the term ‘’Bihari for languages spoken Bihar. This comprised any single language or group of languages such as Magahi, Maithili and Bhojpuri.

Muller (2009) worked on imposing limits of complex predicates in Persian language and focused solely upon its inheritance properties. Chandan (2016) has worked on the various forms and representation of the particle ‘wa’ concerning its different linguistic and sociolinguistic implications in Magahi. Kidwai et al. (2015) have focused solely upon the agreement pattern of Hindi verbs along with its acquisition types.

Priya (2020) has put forward a much-detailed work of the morphophonological aspects and rules of Magahi. Extending her work a bit more towards the sociolinguistic aspects of Magahi she has also worked over the issue of identity, ideology and language shift for the Magahi immigrants. Kumar et al. (2021) have laid their full effort in order to show a detailed and descriptive study of Eastern Hindi, wherein the footprints of allowing Magahi, Maithili, Bhojpuri and Awadhi can easily be seen. Chandan (2018) have tried to portray the possible aspects of plurality in Magahi language with special reference to counting numbers or linguistically we can say that by taking massive nominal subjects. Corbett (2000) tends to

show as to how Magahi as a language can make a three-way number distinction grammatically. This work of Corbett was based mainly on marking the nouns along with its possible numbers and the changes that it depicts in the same for Magahi.

Jespersen (1924:58) through the idea of philosophical representation of word class or order points out that a plenty of existing words definitions are not as precise as it is required to denote its meaning. While defining the meaning of each word, he was the one who found several ambiguities in them, which is due to the absence of proper arrangements which are not uniform and is neither done based on its real class divisions. So basically, his work was to focus the problems of language system with division of word class.

Nilu (2016) through her research was able to easily show the phenomenon of CP formation in Magahi. Though it was a purely theoretical work but she was successfully able to prove that CP formations are very much possible in Magahi unlike other languages but with an exception of change or applying certain variations in its forms and features of verb forms and words.

Therefore, it much than expected have been done in terms of CPs in almost all natural languages, but when it comes to the low-resourced languages like Magahi, one cannot find much in this accord. However, extending the same as per current trends of research, much more have been done in this accord by applying latest aids and technologies which of course have benefitted a lot to the entire linguistic community.

2.3. Complex Predicates in South-Asian Languages

It was the early European linguists or scholars who first noticed any sort of multiword verbal expressions or the construction of CPs in South-Asian languages. These were later carried on through various findings and researches in general linguistics or in the field of COLING through means of latest aids and technologies into different major as well as less-resourced languages. Until mid of 1989 there was no major attempt that has been made to take these concerned researches way forward but, with the flow of time a comparatively great thirst has been seen in recent times to examine and identify such constructions in detail for South-Asian languages. The different major works done in this accord have tried to quench this thirst. The major plot of study for CPs in South-Asian languages till date has been seen only for identifying serial verbs or else towards the construction mechanism and role of light verbs.

Bashir (1993) tends to show as to how compound verbs functions in a language by applying causal chain model which help in amalgamation of a compound verb and also differentiate the same at multiple levels depending on the situation and context. Butt (1995) successfully portrayed the idea of making possible choices of CPs and its existence for meaning on light verb constructions in Urdu. In this she advocated the idea of making accurate choices of light verbs that lays the existential foundation stone for CP formation in Urdu.

Davies (1988) through his research showed that how the construction of CP in Telugu goes beyond any functional control. With this he tried to show as to how verbs interchange their functionality in Telugu to perform as a CP and hence modifies a sentence structure. He also showed as to how beautifully light verb constructions are unified in order to act as CP construction.

Annamalai (1979) through the detailed study also conveyed the central idea which is to show the different variability of CPs and its meaningful relation between the verb sequence in Tamil. Hook (1993) work was entirely based on the aspectual concept of compound verb formations in Indo-Aryan languages. In his work he tried to show the possible contrast of compound verbs in Indo-Aryan languages such as Hindi, Urdu and Marwari with special reference to different linguistic aspect. He thus tried in his work to frame a new paradigmatic category of CP.

In order to study the idea of CPs in some of the prominent south-Asian languages let us concentrate a bit more to the upcoming sections, as few of them have been discussed there in detail.

2.4. Literature Survey on Indian NLP- A Brief Descriptive Sketch

Natural language processing researches for Indian languages have received massive attention in recent past years. Venkatpathy et al. (2005) have worked on relative compositionality of noun-verb collocation technique of Hindi for identifying possible MWEs in the language by applying the same SVM model, taken for this undertaken research. As mentioned earlier, there have been very less researches done by applying technical aids in Indian and modern languages. Other than this there also exists other prominent works in NLP concerning foreign languages like English. In order to measure the idea and nature of compositionality in English major contributions have been given by linguists like Church and Hanks (1989). They both applied techniques of mutual information and Log Likelihood

respectively. The idea of analysing respective distributive frequency and property of objects in English and linguistics in the field of NLP came into existence by applying the method of verb information Venkatapathy and Joshi (2005). They successfully identified the idea of similarity of identifying any sort of objects in a sentence by creating a pair between a verb and an object over the machine. For this they have applied the Latent semantic analysis method.

Mukherjee et al. (2013) have applied the global linear model (GLM), for POS tagging in Bengali. They have also worked on the idea of extracting verb phrases containing CP constructions from English and Hindi language by applying corpus alignment technique. Chakrabarti et al. (2008) have applied the simple automatic technique for extracting the Hindi compound verbs from possible verb phrases. Their designed tool has achieved a highest percentage of accuracy that ranged about up to 98%.

Kunchukuttan et al. (2008) have applied the statistical method for developing a system that helps in identifying compound nouns as MWEs in Hindi. The system designed were so effective and accurate that it reported an accuracy of 80%. Sinha (2009) developed a system tool for mining CP constructions of Hindi by using a parallel corpus of English and Hindi. They applied empirical methods of research while designing their tool which gave an accuracy of 89%. For mining Hindi corpus and designing various tools the field of NLP has been extensively supportive.

Sriram and Joshi (2005) has designed a COLING model that helps in identifying the CPs and its categories through statistical analysis of the given corpus. This model has reported limited success with very less accuracy. Sinha (2009) has beautifully explained several different aspects of Hindi MWEs along with its formation and identification by applying various linguistic features.

Martin et al. (2011) and Knublauch et al. (2004) have computationally focused upon ontological concepts and classifications and have thus developed semantical web applications by applying AI techniques for business intelligence. This has been designed with special reference to MWEs structures. Bhattacharya and Patel (2015) focused on identification of MWEs in Hindi. They have applied the word embedding techniques on wordnet based features.

Chakrabarti (2011; 2014) have used a semi-automated technique for Bengali MWE extraction. He has done the task by crawling a large data of about 393,985 words or tokens

out of which with a number of about 283,533 tokens, he was able to make the raw corpus for the system. For his work he has mainly opted for popular idiomatic expressions along with several collocational terms of Bangla language in order to test and build and evaluate his model. He has chosen such type of corpus because they were frequently used by the people to communicate. He successfully gave a better evaluative result of the system designed by comparing both the data of the model which is the statistical model of evaluation as one with the baseline system of the model as another. With this he was able to extract MWEs from large set of corpora by applying clustering model of the data.

Along with this there also exists prominent linguists who have made major contributions to the field of NLP. Some of the well-known figures who have demonstrated an extreme interest in the topic of MWEs in Bangla includes Paul (2004), Sarkar (2008) and Chakraborty (2021).

With the Head Driven Phrase Structure Grammar (HPSG henceforth) linguists like Paul have built a constraint-based mechanism tool for Bangla NLP related tasks. This mechanism can easily be now used for many Indo-Aryan languages for the identification of verb construction process in COLING. Despite the fact that it is a mechanism that supports all different types of compound verb constructions in Indo Aryan languages, including Bangla compound verb constructions especially which has drawn a lot more attention of the linguists for its sequencing techniques.

In order to extract MWEs in Urdu, Hautli and Sulger (2011), have developed a clustering technique. This is another milestone in the field of NLP achieved for Urdu. The syntactic information of the words or tokens that are unique in Urdu is one of the main foundations of this method. It has been observed that Urdu has a strong postpositional feature when MWEs were extracted from the language. One may quickly identify and pinpoint the MWEs in Urdu by using these postpositions. The Centre for Research in Urdu Language Processing (CRULP) at National University of Computers and Emerging Sciences (NUCES) at Lahore in Pakistan have adopted a bigram extraction method which has been applied to 8.12 million corpora of Urdu newspapers from where the text has been collected to perform the task.

Similarly for Manipuri it was Nongmeikapam et al. (2011) who first spotted the idea of extracting and presenting the use and identification process of Reduplicated MWEs in Manipuri by applying NLP methods. In order to identify the techniques reduplications

created in Manipuri, he used the Conditional Random Field (henceforth, CRF) model. On Manipuri, the results of his applied approaches and models showed a recall of 62.24%, precision of 86.06%, and F-measure of 72.24%. The pattern recognition method or the binary classifications of the information retrieval system or IR system which aids in gaining and retrieving the system's accurate value, comprised the central idea of this entirely designed system. Later, he also used the same CRF model to test his system designed for Manipuri. For Manipuri the results of his applied approaches and models showed a recall of 62.24%, precision rate of 86.06%, and F-measure of 72.24%. The pattern recognition method or the binary classifications techniques of the information retrieval system, aids in gaining and retrieving the system's accurate value. The CRF model was then hence later used by Nongmeikapam to test his system in Manipuri.

According to Katz and Giesbrecht (2006) and Baldwin et al. (2002; 2003) the constituents of compositional MWEs that arouse during the context of a language are comparatively fair to those seen in non-compositional MWEs. For the first time in 2010, Baldwin included all varieties of MWEs in his research, which not only discusses the concepts of MWEs with their specified uses but also their various linguistic qualities that actually defines them as MWEs. After analysing and studying them deeply and thoroughly, he was able to pinpoint several issues that are there in these MWEs which were in terms of structures, patterns, formation techniques, results etc.

Sinha (2009) later have extracted the complex predicates in Hindi to identify only the light verb constructions because of their unique linguistic nature. For this he used English and Hindi parallel corpus system along with some traditional and empirical techniques that have been applied for the evaluation of the system designed. Applying such techniques resulted in an average precision and recall rate of 89% and 90%, respectively. He has made every effort to extract Hindi MWEs, which has not only aided for its identification but has also enabled the creation of an electronic dictionary of these MWEs by collecting and assembling the data applying a semi-automatic methodology for the same.

Renu et al. (2016; pp.18-22) have successfully designed a model to extract and identify the MWEs in Hindi. They have incorporated a little advanced technique of CRF model which is CRF++ approaches and concentrated extensively on this specific domain. Several large-scale datasets for Indian languages have been given by Singh et al. (2016) and Behera et al. (2016). They have made a huge effort to outline the problems and difficulties

that can arise while translating action verbs of particularly complex predicate types from English to up to ten Indian languages wherein Magahi has also been a part of.

Ojha et al. (2015) have successfully designed a POS tagger system tool for a couple of Indo-Aryan languages that includes Hindi, Odia and Bhojpuri. Here in this system, they have annotated the raw datasets by using the popular BIS tagset used by almost all Indian languages. After tagging these, they have applied the SVM as well as the CRF++ algorithm to test these raw data over the machine. For this they have taken approximately 90k data of the concerned languages from different domains. After testing the tagger tool for the concerned languages have achieved an accuracy of 88 to 93.7% for SVM whereas the CRF model ranges its accuracy from 82 to 86.7% as whole. Through this research they have hypothesized that language showing much variations in their constructions are much prevalent to CRF testing as compared to SVM.

Behera et al. (2016) have also dealt with the issue of linguistic divergence of machine translation system with special reference to English and Bhojpuri languages. For this they have considered English as source whereas Bhojpuri as the target language. For this they have applied Dorr's theoretical framework which applies both lexical and syntactic divergence methods along with the possible conceptual structures for giving a proper resolution to the issue of divergences in machine translation. This research by Behera has cleared the road for all structurally constrained languages to create an appropriate and effective machine translation system.

2.5. A Brief Descriptive Sketch of NLP works in few major Indian Languages

The desired mention sub-section of the chapter will discuss broadly the detailed review of NLP works and its literature in few of the major Indian languages such as Hindi, Punjabi, Bengali, Marathi, Magahi, Odia, Bhojpuri and Sanskrit etc. All such works are discussed in the respective upcoming sub-sections of this chapter that will now be followed from here in sequence.

2.5.1. Hindi

Kumar et al. (2015) have suggested a graph-based method for Hindi text summarizer. The primary idea behind this method is to extract crucial information from any Hindi-language paper. In order to determine the relationship between two sentences and the significance of each sentence present in the document this idea has been used. This has also

been done to determine how well the sentences relate to one another. Here in this case, they have induced the semantic similarity technique. As per them, they believed that any statement with a high degree of importance would have the same details but at the same time if any sentence has higher significance, then only that sentence should be included in the summary.

For this they have proposed thematic word-based method which was first proposed by Kumar and Yadav (2015). In this method thematic words are being produced as an output by analysing terms, frequency and also the inverse frequency of the text. This system produces a list of words with a common theme and uses those to build summaries. Later, the technique of Hindi WordNet continues to analyse this generated summary. For this entire process, an algorithm that summarises both Hindi and Punjabi-written materials was proposed by Gupta (2013). The statistical method was the foundation stone for this algorithm. This text summarizer summarizes several basic linguistic entities that includes crucial words, cue words, nouns, verbs, negative words, font features, named entities, sentence positions, sentence lengths, and numerical information etc. in order to test the percentage of accuracy and possible weight of meaning for possible each feature used in the summarizer he has utilized the technique of regression, which is a MT technique of predicting the essential values of words as per its meaning.

2.5.2. Punjabi

In Punjabi it was Gupta and Lehal (2012) who once again presented their own idea of text summarizer. This time it was based purely on nine weighted different text features that includes named entities, title words, and possible keywords of the concerned languages through which they were able to determine the sentences. In order to identify the Punjabi terms associated with text attributes, they utilised rule-based and dictionary-based techniques. Later on, several other methods for summarising Punjabi text have been proposed by Gupta and Kaur (2016) following this. This was a bit more hybrid as it has used support vector machine model and simple text features for its analysis and results. They have utilised an entropy-based method to find key terms in the text which have given a quite valuable response in return.

Gupta (2014) later worked on automatic word stemming technique of Punjabi language. This was an advanced intelligence system developed for this language. As per him stemmer is a basic linguistic criterion for any languages in the world. In COLING a stemmer is basically a morphology-based information retrieval system that helps in reducing the

inflected or sometimes the derivational words from their stem or root part. So basically, it is a morpho-computational approach of identification. He developed a model for Punjabi that incorporates several NLP application developments such as text summarization, keyword extraction, topic tracking etc. In order to build this model, he has implemented a huge list of Punjabi stemmers including some proper nouns and names in MS access tool from backend and in ASP.NET server from front. This dual process of identification has successfully recorded a high percentage of 87.37% of accuracy for the stemmer tool. Kaur and Saini (2015) have proposed their work for the identification and analysis of stop words in Punjabi poetries by applying NLP approaches. The sole objective of this entire work was to automatically classify poems written in Punjabi.

Singh et al. (2015) have used deep learning classification techniques for morphological evaluation and sentiment analysis of Punjabi text. After recognizing a ten-fold cross validation of the data supplied, the average accuracy of sentiment prediction for each of the four classes into which the corpus has been divided was 93.85%, 88.53%, 83.3%, and 95.45% respectively. The suggested approach achieves commendable sentiment classification accuracy of 90.29% over 275 Punjabi text documents that have been supplied to the tool while performing the task.

2.5.3. Bangla

Abujar et al. (2017) have suggested a heuristic method and rules for summarising Bangla texts. In order to receive more profound results, different linguistic principles have been used for the extraction of each text feature. For instance, it has used three different criterion such as criterion of quantity of sentences, number of paragraphs and the frequency of repetitions made etc. The machine has determined successfully the impact rate of each word of the document.

Furthermore, in order to summarise the document in Bangla text, Efat et al. (2013), additionally concentrated on identifying text feature scores. For this he used K-means clustering, Akter et al. (2017) technique popularly undertaken by various linguists today for developing a Bangla text summarizer. According to the scores of their features, they divided the material into two different clusters, and then the best-scoring sentences from each cluster were removed and was then hence used as summary sentences. Later on, in addition to this a Bangla text summarizer based on text feature ratings of sentences was suggested by

Sarkar (2012).

Ritu et al. (2018) have worked on performance analysis of different word embeddings methods for Bangla Corpus. They have applied the newly advanced word clustering techniques for the execution of this task. This is used as it is not only an advanced techniques of NLP rather, it reduces the time taken for processing of the data thus, improving them memory and efficiency of the model designed at the same time. for this task they have applied a total of 5,21,391k word class or tokens which consists of only the Bangla unique words. This later made a cluster of each other thus resulting into a total accuracy of 78.91% as a whole.

2.5.4. Marathi

Kulkarni et al. (2022) have worked on deep learning models for Marathi text classification. Their work seeks to give a detailed and thorough review of all NLP resources and COLING models that are now available for Marathi as per current research trends of COLING. It also aimed at portraying the fact as to how these models are being utilized by Marathi linguists for the classification of Marathi texts. The entire work was based over all possible and available publicly available sources from where these texts can easily be collected. This technique of text classification has popularly used various possible techniques for its processing. Few of them are Convolutional Neural Networks (CNN). This is a deep learning technique used for image processing and recognition. Next in this series is Long Short-Term Memory network (LSTM). It is a type of Recurrent Neural Network (RNN). This is massively used for identification of sequential data predictions and learning issues. This is also a multi-layered networking technique that helps in identifying and memorizing different patterns of a language in order to give possible best results. Sahani et al. (2016) have worked on automatic text categorization system with special reference to documents of Marathi texts. This has used the vector space model for the text categorization that have achieved a high accuracy rate of 95.83%. This means that the model designed was capable of correctly sorting more than 95% of Marathi texts. For all other popular Indian languages and its NLP related tasks platforms like Universal Language Model Fine-Tuning (ULMFIT) and Bidimensional Encoder Representations Transformers (BERT) has been used. These tool and platform are extensively used in order to provide an appropriate comparative study among different sets of languages provided.

Mhaske and Patil (2021) have created a huge and large resource for opinion mining task. This was completely based on Marathi movie reviews. The data was collected from

Marathi movies and somewhat from social media review documents. This was done to develop an automated NLP system for opinion mining. The testing of the data went onto several different phases in order to validate the same up to maximum level. The document level polarity classification method which was based on lexicon creation successfully achieved F-measures of up to 0.75% and 0.56% for the positive and negative classifications respectively. These findings were used later as a motivation for linguists to carry on with this line of inquiry and make some additional efforts to improve the available tools and systems for the language concerned.

Furthermore, Marathi text summarizer based on the text rank technique was also suggested by Rathod (2018). The same was later being used by Mihalcea and Tarau (2004). For this task PageRank algorithm has been utilised on a graph-based technique to determine the importance of sentences. This entire model includes two unsupervised techniques for extracting sentences and keywords from any Marathi texts. Gaikwad (2018) gave a rule-based Marathi text summarising technique that generates a list of questions for each sentence which was entirely based on a nominal pattern of sentences. In this entire process, the answers to the top-ranked questions are then extracted after each question has been ranked in order of priority. The compilation of these questions and its responses is then regarded as the document summary. This entire model has received an accuracy of 82% as whole as its final outcome.

2.5.5. Sanskrit

Saxena and Agrawal (2013) have worked concerning towards creating a dependency parser for the Sanskrit language. as They have together attempted for creating Natural Language understanding (NLU) and Natural Language Processing (NLP) systems. They have used the concepts of Ashtadhayayi, a book of Sanskrit grammar by Panini in order to implement their ideas. They attempted this research for Sanskrit language because they believed that it is a univocal and acute language that always final structures as result that inherits no change further changes in the same.

During the process of making a dependency parsing tool for Sanskrit they have used Deterministic Finite Automatic (DFA) systems for morphological investigations whereas the idea of Utsarga Apavada (UP) were used for relational identification between the sentences. In this entire model they have applied rule-based approach as it helps in easily assigning the attributive values to the input string which then creates a particular environment for some

specific rules of Sanskrit also known as ‘sutras’ to get activated in the system. This entire research was a three-fold experiment that comprised of three basic linguistic databases, which are nominal, verbal and Particle such as prepositions, clitics etc. The entire research has given a tremendous output as the parser worked absolutely fine with all the possible data supplied hence, giving maximum result.

Jha et al. (2009) have also developed an inflectional morph analyser for Sanskrit language. This morph analyser could easily identify and analyses the inflected noun forms or verb forms of any text of Sanskrit which is in accordance or acting as a joint compound. The system developed can easily now identifies and tags each word of Sanskrit compound at three major labels of POS tags in the concerned language.

Jha (2010) have also successfully compiled a book on Computational linguistics with special reference to Sanskrit language. Jha et al. (2005) have developed an analysis system for Sanskrit language by using machine learning techniques. This was a much-detailed work that comprises of several tasks such as developing and designing basic linguistic resources for Sanskrit translation, POS identification module and Reverse Sandhi identification technique for initial segmentation method. the analysis system through this method developed for Sanskrit language performed satisfactorily but, as the research undertaken were in its initial phase therefore, the errors encountered could be overlooked. This error causing methodology has reduced its mechanism thus giving a lower output for the language concerned. However, it was also suggested that the problems or issues encountered could be easily eradicated if such languages can be more prone towards technologies in near future.

2.5.6. Magahi

Mundotiya et al. (2020) have laid their work forward in the field of NLP for developing a dataset along with baseline techniques, involving deep learning methods for three major low-resourced languages of Bihar including Magahi, Bhojpuri and Maithili. This was done to build a Named Entity Recognizing (NER) system altogether. This was built for machine translation techniques that tends to lay a helping hand for the machine to translate from all the mentioned languages to Hindi as whole. This was done by annotating the parts of different sizes of the corpus available for concerned and mentioned languages. The tagset followed have a total of twenty-two different tagging labels as tagsets when designed. The model applied for testing these data was again CRF which yielded a high resulting rate of 96.73% for Bhojpuri, 93.33% for Maithili and 93.04 for Magahi respectively.

Kumar et al. (2017), successfully developed a training model for Magahi MWEs identification and classification. This task was executed by applying popular SVM techniques for automatic identification and classification of possible MWEs of Magahi. In order to execute this task, they have successfully implemented POS annotated data of around 75k word tokens. Inside this huge number they also put 11k word tokens of MWE which were tagged after designing a detailed MWE tagset for this language that comprises nine different levels for its annotation which was adopted from Singh et al. (2016). Once this entire data was put to training and evaluation it performed exceptionally well by giving an output result of 81.57% as its reliability and correctness.

Kumar et al. (2012), have created an NLP based corpus for Magahi and have also discussed the possible challenges that have been encountered in doing so for any low-resourced languages. For this they have collected data from some popular blogs, which contained several stories or biographies written in Magahi. For the annotation of the collected data, they have obtained BIS tagset and have tagged the same at POS level. They have successfully tested these collected corpora on four different model of COLING. These were SVM tool based on Support Vector Machines (SVM), Tnt taggers based on Hidden Markov Model (HMM), MxPost tagger based on Maximum Entropy systems and MBT tagger which is a Memory based model in COLING. For all these taggers they have induced round about 50k Magahi words that were tagged accordingly using 33 different tagsets which were of course obtained from BIS and used by almost all Indian languages. After doing so they finally trained a total of 13k POS tagged datasets of Magahi, which were done upon a frequency-based baseline tagger. Except Maximum Entropy Tagger, none of the tagger performed well due to less availability of much refined corpus. Out of all these the Tnt tagger resulted as 86.09%, MBT gave 86.22%, MxPost resulted into 89.61%, SVM tool 49.61% and Baseline as 71.8% as a whole. They have also developed a computational model for developing resources for some major languages of U.P. and Bihar. These languages include Hindi, Magahi, Bhojpuri Awadhi, Braj etc. During this research they have successfully compiled corpora of diversified length in order to develop a system that can identify the corpus of these languages easily. This model was first ever data-based study to compare closely related languages of U.P and Bihar that are not much prone resources. This model has shown an accuracy percentage of 96.48%.

Raj et al. (2021), have succeeded in developing universal dependencies treebanks for languages like Magahi and Braj. This was developed using the framework of Universal

Dependency Techniques or UDT. It is an NLP technique for consistent annotation of words or tokens across languages. This research tends to elaborate in detail as how the different sorts of dependencies are interrelated between these two languages through treebank graphical representation. Alok (2016) explained the detailed theoretical concepts as to how split constructions are formed in languages like Magahi and Hindi.

Therefore, so far as the question of NLP researches are concerned in Magahi, there has been a lot that has been done till date. In order to continue this trend this undertaken research will thus just add some more major contributions.

2.5.7. Bhojpuri

Unlike all other languages mentioned above Bhojpuri has also received great importance in the recent past years. It is due to this reason that it has been a subject of further research and study not only for theoretical linguists but also towards the area of NLP and COLING. Ojha et al. (2015), have trained and evaluated the POS taggers of a couple of Ind-Aryan languages that includes Bhojpuri, Hindi and Odia. The research was done to evaluate the data of these three languages by applying SVM and CRF algorithms. The central idea of the entire research was to identify the error patterns of these languages and provide solutions if any for the same concerning the mentioned algorithms. It was undertaken with a total of 90k tokens or word class out of which 2k word tokens were taken as training data for each language category. The system designed were successfully tested and have resulted with a high percentage ranging from 88 to 93.7% for SVM whereas it was 82 to 86.7% for CRF model. In this research the CRF model performed a bit lower in case of Odia and Hindi but recorded the maximum for Bhojpuri. This cross lingual study also proved that the CRF model performs better with the languages that has variations in their structures which is absolutely not true in case of SVM.

Behera et al. (2016), have dealt with the issue of linguistic divergence concerning English and Bhojpuri machine translation. With this research they have successfully categorized the possible linguistic divergences such as lexical, syntactic and semantic along with its issues among English and Bhojpuri. Singh et al. (2014), have successfully annotated Bhojpuri corpus buy using BIS tagset for Indian languages. For this they have collected round about 5,300-word tokens from Bhojpuri stories. It was collected in a spoken language form and was then linguistically transcribed. The entire work was based on mainly focusing two different variations of Bhojpuri which is the type of concerned language spoken in

Bhojpur area of Bihar region while the Benarasi style spoken in Varanasi area of Uttar Pradesh. This entire research has successfully paved a way forward for all such low-resourced languages of this kind across India.

Ojha (2019) have dedicated his entire research for English to Bhojpuri Statical based Machine translation system by Karaka model. In NLP the idea of Karaka model evolved out right away after Panini. It is a rule-based model used mainly for Indian languages in NLP. With this work Ojha has tried to take the entire idea of universal dependency a bit forward but not limited to Karaka dependency. Much later with time the same model was used again by the researcher in SMT systems for evaluating the corpus set of two different languages namely English and Bhojpuri by establishing possible relation after meaningful comparison.

2.5.8. Maithili

India is a home for almost 22 scheduled languages as per the 8th schedule of Indian constitution. It is a language that belongs to Indo-Aryan language family spoken majorly in Mithila region that comprised of areas of ancient Videha now popularly known as Tirhut region. Spreading its essence furthermore, it is also spoken now to some parts of Jharkhand and Nepal. Priyadarshi and Saha (2020) have created a POS tagger for development and online resource creation for the language in the field of NLP. For this they have taken a corpus size of more than 52k Maithili language and have trained the same over the CRF model in order to develop a classifier. Their system attained an accuracy percentage of around 82.67%. They have also tested some raw corpus of the language which were obtained from some line sources like Wikipedia and other online Maithili sources in order to train the neural networks module for word embeddings feature. Here also the same CRF platform has been used which attained an accuracy of 85.88%.

Sinha and Jha (2022) have also incorporated their research in this field by training the India language datasets that is being used later for text summarization in some major languages into which Maithili has also been included. With this research they tried to know whether there exist some similarities into these languages in terms of their structural patterns and also to examine the advancements of researches in these languages in the field of NLP with the ongoing flux of time.

Nidhi and Singh (2018) have worked for MT systems of Maithili and English with special reference to divergence. The aim of this research was to develop an MT system which is statistical based for the language concerned. As this task was commuted for divergence

issues therefore, it also measured the possible degree of divergence available along with the possible errors encountered thus suggesting possible solutions for the same.

Kumar (2020) have dedicated his research for developing a Named Entity Recognizer (NER) system for Maithili. Ranjan and Dubey (2016) have developed a system for Maithili language which can identify the isolated word form of this languages. They have developed this recognition system by using HMM model. The system developed have attained an accuracy percentage of 78% as a whole.

Therefore, unlike all other languages, there has been a lot in the field of NLP and COLING has been done so far in Maithili language as well. As this is also a language mentioned in the 8th schedule of Indian constitution therefore, these entire researches have gained much importance than expected which of course have higher chances of corpus availability as compared to other languages of Bihar which are low-resourced such as Magahi.

2.5.9. Odia

Odia has now been declared as one of the classical languages of India Pattanayak and Prushty (2013) Jha et al. (2014). It is also among one of the scheduled languages as per the Eighth Schedule of Indian Constitution. It has its roots belonging from the vast Indo-Aryan language family. This was earlier known as Oriya. This is chiefly spoken in the regions of Odisha earlier known as Orrisa, which is a contemporary name falling under the historical Kalinga empire of ancient times. Unlike all other languages mentioned above in different sub-sections Odia has also been very rich and lucky in terms of receiving much importance and attention of the researchers as a result of which there has been a lot that has been done in this accord and is still continued till date and further.

Jha et al. (2014), have tested two Indian languages including Hindi and Odia on Typecraft platform, which is an online corpus-based testing platform of datasets of several different languages. It is also a platform wherein all natural languages can easily be documented. It was a collaborative project between ILCI group of Jawaharlal Nehru University (JNU) and the Typecraft (TC) group of Norwegian University of Science and Technology (NTNU). Under this project the sole task was to annotate the available data of the concerned mentioned languages thus building a descriptive framework for its constructions. Secondly, it was concerned with the task of data exchange in order to ease the idea of data portability between the two which is TC and ILCI.

Behera et al. (2018), have also worked for identifying possible issues and challenges in developing POS tagset for Sambalpuri which is a local variation of Odia spoken chiefly in the regions Sambalpur, the western part of Odisha. This research has been done by applying a two-way model which is applying CRF++ as one while SVM as another. For this a large size of corpus for around 121k has been collected from different web sources of Sambalpuri and has been thus annotated by using the popular BIS tagset for Indian languages. Later, it has been trained on both the models by applying a data size of 80k and 13k respectively which has resulted as an accuracy percentage of 83% in terms of SVM whereas 71.56% in term s of SVM. Therefore, one can see here that the later module used has resulted into less accuracy percentages as compared to that of the former one. An SVM Tagger was also developed by (Das et al. 2015), for Odia language by applying a training set of approximately 10k word tokens from Odia which reported an accuracy of 82%.

Behera (2015), have laid his entire research concerning for attaining the possible suitable model required for Odia POS tagging. Here also he has applied the same two popular models of NLP namely SVM and CRF++ modules in order to test his experiment. In this research the total result given by the two statistical taggers which is SVM (96.85 and 93.59) as one while CRF++ (94.39 and 88.87) as another shows that the latter performs much better than the former with a laying margin of 2.46 and 4.72 percentage in both the supervised and unsupervised dataset respectively.

Therefore, in this entire chapter one can see as to how the different NLP task has been undertaken and executed in various Indian languages till date and the development is still counting. It is therefore due to this reason the idea for this research has triggered the mind of the researcher in order to take ideas from the same and implement these into his own language which is Magahi and execute it in order to bring Magahi onto a broader platform of technologies by applying latest computational and NLP techniques.

Hence, with the possible reviews of mentioned works done and also portrayed in detail through different sections and sub sections of this chapter with reference to Magahi in specific along with all other South-Asian languages as a whole one can see as to how beautifully and deeply the prominent linguists along with some recent ones have done an enormous amount of tremendous work. This will not only pave the way ahead for current researches but will also enhance and promote other future works in the area concerned.

Chapter 3 Complex Predicates in Magahi

3.1. An Introduction to Complex predicates in Magahi

Sentences in all languages always consist of different basic lexical and linguistic categories. These categories could be of different types, such as derivational morphological or sometimes inflectional. These could also consist of different grammatical categories like nouns, verbs, adjectives, etc. All of these are referred to as the necessary categories, which are very important to form a sentence in a language. On the other hand, there are also some unnecessary categories like suffixes that are just appended to the above categories just to spread their actual meaning. Unlike all other verbs of Indic and South Asian languages, Magahi verbs also function as a lexical unit, carrying all the necessary inflections such as tense, aspect and mood (TAM). Along with this feature, it also exhibits the property of negation. It is obvious to note here that like all other South Asian languages verbs in Magahi also have the properties of inflections to denote the honorifics of a subject. Therefore, it can be said that verbs are the central element of a sentence that contains all the necessary information that is true and applicable to all the languages of the world.

The first section of the chapter gives a brief introduction to the overall framework discussed concerning complex predicates in Magahi. This is later followed by various sections and subsections that cover the various types and sets of CPs in this language. All these sections and subsections of the chapter give a thorough and detailed picture of Magahi's CP, especially focusing on its construction procedures. These include noun-verb constructions, adjective-verb constructions, compound verbs and reverse compound verbs etc.

3.2. Features of Complex Predicates in Magahi

As we all know, verbs play a very vital role in a sentence, as they carry the entire meaning. It is due to this reason that the verbs are known as the nucleus of the sentence. Being the nucleus, it controls the entire structure of the sentence. It is Butt (1995), through whose detailed research on CPs in the Urdu language has brought several atypical characteristics of CPs in a significant role of study. Like all other Indian languages such as Hindi, Bhojpuri, Bangla etc. Magahi does also exhibit some unique features of CPs. This section of the chapter will discuss in detail some of these distinctive features of Magahi CPs through its upcoming sub-sections. These characteristics can be easily understood with some illustrations or examples mentioned in the respective paragraphs.

3.2.1. Scrambling

In linguistics, scrambling is an adaptive feature of any language in which the words of a sentence can interchange its position. However, the property of scrambling can sometimes change its meaning, but not the syntax of the sentence. As we all know, Magahi is a language that follows SOV word order in its sentence construction. This property makes Magahi a verb-final category language. The features of scrambling in Magahi could be easily understood with the help of a few examples mentioned below: -

1. स्याम राम के पनिया देलकइ
sjɑmə rɑm ke pənɪyɑ d̪eləkəɪ
'Shyam gave water to Ram'.

2. स्याम के राम पनिया देलकई
sjɑmə ke rɑm pənɪyɑ d̪eləkəɪ
'Ram gave water to Shyam'.

3. पनिया देलकई स्याम के राम
pənɪyɑ d̪eləkəɪ sjɑmə ke rɑm
'Ram gave water to Shyam'.

Here, in the above examples, it can be seen that in the first example (1) the ordered arrangement of phrases which are subject, indirect object, direct object and finally the place of the verb is shown. But it is obvious to note here that the same example as mentioned in (2) and (3) has changed the word order of the sentence without changing the gist of its meaning. This implies that in most cases the interchangeability of word order does not affect or change the meaning. However, the structural phenomena or the syntax of the same will certainly be affected.

Also, in the second example (2) one can see that the indirect object Shyam precedes the subject Ram and the direct object which is /pənɪyɑ/ or water follows it. Similarly, in the third example (3), the object /pənɪ/ or water precedes the indirect object Shyam and the subject Ram follows the indirect object in the sentence.

Therefore, in the above-detailed explanations using various examples, it is important to note that like other South Asian languages such as Hindi, Bhojpuri, Maithili, Marathi, etc., Magahi also follows the concept of scrambling. These are the languages that mainly belong to the group of the Indo-Aryan family. We have also seen here how Magahi allows this successful scrambling order of word constituents in their sentences without any alterations or changes in their meaning.

3.2.2. Honorificity

Magahi as part of the language has three different honorifics for its verbs. The very first level of the badge of honour tends to determine the relationship between the addressee and the addressee as accidental. In other words, it can be said that this is a marker that represents a similar or friendly relationship between the two. At the same time, the very second level of the honorific in Magahi can easily be seen as the relationship between the addressee and the addressee, with the addressee being quite younger in relation to the addressee. Because of this, the addressee uses a very formal way of greeting the addressee. The markings used for such expressions are set according to gender designations, meaning they may vary for males or females, but the essence and beauty of respect for the elder are not altered.

In the same way, there is also a third type of honour mark in Magahi, where the marks used indicate that the addressee is older than the addressee. This can also be viewed as the reverse case of the second type of honorific. Hence, here with all the above facts, one can easily deduce that the honorifics in the Magahi can be easily classified as non-honourable, meaning that it uses no honorifics while being addressed, medium honourable, showing a love relationship between them the addressee and the addressee, while the third can be identified as the most honourable, who tends to show not only a sense of love but also a sense of respect from addressee to addressee. It is also important to note here that the use of honorifics in Magahi depends directly on the type of sentence. To understand the above facts, let's refer to some below-mentioned illustrations: -

4. तु जाइ थे
   t̪u jɑɪ t̪^{h}e
   ‘You are going’. (informal)

5. तु जाइ थे थी हथ
ʈu jɑɪ ʈʰə ʈʰi həʈʰə
‘You are going’. (informal)

6. अपने जाइ थे थी हथ हथीन
ʈu jɑɪ ʈʰə ʈʰi həʈʰə həʈʰinə
‘You are going’. (formal) .

Examples (4-6) show the inflection of auxiliary verbs based on honorifics bestowed on the addressee. In (4) the non-honorary auxiliary verb /ʈʰe/, is used, indicating agreement with the subject. In (5) the honorary suffix /ʈʰə/, /ʈʰi/, /həʈʰə/ is used to show honour for the subject. In (6) more honourable auxiliary verbs such as /ʈʰə/, /ʈʰi/, / həʈʰə/, /həʈʰinə/ are used, which agree with the second person being the honourable subject. Here the highest grade of the honorific /həʈʰinə/ or sometimes //hələʈʰinə/ is used to show more honour to the subject.

3.2.3. Arguments and Case Markings

The complex predicate in Magahi plays an important role in argument requirements and case assignments. In the case of the compound verb Magahi, the argument requirement depends on the light verb. Choosing a light verb can affect the transitivity of the sentence, as shown in the example below.

The case assignment is governed by the second/light verb in the Magahi complex predicate. Case assignment in complex predicate construction is therefore positional which means, it is assigned by the last verb of the complex predicate. Let us understand this with the help of a few examples discussed below: -

7. राम पिटा गेलई

ɾɑmə piʈɑ gelɑɪ

‘Ram got beaten’. (informal)

8. राम मोहन से पिटा लेलई

ɾɑmə moհənə -se piʈɑ leləɪ

‘Ram(willingly) got beaten by Mohan’.

9. राम मोहन के सीता से पिटबा देलकइ

rɑmə moɦənə -ke siʈɑ -se piʈəbɑ ɖeləkəɪ

‘Ram made Mohan to be beaten up by Sita’.

In example (7), the light verb /gelɑɪ/ goes has only one argument, Ram, while in example (8), the light verb /leləɪ/ meaning take has two arguments, namely Ram and Mohan. Similarly, in example (9), the given meaning of the light verb /ɖeləkəɪ/ takes three arguments Ram, Mohan and Sita. So, the compound verb construction was done with the same main verb, but different light verbs show changes in the argument structure of the sentence.

In the case of the conjunction verb, the nominal host is an integral part of the predicate and plays a crucial role in deciding the number of valences and postpositions associated with arguments in the sentences. Also, in the case of the conjunction verb, the number of arguments and case marking depends on the total noun/adjective and light verb sequence. The transitivity and intransitivity property of the light verb is lost. In the case of a conjunction verb, the nominal host is not the direct object of the light verb.

10. सीता अपन घर साफ़ करी थय

siʈɑ apɑnɑ gʰɑrɑ sɑpʰ kɑrɪ ʈʰɑjɑ

‘Sita is cleaning her room’.

11. रितेश स्वेतवा के पढ़े में मदद करलई

riʈesə sveʈɑvɑ -ke pɑɽʰe meⁿ məɖəɖə kərələɪ

‘Ritesh helped Shweta in studies’.

Here, examples (10) and (11) of the sentences clearly discuss the features of conjunction verbs. In example (10), the words / sɑpʰ/, meaning clean, and /kɑrɪ/, used as the feminine form of the verb do, function together as a single verb. In example (11), the noun /məɖəɖə/ decides what Help means about the presence of three arguments to explain who is helping whom with what work. Here you can clearly see that Ritesh is an agent, Shweta is a patient, and/pɑɽʰe/ is a place. The postpositions /-ke/ and /meⁿ/ are also associated with patient meaning (to whom) and place meaning (whereby), respectively, which is determined only by the word /məɖəɖə/ meaning help.

Hence, the concept of argument and case marking is one of the crucial areas that needs to be studied in Magahi CPs as opposed to other Indic languages in order to analyse them thoroughly.

3.2.4. Agreement in Magahi Complex Predicate

The agreement is a universal feature of languages of the world. It becomes parametric in the sense of the features involved in agreement in different languages. In South Asian languages preferably Hindi, verbs agree with subjects in person, number, and gender. Magahi is different from many South Asian languages in the sense that it exhibits agreement in terms of honorifics. Before we discuss the agreement system of Magahi, let us look at the agreement system of one of the most dominant South Asian languages, namely Hindi. In Hindi, the verb exhibits agreement in person, number, and gender. It does not exhibit agreement with the arguments which are lexically case marked. If the subject does not have an overt lexical case marker, agreement occurs between the subject and the predicate. This was followed in a similar fashion in Magahi. An illustration of the same has been shown in example (12). If the subject is lexically case marked, then the agreement takes place with one of the arguments of the predicate that does not carry a lexical case marker as in (13). There occurs a default agreement on the verb if all the arguments of a sentence are lexically case marked as in (14).

12. रमुआ किताब पढ़े थय

rəmuɑ kiʈɑbə pɑɽʰe ʈ̪ʰəjə

‘Ram used to read book.’

13. रमुआ किताब पढ़ल कय

rəmuɑ kiʈɑbə pəɽʰələ kəjə

‘Ram had read the book.’

14. रमुआ छुड़िया से फलवा कटल कय

rəmuɑ cʰuɽijɑ -se pʰələvɑ kəɬələ kəjə

‘Ram cut the fruit with knife.’

Here, in example (12), the subject of the sentence carries no lexical case markers, so the main verb /pɑɽʰe/, read and the auxiliary verb /ʈ̪ʰəjə/ matched the subject of the sentence

Ram. Similarly, in example (13), the main verb /pl/, read and the auxiliary verb /kəjə/, would agree in number and gender with the object of the sentence (argument of the predicate), since it does not carry a lexical case marker. Also, in example (14), all three arguments of the sentence carry lexical case markers; hence the verb carries the standard convention. Let's examine Magahi a little more closely to understand a more detailed way of its agreement system. It is distinctly different from Hindi due to its this uniqueness.

For Hindi, structural case and verbal agreements are so closely related that it is assigned structural case via Agr.[9] positions (Mahajan; 1990). But that's not the case with Magahi. The Magahi agreement system is always opaque as case markers have no effect due to the agreements in that language.

So far, we have seen that this section begins with the hypothesis that the Magahi agreement system would work in the context of complex predicates in a similar way to that in simple verb forms. (Bhatt; 2012; 2008) reports that verbal agreement goes along with the subject as much as possible, but when the subject does not show features of the agreement, verbal agreement targets the next higher argument. He reports that usually, the noun in the N+V construction agrees with the light verb when the other arguments of the sentence are case-sensitive. But sometimes it doesn't, and in such cases, the noun is not treated as an argument by the grammar and is invisible to a match. Paul (2004) points out that the simple finite verb forms agree with the Bangla personified subject. Deoskar (2006) points out that in the context of the complex Marathi predicate construction, the light verb must necessarily match an uncased subject or object DP.

Therefore, with the above discussions, it can be seen that in the Magahi language the agreement is reflected through inflection which is attached to the constituent particle most possibly a verb, i.e., either the main verb or light verb or else the auxiliary of the sentence. Also, in Magahi complex predicate comprising mostly Noun/Adjective/ combinations, it is the main verb, which always carries the root form or participial form and the light verb carries inflections and gives information about the person, tense, aspect, mood etc of the sentence.

[9] It is a concept in linguistics which denotes the change of position of word from dependent to relational category due to inflection and agreement.

The subject-verb agreement of the Magahi language is reflected through suffixes attached mostly with light verbs carrying the honorific markers of the person concerned while there is no agreement shown for this language based on number and gender.

In Magahi complex predicate construction, the light verb shows agreement with either the subject or object. This agreement is with respect to the person and honorifics or non-honorifics of the addressee component. Magahi shows three unique features in the context of the agreement system. These agreements take place between the light verb and non-overt addressees in the sentence. The agreement in Magahi is also shown and identified between the light verb and the listener to whom the sentence is stated. In the context of the Magahi language, the listener acts as a silent participant in agreement. Case markers are opaque in the agreement feature which means, there is no role of case markers in the context of the Magahi agreement system. These special features make Magahi a very unique language as compared to other Indo-Aryan languages such as Odia, Bangla Bhojpuri etc.

Thus, we can say that the agreement system in Magahi depends mostly on three facts: Opaque case markers towards the agreement, the role of the addressee component, and the degree of honorific factor involved with the listener and speaker. This means that along with the honorifics, the agreement system in Magahi is also based on the role of a person or gender concerned in the sentence.

It is quite evident here from all the facts that the Magahi agreement system functions in the same way in the case of a complex predicate as it has with the simple verb. Therefore, this section provided a brief sketch of the features of agreements in Magahi, elements of the agreements in Magahi, uniqueness of Magahi agreements, and structural representation of these agreements in Magahi complex predicates and has also discussed all of them above in a very detailed manner.

3.2.5. Negations in Magahi Complex Predicates

Negation plays a significant role in determining the structure of Magahi. In this language, it is expressed by negative elements like nɑ/məʈ̪ə. The negative element /məʈ̪ə / used in the imperative is shown only in example (15), while in the remaining cases it is used as /nɑ/ as shown in example (16) below.

15. तु ओहिजा मत जो

ʈu ohɪjɑ məʈə jo

‘You don’t go there’.

16. उ ओहिजा न/मत गेलई

U ohɪjɑ nɑ/ məʈə gelaɪ

‘He didn’t go there’.

Here in example (15) the negative marker /məʈə/ is used in the imperative form. It is used in the sense of a suggestion not to go there. While in example (16) the two negative markers nɑ/məʈə are used simply to negate the affirmation of any ongoing activity in the sentence.

The use of mat in case example (16) will lead to the ungrammaticality of the sentence. Languages like Hindi and Magahi have most possibly these negation particles that have been studied extensively in the literature. Therefore, like Hindi Magahi negation is also of two types. These are constituent or phrasal negation and clausal or sentential negation. (Kumar; 2006) elaborates that the sentential negation takes the scope of the entire sentence as in example (17) and constituent negation takes the scope of the phrase to which it is adjoined, as in example (18) which will be discussed and hence explained in detail in the upcoming below sections.

3.2.5.1. Sentential Negation

Negative markings are usually used to express sentence negations. In Magahi, Negative adverbs and preverbal marks are the most common phenomenon, the latter being either particles that are separate words or affixes and clitic elements that are part of verbal morphology (Zeil-jstra 2004:152). Additionally (Payne; 1985, cited in Zanuttini 2001:512-13) tends to show that some languages like Hindi, Marathi Punjabi etc. have few particles and affixes for encouraging the use of negative verbs for sentence complementation. To understand the idea of the phrase negations in Magahi it is important to study and analyse them thoroughly and in detail, with the help of an example given below: -

17. राम सीता के कहिनो न/मत डंटल कय

ɾɑmə siʈɑ -ke kəɦɪno nɑ/ məʈə ɖⁿʈələ kɑjɑ

‘Ram has never scolded Sita.’

In the above example, it is clear that the negative element /nɑ/ is taking the scope of the entire sentence. It is also clear from the sentence here that it is negating the action of scolding Ram to Sita clearly. The example listed above clarifies the fact that a sentence negation is a type of negation that affects the meaning of the entire clause.

3.2.5.2. Constituent Negation

Constituent negation is marked by placing the negative marker immediately after the relevant constituent. This can be further understood from the following example: -

18. राम सीता के न मोहन के किताब देले हलय

ɾɑmɑ siʈɑ -ke nə mohənə -ke kɪʈɑbə ɖele ɦələjə

‘Ram had given the book to Mohan and not Sita’.

Here in the example above, the negative element /nə/ only occupies the scope of a phrase category Sita. We can see through this sentence that constituent negation is just a way to negate a single constituent or phrase of a sentence without touching the rest of the constituent particles of the sentence. In the above example, it is easy to visualize the cases of supportive negative constituents. These are used when someone wants to agree with most of what is said but wants to question some of it.

Similarly, in example (18) above, the negative particles were only used to negate part of the sentence instead of the entire sentence. Therefore, it is also important to note here that the property of constituent negation only partially negates the proposition without calling into question the entire proposition mentioned in the sentence.

3.3. Verbs in Magahi

Verbs encode a lot of information in a sentence. The occurrence of verbs is complex in nature in any natural language. The complexity of a verb lies at almost every level of the language analysis. Inflections on verb come under morphology, tendency of taking argument as per requirement belongs to syntax, and its meaning and functional aspect comes under semantics. The notion of the verb increases complexities at every level of the analysis. The

verb is the most important part of the sentence as it denotes mostly the action. It is a critical element of the predicate which says something about the subject of the sentence. Verbs express different kinds of actions, events or states of being. In short, a verb is a part of speech and an important element of the predicate which expresses something about the subject related to actions, events or states of being in a sentence.

The various subsections mentioned below will further discuss the various and distinct forms of Magahi verbs along with some best supporting examples.

3.3.1. Simple Verb

When a predicate in a sentence has only one verb, either in the form of the main verb or the auxiliary verb, it is called a simple verb. Simple verb phrases contain only one verb. In the following example /ɖekʰəlijə/, meaning saw, is the verb which behaves as the main verb since it is the only one acting as a verb for the sentence.

19. हम्मे ओकरा देखलिय

həmme okərɑ ɖekʰəlijə

'I saw him'.

3.3.2. Composite Verb

In a sentence, composite verb consists of several forms of verb which are composite in nature like serial verb constructions, phrasal verbs or idiomatic combination of verbs. These constructions are available in abundance in South Asian languages. In these languages, composite verb constructions are compiled of several different sequence of verbs or verb phrases within a single clause expressing simultaneous or immediately consecutive actions. This contains only one single grammatical subject as shown in example (20). But in case of a phrasal verb constructions or idiomatic constructions, the verb deals with those particular construction segments which accommodates few of the idiomatic elements or phrases along with the verb, as explained in example (21).

20. उ भोज खा के चल गेलय

u bʰojə kʰɑ ke cɑl gelɘjə

'He went away after eating in party'.

21. चोरबा दुइए मिनट में रफ्फूचक्कर हो गेलय

corəbɑ d̪uie minəʈə meⁿ rəpʰpʰucəkkərə ho geləjə

‘The thief ran away in only two minutes.

Here, in example (20) the two distinct actions of eating and going away are shown together in a sequence, occurring one after another. However, these two actions are performed by only one subject, which is, denoted by the pronoun ‘He.’ This ‘He’, is not overtly present in the given sentence. In the same manner in example (21), there is the use of an idiomatic phrase / rəpʰpʰucəkkərə honɑ/, which means ‘to ran away’. However, in the later example, the whole sequence of the given phrase is considered as one single unit as it denotes only one action now which is ‘to run’.

3.3.3 Causative Verb

Causative verbs denote the necessary action which is required to denote another action. This implies that it denotes a single event that is comprised of two different sub-events that occurred in parallel. In such constructions one of the sub-events causes the following other to occur. It deals with the process of causation. To denote this chain of events or actions, there exist two types of causal verbs which are direct causal and indirect causal verbs. Direct causal is concerned with one's own or self-involvement in the main action as shown in example (22), while indirect causal is concerned with the involvement of others in the main action which is explained in example (23).

22. उ राम के लड़कीया देखैलकय

u rɑmə -ke ləɽəkijɑ d̪ekʰailəkəjə

‘He made Ram to see the girl’.

23. उ राम से हमरा लड़कीया देखबैलकय

u rɑmə -se həmərɑ ləɽəkijɑ d̪ekʰəbailəkəjə

He gets Ram to see the girl.

Here in example (22), the word /d̪ekʰəbailəkəjə/ shows three arguments, namely /u/, which is He (as causer), Ram as (causee) and /ləɽəki/, girl (as the patient). This verb form falls under the category of direct causative since the act of seeing is performed by Ram and the sub-event of showing task is performed by the subject, He. On the other hand, in the second example sentence, (23) The verb /d̪ekʰəbailəkəjə/ has four arguments in which /u/

means ‘He’ who is the direct causer, and ‘Ram’ here is the mediator who is the indirect causer, while /həmərɑ/ which means ‘me’ acts as the causee, and /ləɽəki/, acts as patient in the sentence.

Therefore, it is clear and evident from the illustrations mentioned here that, this section has successfully shown the causative constructions of Magahi verbs with the help of suitable examples. It has also discussed in detail their related unique properties that helped in its formations. The upcoming next section will deal with the different verb features such as auxiliary verbs, light verbs, helping verbs, and modal verbs along with the transitivity, intransitivity and Di transitivity features of each in Magahi.

3.3.4. Auxiliary Verb

An auxiliary verb follows the main verb in a sentence. The auxiliary marks exist in all Indo-Ayan languages, like in Hindi in the form of (hai ho, ʈʰɑ etc.), while in Bhojpuri they have (bɑ, bɑɽe, həvə.), etc. and for Maithili, it is (cʰi, cʰe, ɑicʰɑ). In the same way, the auxiliary verbs in Magahi come from the root verb ho, hɑ, which means ‘to be’. It occurs in all tense forms of Magahi depending on the nature and demand of the sentence. But as far as the exact and correct forms of auxiliary verbs are concerned for this language, it doesn't occur.

The inflections for auxiliaries in Magahi are marked through the special features of honorifics which are based on the seniority of the person concerned rather than numbers. However, there also exist a few exceptions wherein these markings are based on numbers as well. For the present tense, the word like /hɑ/, /hi/ /hu/ are used along with the possible honorifics and non-honorific markers used in the sentence. These can be seen clearly through the below-mentioned various examples such as (24), (25) and (26). Here the past tense /hələ/ has been used together to denote the degree of honorifics, whereas the non-honorific markers can be seen in examples (27) and (28).

24. हम ही/ हीं/ हकी हियई/ हियउ हियो/हिवा

həmə hi/hiⁿ/həki hijəi/ hijəu hijo/hivɑ

'I am.’

25. तु ओहिजा हले/हलहू

ʈu ohijɑ həle/hələhu

‘You were there.’

26. अपने सब कहाँ हथन/हथी/हथीन

əpəne sɑbɑ kɑhɑⁿ həʈʰənə/həʈʰi/həʈʰinə

‘Where are you all?’

27. तोहनी सब अभी-तक घरवे में हीं/हहीं औ उ चलियो गेलई

ʈohəni səbə əbʰi-ʈəkə gʰərəve meⁿ hiⁿ/həhiⁿ ɑu u cəlijo geləɪ

'You all are still at home and he has already gone’.

28. ओहनी/ओखनी आज इस्कूल जल्दी चल गेले हलऊ/हलथु/हलथुन

ohəni/okʰəni ɑʝə iskulə ʝəlɖi cələ gele hələu/hələʈʰu/hələʈʰunə

'Today they had gone to school early’.

Here, in the above different example sets the use of different forms of auxiliaries such as hiⁿ/həhiⁿ has been shown. This can also be the non- honorific such as hiⁿ/həhiⁿ as non-honorific present tense markers; and the use of certain other forms of auxiliary such as həki/hijəjə /hijo/hivɑ/, /həʈʰənə/həʈʰinə/hələʈʰinə/, as an honorific, polite and present tense markers. In example (28), the forms of auxiliaries used for past tense and non-honorific are /hələu/ which has been changed as hələʈʰu/hələʈʰunə which is honorific forms of the former. This is so because the degree of respect and honour has been changed in the sentence to superlative. Therefore, like Bhojpuri and Maithili, Magahi also has auxiliary markers in its sentences.

3.3.5. Light Verb

Light verbs are one of the major elements that help primarily with complex predicate formation. It is a semantically bleached element. It contains the entire inflectional information about the sentence such as tense aspect, and mood, also known as (TAM). It occurs in Magahi majorly in two forms of CP constructions, such as compound verb and conjunction verb constructions.

Verma (1985) gives a detailed account of about ten to fifteen light verbs in Magahi in which /le/ meaning to take, /ɖe/ meaning to give and /jɑ/ meaning to go have the most

frequent occurrences, which is required for any sort of compound verb constructions. Let's consider some examples like (29), (30) and (31). Here, we can observe that for the construction of conjunct verbs, the light verb /kərənɑ/, means to do, and /honɑ/, which means, to happen occurs most frequently always after a noun or an adjective. This is clearly shown in examples (32) and (33).

Magahi light verb constructions do not form any syntactical unit. This is because there always occur too many different linguistic constituents between nouns/adjectives/main verbs and light verbs. These are most possibly the degree of negations, use of different particles, additional postpositions, etc.

29. अब खनमा खा ले

əbə kʰənəmɑ kʰɑ le

‘Eat food now.’

30. रीना मोहन के जाय देलकय

rinɑ mohənə -ke jɑjə ɖeləkəjə

‘Reena let mohan go.’

31. इ अमवा तू खा जो

ɪ əməvɑ ʈu kʰɑ jo

‘You eat this mango.’

In the very first example here, namely (29), the light verb /le/, ‘take’ occurs with the main verb and /kʰɑ/, meaning ‘to eat’. In the second example, i.e. (30), the light verb /ɖeləkəjə/, meaning ‘gave’, which is the past tense form of the verb ‘give’ occurs with the main verb, /jɑjə/ which means ‘to let go’, and similarly, in the last example, which is (31), the light verb /jo/, occurs with the main verb /kʰɑ/, means ‘to eat’. In all of these examples, it is clear that all three of these sentences resemble compound verb constructions with an imperative character. Now let's look at a few more examples of Magahi to understand the concept of light verbs in this language in a bit more detailed way.

32. उ राम के परीक्षा में बड़ी मदद करलकय

u ɾɑmə ke pərikʂɑ meⁿ bəɽi məɖ̤əɖ̤ə kərələkəjə

'He helped Ram a lot in the examination'.

33. तोरा देख के उ बड़ी खुस होलय

ʈ̤orɑ ɖ̤ekʰɑ ke u bɑɽɪ kʰusə holəjə

'He became very happy after seeing you'.

Here, in examples (32) and (33), the mechanism of conjunct verb constructions are shown. In example (32), the two grammatical elements /məɖ̤əɖ̤ə/ and /kərələkəjə/, are combined together to frame as one single entity. After combining they existed together as one single meaning which is 'did', the past tense form of the verb 'do'. This single unit is formed after merging two different elements which are nouns and light verbs respectively. But in example (33) /kʰusə/ meaning 'happy' and /holəjə/ meaning 'happened' are adjectives and light verb combinations. This also will form a conjunct verb combination, but this time it carries an adjective in place of a noun. However, the magic is not changed and hence this combination also formed a meaningful construction successfully called conjunct verbs. Therefore, it is evident to note that for every light verb construction, Magahi follows the technique of the general verb-complement pattern wherein the complementizer for the desired sentence is always required to make the sentence meaningful and acceptable.

3.3.6. Helping Verb

Another important role of a verb in a sentence is that of a helping verb. These are the elements that come after the sentence and modify the main verbs. Together with the main verb, it forms the verb phrase. One can use several helping verbs in a sentence. These verbs occur in any sentence in three main forms: auxiliaries, modal and light verbs wherein the auxiliaries are also referred to as temporal verbs sometimes. The below-mentioned section on auxiliaries with a suitable example makes it more clear. The auxiliaries help the main verb to indicate tense features, while modal verbs are those that add a bit of mood information to the main verb. These are illustrated below through separate sections and respective examples. Modal verbs generally are the closed classes of verbs because they are limited in number and do not allow any further elements after their addition. A few majorly used modals are words like sɑkɑ, cukə, cɑhə etc. in Magahi. These have no independent meaning. At the same time, a light verb occurs as a full entity when used independently, as shown in example (36).

(i) Auxiliary Verb

34. राम जा हय

rɑmə ɟɑ həjə

‘Ram goes.’

(ii) Modal Verb

35. राम सेब खा चुकले हे

rɑmə sebɑ kʰɑ cukəle he

‘Ram has eaten the apple’.

(iii) Light Verb

36. किताबिया टेबुलवा पर रख देहीं

kiʈəbijɑ ʈebuləvɑ pərə rəkʰə ɖehiⁿ

‘Keep the book on the table’.

One can see that in example (34) the word ‘/həjə/’, meaning ‘is’, functions as an auxiliary verb as it is giving temporal information, which is in the present tense for the sentence. Therefore, here the auxiliary functions as a tense marker. Similarly, in example (35) /cukəle/ meaning ‘finished’, is the modal verb giving the perfective aspectual information along with the auxiliary verb ‘he’, whereas in example (36) /rəkʰə/, meaning ‘keep’, is the main verb marking the action of the sentence and /ɖehiⁿ/, meaning ‘give’, is the light verb giving the additional information of the second person and perfective aspect of the entire sentence construction.

This section has presented the several forms and related features of a light verb construction thus, showing the different types of roles played by the same in different sentence structures. With the discussions above, we tend to understand that a light verb gives us information about the number of arguments that are present in a sentence. It can be present in a sentence in various forms such as a nominal form, adjectival form or else either as adverbial form as well. The next upcoming section discusses the occurrence of different types

of verbs along with their roles and respective requirements as per the different environments of sentences in Magahi.

3.4. Roles and Features of Verbs in Magahi

As we know that verbs play an important role in a sentence. A verb (V) is considered always the head of a verb phrase (VP). There are three major roles played by a verb or verb phrase in a sentence: Intransitivity, Transitivity, and Di-transitivity. In the same manner there exists two major features of verbs in Magahi which are measured on the scale of Finiteness and Non-Finiteness. Let's discuss these all in the below sections or sub-sections with suitable examples of each in Magahi.

3.4.1. Transitive, Intransitive and Di-transitive Verbs in Magahi

Verbs determine the number of arguments required in the sentence. On the basis of the number of arguments it can take, they are categorized as intransitive, transitive, and di-transitive. Intransitive verbs take no arguments; the transitive verbs take one argument; and the di-transitive verbs take two arguments. According to Extended Projection Principle, any clause of a sentence must contain an argument that has to be there in the subject position. Whereas the subject acts as an external argument which is always obligatory for the sentence. Let us understand all these with a few examples mentioned below: -

(i) Intransitive Verb

37.हम्मे हसलियय

həmme həsəlijəjə

'I laughed'.

(ii) Transitive Verb

38. उ गाड़ी चलयल कय

u gɑɽɪ cələjələ kəjə

'He drove the vehicle'.

(iii) Di-transitive Verb

39. हम्मे अप्पन घरवा साफ़ कैलियय

həmme əppənə gʰərəvɑ sɑpʰ kailijəjə

'I cleaned my home.'

Here, in the above examples, such as example (37) / həsəlijəjə/ meaning 'smiled', is an intransitive verb taking no internal argument and / həmme/ meaning 'I' as a subject in this sentence. In the same manner in example (38) /cələjələ/ meaning 'drove' which is the past tense form of the verb 'drive' is a transitive verb which takes one argument, in which /gɑɽi/ meaning 'vehicle', is a direct object for the mentioned sentence. Also, in example (39) / kailijəjə/ meaning 'cleaned' is a di-transitive verb taking two arguments. The two different arguments for the sentence here are 'həmme' meaning 'I' as one and the adjective /sɑpʰ/ meaning 'cleaned'. These both are acting as indirect and direct objects respectively for the given sentence.

3.4.1.1. Finite Verb in Magahi

A finite verb is a verb that inflects for a person, number, or gender. It occurs in an independent clause. It gives information about tenses in the sentence. Let us explain this feature of Magahi with an example, mentioned below: -

40. लइकवा सेव खाई थलय

ləɪkəvɑ sevə kʰɑɪ ʈ̪ʰələjə

'The boy was eating apple'.

Here, in the above example, the verb phrase /kʰɑɪ ʈ̪ʰələjə/, which means, 'was eating', gives many inflectional information about several different hidden meanings of the sentence. This include information like the doer of the action is a third person, singular, masculine and the sentence structure itself denotes that it is in past tense. It is also showing the progressive aspect of the given construction.

3.4.1.2. Non- Finite Verb in Magahi

Non-finite verb group has no tense. They play the role of a noun, adverb or adjective along with that of a verb. They also get inflected for tense, aspect, number and modality. Non-finite verbs occur in three forms in a sentence. These are Gerunds, Infinitives and Participles. Let us see all these three different properties of these three forms of non-finite verbs of Magahi in the upcoming sections one by one.

3.4.1.3. Gerunds

Gerunds are nominal verbs that take the position of nouns but retain their verbal characteristics. Gerunds always take an object or adverbial quantifier for its construction. It occurs in Magahi by adding the suffix /-lɑ/. This is generally attached to a verbal root, as mentioned in the example below.

41. सेव खयला सेहत लगी अच्छा होब हय

sevə kʰəjəlɑ sehət̪ə ləgi əccʰɑ hobə həjə

'Eating apple is good for health'.

Here, in the above example, /khaya/, meaning 'eating', is acting as gerund with the suffix /-lɑ/ which is attached beautifully to the root verb /kʰɑ/, meaning 'eat'.

3.4.1.4. Participles

Participles function as verbal adjectives or verbal adverbs in a sentence. In Magahi, verbal adjectives take the form of /-ɪt̪ɑ/ to denote the progressive aspect, as shown in example (42), and /-ɑl/ to denote the perfective aspect, as in example (43). Verbal adverbs are also formed by adding /-ke/ or /-ɪt̪ɑ/ to the main verb root, as shown in examples (44) and (45), respectively.

42. उ हमरा पास दौड़ित अयलय

u həmərɑ pɑsə d̪auɽit̪ə əjələjə

'He came running to me'.

43. उ चलते गाड़ी से कूद पड़लइ

u cələʈe gɑɽi -se kuɖə pəɽələi

‘He jumped off from the moving vehicle.’

44. बेस से पढ़ल किताब कभियो भुला न हई

besə -se pəɽʰələ kiʈɑbə kəbʰijo bʰulɑ nə həɪ

‘A well-read book is never forgotten’.

45. उ पढ़ के बोल लय

u pəɽʰə -ke bolə ləjə

‘He Spoke after reading’.

46. लइकवा दौड़ित अईलय

ləɪkəvɑ ɖauɽiʈə əiləjə

‘The boy came running’.

Here, in all the above examples such as (43) shows the verbal adjectives /cələʈe gɑɽi/ meaning ‘running vehicle’ denotes the progressive aspect. Similarly, in the very next example, which is (44) the term / pəɽʰələ kiʈɑbə / meaning ‘read book’, takes the form of /-ələ/ with the verb /pəɽʰə/ to denote the perfective aspect. Examples (45) and (46) respectively shows the formation of verbal adverbs wherein /pəɽʰə -ke bolə ləjə/, meaning ‘spoke after reading’, and /ɖauɽiʈə əiləjə/, meaning ‘came running’, are hence formed only after attaching markers like /ke/, and /ɪʈɑ/ respectively to the verb roots.

3.4.1.5. Infinitives

Infinitives are identified as those verbs which occur in object relation with another verb. It is also generated by adding /-ələ/ suffix to verbal root, as shown in the below example: -

47. रिनमा घरे जायल चाही थय

rinəmɑ gʰəre jɑjələ cɑhɪ ʈʰəjə

‘Rina wants to go home’.

Here, in the above example the verb / jɑjələ/, which means ‘to go willingly’, is acting as infinitive with an added suffix /-ələ/ to the root verb /jɑ/, meaning ‘go’.

3.5. Complex Predicates in Magahi

Verbs encode a lot of information in a sentence. The occurrence of verbs in any natural language is always of a complex nature. It is due to this reason that the complexity of a verb lies at every level of language analysis for each language. The property of verb inflection falls under morphology, the tendency to take arguments falls under syntax, while their meaning and desired functional aspects are under semantics. The notion of the verb adds complexity at every level of analysis therefore, the verb is the most important part of the sentence.

Descriptively, a verb is an action word. It is a critical element of the predicate that says something about the subject of the sentence. Verbs can be of different types such as actions, events, or state of being. In other words, it is also a part of speech and an important element of the predicate that expresses something about the subject which is subjected but not limited to denote actions, events or states of being in a sentence.

Languages in general play a prominent role in representing the regional identities in India. This is so because when a language is spoken, a region is reflected or can be smelled well by its structures and tones. A language is not only related to its origin, but also to the cultures and traditions of the people who live in that region as a whole. This section of the chapter will briefly discuss all types of CPs of Magahi with all its relevant examples along with detailed and thorough explanations through the various upcoming subsections.

3.5.1. Conjunct Verbs

Verbs in Hindi and their regional variation such as Magahi are formed by joining a noun or an adjective with a verb. All of these verb-form constructions are known as conjunct verbs. If it is a noun that is adjacent to the verb then it is a nominal conjunct, while if it is an adjective that is adjacent to the verb it is known as an adjective conjunct. In a conjunct verb,

the adjoining adjective or noun must be in abstract form with the verb. In these types of constructions, there must be a proper semantic link between the verb and the x. This x could be a noun or an adjective which exists completely in its abstract form. Once this x becomes strongly associated with the attached category of the word and establishes a proper semantic relationship, it acquires the property of a conjunct verb. The most important thing to note here is that not every sequence of a noun/adjective construction with that of a verb is a conjunct because, it must carry a host, which must be preferably an adjective or noun existing in its metamorphic or in abstract form in order to acquire the status of a conjunct.

For example: -

48. उ दर्जिया हीं जा के कपड़वा छोट करवा लेतय

u d̪ərjijɑ hiⁿ jɑ -ke kəpəɽəvɑ cʰoɽə kərəvɑ leʈ̪əjə

'He will go to the tailor and make his cloth shortened.'

Here, in the above illustration, the adj. /cʰoɽə kərəvɑ/, which means 'to make short' and the light verb /leʈ̪əjə/, together form a conjunct verb construction, wherein the adjective /cʰoɽə kərəvɑ/ shows an agreement with the verb / leʈ̪əjə/, denoting the completion of the task. Therefore, this /leʈ̪əjə/ is thus acting as a light verb for the above sentence construction.

3.5.1.1. Noun-Verb Construction (N+V)

Complex predicates, today are one of the main areas of a research study in almost all languages, especially South Asian languages. These are extremely productive. For this reason, several types and forms of CPs are successively studied and elaborated in a much more detailed way by Butt and Ramchand (2002). In this section, the focus is exclusively on the combinatorial possibilities of N-V complex predicates. In all these combinatorial possibilities, the noun bears all the predication responsibility, and the verb, also known as the light verb, bears characteristic features like case marking, the pattern of agreement, etc. In other words, for every conjunct verb construction, having the combination of a noun with that of a verb, it is the light verb that contributes an additional semantic value to it. This could be in terms of adding information about tenses, aspects, case markings, agentivities, or else the experiencers.

For example: -

49. राम सीता के मदद करलई

rɑmə siʈɑ ke məɖəɖə kərələɪ

‘Ram helped Sita’.

In the example above, it is evident that the combination of the noun /rɑmə/ and /siʈɑ/, along with the two verbs i.e., /məɖəɖə/, which means ‘help’, and /kərələɪ/, which means ‘do’. clearly, denotes the past form of the verb, which together forms a sentence. This type of sentence construction justifies the construction of the kind (N+V) construction thus, giving the sentence a meaningful expression.

In the above example it is also important to note that the first verb also known as v1 which is /məɖəɖə/, meaning ‘help’ carries the main information of the noun, and the second verb known as v2 which is /kərələɪ/, carries the extra-linguistic information of the verb showing inflections as per the sentence. It is this v2 or the second verb that is thus modified or changed as per the demand of the tense, aspect, mood, and honorific markers of the sentence.

3.5.1.2. Adjective-Verb Construction (A+V)

One must not get the idea here that adjectives are the only ones expressing the primary quality of age, size, colour, etc., but it is the idea of the togetherness of both the adjective and the verb that goes into this construction. Unlike other languages, Magahi also supports such constructions. This type of construction is most commonly found in folk tales, but also in the stories or literary forms of Magahi. Unlike other Bihari languages like Maithili, and Bhojpuri, the constructions of the form adjective + verb (adj. + v) in Magahi is also formed by embedding. This means that in a sentence, the adjectives are embedded in the pre-existing structure of a sentence. This is because Magahi uses adjectives mainly for comparative constructions through the process of embeddings. But at the same time, it also uses certain distinct adjectives along with the verb to justify its uniqueness. To further explain the mentioned construction process, please refer to an illustration that follows: -

50. मोहन अप्पन घरबा साफ़ कर हई

mohənə əppənə gʰərəbɑ sɑpʰ kərə həɪ

'Mohan cleans his house.'

Here in this sentence, one can see that the adjective / sɑpʰ/ is embedded in the sentence in order to make it an adjectival phrase. In this sentence one can notice that the adjective /sɑpʰ/ has imparted a new discourse or referents to the sentence which means, it has clearly stated that the house about which the discourse is made in the sentence is clean. Therefore, one can also get the idea here that adjectives in Magahi not only decorates and beautifies the sentence but also adds a new discursive element to it like other Indian languages such as Hindi as well. It is also noted here that the predicating functions in Magahi are not only possible with verbs or nouns but also with adjectives. Therefore, the CP formations with adjectives in Magahi are possible, but to find such constructions in Magahi especially in this proposed work will be a tough task for the author. This is so because, for Magahi it is a huge possibility that not all adjectives could form a successful CP because for this language CP constructions with adjectives depends upon various factors and linguistic properties like semantical, derivational and morphological properties of the adjective that has to undergo in the process of forming a CP in this language.

3.5.2. Compound Verbs

In general, compound verbs (CV) are one of the most commonly used syntactico-semantic phenomena that are most prevalent in today's South Asian languages, including Magahi. This is mostly created by combining the two verbs which is v1 and v2, where v1 is acting as polar and v2 is vector. After successful combination both these acts as a single verb. Regardless of any genetic differentiations, these are easily found in all Indian languages. (Massica, 1976). It is extensively researched today in almost all Indo-Aryan and Dravidian languages because of its most enriching qualities and shared ancestry.

Generally, CVs are the sub-types of popular Multiword Expressions or MWEs, Kumar et al. (2017). This implies that it is not substantially different from multiword phrases, which are made up of more than one linguistic unit, but rather quite similar and identical to them. It is hence asserted that a CV is nothing but a form of multiword compound. Unlike several other Indian languages, Magahi also has compound verbs. It

undergoes with the exactly the same identical development process as it is in all other Indian languages. Let us refer to following instances in order to clarify the concept in more detail: -

51. पइसवा रख

pəisəvɑ rəkʰə

‘Keep the money.’

52. पइसवा अपन पकिटिया में रख

pəisəvɑ əpənə pəkiʈɪjɑ meⁿ rəkʰə

‘Keep the money in your pocket’.

53. पइसवा अपन पकिटिया में से कुरसिया पर रख दे

pəisəvɑ əpənə pəkiʈɪjɑ meⁿ -se kurəsijɑ pərə rəkʰə ɖɛ

‘Keep the money from your pocket on the chair’.

Here in sentence (51), there is no use of a compound verb, as it is the simplest form of sentence. In this sentence someone is simply instructed to keep the money somewhere or at someplace whose direction or place is not exactly specified. While, if we move on to another sentence (52), we can notice that an action over the subject which is ‘money’, is directed and hence instructed to be kept in pocket. In this sentence, the direction and place are exactly specified for the subject. But when we look at the last sentence (53), we notice that an action is not only performed but also denoted much far away from the subject. This denotation of action is shown by the addition of a v2 which is an auxiliary verb which /ɖɛ/. This de-specifies the action of the main verb /rəkʰə/ for the subject which is referred as ‘money’ here at some specific place which is /kurəsijɑ/, meaning chair acting as a location in the given sentence.

Hence, with this one can deduce that a CV is sometimes used as a supporting element in order to clearly denote or sometimes also specif the correct and accurate action of the verb, without which the completion of the above sentence structure is absolutely impossible.

3.5.2.1. Reversed Compound Verbs (RCV)

Reversed Compound Verbs (RCVs, henceforth), are one of the key characteristics of the majority of Indian languages such as Hindi, Marathi, Maithili, Magahi etc. It was first noted by Hook (1974). Most South Asian languages inherits this new distinctive feature into their structures thus making it unique. These are used basically to narrate things like fairy tales, fictional stories, etc. Almost all Indian languages makes an extensive use of such special characteristics of verbs. The RCV construction only requires the proper rearrangements of the two verbs present in the sentence, which together makes up a compound verb construction. Rearranging the places of these two verbs wherein v1, polar and v2, vector, have resulted into successfully creating a very new type of grammatical construction.

In order to understand this a bit more, let us consider an example of this type from Hindi /cəlo ɑo/, in Hindi when rearranged, will appear as /ɑo cəlo/ which means the same but, as we can see that these polar and vector verbs, appearing in the v1 and v2 have now been interchanged from their respective positions, which lead towards the formation of a new structure, known as RCVs. However, the semantic property is still unharmed by this aspect of verb positioning technique. For such constructions, it is also no denying the fact that the situation does not always remains the same as here in the example above the change of meaning of the given expression is both fairly obvious and perplexing.

As mentioned, the meanings of the expressions of the concerned language can occasionally also change after this magical property of reordering. This property of rearranged patterns of verbs in a sentence to what we call as RCVs and its structures are mostly found in the Hindi language. These are hard to frame in Magahi because, the order of reversing linguistic components is impossible in Magahi due to its unique feature of rigidity in word-ordering. However, a construction denoting Hindi RCV can be referred below, to understand the phenomenon in detail. Imitating these sorts of structures only one can frame a RCV in Magahi which may or may not result grammatically correct for this language.

54. उसने शीशे पर पत्थर दे मारा

usəne ɕiɕe pərə pəʈʈʰərə ɖe mɑrɑ

‘He threw a stone at the mirror’

Here in the above-mentioned example, one can see that the compound verb / mɑrə ɖijɑ/ meaning ‘to hit’, has simply been reversed to /ɖe mɑrɑ/. This is a beautiful example, evidencing the magical reversible characteristics of RCV in Hindi. But in Magahi, no such constructions are possible however, simple compound verbs can be framed by applying these magical tricks of compounding. For example: -

55. शर्मा जी हमरा ले डूबलथीन

ɕərmɑ ji həmərɑ -le ɖubələʈʰinə

‘Sharma ji has drowned me also.

Here in the above sentence, the semantic property or meaning of the statement does not get changed as a result of this reversible compounding property of the verb. This also had no detrimental effects on the meaning of the sentence. This is the reason linguistic researchers have now paid a lot of attention today in studying these verbs or these sorts of linguistic constructions. These closed and definite studies will assist not only the standardization of languages like Hindi but also contribute a bit to the further development and analysis of lesser-known languages like Magahi from linguistic perspectives.

Chapter- 4

Research Methodology

4.1. Methodology Applied

This chapter of the work is important in relation to the fact that when doing research on any subject or topic, methodology plays a crucial role. It is a successful approach that is actively used by researchers to conduct their research efficiently. There are many different methodological approaches in research. These are the experimental method, deductive method, inductive method, observation method, etc.

For the proposed work, the author will use the observational research method. This applies, as we know, to the Observatory's principles for analysing the data collected for the research carried out. This method carefully monitors the data collected from online sources and from some native speakers. This is because the native content of the data is much more reliable than any other online data source. However, as mentioned earlier, online sources are also used to collect the corpus.

Based on the observations obtained, the author will continue to work with both data and try to thoroughly analyse the refined data based on his own self-intuitive knowledge. This is equally important as the author himself is from a Magahi-speaking region, so self-comparison is very important for data analysis. This is also important to easily identify and tackle errors accordingly in the later-developed tool, which of course is the main goal of this research.

The methodologies applied while conducting this research can be understood through the following process: -

(i) Research Design: - The aforementioned research is experimental in nature and aims to bring Magahi, a language with fewer resources, onto the platform of technologies to increase the level of prominence and recognition through the latest Machine Learning (hereafter referred to as ML). This is also because Magahi has historically not only been overlooked but also underestimated as a language of study, despite having great history and importance since ancient times. In order to

create an appropriate research design for this study, appropriate computational models for the successful conduct of this research were used.

(ii) Cross-observatory process: - Once collected, the data is verified by some Magahi native speakers, mainly from the areas of the central region of Bihar. This means that the researcher will authenticate the collected data with the help of some Magahi speakers from Gaya, Patna and Jehanabad areas to avoid errors in the data if there are any. This also ensures that the data collected is not tampered with, so when trained with machines and computer systems, this data should always meet the standard of authenticity.

Also, in this process some different versions of Magahi CP data will be taken for the consideration from some other regions as well in order to show some differences if there is any apart from the speaker's orientation of pronunciation. This means that this will help in visualizing the structural or the syntactic differences of the CPs if there is any. As the sole objective of this research is to identify CPs through computational tool developed therefore, in order to perform this task, the above-mentioned steps are very crucial as any error or differences in the data collected will mislead the objectives decided.

(iii) Data collection Method: - As mentioned, this is the process where the data is collected through a one-to-one approach, contacting many Magahi speakers with different versions to test the data. Normal daily discussions of Magahi groups are observed to see how differently they use the Magahi CPs from Hindi in their daily statements. This is done with the help of some blogging websites which are of course the main source of this research. However, this is later identified and cross-checked with Magahi speakers for how correctly they are pronounced or if there are any irregularities in it.

(iv) Population Method: - Considering this is my secondary source for data collection, this process will help the researcher take research methods to the next level. In this process, the data once collected from local Magahi speakers and from blogging website which is magahiblogspot.com[10]. All of this is then later analysed with the local speakers of Magahi as it was done in the previous step. Also, it can be said that the last two methods are related because the last two methods play an important role in the whole research to get the best result without error and even if

[10] http://magahi-sahitya.blogspot.com/

there are errors in the data, all these are taken towards the machine learning approach so that all required steps should have been followed quite obviously, and then the errors found after a two-way process between man and machine. Detailed machine-learning techniques will be covered separately in the coming chapters. Following these basic steps not only leads to the best results with a maximum positive percentage but also reduces the chances and likelihood of occurring errors.

(v) Data-Tagging: - As the next step process for this research, these data will be tagged according to their different CP levels or tags for which a separate tagging system will be developed to which these collected data will later be exposed to be tagged according to their language needs. This whole process of tagging the data after observation will be part of the experimental method. In summary, there are several different methodological approaches that are applied at different steps in this research. For this reason, it can be said that this research implies a multimodal research methodological approach in order to obtain the best results. The application of such methods will advance this work in a very efficient way and in unique way.

Linguistics is therefore a broad field for the detailed study of any language system, while also encompassing several aspects of the study of the language. In other words, we can say that we study sounds in phonology; the formation and composition of the word fall under morphology, the study of the structure of phrases, clauses or sentences in syntax; and the study of meaning in semantics. There are some other aspects such as applied linguistics, sociolinguistics (studying language in the context of society), psycholinguistics (studying the representation and function of language in the mind), neurolinguistics (which deals with language processing in the brain), computational linguistics (studying language in the connection with a machine) and so on. For this reason, it is obvious to note here that this work falls within the field of computational linguistics, which applies a detailed technical knowledge of computer systems and software together with some theoretical knowledge of the language itself.

In other words, this work is a fusion of language and technology. As already mentioned, the data for the research carried out are provided by native speakers of the target language, which is Magahi. For this work, the data collected came not only from some blogging websites but also from some native Magahi speakers from Patna and Gaya regions. This is

because these two cities have been the most famous centres of Magahi and Magadh since ancient times. It is worth noting here that the historical significance of the Magahi spoken in a particular area has nothing to do with the descriptive and syntactical conclusions of the work. The study of the nature and structure of complex predicates does not address the regional language varieties. Most of the data come from native speaker intuition and indirect observation. To maintain the objectivity of research in this area, data from native speakers were cross-checked and verified by the researcher. This was achieved through the use of a thorough procedure for recording both formal and informal interviews. The data collected is cross-checked in multiple stages to ensure the perfect grammatical authenticity of the language.

The upcoming sections and sub-sections of this chapter will now discuss the methods and approach applied on this research in a very detailed manner.

4.2. Method of Data Collection

At the first level, the author collects the data from an online blogging website and then from some native speakers. Once collected, it is thoroughly examined and the CPs it contains are identified and marked. All these tasks are done with the help of some native Magahi speakers from the Magahi-speaking regions of Bihar. So, after collecting data from online sources, the native speakers are the secondary source from which the data is not only collected but also properly verified and annotated. In summary, the author will adopt a descriptive observational approach to the research to advance this work in a fairly efficient manner.

In addition, this section provides a detailed description of the methods used in collecting the corpus or data and the steps involved.

4.2.1. Automatic

The corpus collected for the proposed work comes from an online blog that includes data from all sources such as short stories, autobiographies, poetry and some local news. The idea behind this blogging website is to touch all areas of Magahi linguistically. The mentioned blogging website is also chosen as one of the main sources for the data collection as the style and pattern used in the blog is a daily lifestyle pattern that touches all spheres of the Magadh and Magahi-speaking regions and their people.

As, we all know that the data in this blogging site is quite large therefore, it involves some machine learning approaches in order to collect the same. It is due to this reason, this section has been named as Automatic, as this involves some mechanical approaches in order to perform the task. This implies that for the collection of Magahi corpora in large numbers, some eminent Magahi-speaking societies like Magahi Bhasha Parishad, Patna and popular online Magahi magazines like Magahi Manbhavan[11] which means early morning has been accessed. These were the platforms created by some renowned Magahi scholars who utilised Magahi not only as their mother tongue but also in their literary writings. This is why one can easily find common Magahi terms and expressions here on this website that cannot be found in any other source. Using this website makes the task undertaken more specific as the rough essence of the Magahi language is not overlooked.

4.2.2. Tools Applied for Corpus Collection and Annotation

This section will deal with the entire process of collection of corpora for the research and the techniques used and applied in this process. It provides a detailed description of all the tools and techniques involved in the research.

4.2.3. Indian Language Crawler (IL-Crawler)

Indian Language Crawler or IL Crawler is a tool used to crawl data. It is a web crawling tool applied to all Indic languages to crawl and collect data. Until this tool was introduced in the NLP or COLING field, the data collection processes and methods were pretty nasty. But since its inception, such tedious and trivial data collection procedures have become easier.

IL Crawler is an Indian language web spider or crawling tool that crawls all data available in the World Wide Web (www) domain of the Magahi data collection website. Through this, the data is recorded very efficiently and systematically. To do this, it reads all possible information from the Uniform Resource Locator (hereinafter referred to as URL) provided to this tool to collect the data. It reads all possible links from the HTML (Hypertext Markup Language) page or web page and automatically detects all possible links from which to extract the data. Once the crawler has identified the webpage, it extracts the data from each section and subsection of the page.

[11] http://magahimanbhavan.blogspot.com/

The tool is used only to collect data from online sources and not manual ones. For manual data collection, it is handled manually. So far as the online data is concerned for this research, it is already mentioned that the data is being extracted from a blogging website of Magahi, which carries the words mostly from literary sources or is entirely based on different kinds of literature like short stories, biographies, domains of health, entertainment etc. Each of these data will be of accurate structures and class varieties. The author has chosen this blogging website as the written or literary corpus is very less available in this language. This is due to its non-popularity and less exposure to the general public domain. Since the tool applied for collecting the data is named IL crawler, it is mostly for Indian languages including Hindi, Magahi, Odia, Bangla etc. This is the reason why the author has applied the same tool for the task of crawling and collection of data.

4.2.4. Indian Language Sanitizer (IL-Sanitizer)

This is a tool in COLING that is applied to the data collected from the URL selected for this research. This is applied to the data to clean it. In other words, it is also called the purifier. As the name suggests, it cleans all the raw data previously collected using a crawler and makes it more efficient to work on it. This tool deletes any repeated content or text or tokens from the collected corpus. Likewise, an IL crawler, this tool also works to collect data efficiently. The only difference between the two is that one helps collect the data while the other helps with the cleaning and refinement. After this data cleaning task is completed, it syncs the collected data into a single file in a well-organized manner. In other words, it can be said that this tool helps the author to arrange the data in a proper sequential manner, which assists him to perform the NLP task taken on for the selected title quite efficiently.

4.2.5. Mechanism of IL Crawler and Sanitizer

Web crawling and cleaning is the first step for any internet-based information retrieval system. As already mentioned, a crawler is nothing more than a spider or a web like the structural engine of a machine that crawls through the entire data from all sources on the World Wide Web. At the same time, the Indian Language Sanitizer refines the data and, if present, deletes any repetition from the corpus. Web crawler synchronizes all data such as images, links, HTML code and content from any website while the sanitiser removes unwanted errors and repetitive elements from the data.

Although the mechanism is simple for both crawlers and sanitizers, there are still many different processes in between to make the whole system work. Mathematically, these are referred to as optimization methods. This means that the machine iterates over the entire data, comparing the data internally using various built-in mechanisms, until an authentic and reasonable solution is found.

The web crawlers mentioned earlier in this section use a set pattern of processes that selects each web page only after certain machine requests have been met. This means that the machine comes with instructions to collect only the data needed for a specific or specific purpose and not for any other task. For example, if we need to collect the data for verbs in Hindi or let's say another language or another subject like tourism, medicine or anything else from another field, then one has to provide these instructions from the beginning with this technique. This can also be referred to as topic-specific crawling or focused crawling. Here it is obvious that this crawler only worked for Indic languages like Hindi in general. Therefore, errors may occur when analysing the Magahi data. For this reason, the disinfectant mechanism was used. Using a Sanitizer erases all this data that is irrelevant and provides us with a corpus useful only for our research purposes. However, the chances of getting accurate data are still questionable as it is a mechanism used for standardized Indic languages and Magahi is certainly outside of that framework.

4.2.6. Crawler and Sanitizer Architecture

The idea of topic-based and domain-specific web crawling was first developed by adaptive retrieval agents that select heuristic neighbourhoods for information retrieval. By Menczer in (1997) and then by Chakrabbarti et al. (1999)

The architectural design of a web crawler fundamentally involves five major steps. These are as follows: -

(i) URL frontal step – This refers to the first step of creating the list of different URLs from where the corpus needs to be collected for the task. It will contain all the main sources used for the purpose of data collection.

(ii) Fetch Process– This is the second step where the engine selects and identifies all the web pages from the URLs mentioned in the first step. This means that data

from all web pages located in the URL will be integrated, no matter how long or how big the data source is.

(iii) DNS resolution Process- DNS, which means Domain Named Server, is the next step, which involves the mechanism of translating various internet protocols (hereinafter referred to as IP addresses) into specific domain names. In this step, the DNS mechanism determines the address of the relevant server from which the data must be taken. This means that it identifies the location of all websites embedded in said data collection URL.

(iv) Parsing Process- This module helps in extracting the data from all types of texts contained in all different URLs that are mentioned for every webpage.

(v) Duplicate elimination process- This is the most important step in extracting the data. This is one of the main steps where the precise role of the tool IL Sanitizer comes into play. Using this tool not only removes the duplicate URLs that are already there from where the data has already been retrieved, but also any repeating modules of the data, if any. This implies that it clears the entire collected corpus and make it refined and ready for further use. This is the reason why it is known as Sanitizer. The entire process can be easily understood through the figure mentioned below.

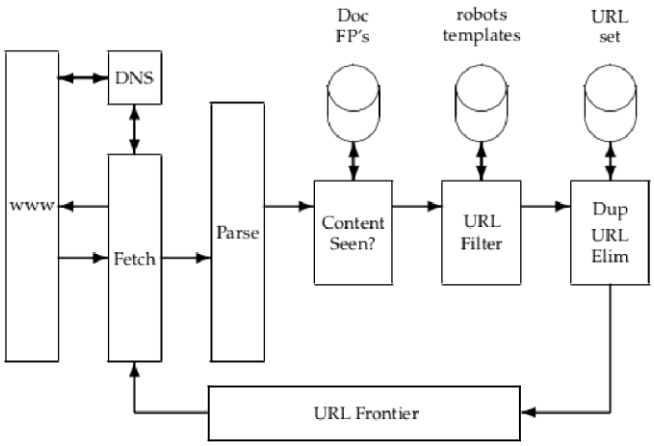

Fig.6. Basic Architecture of a Crawler

4.2.7. TDIL (Technology Development for Indian Languages) and the Digitalization of Indian Languages

As we all know, India is home to many different languages. Out of all these there are only twenty-two who managed to get their place in the Indian Constitution by the 8th schedule. Technology Development for Indian Language (hereinafter TDIL) is a technological platform developed by MECIT (Ministry of Electronics Communication and Information Technology), Govt. of India in 1991. This was started with a vision to secure and protect all Indic languages that are in danger of extinction or have been neglected or overlooked over time. It is a platform that not only provides the latest technological tools to secure such languages, but also keeps records of all these languages to create a huge multilingual resource using the latest technologies.

According to this project, it is said to have been believed that India, which is home to many different languages and language families across borders, the most important of all is the Indo-Aryan diversity. It is the Indo-Aryan language family that as members make up 76.87% of the entire population.

This platform was launched only to develop latest resources or language processing tools using latest techniques to greatly simplify the task of man-machine interaction regardless of language barriers. At the same time, it also promotes the idea of standardizing any language that has not been exposed too much in terms of research studies or in terms of technologies. To fulfil this dream, the TDIL has actively worked with the Central Institute of Indian Languages, Mysore (CIIL, Mysore), the Indian Institute of Science, Bangalore (IISc, Bangalore), several Indian Institute of Technologies (IITs) along with many other prominent institutes which are of national importance and has been able to perform exceptionally well in this regard with their respective researches. It has also continued to collaborate with many other bodies of national or international importance out of which few of them are Microsoft, Google, Amazon, World Wide Web Consortium (WWW), International Organization for Standardization (ISO) etc.

All the organisations and digital platforms mentioned above have collaborated well and are working tirelessly to give all lesser-known Indic languages like Magahi, Odia, Santhali, Bodo, Sambalpuri, Maithili, Bhojpuri etc. a proper status that will help the future generations who wish to work with these languages on the one hand, while digitizing these lesser-known languages in general through the means of technologies on the other side.

4.3. Method of Data Analysis and Annotation

This section contains the methodology adopted for analysing and annotation of the collected Magahi corpora for CP identification. The undertaken task has been accomplished by using the ILCIANN App and following the BIS tagset guideline with the help of which a separate tagset for different classes of Magahi CP identification has been developed. This was based on the common BIS (bureau of Indian Standard) basic tagging guidelines which is usually applied to almost all the Indian languages. The upcoming sub-sections below will discuss all these steps further in a much-detailed way.

4.3.1. Tagset and Guidelines for CPs Annotation

Magahi, being a language with very little resources available today, hence there is no tagset or guideline available for tagging the data of Magahi CPs identification yet. Therefore, for this purpose, the tagset used here to tag the CPs in Magahi is the very first of its kind. This is a bit like the tagset used for Hindi, Singh et al. (2016) and for Magahi MWEs (Sicky;2017) but not exactly the same. However, a very few of the tagging guidelines have been received from the tagset developed for Hindi and Magahi MWEs. This is because Magahi and Hindi both shares the same common boundaries and language families, which is therefore Indo-Aryan and Bihar respectively.

4.3.2. Description of Tagsets with their Corresponding Annotation Labels

This section of the chapter will discuss the tagsets designed and obtained for the CPs annotation in detail along with its suitable examples.

4.3.2.1. Conjunct Verb (CONJ_V)

This is a tagset category wherein only verbs with conjunct properties are being tagged. It is a tagset that identifies those verbs where a noun or a nominal subject come along with a verb (Majhi, 2007). In general linguistics verbs are classified mainly in two major types in complex predicates. For example: -

56. रीता अपन घरवा साफ़ करलई (A+V)

riṯɑ əpənə gʰərəvɑ sɑpʰ kərələi

‘Rita cleaned her house’.

57. रीता अप्पन लइकन के ध्यान रख हई (N+V)

riʈɑ əppənə ləikənə ke ɖʱjɑːnə rəkʰə həi

'Rita takes care of her children'.

Conjunct verbs in general allows two different set of possible structures. These are as follows: -

4.3.2.2. Adjective +Verb (CONJ_JJ)

This is a sub class of tagset incorporated with conjunct verbs. Here in this sub category only those CPs are considered which are mainly comprised of two elements which is Adjective and verb. These are known as adjectival conjuncts. Let us see an example: -

58. तू अपन रूमवा गन्हा देहीं

ʈu əpənə ruməvɑ gənɦɑ ɖeɦiⁿ

'You keep your room dirty'.

4.3.2.3. Noun +Verb (CONJ_NN): - It is another subset of the tag that is incorporated in the conjunct verb category. As the name suggests it is designed only for those words or tokens of CPs which are formed with the combination of noun and verb elements. This is also known as a nominal conjunct. Under this tagset frame only those corpuses will be taken which will primarily have only these two specific elements mentioned above. A nominal conjunct may or may not require any connecting agent for its clarification. In order to understand this in detail let us follow some more examples below: -

59. हम्मे ओकर मदत कर देलियै

ɦəmme okərə məɖəʈə kərə ɖelɪjɑɪ

'I helped him'.

60. लइकन हल्ला करे थय

ləikənə ɦəllɑ kəre ʈʰəjə

'Kids are making noises.'

Here in the examples mentioned above we can see that the sentences formed with conjunctive verbs easily follow a certain morphological approach, which is an adjective + verb or noun + verb construction. Also, here in sentence (59) the conjunctive verb

construction of /ɖenɑ/ and /kərənɑ/ functions as transitive verbs to denote full function or action of the sentence.

One can also see that how the noun or adjective part of the conjunction verb is changed or transformed to make the main verb consistent with the given subject in the sentence. It is therefore possible for the conjunction verbs to change its form as per the requirement and nature of the subject it takes. However, this may not be true in all cases of Magahi while forming a nominal conjunct

Therefore, in the examples mentioned above the annotation method and the tagset applied is decided by applying not only computational but also concerning morphological approaches of the language. The set of tags defined and designed for this category is characterized by the corresponding canonical annotation forms mentioned beforehand of this class or category.

4.3.2.4 Compound Verb (COMP_V)

It is a class of tagset that identifies CP verb types that are often supports and exhibits a combined construction involving two verbs. We all know that in a compound verb construction, the first element, denoted as v1, does not carry any inflection since it always occurs in the root form, while the second verb element, known as v2, carries all sorts of inflections like tense, aspect or mood. This v2 also contains information and details about the modality and honour markers in Magahi. In a compound verb construction this v2 is known as a light verb. To understand this a little better, let's look at an example: -

61. राजू किताबिया फार देलकय

rɑju kɪʈəbɪjɑ pʰɑrə ɖeləkəjə

'Raju teared the book'.

Here in example (61) the main verb v1 is /pʰɑrə/ and the second verb which is v2 or the light verb is /ɖeləkəjə/ together form a single verbal unit as /pʰɑrə ɖeləkəjə/. This is the standard representation of compound verb construction in Magahi. Therefore, this tag set designates only the compound verb constructions of the data.

However, there are also two other subclasses in this category for which separate tagsets has been further classified for each. These are Explicator Compound Verbs (ECV)

and Reverse Compound Verbs (RCV). Both of these subsets along with their respective tagsets are discussed in below in detail: -

4.3.2.4.1. Explicator Compound Verb (COMP_EXPL)

Explicator Compound Verb or ECV, as already mentioned, is a sequential arrangement of two different verbs, where the main verb carries the core meaning, and is thus, accompanied by a second verb, v2. This second verb or v2 gets delexicalized later in the construction and plays no role in transmitting meaning for the sentence (Abbi and Gopalakrishnan, 1991). The designed tagset mentioned above in this section is hence used to only tag ECV verbs in Magahi.

For example: -

62. उ आ गेलै का

u ɑ gelɑɪ kɑ

'Did he come'?

Here, in this sentence, the verb /-ɑ/ meaning 'to come' and /gelɑɪ/ meaning 'to go' are the two different verb forms that have been used. At first glance, one can feel and believe that the addition of /kɑ/ was made to create the construction in question format, but the moment we analyze this construction linguistically, it is found that this addition of /kɑ/ standing at the end of the sentence, is not only to create the sentence in question format but it also acts here as an explicator, as this /kɑ/ carries various information about the sentence, such as tense, aspect and mood.

It should be clearly noted here that it is the addition of which made this sentence a meaningful question, since it is used to doubt a person's arrival. The addition of this /kɑ/ denotes the grammatical feature of the sentence as it marks two different aspects here wherein one is uncertainty while the other is perfectivity. As per the discussion, it is the tagset which only marks those corpuses that have properties of this category.

4.3.2.4.2. Reverse Compound Verb (COMP_REVS)

Languages, in general, have unique properties to represent actions in a variety of ways. To specify this statement in context of Magahi, this separate category was decided. In

addition to a few different verb forms such as prepositional verbs, phrasal verbs or inflectional verbs, there is a possibility that Magahi might uses these special forms of the complex predicate structures, to make the purpose of communication easier to understand. The idea of reverse compound verbs has been first noticed and pointed by Hook (1974). It was him who noticed that there are certain sets of compound verbs, where polar and vector classes can be reordered or rearranged in Hindi to formulate new structures. This was then later followed in many South-Asian Languages like Marathi, Odia, Magahi etc. The forms of RCVs are quite new in Linguistics. However, after Hook who first noticed this, there were also other linguists who have worked in this accord. Kulsum (2014) and Poornima (2008) have also made an effort in this direction in order to identify this sort of construction. Till date there is no concrete reasons available as to why, when and how the verbs are being reversed but still only one fact relies that it a miraculous possibility and characteristics of some languages which allows this sort of construction. The tagset designed under this category is used to denote all sorts of reversive constructions of compound verbs.

While doing this task in the research undertaken, we realized that this special feature of reversal in Magahi is very hard to find. During this research, the existence of RCVs was thoroughly tested according to the linguistic standard provided by Das (2006, 2013 and 2015). Even after doing a thorough and deep investigation of the corpus we found that the occurrence of RCVs in Magahi are not only tough to find but also not acceptable. However, a few instances from the data collected was somehow tagged with the designed tagsets of this separate class. this was done after comparing Magahi with Hindi and wherever, similar construction was found, they were tagged under this category. This was done to investigate if the tool shows some overlapping features or not.

The reordering structure of verb order in Magahi could be understood through some examples given below: -

63. श्यामवा ठंडा के चलते नेहाली में घुसल हय

sjəməvɑ ʈʰə̃ɖɑ ke cələt̪e neɦɑlɪ meⁿ gʰusələ ɦəjə

‘Shyam went into the blanket because of cold.’

Here, in this example, you can see that it is a sentence with normal arrangements of compound verbs, with v1 and v2 in their respective places. But at the same time, if we try to

rearrange the sequences of these two verbs, does that change or affect the meaning of the sentence? To understand this mechanism, let's do it with the same illustration, rearranging the positions of these two verbs. After the rearrangement, let's see this again: -

64. श्यामवा ठंडा के चलते नेहाली में जा के घुसल हय

sjəməvɑ ʈʰɔ̃ɖɑ -ke cələt̪e nehɑlɪ meⁿ jɑ -ke gʰusələ həjə

'Shyam went into the blanket because of cold'.

We have seen here in example (64) that even the rearrangement of verb positions does not make any changes to the core meaning of the sentence. However, this reordering feature in Magahi requires an additional element to complete its formation. Here, in this sentence the addition of two additional elements such as /jɑ/ and /-ke/ were required to complete this construction meaningfully. Therefore, one can understand that such constructions in Magahi are conditional. The collected corpus for the undertaken research does not show many instances of this type, but whatever instances of this type were found, the define and assigned tagset mentioned here for this class will tag each of them after thorough examination of the data.

4.3.2.5. Serial Verb (SV)

This particular set of tags is designed and intended to denote only those sets or subsets of Magahi data that are formed by stacking of verbs. This set of classes will contain only those instances that will have verbs in continuity. Furthermore, it is a set of tags that help the machine to identify special syntactic characters where two or more verbs are joined together in a string format. It will denote those CP types that are bound together in a single clause. Such constructions are easily recognizable in many foreign languages as well as in few Indic and South Asian languages like Hindi.

This class of designed tagset will only denote and tag all SVC constructions which are encountered by the machine. However, there are also four different sub classes of tagset namely serial verb1 (sv1), serial verb2 (sv2), serial verb3 (sv3) and serial verb4 (sv4) has been created. However, there exists only one main class of this type while the other three will denote the other serial verb constructions which are in continuity and that are also present to denote different serial constructions of the sentence.

Languages like Hindi, Marathi English and several others, Magahi also have serial verb constructions. It is the reason why this exact set of separate tag was developed to deal with all such types of data having serial verbs in their structures.

Linguists like Butt (1995), Bowern (2006) and Seiss and Bukhari (2009) are some prominent names who have worked on this convention to point out some key differences between a complex verb construction (CVC) and a serial verb construction (SVC). According to them, these two shares the same characteristics as they both contain two or more verbs that follows a proper sequence. But it is their semantic aspect that sets them all apart from each other. One of the core differences between these two constructions i.e., SVC and CVC which both these linguists pointed out is that a SVC denotes two different events since it requires two serial verbs for its constructions whereas a CVC denotes a single action or work which uses two separate set of verbs for this task. Therefore, the process of SVC is exactly the opposite to that of CVC.

Hence, keeping these peculiarities in mind, the task of tagging the data must be done carefully as any mistake committed in the same can distort not only the selected corpora but also the entire process thus finally affecting the output of the tool. An example (65) has been mentioned below to understand as to how a serial verb construction (SVC) tagset actually works: -

65. रजुआ अपन घरे हमरा खाना बना के खिलैलकय हल

rəɟuɑ əpənə gʰəre həmərɑ kʰɑnɑ bənɑ -ke kʰɪlaɪləkəjə hələ

‘Raju cooked the food and fed me at his home’.

Here, in the example mentioned above, we can see that it is an example of SVC where there are three consecutive verb forms in a single sequence. This is not to be confused with the term /-ke/ as this is not a marker but an attached verb meaning "to do" in English and /kərə/ in Hindi. This helped tie the two separate sets of verbs together, allowing them to form a meaningful construction. Similarly, the idea of using /hələ/ in the last sequence of the sentence is to denote the timestamp of the given sentence. Here in this sentence, it is used to indicate that the work has already been done. It is used here to indicate that it is a past event. SVC has four other distinct subsets of markers that are decided based on the need for the data. These are assigned the annotation symbols sv1, sv2, sv3, and sv4, respectively. Let's

understand both with the same example of Magahi. The four other distinct subsets of SVC markers have been designed on the basis of the requirement and need of the data. These are assigned with the annotation symbols as sv1, sv2, sv3, and sv4, respectively. Let's understand this with the help of same example of Magahi.

66. रजुआ अपन घरे हमरा खाना बना के खिलैलकय हल

rəɟuɑ əpənə gʰərе həmərɑ kʰɑnɑ bənɑ -ke kʰɪlaɪləkəjə hələ

'Raju cooked the food and fed me at his home'.

This is a step ahead process of the previous example. In this similar example here, the presence of constituents like /-ke/ and / hələ/ are used as adjacent verbs and tense characters, respectively for the sentence. But at the same time, is it not surprising how a tense can be marked without the presence of a verb? Therefore, considering this criterion only the other three subsets of SVs emerged.

Now, considering this phenomenon, let's understand only the last half of the given sentence, which is /bənɑ -ke kʰɪlaɪləkəjə hələ/. Here one can observe that the sentence is formed by using three different sets of verbs, including /bənɑ/, meaning to cook, as one /-ke/ as second part, which is the modification of the Hindi verb /kərə/ meaning 'to do' in English and /hələ/ which is the third set. This is the another modified past tense form of the verb 'is' and 'am' in English and /ʈ̪ʰɑ/ in Hindi. When all these three are linked together, they successfully created a meaningful sense of the sentence and thus completely framed a entirely complete set of serial verb constructions putting all these three sets in continuity to produce a meaningful serial verb. Likewise in Magahi there can be instances like /kʰɑ lelɪjo hələ/, /suʈ̪ə gelɪjo hələ/ etc.

With this situation in mind, these three different subsets of serial verb tagsets were designed. This has not only made the tagging task easier, but has also helped in identifying such cases separately in the data with maximum accuracy of each set along with the possibilities of its ambiguities if any, once the model is ready. A detailed outline of the defined set of tag sets can be seen and understood from the table given below along with the appropriate examples.

S. No.	Category Top Level	Subtypes (Level 1)	Categorical Level	Annotation Convention	Example	I-Trans Description	English-Translation
1.	Complex Predicates		CPs	CP			
1.1	Conjunct Verb		CV	CP_CV	उ हमरा पसंद कर हय	u həmərə pəsⁿɖə kərə həjə	He likes me.
1.1.1.		Noun Conjunct	CONJ_NN	CP_CONJ_NN	रीता अप्पन लइकन के ध्यान रख हई	riɽa əppənə ləikənə ke ɖʰjɑːnə rəkʰə həi	Rita takes care of her children.
1.1.2.		Adjective Conjunct	CONJ_JJ	CP_CONJ_JJ	रीता अपन घरवा साफ़ करलई	riɽa əpənə gʰərəva sɑpʰ kərələi	'Rita cleaned her house.'
1.2.	Compound Verb		COMP_V	CP_COMP_V	राजू किताबिया फार देलकय	rɑju kɪɽəbɪjɑ pʰarə ɖeləkəjə	Raju tears the book.
1.2.1.		Explicator Compound	COMP_EXPL	CP_COMP_EXPL	उ आ गेलै का	u ɑ gelaɪ kɑ	Did he come?
1.2.2.		Reverse Compound	COMP_REVS	CP_COMP_REVS	श्यामवा ठंडा के चलते नेहाली में जा के घुसल हय	sjəməvɑ ʈʰə̃ɖɑ ke cələte nehalɪ meⁿ gʰusələ həjə	Shyam went into blanket because of Cold.
1.3.	Serial Verb		SV	CP_SV	चलल आ रहले ह	cələlə ɑ rəhəle hə	Is coming
1.3.1.		Serial Verb 1	SV1	CP_SV1	चलल	cələlə	Walk (-ing) form
1.3.2.		Serial Verb 2	SV2	CP_SV2	आ	ɑ	come
1.3.3.		Serial Verb 3	SV3	CP_SV3	रहले	rəhəle	Is am are (-ing form and tense marker)
1.3.4.		Serial Verb 4	SV4	CP_SV4	ह	hə	Is, Am (perefectivity marker)

Table 2 Annotation and tagging guidelines for Magahi CPs with appropriate tagsets.

4.3.3. Tools Applied for Data Annotation

This section of the chapter discusses the tool used for data annotation. The Indian Language Corpora Initiative Annotation Tool (ILCIANN) has mainly been used for this task. The next subsection discusses the tool and its functionality in detail.

4.3.3.1. Indian Language Corpora Initiative Tool (ILCIANN App V2.0)

Indian Language Corpora Initiative Tool (ILCIANN App V2.0) is a tool that is being extensively used for performing the annotation task of almost all Indian and South-Asian languages. This tool is generally used for not only annotation task and resource-creation of Indian languages in NLP but also for the mass level of crowd sourcing in all ML and NLP related works.

So far, a lot more corpus has been collected through this app not only in Magahi but also in many other Indian languages like Manipuri, Hindi, Punjabi, Odiya etc. Along with this there are also languages from Bihar such as Bhojpuri and Maithili to which this been seen as quite resultative and positive. This tool is a server-based web application that is being used for all sorts of annotation task such as Parts of Speech tagging (POS tagging henceforth) Multi-Word Expressions tagging (MWEs) and many other NLP related operations. Since we all know that it is a tool that functions on web-based applications therefore, the server upon which it works is known as Java Servlet Pages (JSP) that is based on another application of computer science known as Java. All these apps and pages runs on one common server which is known as Apache Tomcat.[12]

The tool was first used for Hindi language in order to create and build large annotated corpus that could be used extensively in many other domains apart from linguistics such as tourism, medical and many more (Jha, G.N. 2010). As the primary task Hindi was the language which was tested at first on this tool showing favourable results. It has an ability of auto tagging the data with the appropriate tagsets supplied. However, there still is a strong possibility that it may tag the wrong tokens or the incorrect tagsets to a particular data which can then be manually corrected. The later method applied in this entire process can help in increasing the efficiency of the tool which can help the researchers in obtaining maximum accuracy. The process of tagging the corpus in this tool is entirely based upon the linguistic categories and task that has been undertaken, which means the categories of tagging and aim

[12]Refer section 5.11.1.2 of chapter 5 for details

of the task must be specified in the very first step in order to avoid any errors. This is equally important because this tool is very commonly used for all sorts of NLP related tasks in Indian languages.

Fig 7. ILCIANN Tool App V2.0

4.3.3.2. Training, Testing and Evaluation of the System

This section of the chapter deals with the data training, testing, and evaluation mechanism. This involves the duo process of manual and machine learning techniques, where all the different tags of Magahi CPs in are further evaluated by comparing the machine-provided tag results with the manually generated tag sets. This is an important step of the research as only this step will help in creating a gold data set[13] to be used for further and final testing, which will then be followed by interpretation. In this step, the training data set differs significantly from the data set that is ultimately to be used for the evaluation. It should be noted here that the automatic tagging evaluation is carried out by the Support Vector Machine (SVM) tool.

This makes it clear that this section is about the SVM tool which has been used as a language research model for evaluation process of this task. The testing and evaluation task was performed with more care on the data so that there must be no errors in the tool, but it was obvious that some errors did indeed appear in the final process. All these errors and their proper assessments will then be discussed in detail separately in the upcoming chapters of this research with many different levels like problems and challenges, correct vs. inaccurate,

[13] Type of dataset which is annotated and tagged for analysis to obtain desired output

ambiguities vs. non-ambiguities and known vs. unknown. Here it is important to note that the two distinct classes of errors mentioned above, namely known vs. unknown, are those that occurred due to the use of gold[14] vs. raw datasets.

[14]ibid

Chapter-5

5. Natural Language Processing and Computational Framework

This chapter discusses the NLP and computational framework developed for this task. The sections and subsections of this chapter deal with a detailed structural presentation of computer and NLP tools and techniques.

5.1. Theoretical background of Natural Language Processing (NLP)

Languages or speeches are the best method of communication which uses all types of communication systems such as reading, writing and of course speaking. But in the field of COLING and for all NLP-based tasks, these are performed by artificial intelligence or AI-based applications. In order to make both the human communication system and the computer applications workable for humans at the same time, it is only the area of AI through which the communication between the two can be well established. Therefore, in order to carry out all these tasks, we need to know the basics of AI, of which NLP or else the applications based on these systems are a sub-field.

Technically, the only task of NLP-based applications is to perform the related or the concerned task and make the entire data workable as per the requirements of NLP applications required for the concerned field. It does this by analysing and processing large data sets collected with the help of NLP applications. In NLP, as the name suggests, the human language dataset is transformed into computer-based datasets that are easily understood by the machine and is later being used for easy understanding of the machine to perform the given task.

In short, it transforms the human-readable language datasets into an information-based electronic record for computers.

5.2. Historical Background of NLP

Unlike all other areas and subjects of language and learning in general linguistics, NLP also has a certain history that has helped all of us to develop better and better over time. The idea of research in the field of NLP first arose in the early 50s when it was first used as a means of machine translation. It was also a period of resurgence and excitement when the idea of speech recognition was first introduced by using the NLP platform. During this period,

investigations and identification processes were first performed by machine translation methods Boost and Richhens (1949). Later this was further developed and in 1954 an automatic translation system from Russian to English was introduced. It was a research demonstration successfully conducted during the IBM experiment in Georgetown University, USA. It was in this year only that the publication for machine translation was presented for the first time.

With all these ongoing developments in NLP and ML in general, the publication of journals in the field of ML was subsequently also started. The first publication was done in 1952, which was then followed again in 1956, resulting in the second publication for the same. Later, in 1961, it was the work of the Teddington International Conference on Machine Translation of Languages and its Applied Analysis (TICMTLA) that turned out to be a breakthrough for the entire NLP field.

Mentioned above are the most important developments that have taken place in the field of NLP and ML, which of course are now considered as the phased manner of different fields of NLP developments. However, there are several other phases as well that came into the picture later and have contributed a lot to the development of the NLP as a separate and emerging research field from time to time. Some of them are discussed in detail below: -

(i) Second Phase (Late 1960s to 1970s): - This was basically the AI influence period. The work done during this period in the field of NLP was mostly entitled towards meaning representation and not only towards general orientation of NLP. This implies that it was a period that focused primarily on knowledge representation of all the tasks that were undertaken in this field. Until this period the field of NLP developments were now capable of addressing the issues based on knowledge and data representations by using AI platforms. It was this period only wherein the idea of analysis and evaluations into the systems of NLP were developed. Later on, the systems of NLP representations were then advanced to question- answering through which the knowledge representation of the machine is being tested through the modes of language interpretations and data inputs (Minsky;1968).

(ii) Third Phase (Late 1970s to 1980s): - It was also referred as Grammatico-logical phase. As the name suggests, it was the period where the idea of NLP advancements developed a little more by testing the knowledge of the

machines with the use of logics and reasoning through in AI platforms. This use of knowledge representations in NLP mode has helped people in developing systems like sentence processors to what we commonly call as parsers[15] today. All such developments during this period in NLP field took the researches of NLP a little further towards developing an online lexicon for any language, which was then advanced later in the year 1980.

(iii) Fourth Phase (The 90s period): - until this period the systems and developments of NLP and its related field were capable of identifying the lexicons through corpus identification. It is due to this reason that it was also known as lexical and corpus phase. Till this period the idea of NLP and its uses were so advanced that it can be used quite extensively for the purpose of language processing and several other language related NLP tasks by using machine learning algorithms and techniques.

Therefore, it is evident to note here that the step-by-step process of NLP and AI developments has shown an increasing influence into the masses which helped in creating a revolution in this field.

5.3. NLP as a Field in Computer Science and Linguistics

Natural Language Processing or NLP is a computationally and technically advanced technique widely used to analyse text of any language at one or more different levels. These levels of analysis can be linguistically divided into many different forms, such as phonological, morphological or even syntactic. In linguistics, the field of NLP is seen as a link between language engineers, what we now call linguists, and computer scientists. With today's idea of NLP together with linguistics, the demanding task of annotating human speech, its representation on the machines, as well as the successful processing of the data to obtain valuable results is solved quite well using a variety of methods NLP applications. This means that a wide range of NLP applications are widely used not only for data mining but also for the purpose of data modelling and processing to develop many different language tools such as morph analysers, sentence modelling, text summaries and much more the linguistically oriented NLP tasks much easier.

[15] Tool to sperate text into small groups or linguistics items as per grammar rule.

5.4. Developments in NLP

Studies of the past show that several major works have been carried out in this concerned field. The first major development in the field of NLP took place in the late 1940s. It was the time when two of the most prominent researchers, Booth & Weaver, propagated the idea of research and projects not only in the field of NLP but also in the direction of MT. This type of idea was first developed in 1946. During this time, linguists were able to use the idea of cryptography and information theory techniques for information retrieval and language translation. After that time, there was a major breakthrough in NLP research around the world, especially in the United States (USA). A few years later, the idea of using linguistic syntactic modules to identify sentence structures was introduced by the famous linguist Chomsky. It was Chomsky who came up with many new syntactic theories during this period. This paved the way for several new projects and ideas in many different languages using NLP and AI.

After the idea of sentence structure identification had been well realized during this period, it was now time to deal with the question of deriving the meaning of sentence structure. This was the time when NLP tasks mostly focused on semantic questions. Therefore, the idea of traceable meaning was developed, which was entirely dependent on the proposed computational models and the solutions achieved with them. It was Chomsky who established the language transformation model in 1965, used primarily for linguistic competence. In addition, many other prominent linguists worked for this agreement. Some of them are Fillmore, Quillian and Schank. All of these linguists mainly focused on transformative generative concepts of the grammar of languages. Her work focused on semantic networks in NLP, case structure grammar and conceptual dependencies of languages. All these linguistic developments in NLP contributed to developing the idea of semantic representations of languages in the field of linguistics through the use of NLP techniques.

With all these ongoing developments from the 18^{th} to 19^{th} century and then later to 20^{th} century, one can see that the NLP developments, with the help of computer techniques and technologies, went a little further towards several linguistic areas such as semantic phenomena, syntactic ideas, and many others. This helped achieve communicative goals for many different languages, most notably Indic, which then further helped to address many different language-oriented problems and challenges through NLP and COLING tools and

technologies. With all these developments, the task of making available an enormous number of electronic texts for languages such as Magahi and several other underserved languages that are still overlooked in the field of research and literature was made available.

Therefore, the increasing availability of online language resources over the past two decades through the use of NLP techniques has proven very helpful and resourceful for the field of linguistics. Also, the mass development of the internet has helped researchers to develop resources and techniques not only for foreign languages like Arabic and English but also for many other Indic languages like Hindi, Odia, Bengali, Marathi along with several other less recognized languages like which is at the marginal graph of extinction.

5.5. Different Linguistic levels of NLP Applications

In the field of science, the complexity of human language can be represented by many different levels. Each level of complexity of human languages can be easily approached through several different stages depending on their complexity types, like from lowest to highest or vice versa. The current research task will follow the same procedure of subdividing the different linguistic layers of NLPs for data processing. It will follow a set pattern of linguistic hierarchy to easily commute this task without error. This follows a detailed three-way process to perform the task at three different and basic language levels for the following reasons: -

(i) Any task undertaken for the research in NLP does not requires any higher representation beyond linguistic concept because, if this happened the original essence of the task, which of course is to identify any linguistic element or else to develop any NLP resources may get missed for the concerned language.

(ii) In order to make the application functional and also to have a better outcome of the same, the data of the concerned language must be checked and tested with the tool and for this task a thorough process of rigorous cross checking must be undertaken linguistically at different level for its authenticity.

(iii) The undertaken linguistic data collected must get tagged in smaller units as per the task required for the research. This can be done at the level of word-tokens and not on the sentential level. If this is being done on sentential level, it might have made the task a little hedgy and trivial for sure.

The main objective why these three crucial steps have been taken into account when performing an NLP task is that any tool or NLP application one wishes to develop must first be fruitful and relevant to the language in question while conforming to a rule must follow - governed approach as opposed to general linguistic grammar on the other hand. In order to make the latter process effective, there are certain distinct linguistic layers upon which the entire NLP rule-making method for each language depends. This is the very first level of linguistics on which the whole process of creating rules for the tool depends. The single function for each of these linguistic levels is the same as in general theoretical linguistics. Each of these different language levels will be discussed in detail as follows: -

(1) Phonological Level: - In linguistics Phonology as we all know are the field of linguistics that deals with the speech sounds of any language. Unlike general linguistics COLING researches concerns phonological principles and rules, but the only difference that implies in COLING with phonological researches are that it uses the digitalized platforms of NLP systems to determine and interpret the signals and sound waves of languages.

(2) Morphological Level: - As we all know that the field of morphology in general linguistics is nothing but a branch that delas with smallest linguistic unit of any language to what we call as morphemes.[16] This is a process where the root word gets adjoined with any smallest linguistic elements such as infixes[17], suffixes[18] and prefixes.[19] Once these elements gets adjoined with these smallest units, it tends to form a totally different linguistic word forms or elements that may be old as a linguistic form but may also be new at times for a language that never existed before in the same.

(3) Lexical Level: -Under this category or linguistic level the orientation of meaning concerning to a specific lexical item is dealt with. For the field of COLING the task is mainly directed towards the idea of POS taggers. This is being done on the basis of either the word level or else the sentence level. Here, the researcher has to decide as to what linguistic element one refers as one lexical item. This means, it could be either a word or else a sentence at times, which is being regarded as one lexical item. This decision also varies as to what idea or task one is going to deal with.

[16] smallest grammatical unit of word
[17] affix inserted inside a word stem
[18] affix inserted after a word stem
[19] affix inserted before a word stem

During this entire process the processing of each lexical item is done by keeping in view the idea of lexical ambiguity[20], which of course is a possibility. Later on, one can choose to discuss all these ambiguities further if needed or required, just to avoid the error of the language tool that has been made for a specific undertaken task of a concerned language.

(4) Syntactic Level - In general linguistics as we all know that syntax deals with the grammatical structure of sentences of any language. It is represented through the structural dependency relation between the words and the sentence pattern of any particular language.

In the field of COLING or NLP the syntactic level relations are being represented through parsing methods. In this the structural patterns of sentences are represented through parsers. This means that, it determines the structural pattern of sentences or texts through immediate analysis of the constituent words, on the basis of the concerned grammar of that language. Through parsing techniques one can analyse the strings or symbols of any language with the help of the grammatical rules, a language follows and if there still is an error in the entire parser, the tool automatically detects the same and reports it as a syntactic error. A parser in NLP alerts from occurring the same error time and again, which helps in further enhancing the efficiency of the machine, in order to have better results with reduced efforts.

(5) Semantic Level: - This is another next advanced level in COLING after parsing in syntactic level. This is so because, as we all know that since semantics in general linguistics is a field associated with meaning identification, therefore, here in NLP it acts in the same manner. The only difference here is that it just uses the advanced technological aids to do so.

In this level the COLING based applications not just focuses on the human interactions and the occurrence of words but also towards its meaningful essence as well. In COLING all such levels are identified by tackling and preparing an auto list of ambiguities[21] and different co-ordinations a token or word has in any language. In COLING the NLP techniques simply helps in establishing the logical relation between the word and the sentence with the help of semantic analyser, which establishes the relations between the two in the form of numerical codes and once these codes are functional, they are being tested on

[20]Lexical ambiguity is the idea where the meaning of an individual word or phrase are being used to several different contexts. It may have more than one meaning depending on the context.
[21]Instance of having understood something in more than one way

the machine as to how many different permutations and combinations a token or a word can go with along with the logical relations between the duo that justifies the connection. All such operations are being performed in NLP through different logical operators such as 'AND', (&), OR, NOT, XOR (exclusive or) [22]etc., to what we call as BOOLEANS[23] in machine learning. The only challenge that relies in the entire task here is to tackle the different sorts of ambiguities which exists with the words that has many different meanings or senses.

The idea of semantic level systems in NLP or COLING is broadly used today for developing and designing automatic question answering systems in the field of AI.

(6) Discourse Level: -Under this level of NLP, the linguistic structures of any concerned language get revealed with several different applications, that supports many different tasks of NLP. This basically deals with the idea of making sensible connections among different words and sentences of languages. The discourse level in NLP generally tries to makes possible connections between the words and the component sentences to which a word is being attached with. At this level the field of NLP basically applies the idea of anaphora resolution wherein, the linguistic entity of any language is generally referred through an anaphor, which is most commonly a pronoun. However, this is not only limited to a pronoun but can also be some other and different grammatical category.

The NLP techniques used in the discourse level mainly gives birth to machine techniques like chatbots[24], wherein the applications like chatbots are so well trained that it can make your task much easier while interaction through human language systems. These chatbots basically functions on the coherence of the concepts. For example, if you ask the chatbot to help you in choosing a product at a departmental store, it can guide you better, but if in the same store setup, you ask the chatbot to make an option for a book choose, it may fail because, the training of the chatbot has been done for that particular set up only to which it is going to exposed to.

Therefore, it is evident to note here that when any machine learning chatbots that have been developed through discourse modelling in NLP or AI is used for a particular field or set-up it will respond only as per that setting as the discourse structure modelling or text

[22] Different forms of boolean operators to perform functions
[23]Method of system denotation that has tends to give one out of two possible outcomes
[24]It is an AI based machine learning technique based on computer programming which communicates through text or sound methods. This is used generally in phone and customer services fields to support queries.

processing of that particular AI machine has been done and trained for a specified field only based on the inputs of a particular setting, wherein it is going to be used. However, the improvements in the same are still to be made to make it much user friendly for across the fields of normal social setting.

(7) Pragmatic Level: -Being the last and the final linguistic level in NLP, the sole focus of this level is upon the correct use of any language systems in a specified situation. Under this level the NLP machines solely focuses upon the contexts rather than the contents for making quick reactions of the developed applications. This level of NLP application development in linguistics is concerned primarily with the purposeful communication of language according to different prescribed situations. It gives more weightage to the contexts instead of contents. This means that it focuses mainly on the level of understanding. This applies to more practical knowledge such as understanding intentions, motives or goals as prescribed in any text of a certain concerned language.

For the field of NLP and COLING the pragmatics level is mainly not only used to develop the information retrieval systems but also to improve the chatbot based question answering techniques. This is so because, till discourse level we are only capable of developing the question answering techniques in AI but, as pragmatic level in NLP is a bit advanced one, therefore it will help in improving the same by one-to-one mapping technique and logical understanding of the same. During this entire process, the AI based chatbots are so well developed and trained that it cannot only understand the sole goal as to why and for what a certain query is being made but also the related context of the same along with its time and location if given.

The pragmatic level in NLP and its AI based application development gave a major breakthrough in the field of NLP as this is not only creates a link between information retrieval systems with AI applications, but also clearly deal with the sole purpose of meaningful and logical elicitation of the same as targeted or planned. However, there are still many other different ongoing linguistic challenges at every point that are there and which also needs to addressed thoroughly and systematically. All these could only be tackled by future researchers through several important and major steps such as undertaking absolute data structures, applying right algorithms in application-based task and last but not the least is

to mapping the entire process of the AI application from time to time to make it more functional, accurate and efficient.

5.6. NLP Approaches

This section of the chapter will discuss in detail several different approaches in the field of NLP upon which the entire AI system is being developed or functions. These are categorized in many different levels such as connectionist, Rule-based, symbolic, hybrid, statistical and many more. All these approaches to NLP systems were developed significantly with time. For the very first level, two among all co-existed already in NLP since beginning. These are symbolic and statistical. Later, in 60s, connectionist approach was further developed on the basis of which a few AI based communication systems were developed. After this, it was the time of symbolic approaches that ruled the AI era for some time.

With the flux of time and concerned recent developments in the field of AI the COLING resources were now able to develop some more advanced resources in order to deal with more critical concepts in the field of languages and computer science. This idea of dealing with some more advanced concepts of the real world through AI means gave birth to a more advanced level called statistical which received massive attention and popularity around 80s. In connection with this, it was the connectionist approaches which got evolved till this time and which took massive attention of AI learners in COLING field. Among all these levels of NLP the statistical approaches deal with lower levels of analysis whereas the symbolic approaches deal with the maximum or the higher levels on the other hand.

5.6.1. Symbolic Approaches

Languages in general are learned and understood through logics. As, mentioned every language in the universe do have some contexts and all contexts across languages has different logics. This approach of NLP generally uses the ‘deterministic’ idea in order to make the machine capable of understanding any human languages just like the same way as humans do. For the field of AI and NLP the symbolic approaches are massively used in order to develop the chatbots, which makes the machine understandable of human languages as to how and what to read write or understand. The idea of symbolic approach basically works on the basis of rule-based learning and language inputs along with the possible human

intervention in order to check those rules and imply the same for coding and programming the by establishing a meaningful co-relation between the input and output given.

In order to develop any NLP based application for any language-oriented task in COLING the applications are being trained manually through a set of identified rules of that particular language. Symbolic approaches are being used extensively at every stage of NLP mostly after 1960s for several different research areas and application developments such as text categorization, lexical acquisition, ambiguity resolution etc. In order to develop several NLP based applications based on this approach, it uses several techniques such as logical programming, decision trees, conceptual clustering, ambiguity resolution methods etc. All these use a specified and identified algorithmic approach in order to attain the best positive result in this approach.

Therefore, it is evident with the fact here that a symbolic approach in NLP always provides a deep analysis of all tested linguistic parameters on all levels of NLP in order to have the positive outcome with maximum accuracy. Symbolic approach in short, uses a very old and traditional idea of developing the AI based NLP systems.

5.6.2. Statistical Approach

Statistical methods or approach in NLP is one of the oldest and empirical approach in COLING. It is one of the prominent approaches in NLP that have paid much attention towards the deep learning and analysis of any language and its grammatical patterns. This method of NLP basically involves in collecting and assembling data with statical inference by studying the patterns of any language. This means that it takes the data of any concerned language with unknown probability distribution and then validating it by applying some inference of the same language family or the same language group. The statistical approach of NLP strongly focuses upon deep learning techniques, which helps in developing a proper neural network[25] in order to perform a duo task such as taking inference from a similar language on the one while, developing an end-to-end parallel system on the other.

Through this approach one can extract data or information from a large text or large files. This is widely used usually to develop a generalized model of language systems which does not requires any significant knowledge of language patterns, which means this approach uses the observatory methods for data processing and evaluations. Out of several statistical

[25] Computational model developed on the basis of human brain and AI techniques, refer section 5.6.8 for detail

approaches in NLP or COLING, HMM (Hidden Markov Model)[26] and SVM (Support vector machines)[27]are the most widely used. The approaches of statistical models are widely used today for plenty of NLP operations such as Speech recognition, Lexical acquisition, Parsing, POS tagging, Machine translations, and Grammar learning approaches etc.

In short, the statistical approaches seem to have the backbone of the entire COLING field, which helps in developing several NLP based language systems across languages.

5.6.3. Rule-Based Approach

As the name suggests, it is an approach that follows a certain set of defined and identified rules that are man-made. These rules are then being applied to machines in order to manipulate and train data sets in order to reach the desire goal in NLP applications.

It is a bit similar to statistical approaches in NLP, which focuses solely upon the grammatical rules and patterns of a particular language. This includes all sorts of extraction methods which comprises of several stages of evaluations of the data sets of a particular concerned language. The evaluations of these data sets are being done by the help of grammatical categories of that particular language. As stated by Brill (1992), the rule-based approach in the field of machine learning and NLP is just to master the limitations of rule-based approaches which creates hindrances in identifying and processing any natural language through the means of machine learning.

One major disadvantage while using the rule-based approach of machine learning in the field of NLP is that the final result received after the entire process of language processing through rule-based may be incorrect or may be not acceptable sometimes as this method follows manual evaluation techniques in order to identify and process the raw data, which may require corrections and continuous checks for correct evaluation of the same. The entire process in this method is a long run process in order to obtain accurate results, therefore it is not being used on large extent today as it is much time consuming. Another important reason, why the rule-based technique is not being used extensively today because, it is too much expensive, and also much vulnerable of committing errors in analysing data sets. Therefore, today with the onset of time this method, has been taken over by several different advanced techniques, but at the same time this is being used for several other traditional languages like Sanskrit, wherein it works fine with certain limitations defined for

[26] HMM are statistical based model used for machine learning and NLP tasks refer section 5.9.1
[27] NLP based supervision-based AI learning model refer section 5.9.2

the same, which needs several long processes of manual data set analysis, wherein the possibility of having errors cannot be overlooked.

5.6.4. Connectionist Approach

The field of NLP requires high level symbolic and processing capabilities which includes a large scale of programming of data sets by means of manipulation of not only data sets but also the constituent structures. The connectionist approach in NLP works on the sole idea of networking wherein it interconnects the simple units of any languages that has yet not been processed with the data sets of such language connections upon which a particular language has to be trained. This means that the raw language data is being connected and processed with data of similar language that has been adopted as a model for that language to be trained with by establishing possible connections among the two languages. Out of these two languages one is considered to be as pre-stored on the basis of which the other is said to have been trained by means of knowledge stored in previous one. This entire process of comparing and training language data sets in the field of NLP is known as computation. Connectionist model in NLP systems is a bit identical to statistical approaches, wherein it develops new language models from pre-existing linguistic data-sets on the basis of which all new data sets of concerned language groups are to be tested.

The only sole difference in connectionist approach from statistical approach is that the later model ought to compare and combine the data sets by applying various different theories and principles and also allows some sort of modifications and transformations by means of manipulations in their logics or formulas. This approach of NLP is broadly suited for AI based applications like domain specific translation tasks, syntactic parsing, developing associative-retrieval systems etc.

5.6.5 Traditional Approach

In NLP related task the traditional approach is a step ahead approach to statistical approach. It follows several sequential key steps at basic and advanced levels to solve and handle NLP related tasks in COLING. It is the traditional NLP method that comprises of basic steps involving several distinctive tasks such as removal of unwanted corpus by pre-processing, featuring the refined textual with numerical representations. The numbering or ordering of the data is important as this will help the machine in identifying and differentiate between raw corpus and refined corpus.

This helps in better training of the system developed and yields novel and desired positive outputs. Though this entire process is quite time consuming but, at the same time is very crucial thus helping to obtain much better performance for any sort of undertaken NLP task. With this traditional approach one can easily reduce the instances present in the data supplied to the tool. This means it removes all sorts of unwanted elements from the supplied corpus such as punctuations or vocabs that are not at all required for the tool.

This implies that though it is a time taking process but it also reduces much of the human effort by reducing the size of the data by removing such extra linguistic elements, thus supplying the tool with only important information of the concerned language required for the task. This approach is widely used for several linguistic oriented purposes such as text summarization, thesaurus creation, creating good online external resources etc.

5.6.6. Corpus-based Approach

In the field of machine translation or COLING it is quite obvious to have a large text containing huge corpora of different domains. A close and thorough analysis of all these different corpuses has given an insight to identify these linguistic corpora data with different ideas by applying new approaches that get facilitated with the passage of time. This not only help in getting the corpora of the specified domain but also helps in working on the same by applying latest technologies.

On the basis of this Corpus-based approach Brown et al. (1988; 1990) has used stochastic methods to identify parliamentary debates and speeches in languages like English and French. Through this method he was able to arrange and align sentences of speeches in order to evaluate the probability of any word or sentences that may have occurred twice, once or else have been repeated several times. Initially it resulted with the analysis percentage of 48% but later, got improved to increase some more percentage by applying several statistical techniques and also by giving more data from inflectional morphology background, that helped in syntactic transformation of the tool, and also in dealing with some unsupervised sentence structures. Apart from this there were also several other approaches to corpus-base systems have been developed and studies such as WordNet by Miller (1998), VerbNet by Levin (1993), PropBank by Kingsbury (2002) and FrameNet by Ruppenhofer et al. (2006) etc.

Therefore, it is evident to note here that in a corpus-based approach system in COLING or MT data disambiguation at several levels can be done in order to achieve maximum accuracy of tool developed.

5.6.7 Hybrid- Approach

This is one of the modern approaches towards MT in the field of COLING. Here, the system set up is closely monitored in order to see its instances as it is a mix up of several modern techniques. Here in this approach the sole aim is not adopt the exact correct translation or to develop the exact and accurate system for translation rather, to transfer and obtain crucial and important elements in order to build the system and then process the task as required. It is most commonly used in techniques like scientific communications, technical writings, building bio-informatics systems etc.

5.6.8 Neural Networks

In machine learning or AI techniques, neural network is regarded as series of algorithms that tends to help in identifying the possible and meaningful relationships of any language undertaken with the data sets that exists already in the domain. A neural network operates in the same fashion as a human brain, wherein it mimics the already existing data-sets just like the brain does with the pre-existing set of information in order to process any new information. This means, in neural networking the processing of information is being carried out simply by imitating the pre-existing set of information that are there in the tool. In case, it gives any error in the process, then that particular set of data is being addressed and dealt separately in order to resolve the issue. The only exception in neural networking techniques while processing and analysing the data sets is that it is adaptable to change the data inputs any time. This helps in giving the best output results with no need of redesigning the data in between in order to get the output. This uniqueness of neural networking techniques of modifying the data sets in between the process not only reduces the human efforts but also helps in achieving the best possible results at any time.

Neural networking-based AI techniques are popularly used in trading platforms, financial services, marketing research, fraud detection systems and risk assessments. All these AI based neural networking techniques uses deep learning algorithms, which have shown massive growth in today's 21st century.

5.7. Brief Descriptive Sketch of different NLP Applications

This particular section of the chapter will briefly discuss all the possible applications developed through the AI techniques by applying NLP aids. It will also discuss the further ongoing developments of the same in this accord.

5.7.1. Information Retrieval

Information Retrieval or IR in AI is concerned basically with the sole idea of searching a specific document from a pool of files of documents wherein a large amount of data file is stored. It is also concerned with searching a specific information from the documents by creating metadata[28] of the documents. It helps the humans in making the search easier by creating relational databases which helps in searching the right and appropriate information from the data supplied.

For information retrieval systems World Wide Web, WWW henceforth, applications are the most legitimate and authentic AI based application which not only helps in searching and identifying the right content but also helps in getting the same with the help of authentic latest technologies. Until today, WWW are the most trusted search engines which are the most accurate and authentic IR based applications. It simply takes information or inputs from human and then gives output results by crawling hundreds of web pages as per the information received. One these web pages are available in their specific domains it then searches for the particular information or the pages that has been given as an input in step one in order to give proper, concerned and accurate results.

All these mechanism works on spider web like bot structures, which not only helps in identifying the correct information but also in downloading the same concerned web pages of the information supplied, which one can access later through the concerned links, in order to download or read in the web page mode. The same mechanism also functions correctly for many other information-based application systems like text retrieval, document retrieval, data retrieval etc. the only difference is all these have their own set of information or codes supplied which enable an authentic and accurate search by applying specific technologies.

[28]It is the by-product of the data collected as it describes the basic attributes of the concerned data collected through AI technologies.

5.7.2. Information Extraction

It is one of the most recent developments in NLP. One can say that it is one step ahead process in data and information extraction in NLP. Today as we know that data is a kind of capital that is going on increasing in massive numbers every day. It helps in creating new digital source of information with the help of new digital aids and products. This is being done in the form of news, information, corporate files, medical records etc that is being maintained and updated on a daily basis which of course gets overflooded with much of information every day. Therefore, it is quite difficult to categorize and extract the exact data from the sources required when needed. Hence, it is due to this reason, thew idea information extraction has been introduced which of course makes the lengthy human task of collecting the same much easier.

NLP today, makes this task easier by using computational aids and methods in order to process information from the data available which may be in written or spoken form, which are used today by the humans as a large source of communication, Assal et al. (2011). Apart from these mentioned above there are also several other processes that are involved in NLP to perform this task. Information extraction today, is one of the most recent NLP based application which can be used with a wide range of NLP based tasks such as question answering, data mining, visualization etc. this focuses solely upon extracting important information from the data collected by specifically tagging the vital key elements of the data. For e.g., here, in this undertaken task the CPs of Magahi are being tagged with their appropriate tagsets in order to identify them correctly once they are being trained and tested in the system at a later stage. Therefore, mis-tagging the data may lead here to load the result of the tool correctly.

5.7.3. Text Summarization

As mentioned earlier, a lot more tremendous information is available today in the world which are being updated, changed or modified in every second. Therefore, it is quite tough to maintain records of the correct information from accurate sources. But, with the advancements of technologies in the field of COLING this has been made easier by the help of text summarization.

Before we move on to its mechanism as to how a text summarizer works by using NLP aids, it is important to know what exactly summary means. Summary as we know is nothing but an idea of giving the gist of information by shortening the large without missing

important elements from the same. The sole aim of a text summarizer in NLP is to produce and present the original source text in to smaller version without missing any important and crucial semantical elements of the same. One of the most important reasons why this idea of text summarizer has been introduced today is to reduce the human effort of reading a large text thus saving time.

Text summarization methods in NLP can be well classified in two steps which are extractive[29] as one while abstractive[30] summarization as the other. Apart from this a text summarization process in NLP involves two different group stage of sharing information which are termed as inductive and informative. Here, inductive stage represents the core information or the central ide of a text which covers around 5 to 10 percent of the entire text while informative summarization of text on the other hand gives concise information of all the information in the text, which covers around 20 to 30 percent of the detailed source text.

All these processes of text summarization in the field of NLP involves four major steps which are as follows: -

(i) Topic identification- It is one of the most important steps involved in text summarization. The main aim of this step-in text summarization process is to identify the prominent and crucial information from the entire text. This step involves major techniques of identifications such as text positioning, identifying frequency of words, determining Cue phrases etc. out of all these Cue phrase identifications is one the most crucial step as this identifies the positioning of all the phrases used in the text in order to make the task easier and wiser and also to avoid the probability of repetition of the same if there is any.

(ii) Interpretation- This again is another crucial step involved in the process. In this step several important subjects and information of the text are being incorporated in the text in order to receive a well generated content, without missing any sole information.

(iii) Summary Generation- this is the final stage of the entire summarization process. The text summarization tool identifies all the necessary elements of the text and formulize the same in well- ordered or arranged manner in order

[29]Extractive Text Summarization- This consist of collecting and summarizing important information from important paragraphs and sentences that carries most crucial information of entire text.

[30] Abstractive Text Summarization- It is the technique of understanding the sole concept from a large document source and then expressing them all in a clear understandable natural language form.

to obtain a suitable and summarized text, from the machine but, before accepting it as a final one, one must go through all parameter checks in order to avoid any missing elements or mistakes in the same, so that it could give more suitable and accurate result in the machine thus reducing any error probabilities for the same.

(iv) Evaluation- During this entire process of text summarization, the evaluation process of the same is very crucial as in this process all relevant features of sentences, words or tokens used are not only decided and calculated but also being weighed properly and then an accurate value are being assigned to each token, words or sentences in order to obtain a final summary.

All these, are being performed using weight learning method, wherein final score value of each token is being picked up and is being used as per the requirement of the text. Evaluation step of text summarization is one of the crucial methods as this helps in picking the top ranked sentences for the text which are to be used in the final summary creation.

It is also evident to note here that for value assigning and picking the accurate word token for text summarization during the evaluation step two broadly important aspects are being used. These are extrinsic and intrinsic aspects used during evaluation process. Through intrinsic evaluation mode one can measure the quality of the summary using human evaluation methods which means this will be a manual task while, in extrinsic methods the same task is being done through the machines by using task- based performance method wherein, it uses the idea of information- retrieval techniques of machine learning.

Today, the most commonly used text summarization methods in machine learning are standard keyword method, Cue method, Title method, Location method etc. which are used wisely in machine learning techniques for not only summarizing the text but also to weigh its sentences accurately with maximum accurate percentage.

5.7.4. Machine Translation

The idea of Machine translation MT henceforth, in COLING delas with the application in computer systems for the task of translating one natural language text to another by applying COLING techniques. It is one of the oldest techniques of NLP which go

through all levels of NLP and uses all its techniques starting right away from smaller word-based approaches to higher levels of thorough analysis. Today a lot of different level researches has been done in this field. But, despite this fact it is also evident that it still has a lot of issues and challenges for the NLP tasks for many languages pertaining to general domain analysis but at the same time some success has also been registered under its name for several domain specific translations.

At present the MT research technology generally applies many different approaches in order to achieve its goal and all these approaches are directly dependent upon different pairs of languages involved and undertaken for MT task. However, as per Siddiqui and Tiwary (2008) and Hutchins (1993), there also exist several other MT approaches today, that helps in making this trivial task a little easier.

5.7.5 Sentence Understanding

It is one of the simplest tasks performed by NLP applications. Here in this stage, the developed NLP applications helps machines to understand and analyse natural human language through NLP aids. It helps the computer systems to make the human language understand no matter in what form it is such as written or spoken. In this process, the machine is made capable of understanding natural human language sentences through continuous and rigorous model training, which then helps in understanding its meaning with reference to preceding sentence and then relating the entire structure of sentences with the appropriate context in order to adopt a complete meaningful text. Once the entire process is completed it then moves on to another next level of sense disambiguation (see 5.7.8 for detail), wherein the meaning of the entire sentence is being thoroughly examined and compared along with the concerned text thus, giving a complete understandable text. All these entire goals of sentence understanding task is being achieved in three major steps which are semantic parsing, semantic classification and discourse modelling.

There exist several NLP applications that has been developed so far with the idea of sentence understanding and many more are still in the pipeline as the ongoing researches are still continuing. Some of the major NLP application working on this sentence understanding module are document classification, spam filtering and most importantly sentiment analysis.

5.7.6. Multiword Expressions (MWEs) Taggers

Multiword expressions MWEs henceforth, as we all know are one of the crucial areas of NLP identification system. It is a kind of linguistic element that occurs frequently in any language. Apart from the written source these can also be popularly noticed during spoken languages, but are being rarely seen in any formal texts.

In general, linguistic terms MWEs are refereed as linguistic or lexical items that can be grouped in to several simple linguistic elements containing different categories for each Sag et al. (2002). This means that every MWE contains and displays numerous multiple linguistic units such as syntactic, semantic, pragmatic etc. For general understanding one can refer compounds, phrases and many other different linguistic terms as MWEs. MWEs taggers have been developed in many different languages by different researchers. All of these have applied different techniques for the same. Kumar (2017) has applied rule-based technique for identifying Magahi MWEs from a large source of annotated corpus collected from different sources. The tagger developed by him has shown a maximum percentage of accuracy of about 81.57%.

In MWEs tagging a large amount of dataset has been converted into different list of words wherein each tuple of word has been assigned an appropriate MWE tag as per its requirement, nature and origin. For e.g. A total of nine class of tagset and sub-tagset has been assigned to the data collected for the data annotation, which is then annotated following a prescribed guidelines for the same, Sicky et al. (2017). Once the labelling of data set is done manually it is then being trained with an adapted computational model SVM classifier (Support-Vector Machines)[31]. Once this SVM tool auto-tags the MWEs data then again, the unsupervised data that has been left untagged by the machine is tagged manually in order to avoid any discrepancies at last. Therefore, a rigorous three-step technique has been followed in order to avoid any mistakes and also to ensure maximum accuracy of the tool. This has not been done only aiming at building a MWE identifier system but also to develop a successful model for languages like Magahi that are still under-resourced and are not much exposed to researches and technologies. This also helped in maintain an online corpus for the same, which would help the future researchers to fulfil their needs in terms of COLING researches of this kind.

[31]Refer section 5.9.1 of the chapter for details

5.7.7 Parts of Speech (POS) Taggers

POS tagging in the field of NLP or COLING deals with the centra idea of assigning part of speech markers to each of the word or corpus of a language. It is the lexical item that is assigned to each word or token collected by Nainwani et al. (2004). In other words, it is the process through which each word, token or corpus is being assigned with a text label in order to confirm that grammatical status of that particular token as per their respective morphological or syntactic categories, Hardie (2003).

In general terms POS tagger is the NLP based software system that helps in assigning the appropriate category to each word or token that are available in large amount. It works in such a way that it first reads the text automatically and then assign the desired POS tags to each token as per the requirement. The tagset assigned to the tokens are possible Noun, Verb, Adverb, Adjective etc. Toutanova et al. (2003). The tagging software helps in assigning the relevant and specific tag set to each token and passes the data to another lever for further processing and validation. It assigns the tagsets to each token or word in such a way that each tagset becomes unique and sometimes ambiguous in its own, mostly having no similarity of one category with the other. This implies that if a token has to be tagged as noun, then the software will tag each token of the same kind with different tagsets of noun only and not with any other tagset. The only evident thing to note here is that one must, clarify and identify the tagset to be tagged with token much before it is being supplied in the tool for further processing.

In NLP there exists several different taggers such as rule-based, stochastic and hybrid. POS tagging in linguistics is one of the most suitable ways of analysing the category of word tokens and assigning the same while deciding its characteristic feature. This is not only a detailed grammatic process rather also a semantic one as it helps in analysing the exact meaning of the concerned tokens, which of course makes the trivial task quite easier. Schachter (1985) and cited through Mitkov (2003). Till date a lot more POS taggers have been developed for almost all Indian languages, but the same lacks much behind when it comes to the regional languages that are not much exposed yet beyond its local boundaries. A few of them have already been discussed in chapter 2 of this research work.

So far there exist a lot more major NLP based POS tagger applications. A well-developed POS tagger tool can also be used as an initializer for too many NLP applications (ibid). POS tagger systems can be used widely in information retrieval (IR) NLP Systems. This helps majorly in indexing the textual data as per its grammatical requirements such as

Noun, Verb, Adjective etc. it is also used extensively for processing the audio or the speech data wherein one similar word with different meanings can easily be identified. For example, it can easily classify the word 'that' which of course is a pronoun with the word 'that' which is a conjunction here, but this can only be categorized through the syntactic structures of data along with its pragmatic relation with the text undertaken. Other than this it can be used extensively for shallow parsing, structure transfer, building chunkers, language parsing etc.

Therefore, one can understand here that how crucial is the idea of POS tagger is in the field of NLP and COLING, in determining and refining the data correctly and accurately. Though, its need in Indian languages was understood quite late, but today with the passage of time it has made much of the trivial NLP task easier with its uniqueness and efficiency. In the recent past years, the POS tagging tool has been developed in quite a well manner as before this the data collected was tagged manually, which still can be seen today in most regional languages that does not have POS tagger yet. For all those languages the entire POS tagging procedure is being done manually through the means of identifying each token by understanding its lexicon and paradigms. In recent past years a lot more POS taggers have been developed for many Indian languages like Hindi, English, Punjabi, Bengali, Marathi, Magahi, Maithili etc. All these have been modified and also trained with new possible technologies whenever required in order to enhance the performance of the concerned POS taggers for any concerned language.

5.7.8 Word-Sense Disambiguation

A word is used in many different ways in a human language. Word sense disambiguation (WSD, henceforth) is one of the prominent NLP based application that makes this understanding task easier. It is the method through which the meanings of the word concerned in a text is identified differently and correctly in the field of NLP. In NLP system the identification of correct meaning of a word is a huge challenge which is somewhat tackled by WSD. It basically helps in solving the issue of ambiguity that are there in a word as one single word can be used in several different situations in order to propagate different meanings.

The WSD system works in such a way where it is supplied with a list of large number of words along with its different associative meanings. For e.g. It selects a large number of words with its different senses and creates an online Wordnet,[32] which is considered as a type of online NLP dictionary. It then classifies and adjust the meanings of the each concerned

[32]It is a lexical database of words with several different languages with its meanings having different grammatical classes that helps the database in adopting the exact associative meaning for each word class.

word of the same kind with its respective and associative context. This means that a WSD application trains every classier for each word in order to make it fit for a particular context. The only challenge in this task is to differentiate between the ambiguity classes some times. which means sometimes it fails to identify the correct class of the word and make it fit for a particular context. A WSD is NLP is generally used for designing online dictionaries, tuples, Thesaurus etc. A WSD approach works on supervised machine learning method, wherein it takes the supervised and purely identified appropriate text or word classes in order to perform this task correctly. A WSD is widely used in designing an online lexicography or dictionary system.

5.8. Machine Learning

Machine learning (ML, henceforth), is a branch of AI and NLP which imitates the human learning techniques through means of data algorithms. It gradually improves its learning in order to improve its accuracy. It uses datasets as its input in order to give output values. The output value changes every time once the input value of the same gets changed. It also helps the machines in determining the accurate predictions on the basis of noted past observations.

The sole focus of an ML system is to extract possible information from the data collected by applying several computational and statistical methods. ML techniques are used broadly today in solving several NLP based trivial tasks. This includes speech recognition, document categorization, document segmentation, POS tagging, named entity recognition (NER), parsing, transliteration etc.

A ML technique primarily involves two main tasks which are of learning and training the system as one while making the system able to predict as the other. In ML the system is supplied with a set of training data that consists of multiple domains but it should always be kept in mind as for what reason the task is undertaken. This implies that in machine learning it is always mandatory to identify prima-facie as for what purpose the ML technique is being used. This is important as it will help in achieving the goal correctly. Therefore, it is evident to note here that each ML system should be pre-defined before undertaking any task in order to obtain and acquire an effective and accurate model for the concerned task.

Along with this, the ML system must also be supplied with the appropriate domain knowledge while inducing the training dataset in to the machine along with the specified characteristics of the data so that the machine can easily identify and classify as to what reasons and goals it has been undertaken for. this entire step in ML process in termed as inductive learning. In general terms this can also be termed as conceptual machine learning.

In machine learning technique every undertaken raw data is being supervised wherein each dataset or data inputs are being mapped with its target values, which means they are being supervised with certain set of data which are already taken as their source processing data. This implies that if a dataset in ML technique is supplied for CP identification, then it must have adopted a pre-supervised model or data sets for its accurate processing in the machine. This is also important in order to reduce the chances of any errors however, encountering the same cannot be over looked.

Another important phase of training the machine is by applying the prediction model. Here, in this process a larger set of input data are being mapped with corresponding target values. The only challenge here is to adopt a good learning model which shows better prediction results based on the data and the language domain supplied. In other words, one can say that the model undertaken must be good in terms of structures and grammatical patterns that can help and ease the work of the undertaken language data.

In machine learning there are several different leaning techniques that are categorized in to different sections as per the needs. Some of the common and all round used machine learning techniques are Supervised learning, Unsupervised learning, Semi-supervised learning, Reinforcement learning and Transductions. But here, in this undertaken research only two models of machine learning have been adapted which are supervised and unsupervised learning. It is these two techniques only which are being applied for the undertaken research.

The sole reason why we have used supervised model of learning here is just because it can easily detect the target function and trains the data upon the same model thus, giving a labelled output for all the data as per the need and requirement. In final output, if the label assigned to the data is concrete and accurate then it is known as classification however, for all other labels which have not yet been identified or which shoes any hindrances for the task to complete are termed generally as regression, which means it has some issues which needs to be addressed and taken care of separately.

With all the above discussions and facts one can deduce that learning in machines are not only concerned with labelling or remembering or assigning the tagsets to the data in machines but also, the idea of generalization to the data sets which are yet unseen or still unsupervised. Any changes, further that are being made to the prescribed learning model can be seen as acquiring new knowledge. It is due to this reason that the idea of machine learning in NLP has been further categorized into four different stages, which are discussed as follows: -

- Model Learning: - This is a machine learning task based on prediction. In this process of machine learning the system tries to predict the values of unknown functions. Later, these unknown functions are being trained with the known functions which have already been supplied in the machine in order to train the unknown dataset. Here, the unknown datasets also try to predict the patterns of datasets supplied already as a source and hence, trains the raw datasets. This entire process is based on the statistical model of training. But in the same model if the function is discrete which means each function of the supplied dataset is different in relation to the other and not having no corresponding values then the function is called classification. Apart from all these there also exists a regression model learning wherein each training dataset is being investigated in terms of its relationships between independent variables to that of a dependent variable in order to receive outcome.

  Therefore, it can be understood here that model learning techniques in machine learning is used for predictive learning in order to obtain the desired outcomes.
- Concept learning: -In machine learning concept learning refers to the method of learning by means of boolean-values, over large set of training data. This means that in concept learning a data module is being trained on machine learning platforms in order to classify objects as per the information and requirements supplied initially. It helps in classifying objects from large datasets taken as an illustration for learning process thus imparting each a separate word class or desired labels as required. The entire learning process in this module is based on observation wherein the larger datasets observe the patterning of the datasets thus helping in training the raw data further based upon the given module.

  Therefore, it is understood that a concept learning is the easiest step followed during machine learning wherein the machine acquires the pre-supplied concepts or classes or objects in order to train the raw datasets. It is thus also termed as supervised model of machine learning.
- Explanation-based learning: -In NLP or machine learning it is related with the idea of training ability that uses the single training instance. This means that here the machine algorithms learn with only one set of examples instead of taking several. The idea of explanation-based learning works on the principle of Explanation based generalization (EBG). In this principle, there are two steps involved which are explanatory method as one while generalization method as the other. In the very first step the machine identifies and puts aside all the unimportant datasets which are not

suitable and is different from the datasets upon which all these unsupervised datasets are to be trained. Later, the rules are being generalized for all these unsupervised data which of course has been taken from the supervised model or from the datasets which are being used as source for the training purposes. This is important because it will help in generalizing the training rules for all unsupervised datasets and will also help in identifying the supervised model of the data as the key point of the entire training process. One can also term this step of learning as the statistical module of learning. Fig (8) explains it in a better way.

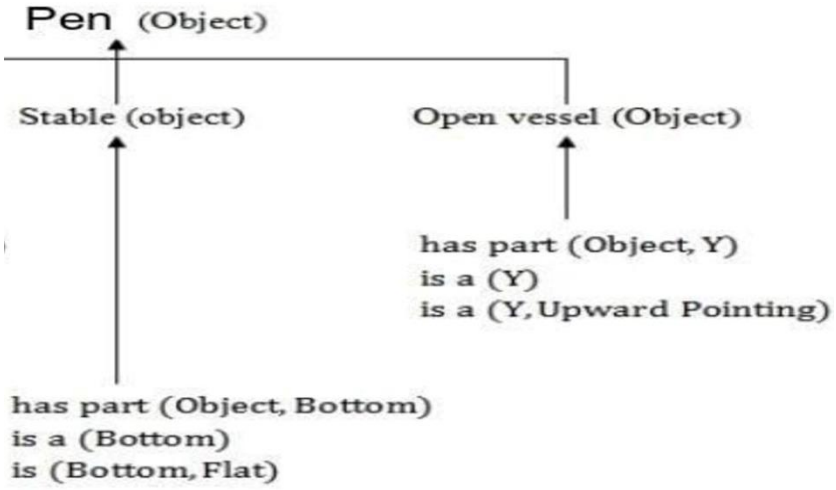

Fig.8 detailed description of concept-based learning

In explanation-based learning the instances of correct and incorrect information are being noted along with their performances in order to make the machine learn. This means that through this the system ought to learn concepts along with the rule formulating processes for machines in order to make it more efficient, accurate and valuable for all other unseen tasks so that each one of them could be trained in the same order.

- Case-based learning: -This is also termed as exemplar-based learning process. This means that it helps in training the AI machines on the basis of examples. In this module the machine tends to fit in their memory the subsequent and different examples from concerned language correctly and classifies its data along with its rules in order to use them wisely and makes it productive with its accurate performances. In other words, one can say that it works on the principles of making analogies to process the untrained data. This means that it tries to develop a meaningful and orderly relation with the processed and supervised data in order to produce it as a model for unsupervised data.

In this process of case-based learning or CBL henceforth, it uses the database of problem solutions in order to solve further news problems. This means that it refers to the stored tuples or cases for solving new problems. This is why it is also termed as complex symbolic method of problem solving. It works on the technique of problem solving on the basis of instances being taken from the previous data. For e.g., if in a sentence or a word a new case or else a new class is said to identify, then the CBR will prima-facie check whether the machine has any identical training case similar to that of tested data then the similar solution for the same is being returned. But later if there is no any identical case, instances or word token are found, then the CBR here, will now start searching for all other training existing components or datasets that contain similar instances as compared with that of the new ones. This means that it works on the analogy system between the datasets wherein the possible analogical pair of the data are being identified and then trained accordingly.

With this idea of sharing the solutions from seen to unseen data it is seen that a CBR module works on the principle of combining the solutions of previously supplied and existing solutions of tested data to that of unsupervised or raw data in order to obtain possible solution for the new ones that still has no solutions until this testing and cross-checking was done. Here in this entire process of CBR it might be a possibility that any set of data might look for compatibilities which is only possible through individual solutions therefore, here in such cases the logic of back tracing might work as this pave the way a little broader to look for some new possible solutions to the identifier which never existed in the tool or the data previously when supplied and tested.

The only challenge to this entire task of CBR is that it to identify for a good and appropriate similarity between the data in order to have a possible similar solution for the same. Looking for proper indexing by comparing the salient features of datasets seems to have another vital challenge in CBR as this is a very trivial task in identifying and classifying the features of each class of words or tokens separately which of course is a time taking phenomena and encountering errors in this method cannot also be ruled out here.

In ML there exist plenty of AI applications based on this such as customer service help desk system wherein the solution is obtained on the basis of diagnosis

of product along with its relation to word types, class, structures etc. apart from this it can also be seen broadly in fields like engineering and law, wherein it can obtain solutions from technical designs and legal rules respectively. CBR is also massively used in fields like medical diagnosis and education wherein history of each older patient is being identified along with their case histories and later the same is being used in order to treat and diagnose new cases and Patients with pre-existing possible solution or remedies. However, encountering error by tool in this type of ML process can also be very crucial as well.

5.9. Computational Model for Magahi CP Identification System

This section of the chapter will discuss model for the undertaken task, along with its advantages, and disadvantages if there is any. It will deal with a short precise introduction of the applied statistical model which will be used further as a model for developing the CP tagger.

5.9.1. Support Vector Machines (SVM)

In machine learning techniques support vector machines, SVM hence forth are the best supervised model for associative learning by applying possible algorithms that helps in analysing data and identifies pattern of the same that are involved in due process of 'classification and analysis task'. There have been several different interpretations of SVM as given by Sober and Benedito (2001), Cortes and Vapnik (1995) and Edgar et al. (1997).

For them SVM is a classifier tool that uses several popular methods in order to build a hyper-plane based structured classifiers, which processes the raw datasets and trains it accordingly as per the requirements in order to have excellent resultative performance which becomes compatible for variety of NLP applications. It is a tool that is dependent completely upon the statistical techniques and modules of learning as developed by Vapnik (1995) and his associates at AT & T Bell labs. Ever since after the technique of SVM has been discovered it is now seen as an alternative to develop several other linguistic classifiers such as Radial basis functions and multi-layered perception classifiers. SVM classifiers are very latest approach for supervising the pattern classifications which gets applied to a large number of data which makes the task easier for pattern recognition within a wide range of multi-domain datasets.

Today, this has become immensely popular because of its proper and intensive learning theories which are primarily based upon strong statistical foundations of mathematical approaches thus showing a very powerful result and effect. Due to its immense accuracy and suitability in almost all languages it is one of the most widely used techniques for machine learning algorithm not only in tasks related to foreign languages but also to Indian languages as well. For instance, there were several language problems that are encountered in other languages as well apart from Hindi such as Tamil, Odia, Telugu, Maithili, Bhojpuri or Magahi etc. All these are effectively and efficiently solved by using NLP techniques thus applying the SVM learning algorithm. Apart from these there are also languages that belong to Dravidian language family such as Kannada, Telugu, Malayalam, Hindi, Sanskrit, Bhojpuri, Magahi, Maithili etc., and the problems of these as well, are well addressed, undertaken and supervised by SVM. All these languages belongs either to Dravidian or else the Indo-Aryan language family, therefore they might show some sense of resemblance among each other either syntactically or semantically on several different language parameters. Therefore, it is due to this reason that we have chosen the SVM module for tackling this undertaken task.

Support Vector Machines (SVM) is one of the best machine learning algorithms, which has achieved best performances above of all on several machine learning tasks especially with Indian languages. It is a simple way to build a binary classifier which constructs a hyperplane like structure that later helps in separating input classes for the machines which then later helps in separating different class members of one kind with that of another non class members and thus builds up an adequate training set. For all the inseparable classes of data the machine needs to create a higher dimensional space than defined earlier in order to fit those data class in that space thus creating a separate hyperplane for all those datasets that are still unidentified. This higher sperate class of datasets is known as feature space. This feature space is very different as compared to that of all those input space that has been pre-occupied by all the datasets that are used for raw data training.

Support vector machines or SVM tool is one of simplest and most effective classifier and generator that picks up all the suitable data required for the task. It is one of the most reliable and trusted classifiers for almost all practical NLP applications. It is also robust and flexible for data feature modelling as it helps in processing the data in a much faster way than any other statistical tool, in order to perform the task of automated tagging. One can say that a SVM tool used for data tagging and sequencing is the most beneficial one out of all other as

it performs the task of automated tagging of the datasets much faster. The only thing that is required as a pre requisite for better functioning of this tool is that, more the amount of the datasets it has, more accurate and authentic results with maximum percentage it will give. In short, it performs much better when supplied with huge data volumes. The SVM tool with English language performs very well with an approximate high percentage of accuracy which is somewhere at around 95% and even beyond this and is thus regarded as one of the best and suitable taggers as compared to several others till date, Marquez and Gimenez (2006).

SVM in machine learning as stated by Joachims (1999), are the supervised learning statistical model that analyses and processes the data by recognizing patterns of the previously supplied data, which is being taken or regarded as the learning data modal for the undertaken task. The training of this supervised learning model is based on a given set of N training examples {(x1, y1) …., (xN, yN)} in which each instance of xi represents for vector RN of which the representation of the class label is yi∈ {−1, +1}. The linear hyperplane created by SVM helps in identifying the positive set of examples and separates them aside from that of the negative ones with maximal margin, Gimenez and Marquez (2006) SVM Tool Technical Manual v1. 3.

For every non-linear classifier it is denoted through the value f (x) = Sign (g(x)) and for every input vector it is denoted through f(x) = +1(wherein x denoted the member of given class or category, whereas f(x) = -1(here in the value of x with negative value denotes that it is not the member of the given class). The value of g(x) is proportional to m whereas zi is denoted as support vectors and K stands for the representation of Kernel (refer fig.9 below)

$$g(x) = \sum_{i-1}^{m} w_i K(x, z_i) + b$$

Fig.9. SVM Algorithm used for classification

Hence, it is deduced from the discussions above that the system for the undertaken task has been developed by applying SVM techniques which helps in proper classification. This entire process of classification is performed by creating a N-dimensional hyperplane structural model which separates each class of data categories into their respective classes as defined.

5.9.1.1 Geometrical Interpretation of SVM

The geometrical representation of SVM classifiers gives a brief descriptive sketch of the tool. It not only helps in detailed explanation of the functioning of the tool but also in understanding the mechanism of the same. All these are sorted through the geometrical optimization of the tool thus, leading towards the problem-solving techniques for all sorts of undertaken tasks as specified or being undertaken.

During this geometrical representation, the SVM tool is thus represented with a set of training examples as specified according to the task. These examples here in the tool can be represented as (x_i,y_i). Here the interpretation of the value x_i is being denoted as the real data or the original instances of the corpus whereas (y_i) is represented the different labels, sets or classes each data sets represents. If any classification issues or problems is to be tackled through this geometrical representation then the recognition issues of the two different sets or classes which are represented through a specific pattern wherein the positive sets are denoted as yi = +1 or else yi = -1. This means that it has two different data training sets examples undertaken for the result processing. Here, the training set example is considered positive if it has taken training example with the value of yi = +1 and negative otherwise if the value of yi = -1. Both these datasets are being separated into two different classes or categories through a hyperplane structure. This is done in order to obtain the maximum separation among the two different classes of datasets. This is shown in detail in Fig. below.

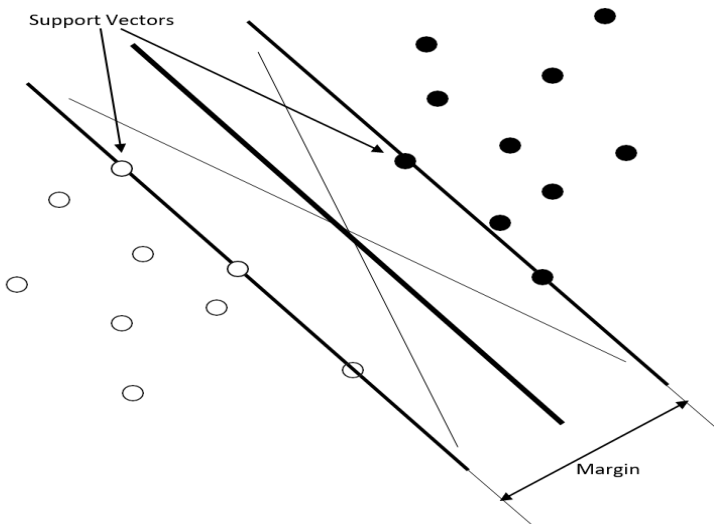

Fig.10. Geometrical Interpretation of SVM

Here, in the fig (10) given above the separating classes shown with large margins minimizes the problem of errors encountered in the data sets. In other terms, the minimum generalization shown here in the fig means that whenever any dataset having unknown class values are supplied to the tool for classification task then the possibility of having an error with given class or sets of data will be minimum.

Similarly, will be the case with the maximum separation-margins of the given dataset in the SVM tool. To understand this in detail please refer to fig. above which also explains the concept of maximum margin of the tool. In order to understand this in detail, we can clearly see here that there are two different planes that are parallel to the given classifier which passes by more than one points of the given dataset. All these parallel planes are known as bounding planes. The measuring distance of all such bounding planes that surpasses through these points are known as margins. There also exists a SVM learning technique which maximizes these margins by creating a hyperplane structure. On the other hand, the meeting points of the datasets shown in the graph which fall on these bounding planes are called supportive vectors, which plays a vital role in the entire process and because of which it has been named as support vector machine method. In COLING the term method means the use of algorithmic process obtained for the task undertaken. Considering the situation on a different note such as if the training data sets are to be separated without errors through the same hyperplane modal, then the supposed error ratio of a training data set equals the ratio of support vectors with that of training vectors. But, since this a supposition undertaken for enlarging the dimension of the errors just to receive better results, then a small set of trained datasets can then be induced inside the model for eradicating such issues and also to have better generalization results.

At the same time, there can also be situations when datasets are possibly not classified or identified then, to maximize the margins of the tool shown in graph, the number of unidentified datasets can be minimized by supplying the tool with a bit of correct examples, just to obtain some better results in less period of time. In simple terms, it can be understood that sometimes the SVM training algorithm also permits the manipulation of data, which may or may not be vulnerable of obtaining an error.

Therefore, one can understand and see that how SVM are user friendly not only with the datasets but also, with the users who handles the task. It is due to this reason that the SVM algorithm is being massively used today for almost all data driven tasks.

5.9.1.2. Properties of SVM Tool

This section of the chapter will briefly discuss some unique properties of SVM tool. Few of them are discussed as under: -

(i) Simple- SVM tool is easy to handle, configure and also to train with the datasets. All such tasks are easily tackled through easy and simple configuration file. It gets easily tuned not only with one rather, with multiple datasets or files. This feature of the SVM tool makes the tool easy to access for the task assigned such as classification, tagging which is done through means of understanding the standards of datasets thus, providing valuable outputs. It works on the principle of embedding which means, it embeds any pattern of supervised or unsupervised datasets as per the requirements and performs the task.

(ii) Flexibility- SVM as a tool is so flexible that the shape and size of its context features can be easily moulded in to any different shapes and sizes. This means it can be easily adjusted to different forms. This flexibility of the tool is one of the vital features that helps in easily detecting the POS taggers along with the different ambiguity classes of words. The flexibility feature of this tool helps in easily allowing different combinations of classes or datasets in order to obtain different sorts of results of the tool. However, this difference of results does not affect the accuracy percentage of the tool.

(iii) Robustness- This feature of the tool is well defined by well adjusting the parameters of the tool. The robustness feature is generally used in order to perform sentence-level analysis task. This feature of SVM tool uses several different strategies in order to increase maximum results with less errors which obviously increases system effectiveness.

(iv) Portability- This tool as we all know is independent irrespective of languages. This has been successfully applied to several different languages such as Hindi, English, Odia and now Magahi of course. Other than this, it can also be used to unsupervised corpus across languages irrespective of different language features.

With the onset of time and advancements in the tool, SVM can also be used widely for any unsupervised data or else languages which faces scarcity of exposures in order to have more language resources in terms of research. Use of Magahi language for this tool in the undertaken research is one of the better good examples concerning this. Therefore, it is understood from the facts above that the

portability feature of this tool makes it easier to use across languages and also make it learn from any unsupervised piece of data by the means of involvement of non-ambiguous words, thus creating a dictionary of learning as an additional help, which later helps the researchers in creating an e-dictionary of its own kind based on morpho-syntactic category.

(v) Accuracy- The SVM tool as stated earlier exhibits maximum accuracy with high percentage. It does so when it is supplied with rich amount of corpus along with proper sets and defined features as required for any undertaken task. This supply of corpus allows the tool to model the training of the corpus for both the unsupervised and the supervised datasets, which later helps in achieving the most precise results required. Along with this, the SVM tool is taken as one of the appropriate ways in order to obtain more detailed, efficient and accurate results having less issues, if any.

(vi) Efficiency- SVM as tool is also more efficient for training any language model in the field of NLP and COLING. It is due to its performance and the task of tagging that the tool executes quite well and also in an ordered manner. Likewise, all other features of the tool, the efficiency feature of the tool depends as to what feature, or dataset size, along with the schemes of tagging one has selected. It can go up to a tagging speed of 1500 words or token per second. This is because of the use of the kernel method of the tool, which makes the tool to work, train and learn efficiently at a higher rate with maximum speed. At the same time this speed of training and learning of the tool also varies, as it depends totally upon the nature and the size of the corpus, being supplied.

5.9.2. Experimental Set-ups

This is one of the most vital and important sections that are very much require for the entire computational setup. This section includes a detailed and brief introduction of all the computational experimental setups of tool including feature extraction, configuration, training, and testing of the data for the applied model in the undertaken task.

5.9.2.1.Feature Extraction

It is one of the major steps in training the module. For this proposed research several simple features from the dataset have been selected and is being set to default mode. The reason behind doing so is that it contains all sort of linguistic features of Magahi words. All

such features are being encoded in the form of different forms of words or verb types such as compounds, conjuncts, serial verbs etc. The three mentioned here are taken as the main types of feature encodings while there are also several others subsets such as serial verbs1, serial verbs2, reverse and explicator compound verbs etc. that is supplied to the tool as another dataset types in order to classify the raw datasets, thus helping the same in giving an efficient result of the tool.

The reason behind selecting these specific features for the tool and also for the datasets is because, it can make the machine identify the data types quite easily while processing because, all such basic types of data and its types have already been dealt with in detail concerning to all its desired and requisite linguistic properties. This was done as it was also required for proper data annotation, which also has been completed beforehand during the annotation phase.

5.9.2.2. Configuration

The SVM tool uses a certain defined structure of configuring files that are being used extensively for the process of system learning. These configuration files in SVM contains files of medium verbose types such as (-v2), while all the rest learning and tagging modules inside the system are being set in a set pattern of direction following the start point from left and then finishing the same towards right. This means that the entire module of verbose system follows a definite and defined directional pattern of left-right-left (L-R-L) combination.

Apart from these mentioned above, there also exists several different processing files that are being used for all other major configuring procedure of the entire system. These consists files like C parameter tuning, compression of models, setting up example features for desired model, filtration of features comprising words or tokens that are needed for the purpose of system making. Along with this the unigram feature model has also been introduced for this model. During the training process the same unigram featuring will be taken for successfully testing the designed model.

For any unigram designed feature modelling first of all the unigram template features needs to be configured. In this featuring model every line or row or column must represent only one template feature of the concerned language or the language undertaken for developing the model. In each row or column, the input of corpus or word tokens must be

represented. This must be aligned right Infront of every single line or row or column as failing this will not help in getting the desired output, hence crashing the entire tool. Therefore, the absoluteness of each position in this unigram template featuring model is essentially required. One can understand this mechanism with below mentioned figures wherein the first one denotes the unigram featuring template whereas second denotes the details of sentences supplied to the featured template, while the third denotes the details of functions supplied to the unigram model.

template	expanded feature
%x[0,0]	the
%x[0,1]	DT
%x[-1,0]	rokens
%x[-2,1]	PRP
%x[0,0]/%x[0,1]	the/DT
ABC%x[0,1]123	ABCDT123

Fig.11 Unigram feature Template

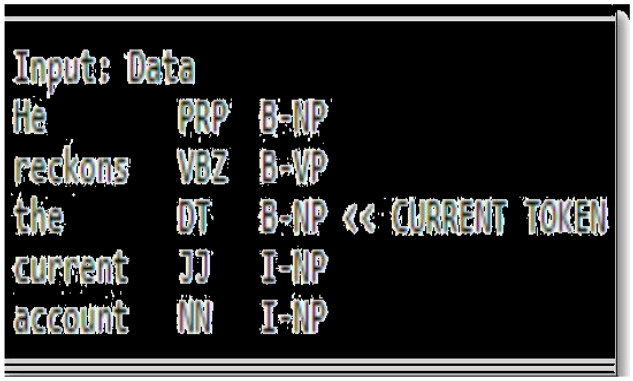

Fig. 12 Details of featured templates used by unigram

```
func1 = if (output = B-NP and feature="U01:DT") return 1 else return 0
func2 = if (output = I-NP and feature="U01:DT") return 1 else return 0
func3 = if (output = O and feature="U01:DT") return 1  else return 0
....
funcXX = if (output = B-NP and feature="U01:NN") return 1  else return 0
funcXY = if (output = O and feature="U01:NN") return 1  else return 0
...
```

Fig.13 representation of functions of unigram model

All these are setups are then being as default mode. This default setting of the entire module includes illustrations from trained or untrained or else known or unknown data file types, Behera (2015). It is done so because, this helps in identifying categories of each word or tokens in a quite efficient manner. Apart from these configuring files, there also exists several other templates or word tokens that are different or else does not matches with the tokens or words that are identical to the tool, to what we call as unsupervised tokens or words. These are also being used by the machine for the configuration procedure of the tool. This is equally important in order to obtain a better comparative result of the supervised as well as the unsupervised data sets.

The effectiveness and response of these configuration files depends directly on the type of raw words or tokens supplied to the machine. This means that the accuracy and the response percentage of the tool is dependent upon the configuration template supplied to the tool. Mentioned here under in the given figure shows, the featured templates that are being used by the SVM tool for its further configuration process. These configuration files of the tool include all types of data sets such as known vs. unknown, supervised vs. unsupervised, ambiguous vs. non-ambiguous etc. This is so because missing of any of the above class of da sets can affect the result badly thus, affecting its final output. Hence, it is required to add all the above parameters in order to have best possible outcome from the tool developed.

```
--------------------------------------------------------------------------------
--------#ambiguous-right [default]A0 = w(-3) w(-2) w(-1) w(0) w(1) w(2) w
(3) w(-2,-1) w(-1,0) w(0,1) w(-1,1) w(1,2) w(-2,-1,0) w(-2,-1,1) w(-1,0,1)
w(-1,1,2) w(0,1,2) p(-3) p(-2) p(-1) p(-2,-1) p(-1,1) p(1,2) p(-2,-1,1) p
(-1,1,2) a(0) a(1) a(2) a(3) m(0) m(1) m(2) m(3) z(2) z(3) z(4) ca(1) cz
(1)A0unk = w(-3) w(-2) w(-1) w(0) w(1) w(2) w(3) w(-2,-1) w(-1,0) w(0,1) w
(-1,1) w(1,2) w(-2,-1,0) w(-2,-1,1) w(-1,0,1) w(-1,1,2) w(0,1,2) p(-3) p(-
2) p(-1) p(-2,-1) p(-1,1) p(1,2) p(-2,-1,1) p(-1,1,2) k(0) k(1) k(2) k(3)
m(0) m(1) m(2) m(3) a(2) a(3) a(4) z(2) z(3) z(4) ca(1) cz(1) L SA AA SN CA
CAA CP CC CN MW#
--------------------------------------------------------------------------------
--------REMOVE_FILES = 0
```

Fig. 14 configuration files for SVM

5.9.2.3. Training the System

The training of the designed SVM model is done through a given set of classifiers which contains examples of each class such as annotated or unannotated. In other terms it is designed in such a way that it can tackle each set of words or tokens required for the system. All these sets comprise of a defined set of classifiers. The training of each of these sets are being done by applying a certain specific SVM classifiers. Here, in this task it uses SVM-light 7, Kumar and Behera (2017). The SVM-light 7 which is being used here is one of the latest versions among all other SVM classifying techniques that are used for identification or classification purposes. This technique of SVM-light 7 was first developed Thorsten Jochaims and was thus implemented by Vapnik in C parameter for training purposes in 1995.

In order to commute this task a total of 5,31,135k tokens have been taken. Out of all these a total of around 45k tokens were taken to train the model for CPs identification and analysis process. All these tokens have been used in different training phases staging from training stage 1 to training stage 3, wherein the amount of data set mentioned above have

been used in short chunks and also was evenly distributed. This resulted in the phased manner training of the system.

To train the model, the raw data must flow into the tool column by column. This means that the data must be entered into the tool as one token per line of the corpus. This should be done word by word. It is only necessary to enter the data in the manner mentioned, as this gives a clear representation of the tokens in the first row, while the corresponding tags associated with them are given in the second row. Apart from these two columns, all the other columns of the tagging format need not to be mentioned as all of these contains the various additional information which are not required for the task. The first two corresponding columns contains the tagging labels, of the concerned POS tags and CP tags, required according to the data, Marquez and Gimenenez (2006).

5.9.2.4. Testing the System

Testing the system is the next progressive step involved for the undertaken research task. The same tagged data set mentioned above will now be used for another further process which is testing. It is a one step ahead method required for tool development. In this process the undertaken annotated datasets will be checked through machine whether it has been assigned with the appropriate tagsets. This means that it a process of rechecking the tagset with the corresponding annotation labels. The task is automatically performed by the system tool based on the file which should be taken as a predefined model file also known as golds file. This gold file contains an enormous amount of pre- tagged data as per the requirement of the tool and the task undertaken. This entire process is considered as automatic annotation process wherein the expertise of machine is extremely important.

At this stage, it is quite possible that the tool might commits some mistakes while tagging or may overlap some features during selection of the data. If in any case this happens, then this entire task is repeated at different intervals and is being corrected in several different stages manually in order to obtain maximum accuracy. Therefore, in other words one can say that this process is an amalgamation of both machine and human techniques (ibid). Here in automated tagging the same technique of tagging the dataset will be followed which has been done while manual tagging which is tagging each single token or word with their respective corresponding lines. This is so because, it makes the desired machine tagger to easily recognize and understand every single token or word very easily thus giving maximum accuracy while testing.

As this entire process involves the automated annotation data techniques, therefore it is also known as a machine learning approach f the system. In this machine learning data driven tagging approach there already exists a single window containing too many tagsets of different POS labels, as required by different datasets. It is evident to note here that all these tagging POS labels will remain the same for almost all Indian languages and require no change. If at all any change in the same is required, then the same will be done manually only in the undertaken dataset of the prescribed language format and not in the tool. However, the datasets which needs to be tagged is then later exposed towards the machine learning window of the tool. Here all the required or desired tagsets for the undertaken dataset inside the tool will automatically be selected and assigned accordingly as per the requirement of the supplied words or tokens from its dropdown menu which finally assigns the appropriate tagsets to the supplied data. This entire above-mentioned process follows a standard input/output system wherein the word or sentence token is being represented in first column and the corresponding tag will be represented by the second one.

Keeping in view of all such trivial functions of the tool, it is clear to understand that the entire testing procedure of the tool follows a detailed online process which is quite challenging sometimes. It is also difficult as the tool needs to be in systematized and arranged fashion wherein proper arrangement of each word or tokens must present there in an organized fashion along with its accurate and desired corresponding tagsets.

5.10. Developing Linguistic Resources

This section of the chapter will discuss the different steps, methods and process involved for developing proper linguistic resources for the undertaken task. This involves two major steps such as validation and tokenization which helps not only in modelling the system tool but also in making the datasets a bit more refined one so that the machine could not face much issues while functioning and could work more efficiently. Apart from the two mentioned above, it also involves few more minor processes, which is not necessary for the undertaken task and hence, not discussed here in detail.

5.10.1. Annotated Corpora

For the training phase a total of approximately 50k words or tokens has been taken for experiment of the tool. This consists data types of all domains including entertainment, literature, general surroundings, health, stories, fictions, agriculture etc. these datasets are

also comprised of both nomenclatures such as seen and unseen. It is natural that during the process of manual annotation some the annotators have committed certain errors but at same time the possibility of the same got reduced once they got continuous exposure of the entire process. Also, at the same time a huge number of data were annotated in order to test the same with the help of tool which of course have increased its efficiency and has minimized the possibility and risk of encountering errors. It is quite possible that some errors are still encountered then this could be due to the nature of the data collected as this has been collected through different sources and speakers and also due to its different annotation patterns. This means that the instances of the collected data may vary not only from sources to sources but also from person to person. In connection to this there also be a strong possibility that some instances might not receive any tag or is left untagged because of the reason that under what POS category it should be kept into or else as to what proper tagset it should adhere to.

As mentioned, the model adopted for the training of data is statistical one, therefore the quality of the same must be maintained at its best failing which will hamper or impact the efficiency of the desired tool. But in case if there is any error which are still encountered even after this thorough rechecking, then they are handled and addressed accordingly. Therefore, in order to minimise this risk of encountering errors in the tool and address them accordingly validation process is being followed, which is a one step ahead method for data sampling and refining for final testing.

5.10.2. Validation

This is a bit more refined process of data collection. Here all annotated datasets get validated after being thoroughly checked. This is done in order to certify that there exist no more extra linguistic elements except those which are required for the tool to function. This also implies that it refines the data up to the extent that it contains no further errors in undertaken language corpus. At this point the decision of validation of datasets are based on the judgement of the annotators which are almost correct but, in case if at all there is still any discrepancies with the same then it is being decided on the basis of the pragmatics or the context of the entire word token or sentence. This step of pragmatic judgement of the context is said to have been given the utmost priority for judging the accuracy of the tags, which is considered as final during this entire validation process. It is also evident to note here unlike

all other steps mentioned-above that if any word or token is being wrongly tagged by the machine then the same has to be corrected manually and then being put for final testing.

5.10.3. Tokenization

This is one of another major steps involved during the entire process of developing linguistic resources for any language in NLP or COLING. It is the stage where the annotator thoroughly checks each and every word token or class by closely examining the same. Here the annotator ascertains that the datasets that are being used for the experimental purpose must be of qualitative range and also affirms that each and every word token or class have been separated with a tab or white space. This is crucial as it will help the SVM to identify and closely examine each word token or class clearly without fail. This entire tokenization process is being carried out by the help of Java Class Tokenizer.[33] This Java Class Tokenizer may tokenize the supplied or the infused data wrongly if there are any irrelevant spaces or characters or even there is absence of single space between the tokens. Therefore, fir this step one has to be very careful and attentive while marking with punctuations in the data file.

5.11. User Interface and Architecture

As we all know that the entire work is based on technological platform hence the tool developed for the identification process is entirely a Web-based platform which functions fully on JSP codes and runs with the help of Apache Tomcat on the internet-based servers. It is advanced and well equipped with SVM classifier tool which has already been used for its training purpose. There are several systematic stages that have been followed thoroughly for the extraction of the desired output which is the CPs in Magahi. At first stage which is the input stage, the tool receives the UTF-8 encoding file and all the other extra undesired file of the text are rejected. Secondly, the tokenizer which is present there in the tool tokenizes every class of words or tokens as per their desired characteristics along with required suitable punctuations in single-line-per-word manner. Afterwards the SVM based CP identification tagger tool tags every token or word class along with its possible desirable tags. Once all the above-mentioned three function of the tool is performed accordingly then, the CP identification extractor will now look for every possible word class or tokens that are there present in the training database and tags each one of them with their desired annotation tags.

[33] Type of string tokenizer which permits the SVM tool to break large strings into small tokens.

Later, as final stage of data processing in the tool the SVM based Magahi CP tokenizer de-tokenizes the CP data in an ordered manner which is word-by-word style format thus yielding the final output. For detail understanding refer model diagram in fig 15 below.

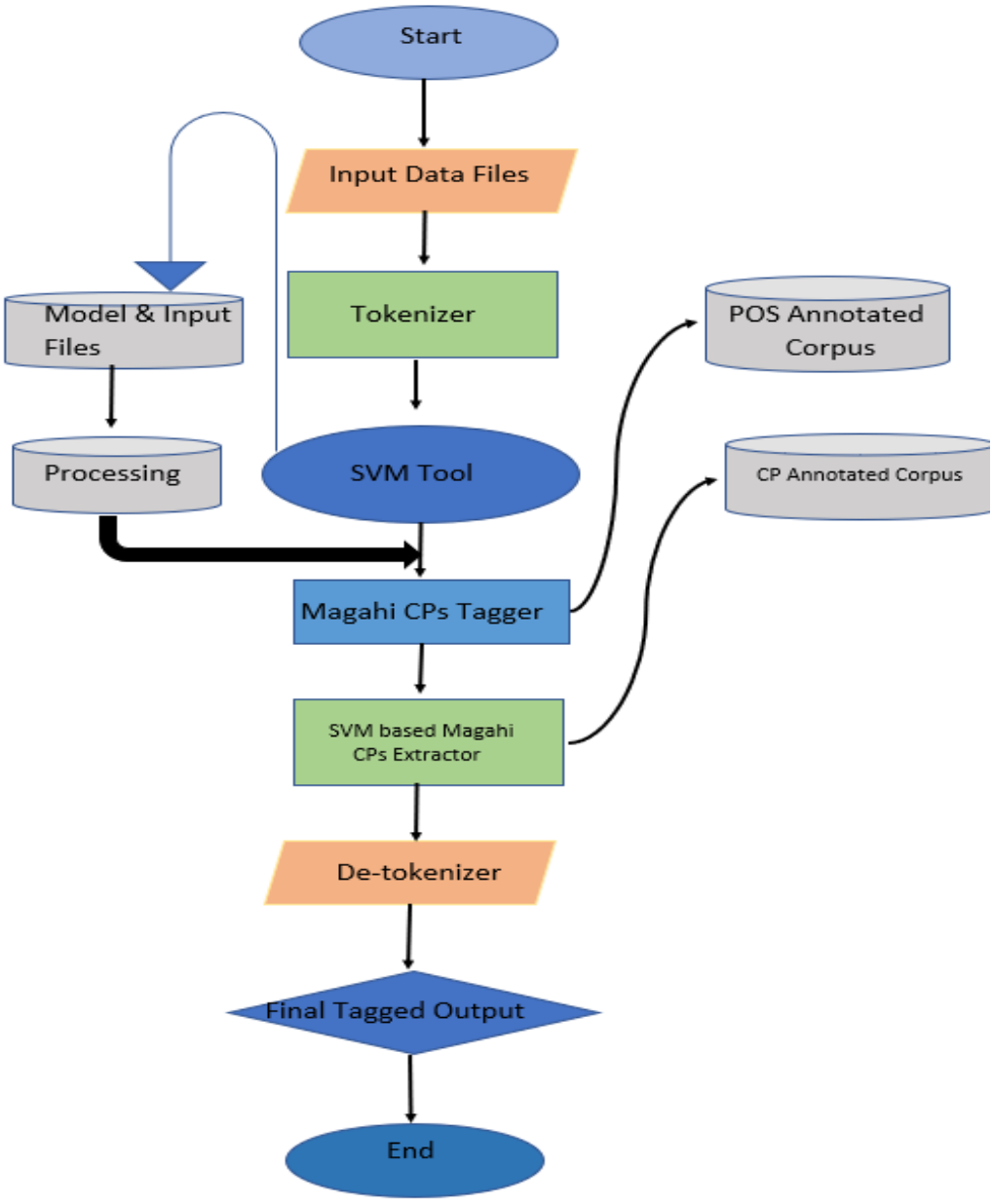

Fig. 15. User interface architecture

5.11.1. Technology Applied

For the undertaken research a wide range of different computational technological aids have been used. The raw data for the application have been developed by using servlets which is a Java server page. The data as mentioned earlier have been collected from a popular online Magahi blogging website called magahiblogspot.com. The hypertext markup language

(HTML henceforth), link of this blogging page is then later exposed to a web-based crawler for crawling the data. This web-based crawler runs on java server and this java server runs on JSP module containing UTF-8 (Unicode Transformation Format) Unicode coding which is common for almost all Indian languages.

After finishing all the above-mentioned primary procedures, the second stage of technology application begins, which is setting the right path for JSP (Java Server Pages) servers on the system. This is an important desirable step for the Java platforms in order to make the tool functional. As soon as all the basic setup of the tool over the system gets finished then at the later stage the webpage link of the concerned website is brought forward. This link of the concerned webpage is then exposed towards the Apache Tomcat Server of version 4.0. It is a web-based server platform which runs almost all sort of java applications. Another most important reason as to why only this version of Apache Tomcat has been used for the task is because it clearly reads out all sorts of JSP pages and hence also creates an auto-backup of the entire data through crawling process and stores it in a separate file.

The URL (Uniform Resource Locator) mentioned above from which the corpus has to be collected can easily be accessed through any of the compatible browser of the system. The given URL unzips all encrypted data file with the help of the selected browser and runs successfully on the designed system tool with accurate programming inputs and path setups. This means that the path which has been setup for the selected program must not only be compatible for the system undertaken but also for the desired server platform over which it performs this trivial task. The browser selected for this purpose must be able to recognize this infused Magahi jsp encoding file which can only be possible over the web-server platform of java. The detailed function, mechanism and procedure mentioned-above must be followed carefully step-by-step so that the tool does not misses out any function thus giving maximum accuracy with possible absolute results without fail. In order to make general masses understand the detailed functions of the desired tool the steps are hence summarized as follows in a nutshell: -

(i) For the desired task Java as programming platform must be used for all set up process.

(ii) One must use the jsp format of the tool in order to make the data run which has been refined through a prolonged process of crawling and sanitizing.

(iii) The desired server used for all such COLING applications for Indian languages is Apache Tomcat with its possible latest version of that time. For now, we have used the Apache Tomcat version 4.0.

5.11.1.2. Apache Tomcat 4.0

Apache Tomcat 4.0 is a web container that permits the users to run the servlet technologies and Java server pages as per their desired applications. It is built up for servlet version 2.2 and JSP specification 1.1 having no change applied in the same. It is widely used for all sorts of java-based applications applying latest related technologies what so ever. This is an open community-based server application allowing both open and participatory domains. The tomcat server used here in this task is a licensed one with Apache license projects. This is made under collaboration of all premium server developers across including national and international technological groups the globe. This is the reason why it is most reliable platform for handling all trivial and massive scale web applications. Unlike all other software this also requires a server platform for its proper functioning which is the local host server platform, Chowdhary (2006).

5.11.1.3. Java Server Pages (JSP)

Java Server Pages or JSP are a platform of collection of technologies that tends all the web-based software developers to design new web-pages of HTML, XML or SOAP types. This was first designed by Sun Microsystems in the year 1999. JSP as the name suggests always uses Java servers as its base for operations of programming languages. In order to run this page proper networking system is very much required. The servers required for its smooth and accurate functioning is servlet containers, which contains all sorts of necessary programming components required for the servers to interact with java servlets. The java servlets are the programming platforms for java applications which has the ability and capability of making these servers successfully run on its platforms. The popular sever on which these JSP pages are made to run successfully is Apache Tomcat with its possible latest version available for better results. The JSP pages are used extensively for creating all sorts of web-based contents by applying Java-written XML properties and scriplets.[34]

[34] http://www.serverwatch.com/news/article.php/1125001/Apache-Tomcat-40-Final-Released.html

Therefore, JSP or Java Servlet pages are those platforms which utilizes java objects lodged in with the HTML codes.

5.11.1.4. Java Servlet Technology

It is a java-based programming application that runs on Java based web or application servers. It is used for performance enhancements of webserver. The most important reason for using this quite widely is because of its user-friendly and platform independent nature. It incorporates many advanced features of Java programmes such as portability, protection, can be re-used from time to time and application to application, good in performance etc. this is an application that enhances the working performance and abilities of a server. This can respond to any type of request made during a programming process. This can only be used with online java servers. Java servlet technologies can easily accept and collect a programming request and formulate immediate responses in return based truly upon the nature and size of programming and coding infused.

5.11.1.5. Google Collaborator (Google COLLAB)

Collaboratory, or Collab in short, is a product by Google Research. It allows anyone to write and run any Python or java jsp code from the browser and is particularly well-suited for machine learning, data analysis, classification and identification purposes. It is an online and easiest mode available for different machine learning and coding modules. Technically, a Collab is a hosted Jupyter notebook service that requires no setup to use and provides free access to computing resources and build applications. More specifically, Collab by Google is a free online technical testing environment that runs entirely in cloud format. For this no setups are required and the notebooks we need can be easily imported along with the files that we need to work on. For this research, we successfully imported the Java and python files and successfully completed the task. This also has a special property of easy edits. This implies that anyone can easily edit the module and not the file depending on their own desired uses. The file once created can only be edited if one has a granted for the same. This Editing is done just like we edit documents in Google Docs. Collab supports many popular machine learning libraries that can be easily loaded and imported. In our case, we swapped the task by importing the Python notebook and java in the same[35].

[35] https://colab.research.google.com/drive/10p_B6KkKX6Y_72cyRDQuL7D_zA8mp_Cr

Chapter- 6 Evaluation and Error Analysis

6.1. Evaluation Schema

The evaluation section includes the most important part of the entire work, since it contains the most important information about the designed scheme along with its percentage of performance. This was done in a two-way approach, involving evaluation of not only the token-level tag sets, but also the tag sets denoting sentence-level tokens. This section is very important because without this the task to measure the value of the effectiveness of the designed tool is incomplete. This follows a two-way methodology, where on the first level the respective assessment, which includes sentence-level and token-level tagsets, is assessed, while on the other side, the entire scheme used for identification and analysis when mapping CPs developed by Magahi through NLP-developed tools will be evaluated further.

6.1.1. Evaluation of the Overall System

Giminenez and Marquez (2006) believed that each result and score in NLP or COLING is not only consistent, but also scored against the Most Frequent Tag (MFT) reference. Therefore, in this case, the evaluation percentage obtained for the entire tool includes only this fact. The CP identification tool accuracy along with its percentage given here is based entirely on the automatic SVM tools evaluation pattern used for the SVM-based statistical CP tagger. The assessment task was carried out in three different phases. A total of 45k raw data tokens were trained for each phase using approximately 2,000 gold standard data sets that were manually checked and corrected for their respective CP tags. The same process was repeated thrice and after considering all the stages with maximum results, the whole system developed for the CP identification mechanism achieved an overall accuracy percentage of 64.57%, which was the maximum among all. The same procedure was also performed while evaluating the tool at sentence levels.

```
[ ]  df = datasets.load_iris()
```

```
[ ]  from sklearn.linear_model import LinearRegression
     reg = LinearRegression()
     reg.fit(x_train, y_train)
     y_pred = reg.predict(x_test)
     print("R^2 Accuracy: ", reg.score(x_test, y_test))

     R^2 Accuracy:  0.6457291807728918
```

Fig.16. Accuracy percentage

The reason for achieving such low accuracy of the entire system is that it most often encountered unknown tokens. These unknown tokens were those which were untagged in the raw data just to see and observe how well this can go with such types of data. This is done to create a big line of difference between known and unknown records. The issue of ambiguous sentences and tokens was another major reason behind the tool getting such a low rating percentage. Magahi being the least resourced language among all Indic languages and the nature of its resemblance to Hindi standards is the final and the ultimate reason for this. The tool also showed several ambiguities in its performance which hampered the overall evaluation up to some extent. This overall evaluation of the system was performed after evaluating and examining all types of accuracies, including the accuracy of the tool, the evaluation of several word tokens consisting of known, unknown, ambiguous and unique CP tokens distributed at both the sentence, and the token labels, etc. The percentage of ambiguities at the sentence or word or token level, together with its accuracy, will be further discussed in the later sections with its possible evaluative graph representations.

6.1.2. Evaluation of Tagsets at Tokens or Word Labels

In this section, the evaluative percentage of each word or token label will be discussed. This will then be represented further in detail with the help of a graphical representation. In this graph, the accuracy percentages of different tagsets will be shown. In order to discuss the evaluative percentages of respective tagsets at a single token or word labels let us refer the graph below: -

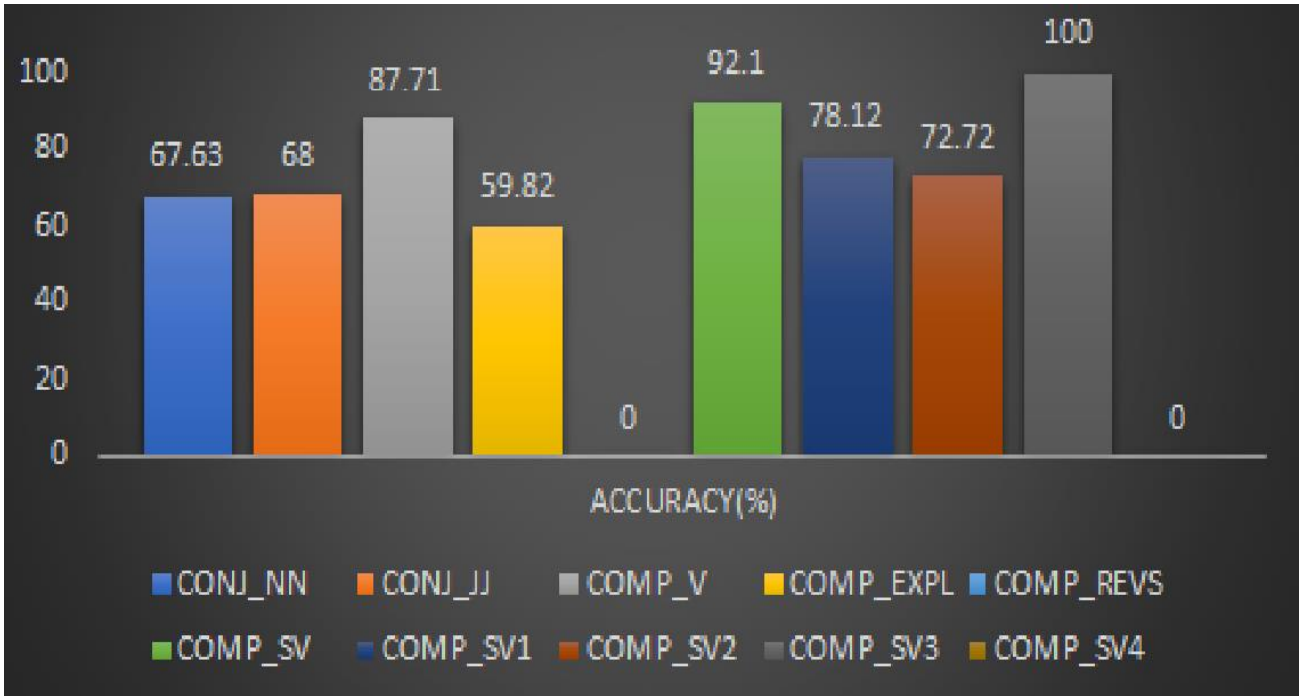

Fig.17 Accuracy distribution percentage across all token labels

In the above graphical representation, the accuracy distribution across all labels has been shown for each CP label encountered while training the dataset. The highest level of accuracy has been seen in the verbal categories. Out of all the tagsets evaluated the maximum was recorded for compound verbs and then it is being followed by all the serial verbs. Considering the secondary labels, the accuracy percentage was high for adjectival conjuncts which are denoted by CONJ_JJ and then followed by nominal conjuncts which are denoted with CONJ_NN. The least accuracy was reported with the explicators in Magahi. This was so because the data tagged does not have much of the explicators present in them and so was the case when it was trained and tested on the machine with the seen vs. unseen datasets. There were also tagsets that have shown no percentages of accuracy with the tool, these were mostly related to the classes of serial verb four, which is denoted with SV4.

This evaluative analysis of the datasets has been obtained and registered respectively after training each of these with their respective gold tokens along with the raw data. The accuracy graphs show such a huge contrastive dissimilarity because the data collected has the highest number of verbal categories as compared to any other category. It also has the categories of nominal conjuncts and adjectival conjuncts the most, therefore due to this these were the second highest to register their accuracy accumulative percentages. These categories

were the most easily detected token on both the token label as well as the sentence label. This is so because the language undertaken for analysis shows an enormous number of applied uses in the data which of course was obtained from a literary source containing multiple domains.

From the graph mentioned above, it is clear that CPs in Magahi are most frequent in their uses while framing verbs or phrases. Therefore, due to this, the accuracy percentages in all the verb labels were recorded as maximum. However, the slight dip and downgrade percentages in some major labels such as nominal conjuncts, adjectival conjuncts, and uses of explicators in the languages cannot be overlooked. Along with this, it is also clear from the graph that the data collected does not have any instance from the category of the reversible compounds, however, uses of the same in the concerned language cannot be overlooked. Therefore, in order to increase the percentage labels of all. Such classes can be created in near future by promoting the frequent usage of more and more core Magahi language data through online websites or through Optical Character Recognition (OCR) mapping must be encouraged in such languages which do not have enormous datasets for promoting COLING researches concerning these languages.

6.1.3. Evaluation and Error Analysis of Tagsets at Sentence Labels

This section will deal with the accuracy and errors of the tool measured at the sentence level. This also has been monitored taking both seen and unseen data. The graph mentioned below will briefly show the accuracy along with the error percentage of the tagsets, noted at each sentence level. This evaluation was needed because the data which we have collected for the task was not accurate and precise. Therefore, in order to make the tool more effective and of usable format, this regressive accuracy module was adopted. For a more accurate analysis of the tool's performance let us refer to the graph below and discuss the same in detail: -

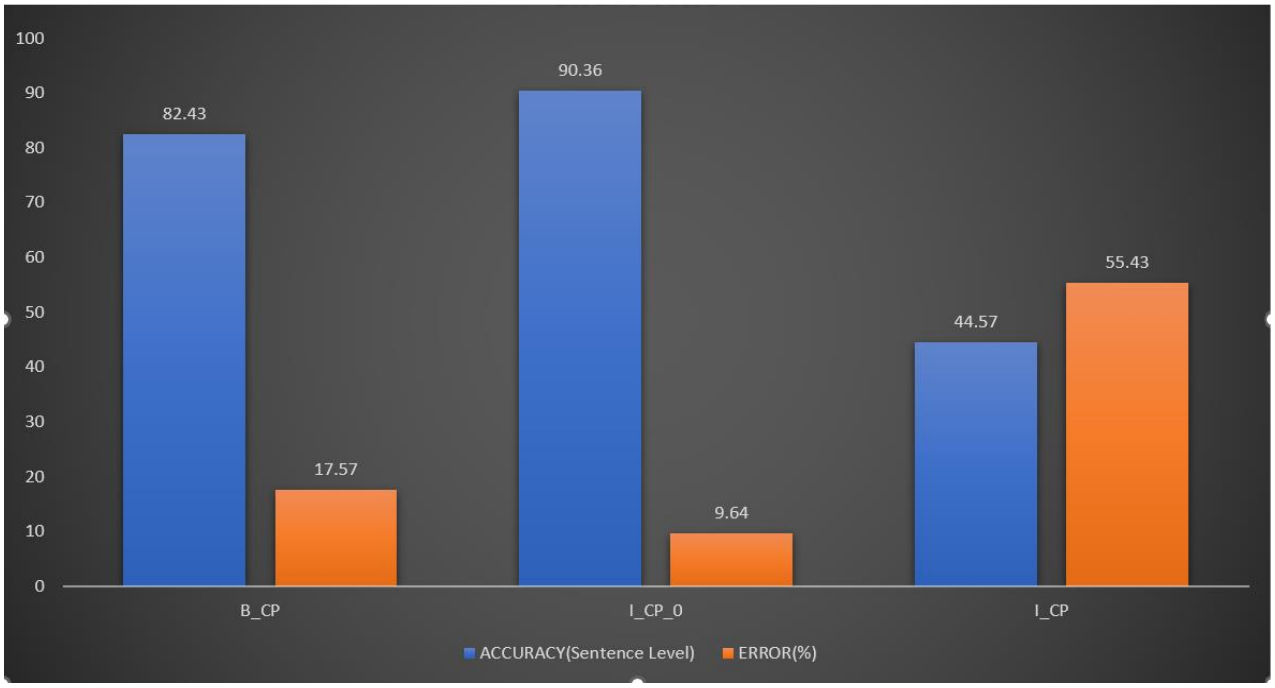

Fig.18 accuracy and error percentage at per sentence label

Here, in the above-mentioned graph, we can clearly see that at the sentence label the accuracy graph of tagsets has resulted in a much better way than the tagsets at a single token or word label. The tagsets denoting with B_CP denote the beginning of the sentence whereas the tagset with I_CP denotes the internal part of that sentence containing CP structures. There also exists a third category of the tagset which is denoted with I_CP_0, which means those instances were not tagged and have been left empty because most of them have not shown any close proximity between the sentences along with the tokens of CPs or else vice-versa.

In this graph, we can see that the highest rate of accuracy of CPs in Magahi has been predicted by the machine at a situation where they are attached at the intermediate level. The same has been denoted by the tagset I_CP_0. Here, the numerical value zero marked the end of the sentence either instantly at the same point or else at some other places denoting no such constructions anywhere after that. This has reported a higher rate of accuracy with 90.36%. Following the same, criteria the second highest level of accuracy reported by the machine at sentence-level tagging is for the class of B_CP which reported a total accuracy percentage of 82.43%. This shows the beginning of any sentence. Here, one must note that at all such labels the CP constructions were tagged from top to bottom, thus covering the entire structural formation of the sentence. It is also evident to note here that this type of evaluation

has been done with a tool, by applying a descriptive grammar approach, wherein the machine was made learnt with all possible structural rules of the sentence of the concerned language. This was the most trivial method of evaluation as this required a lot more time to construct rules for the machines and make them understand the same with continuous training.

The errors encountered by the tool with such constructions at the sentence level have also been denoted in the same graph. This was measured and identified by applying the deductive and observatory approaches. Here, when the tool automatically tagged the sentence once after training, then the structures and the connecting matrix of the sentences along with their nearby elements were thoroughly examined. After examining both parameters, it was observed that there were a few instances wherein the machine tagged the wrong sentences with the wrong following linguistic particles. This was not only with the case of verbs but also with several other following linguistic elements of the sentence such as adjectives, nouns, adverbs and several other immediate linguistic constituents. This was most frequent in the case of phrasal verbs and during the creation of phrasal integrities. It is due to this reason the intermediate label tag which is I_CP has received an excess of error percentage rates with around 55.43% as a whole.

In order to eradicate such issues in the tool in near future, it is evident not only to have proper tabulated data of the concerned language rather, learning of structural rules for the ordering of sentences should also be induced during machine training. This structural approach of manual correction will not only enhance the functionality and percentage of the tool but also will make the tool cohesive in linking the same, thus reducing the chances of committing any further errors of that sort.

Therefore, with the graph mentioned above one can easily detect and decide not only the types and labels of errors at sentence levels but also, can quickly decide the best possible structural solutions to eradicate or else reduce the number of the same.

6.2. Accumulative Error Analysis of the Tool

This section deals with the overall accumulative error percentage rates encountered by the tool. This shows as to which class or category reports the most issues and has also posed a huge challenge in designing the system tool. For any COLING-related tasks, it is important to report errors if there are any, as it becomes more vital in correcting the tools for any future approaches of the same kind. The advancements of the tool in this accord will not only help in eradicating the issues and errors encountered for the current research as well as for all

other future research of this kind. the below-mentioned pie graph shows the different sorts of errors encountered by the tool. each of them has been represented by its specific percentages along with its nature and class.

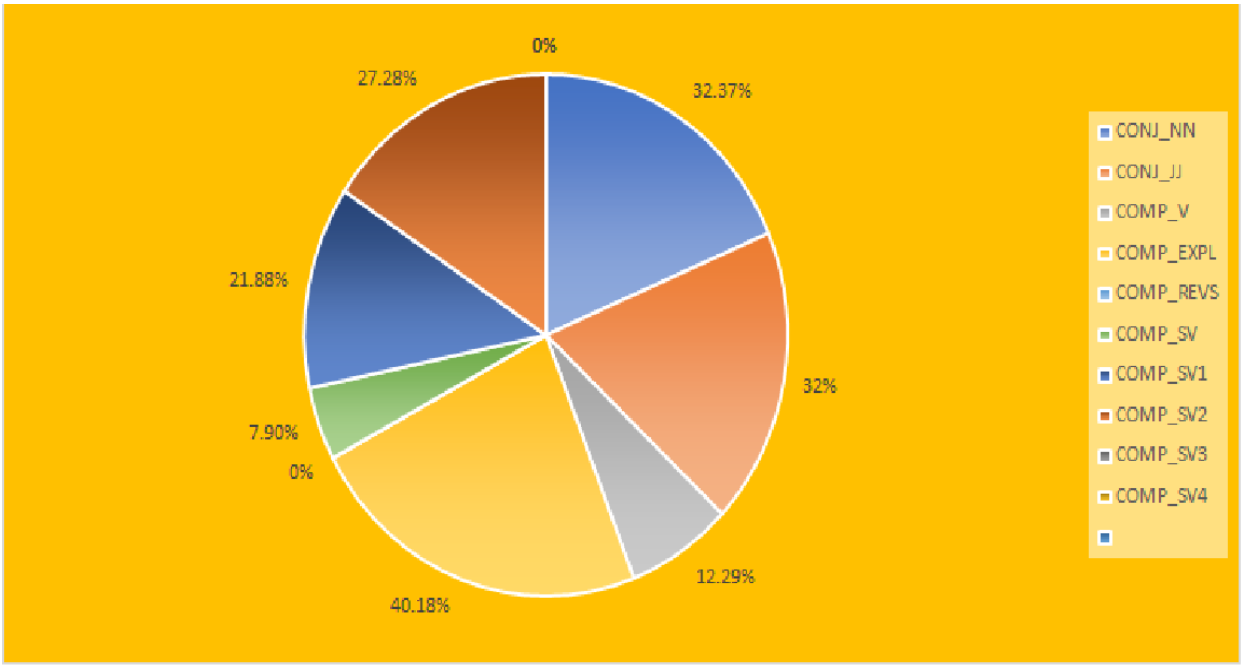

Fig.19 distributive pie diagram showing error rates at all labels of designed tagsets

In the above-mentioned graph, the detailed analysis of errors encountered by the Magahi CP identification and tagger tool has been shown. The highest percentage of error rates has been shown by the explicator compound verbs category with a total percentage of 40.18%. This is so because, the nature of the data collected, contained most of the supportive constructions of this sort and a majority of them have been taken from Hindi. It is due to this reason the machine tool has functioned and resulted mostly considering them as the instances from Hindi. Another important reason is that most of the explicators also functioned as serial verbs, therefore the machine found it a bit difficult to analyse the same and got intermingled with them thus, affecting its performance in this category.

Secondly, it is the nominal conjuncts that have reported the maximum error percentage rates with a total error percentage rate of 32.37%. It is due to the reason that most of the identifications of this sort have overlapped with adjectival conjuncts, which is another separate category of the conjuncts in Magahi. The machine found it tough to identify and analyse these two separate classes of compounds, hence reporting the second topmost errors of this sort. On a third range lies the errors of the serial verbs. This is because the tool found

it hard enough to differentiate between the different classes and subsets of verbs in this category. It is due to the reason that the corpus supported much of the Hindi constructions and not that of core Magahi. It is due to this close contrastive reason that the machine was unable to create a distinction between the two.

On a third range, the least error was reported by the classes of serial verbs, wherein the meaning and the classes got overlapped due to close inference of the two. The classes of serial verbs 1 and serial verbs 2 were the last two categories of this sort. They both reported error per cent of 21.88% and 27.88% respectively. The error with sv2 is huge in this category because most of its instances got intermingled with the sentential level and behaved as a separate class or category which were left unidentified by the machine.

The above-mentioned distribution of error rates in CPs identification of the tool is hence directly connected and contingent to the hits and trials it covered on the unseen data of the concerned language. The machine shows this much of errors because most of the word tokens or the sentences were of open class categories. This means that it was those datasets which have not been assigned with any tagsets and were left blank. These words altered their class or categories while testing thus allowing them to add new tagsets adjacent to their data types. Also, there existed much of English lexicons which were written as it is the supplied corpus. this created a confusion rate for the machine thus, allowing them to commit errors while auto-tagging through machines. There also existed the closed-class categories of the lexicons that permits no space for adding any tagsets adjacent to them. This created a huge problem as it does not allow the tokens to get auto-corrected by assigning auto-corrected tags. Therefore, one can understand that while identifying errors both situations were responsible one the open class categories which allowed the tags to get induced in the datasets, which of course got wrongly infused and the closed class categories which does not allow the categories to be infused with appropriate auto tags because of the insufficiency of spaces adjacent to them. Hence, these both situations can be seen in proportionate nature.

Also, the machine encountered most of the errors due to the possible gap between the two sister languages which are Hindi and Magahi. This created a huge lexicon gap between the two thus, creating challenges and issues for the tool in identifying and assessing them correctly. This means if any word has appeared for the maximum time in the data and if it belongs to any of the mentioned sister languages, then it has auto-tagged the instances belonging to that category only. This was hit by the machine depending on the number and

nature of occurrences that have occurred in the datasets for that particular class or category. This created a huge gap while performing evaluation tasks.

There also existed situations wherein we found it difficult to assign a particular class to a particular set of data, hence leaving those unattended. This unattended data was auto-tagged by the machine with what so ever tag it gives and while evaluating it created the maximum issues, thus resulting in more errors. The availability of pure and consistent data for the concerned language is one of the most vital reasons to face such huge errors. The machine encountered huge errors in analysing the sentential structures as the data for training also known as gold data were short in numbers and were of confusing parameters with each word class or category thus posing a challenge of ambiguity types for both the classes which is sentence or word class labels. This made the machine inconsistent in analysis and identification tasks thus, reporting maximum errors as a whole.

With the nature and types of errors mentioned and discussed above it is evident to note that for evaluation tasks one needs to have ample data only concerning the core language that has been undertaken for the task and not others as this will not only create confusion between the two but will also affect directly to the evaluation patterns of the system or the tool developed.

6.3. Accuracy of Tagsets at Different Ambiguity Labels

This section of the evaluation chapter will briefly discuss the two different sorts of accuracies at two different ambiguity labels. One is at the word or token class as a whole and the another will comprise only the tokens at sentence labels.

6.3.1. Accuracy per class at Ambiguity at Word or Token Labels

For any language, it is natural that there exists some sort of ambiguity. But through linguistics, these nature and classes can be classified, and assessed and can be solved with proposed solutions. Unlike all other languages, Magahi also showed some ambiguity percentages while testing the tool. The designed tool also posed some sort of ambiguities with few words or tokens collected and tested. One faces the issue of ambiguity when a specific word form shows its existence in various contexts, reflecting different grammatical categories. Here in this work, the issue of ambiguity was encountered when the same word or tokens received different tags while auto-tagging, containing different contexts. It is important to note that in this section only the ambiguities at the word or token labels have

been discussed. The ambiguities at sentence labels have been discussed separately in the next upcoming section of this chapter. In order to understand the ambiguities at the word or token labels let us refer to the graph below: -

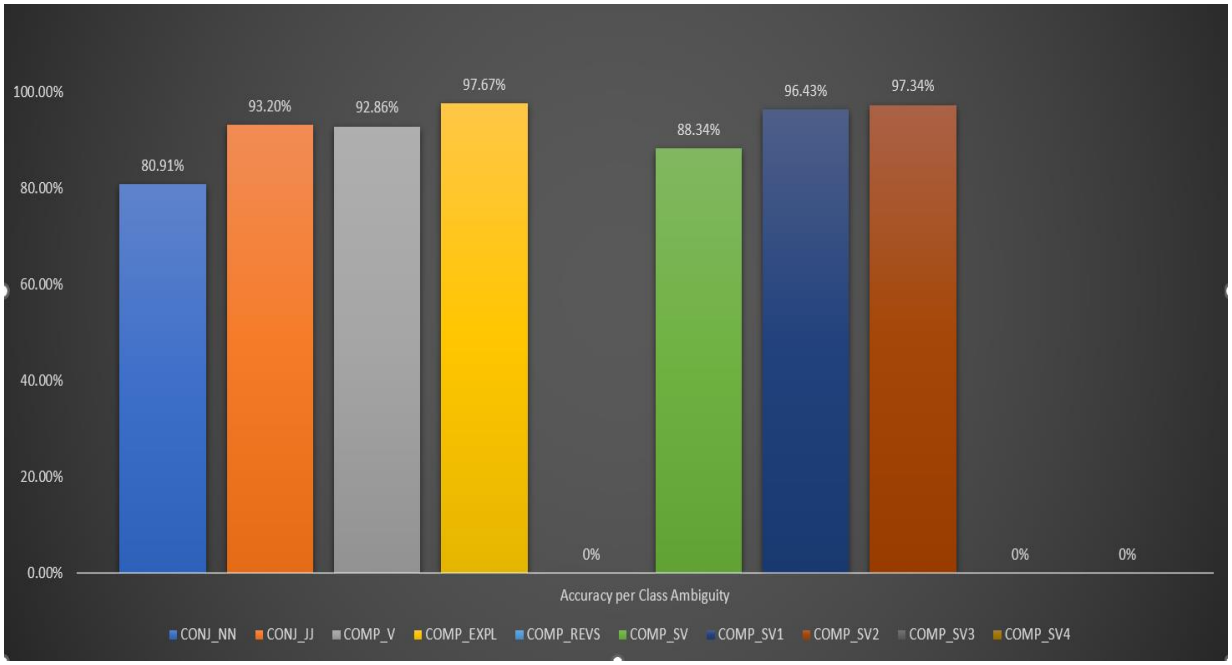

Fig.20 Accuracy per class ambiguity at all token labels

In the graph mentioned above, the maximum percentage of ambiguity was faced and reported mostly by the verbal categories. Out of all the maximum ambiguity is reported by explicator compound verbs with a total percentage of 97.67%. the other two categories of verbs that have reported the most frequent ambiguities are the serial verb classes belonging to serial verb 1 and serial verb 2. This is because most of them occurred in continuity in the tokens, bearing almost the same meaning. They both reported an ambiguity percentage of 96.43% and 97.34% respectively.

Apart from these, the tokens of adjectives and noun conjuncts have also reported much of ambiguities. The adjectival and nominal conjuncts, bearing the tagset labels of CONJ_JJ and CONJ_NN respectively report ambiguity percentages of 93.20% and 80.91%. This is so because most of the time it was difficult to decide as to what notation or tokens a single data should receive. There were also instances, wherein it was hard to assign any tagset to those tokens as a reason why they were left empty. These emptied tokens when left for training, evaluation and testing automatically got tagged by the machine tool on a majority depending on two vital reasons wherein the use of most frequent tags for that sort of tokens

as one while establishing the close parametric relation between the two is the other. This was due to the lexical ambiguity among the closely specified labels of the identical types.

Therefore, in order to reduce this large number of ambiguity types in the data it is important first of all, to collect a bit accurate and relative data of concerned language, which does not overlap with its identical languages in terms of meaning. Also having a specified corpus collection system for any concerned language undertaken for the research is the other. But, so far, the case of Magahi is concerned with the lack and availability of the datatypes in the concerned language is the most vital reason why it has been reported with such a huge percentage of ambiguity, which can also be seen as overlapping its structures and writing system which is identical to that of Hindi and also because the data has been collected from a social media blogging website, therefore, the absence of use of core Magahi terms cannot also be overlooked here in this case. However, the accuracy percentages of the graph above have shown a much positive result as the data collected were from the identical twin set of languages which is Hindi and Magahi, which supports more or less the same structural patterns in their formation process.

6.3.2. Accuracy per class at Ambiguity at Sentence Labels

Ambiguities can be classified and identified at different labels such as sentences, words, tokens or any separate linguistic or grammatical category Crystal (1988). Ambiguity at sentence labels means the idea that exhibits the property of having more than one interpretation of one single sentence.

Unlike all other languages, Magahi also showed a huge amount of ambiguity at sentence labels when being tested on the machine. A detailed analysis of such ambiguity types in the collected and tested data can be understood through the detailed explanation of the below-mentioned graph: -

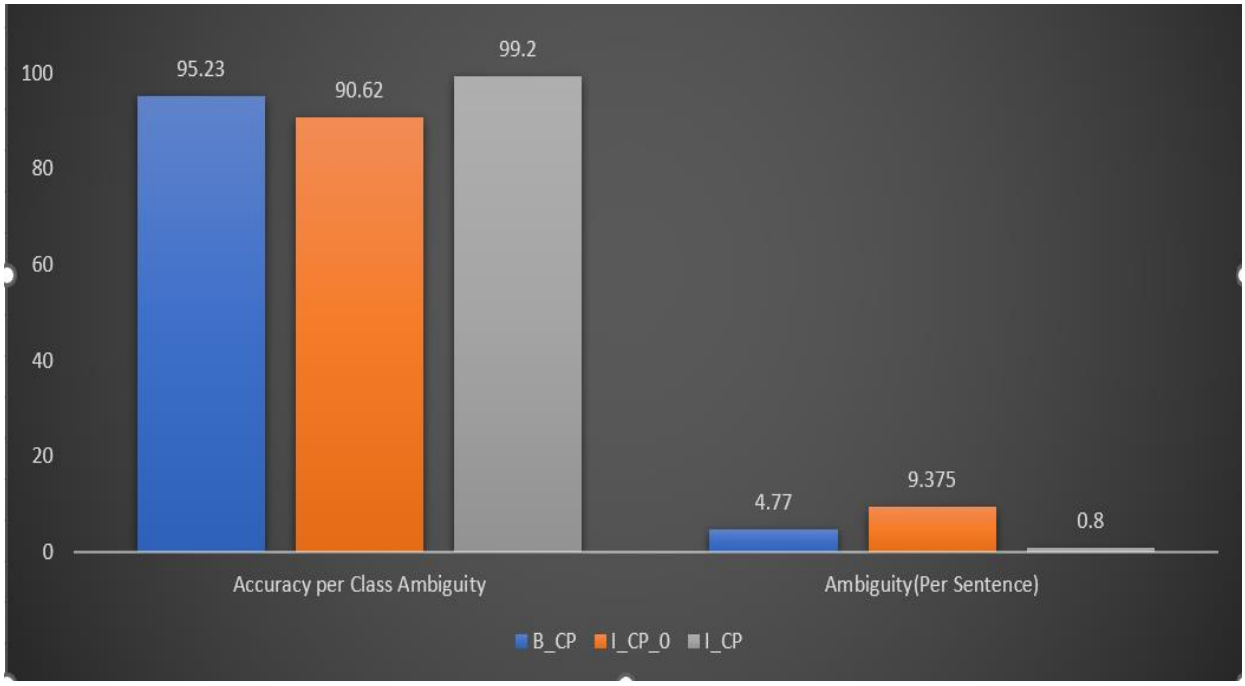

Fig 21 accuracy per class and ambiguity per class at all sentence labels

As mentioned, and depicted from the graph it is evident that the accuracy of ambiguous sentence for each class is not only different but also fluctuating. This fluctuation is directly dependent on their nature and classes of occurrences it encounters with the words or the classes it is adjoined with. The highest level of ambiguity along with its accuracy at sentence label is with intermediate elements of sentences, also known as the inner elements of the sentences. This means that the classes which were left untagged and unspecified with any labels in the sentence have shown much accuracy in terms of their ambiguity. This is because they were not only left free to tag with their specified or desired tags, but also to receive the appropriate tags of ambiguities depending on the nature and uses of such types. It is due to this reason the tool has registered a higher percentage of such types of sentence labels with around 99.20%. these were basically the intermediate phrases which can easily be attached to any sentence in an ordered way in order to to give possible meaningful senses. These were mostly the sentences of short types denoted in the graph as I_CP, meaning intermediate constituents of complex predicate elements.

At the second stage the beginning elements of sentences have ought to show the most ambiguities. This is because, there are possible shortest phrases, which were not only independent but also meaningful bearing valuable structures of Magahi CPs. It has registered an overall accuracy ambiguity of around 95.23% as a whole and is denoted in the graph with

denotation of B_CP. The reason behind giving such a huge positive result of ambiguous sentences for this class is because such types of sentence structures have occurred the most.

On the other hand, the ambiguity of each sentence has also been registered by the tool. Here again the intermediate sentence particles have shown the maximum percentage of results whole testing while, the least has been recorded by the beginning elements of the sentence followed by the final elements of the sentences where the instances have been finished. Adjoining all these classes the ambiguity at per sentence label has been recorded differently for all the three different categories. As shown in the graph a maximum of 9.375% which is around 9.4% has been reported by the classes denoted with I_CP_0. This is because most of the sentences either ended up here or else have been unable to find any appropriate tag for the required data which were left unattended or untagged. The second most highly recorded ambiguity at each sentence level tagsets was with the beginning classes of the sentences. These are denoted with B_CP having a total of 4.77% and the least was recorded by the intermediate tagset labels denoted with I_CP. This has a total ambiguous percentage of around 0.8% or around about 1%.

The accuracy of ambiguities at each label of sentences was recorded this low was only due to the lack of proper data. This implies that once the SVM tool receives the stipulated and concerned data of only the concerned language, it will definitely receive higher accuracy rates. However, the issues of ambiguities can be solved by the solutions suggested in this work and also by the ample use of Magahi by Magahi-speaking people. This will not only create lots of data but also will give a proper rule, that will help in decoding these languages syntactically and computationally in a quite well manner. The study of grammatical patterning and orders will also help in decreasing such huge ambiguity issues, thus making the tool and future NLP research of this sort will be quite effective for any less-resourced languages. This will also be helpful in easily contextualizing the terms of the language hence, giving the most positive, possible and desired outcomes thus, reducing the possibility of any sort of ambiguities from the languages concerned. Most of the ambiguity issues have been encountered in the corpus due to the pattern identification. Therefore, in order to eliminate these pattern ordering phenomena, it is high time for Magahi to shift to its own grammatical patterning and applied rules in order to overcome such ambiguity issues which is either at per class of word or token labels or else per sentence labels.

6.4. Issues and Challenges

This section of the chapter will briefly discuss the possible issues and challenges we have encountered while developing the tool for CP identification and analysis in Magahi.

It will discuss some theoretical as well as a few COLING issues and challenges through its different sections and sub-sections. It will also discuss a few constraints either theoretically or computationally that have been encountered during this research.

6.4.1. Theoretical issues

This section will deal with the possible theoretical issues encountered during corpus selection and human annotation.

6.4.1.1. Derivational Issues

For both theoretical and practical studies of language, complex predicates have received a lot of attention. This is because of their morphological, syntactic, and semantic behaviour. The structures of complex predicates are fascinating. The internal and external occurrence also known as the morpho-syntactic characters of various kinds of Complex predicates has been one of the research concerns of this area. The study of complex predicates has made it more intriguing and difficult as this derivational occurrence of complex predicate constructions has made it quite distinctive in this field.

It is no denying the fact that this detailed study of CPs formation in Magahi along with its construction and identification process through NLP techniques also examines the distribution of many of its derived forms syntactically. While constructing any CP, it has been observed that the study explored many derived forms of both the main verb and the light verb. This is concerning mainly the derivational process that leads to such magical creations in Magahi which undoubtedly advanced its typological research of the languages concerned along with its respective region wherever it is spoken.

The idea of derivation denotes the creation of a new term or word class. unlike all other languages, Magahi also has this very crucial property. Studying such concepts in Magahi has made it more engaging and exciting. The derivational issues of complex predicates have been further classified and explained under two broader sub-categories, which are as follows: -

6.4.1.1.1. Morphological Derivation

Languages like Magahi exhibit a large number of derivative verbs. These basically come largely from nonverbal roots in Magahi. In Magahi language, there exists certain verbs which have Nouns or Adjectives as their roots. These constructions tend to derive the construction of complicated predicates.

For example: - 67. भैंसिया दिन भर धूपवा में रहे से तता गेलय

bʰaiⁿsijɑ ɖinə bʰərə ɖʱupəvɑ meⁿ rəhe se ʈəʈɑ geləjə

'The buffalo dried up standing in the sunlight for the entire day.'

Here, in the example (67) above one can see that Magahi has instances of CP constructions like /ʈəʈɑ geləjə/, meaning 'dryied up', which needs the addition of some nominal suffixes like /-jɑ/ in order to make it a proper verb combination and also thus framing a CP construction. In the same manner, there were also instances in the corpus which resemble similar cases. Few of them are /ləʈijɑnɑ/, /ɖʱusijɑnɑ/, /cicijɑnɑ/, /gʰigʰijɑnɑ/ etc., which need a noun suffix /--jɑ/ to form a successful verb form. For these constructions, they also need a second verb to explicate its meaning. The light verbs attached to all such main verbs in Magahi are similar to Hindi and thus they also allow instances like /lenɑ/ or /ɖenɑ/, to complete their CP formations. The instances of the verb mentioned in the examples above successfully frame a CP resulting in formation of a single verb altogether, thus denoting quick action. It is also observed that the CP formations in Magahi suffer combinatorial restrictions sometimes. This was true in both the cases such as for nouns as well as for adjectives as well. The extra suffixes attached to all such verbs becomes very difficult to identify and analyse by the machine tool. It is due to this reason that such types of instances have received the wrong tag possibly. This was sole reason why the machine tool was not able to identify these combinations, even after making manual corrections which finally affected the result and final evaluation of the tool, thus giving a high error percentage of such kinds.

6.4.1.1.2. Syntactic Derivation

This was similar to that of morphological ones. Here, the base word was used in a different syntactic environment, maintaining the essence of its meaning. However, a change of meaning sometimes cannot be overlooked. This seems to have a great theoretical issue

while framing and identifying the corpus. During the CP formation in Magahi, there exist different sorts of light verbs in this language that actually serve most of the time as auxiliary markers.

For example: - 68. उ जाइये रहल हे

u jɑije rəhələ he

'He is going'

Here, in example (68) we can see that the two simple verbs which are /jɑnɑ/ and /rəhənɑ/ were used not only to mark the action but also the continuity of the entire sentence. This means that the sentence itself implies that the task of going has not been finished yet thus, forming a case of serial verbs. But the machine found it a bit difficult and challenging in identifying such instances despite, giving some human annotations into the data for the same. The machine time and again tags it only under the general category of verbs such as compounds or explicators and not as the sub-classes of serial verbs. Therefore, these sorts of instances into the data were not tagged correctly by the machine thus, needing more human efforts in order to correct the system hence, providing the tool with a set of metadata for better training, which means the more manual correction it involves, the better output of the result for the tool we will get.

In other words, one could also classify such constructions as the syntactic issue of complex predicate constructions. This is because such issues are encountered mostly with the cases of nouns, modals and participles, which falls under the broad category of conjuncts. The training of machines to deal with such constructions in Magahi CP identification process poses many difficulties pertaining to several other theoretical issues like this. Hence, we can conclude that the enhancement of such characteristics in the raw data will surely increase the accuracy and functioning of the tool, thus making the study of such concepts in Magahi a vital area of research applying modern aids and technologies.

6.4.2. Imitative or Echoic Construction Issues

The words of any language mimic a sound while framing imitative or onomatopoeic constructions. Languages like Hindi, Magahi also has this feature of construction. Most of such types of constructions in Magahi function like CPs. Here these constructions function as the primary verbs or the main verbs.

For example: - 69. ओकर बतिया सुन के हमर दिमगवा झनझना गेलय

okərə bət̪ija sunə ke həmərə d̪iməgəva jʰənəjʰəna geləjə

‘Listening to his/her chats blew my mind.’

70. मास्टर के डांटे से रमुआ थरथराये लगलय

maʂt̪ərə ke ɖaⁿt̪e se rəmua ʈ̪ʰərət̪ʰəraje ləgələjə

‘Ramu trembled after the teacher scolded’.

In the examples above which is (69) and (70) constructions like /jʰənəjʰəna geləjə/ and /ʈ̪ʰərət̪ʰəraje ləgələjə/ which means ‘to temble’ and ‘to tremble’ respectively are created through the imitation and reiteration of words like /jʰənəjʰəna/ and /ʈ̪ʰərət̪ʰəranа/ respectively. Such constructions act here as verbs, but for the machine, it was difficult and hard to identify these terms due to which the tool tagged it wrongly quite at some instances, which of has directly affected the final output percentage of accuracy.

6.4.3. Issues with Reduplicative Constructions

Reduplicative construction is a type of word formation process in linguistics. In this process sometimes the root, stem or even the entire word is repeated either exactly or with some small alterations to it. Unlike all other Indian and South-Asian languages in particular Magahi also exhibits this special property. In Magahi language there exist certain verbs which occur as the main verb while being reduplicated thus, resulting in a CP formation. The reduplicated constructions in Magahi mostly have shown issues at three major levels such as compounds, partial reduplications and partial corelative reduplications. Each of them is discussed in detail here under: -

6.4.3.1.Compounds

These are partially reduplicated structures in Magahi. As Magahi is not a major language hence, the possibilities of having meaningful reduplications are quite less in this language. The compounds in Magahi as already discussed are paired components wherein the later component of the construction bears little meaning and resemblance to the former but is not exactly a precise repetition of the former.

For example: - 71. मोहना के आवे के कुछो सुन-गुन न हई

mohənɑ ke ɑve ke kuchʰo sunə-gunə nə həɪ

‘There is no news of Mohan’s arrival’.

Here, in example (71) above the constructions like /sunə-gunə/, meaning ‘information/hearing in English are formed actually with reduplications. But for all such constructions the tool marked the first instance as the main verb while the other as an adjective despite being the fact that it is actually a element of the reduplicative compound construction. In this work, all such constructions posed a great challenge for the machine to identify its respective and definite category, thus giving them a wrong tag, which is another vital reason for the decreasing level of tool’s accuracy. Such issues are also termed as issues of semantic reduplications in COLING.

6.4.3.2. Incomplete Reduplications

These reduplicative constructions are being used to describe the nature of an occurrence in Magahi. Complex Predicates in this language are being constructed by reduplicating a noun or verb base by simply adding /aa/or /ii/. This implies that the action is carried out in a group or between two or more people.

For example: - 72. लइकन खेल-कूद में मारा-पिटी कर लेलके हल

ləikənə kʰelə-kuɖə meⁿ mɑrɑ-piʈi kərə leləke hələ

‘Children started fighting while playing’.

In this example (72) one can see that this is a construction which involves two different set of events. Here, the first part of the constructed CP is half reduplicated. This is so because the meaning of the first compound is not clear until it is not attached to the second one. Such CPs could not be found in Magahi without the existence of a verb form called /kərə/ meaning ‘to do’ or the v3 form of this verb. Therefore, the machine tends to identify such constructions of CPs as serial verbs, wherein it identified half of the sentence part as serial verbs. Here, the machine identifies /mɑrɑ-piʈi/ as a class of serial verbs whereas it is actually a reduplication compound. Such wrong predictions by the tool hence, created a hindrance for the machine in identifications and analysis task.

6.4.3.3.Complete Reduplications

This sort of CPs formation in Magahi exhibits the tendency of repetitions. It is formed by adding the suffix /-aa/, which is adjoined with a simple verb in order to create a CP construction. Such structures of Magahi CPs indicate that the task undertaken needs to be completed or might involve some other agents or people to finish the same. In order to put it in another way, we can say that it is used for showing the divergences in action verbs.

For example: -73. ओहिजा जाय से पहिले सब समान मिला-जुला ले न तो बाद में दिक्कत होतउ

ohiɟɑ jɑjə -se pəhile səbə səmɑnə milɑ-julɑ le -nə ʈo bɑɖə meⁿ ɖikkʈə hoʈəu

'Check all your luggage before going there otherwise it will be a problem, once you return'.

Example (73) is a clear illustration of complete or full reduplication. Here the reduplication construction /milɑ-julɑ/ meaning 'to check' shows the right and accurate meaning in the sentence. Such reduplications could be of any POS category before being framed together, but once framed, it will only form a verb. For all such constructions, the machine tool not only find it difficult to identify but also it acted as a hindrance for identifying their CP labels. Here in this example the tool marked it as verb for the first instance whereas noun for the other. Hence, all such constructions like /milɑ-julɑ/, /hilə-milə/ etc are marked differently by the tool, as the it found such instance to be a bit challenging while identification and classification, which has directly affected its overall performance.

Apart from all the theoretical issues mentioned above in detail, there also exist a few more challenges for the machine tools which has been developed to identify the CPs in Magahi. The issues pertaining to machine learning for the taggers are discussed below in detail: -

6.4.4. Issues in Framing Tagsets

While framing a tagset for any natural language, there are numerous issues faced. Out of all assigning the right CPs to the tokens is the most vital. This becomes even more applicable when it comes to handle the case of low-resourced languages like Magahi. For a language like Magahi designing a tagset is actually a trivial task. This is due to two most prominent reasons which are, it does not follow its own script for writing as it is similar to that of Hindi as one while the relation of these languages to the same language family is the other.

As we all know that in order to perform any task which is concerning NLP or COLING applications, we need to create the corpus first. This could only be done by tagging the data or the corpus, which is the first stage of corpus creation. For this, we also need to tag the corpus with their desired tagset, which is also the initial step of data annotation. This is required in order to convert the corpus into the usable format of the machine. But this was not true for Magahi as this does not have its own set of tagset which can be used for the desired purpose. As a result, to this, the BIS tagset of Indian languages has been borrowed, which is the most standard tagset for all standard Indian languages.

It is due to this reason, that there existed some instances of Magahi which cannot get tagged appropriately by the BIS tagset as the tool cannot identify its correct POS category. Another reason for this is the borrowed uses of Hindi in this language, which have created a huge hindrance for the tool in order to decide whether it has been annotated correctly or not. Therefore, the issue of tagset creation is one of the most prominent reasons as to why the tool was not that accurate and precise in its performance.

6.4.5. Corpus-Related Issues

There exist multiple challenges to corpus collection. This includes lack of Unicode coding for languages like Magahi, its substratum status, use of different writing scripts, and similarities with Hindi. These pose a great challenge for refining a corpus. There has been a massive development in corpus collection area for languages like English, Hindi and many other foreign languages like German, French etc. But when it comes to lesser-known languages like Magahi the situation seems to have been quite disappointing. This is also due to the reason that the govt. is not giving much importance to these languages due to which many of them have been overlooked until today. The corpus collection task has been successfully undertaken and accomplished by several prominent research institutions like IIIT Hyderabad, ILCI-JNU, CIIL Mysore, TDIL etc.

6.4.6. Lack of Unicode Encoding

As mentioned, the low-resourced languages have not received much exposure towards the NLP aids and latest technologies therefore, it is quite obvious that there is not much software or latest aids available for their text encodings. For any language in the world and its NLP-related tasks, Unicode encoding is an NLP-based technique or information technology standard that makes the task easily readable, acceptable and capable for the system tools to

read. It is of fixed length and character that encodes any natural language of the world, with the help of special characters present in it.

As Magahi is not much studied in terms of NLP-based tasks and techniques, therefore, there are no specific Unicode encoding techniques available till date for this language. It just uses the UTF-8 encodings, which is being used by most of the Indian languages of which Hindi is of utmost importance. As this is also an encoding character of standard Hindi therefore, it is possible that it might create some issues or errors while encoding the characters of Magahi, which later created an issue in identifying the same by the machine. It is due to this reason that the entire character of Magahi has been first converted to a UTF-8[36] encoding character and then the same has been inserted in the tool for the automation process of tagging, wherein certain special Magahi characters were not easily identified. Apart from this there also existed certain phrases, sentences or tokens or words that got intermixed with standard Hindi due to which the tool has identified them as per the standardized Hindi norms, which has created a huge issue for the tools to identify clearly. It is due to this reason that it required a lot more human effort, in the end, to correct those instances of Magahi manually, which has been wrongly encoded by the machines as per the Hindi standards.

6.4.7. Non-Classified usage of Words or Expressions

Magahi being spoken in different regions of Bihar and also in some parts of its close neighbouring states like Jharkhand, poses different usage of a single expression or word. It is not only spoken differently but also written and understood vividly across all different Magahi-speaking regions. Due to this variance in its uses, it has many different dialectical variations in and around the regions of Bihar. For example, in some regions of Bihar like Nawada, Patna and Jehanabad the word /nə/ meaning ‘no’ is being spoken as /naɪ/ whereas the same term is simply spoken as /nə/, in all other regions of Bihar especially the central region such as Gaya. It is spoken with strong stress over the word.

Similar is the case for another term / kocɪ/, which is a Magahi word meaning ‘what’. This is being spoken as /koccɪ/ or else /koⁿcɪ/ in some regions of Bihar such as Gaya, Patna, Jehanabad, and Nawada. Here it is evident to note that the use of different convention for /nə/ or /naɪ/ does not make any major difference during identification but, the instances of using

[36] It is a unique transformation format for encoding the common characters of Indian languages used by world wide web format.

/kocɪ/ which changed its form to /koⁿcɪ/ made a huge difference to the tool's performance posing a huge challenge to identify the changes and thus tagging it to a different tag. Similarly, there are several pronouns in Magahi such as /t̪e/, /t̪u/, /u/,/uhə/, /uhɪ/, /ɪhi/,/ehə/, /ehɪ/, which is used in place of this, that, he, she it, they etc. are pronounced differently in different regions of Bihar.

Here it is evident to note here that the meaning of such instances does not alter the meaning even after being spoken differently, but has created a huge challenge for the tool to make clear distinction between the two, wherein one is identified as Wh entity, while the other was left blank as unidentified. At the same time, it is also important to note that due to this huge and high variance in its usage from one part of Bihar to another these different terms and patterns of Magahi, made it complex for such data instances to get easily identified by the tool. All such changes of use of words were kept under non-classified category by the tool, as it was left unidentified, even after testing and evaluation, thus hampering overall performance of the tool.

6.4.8. Difference in Writing Systems

Magahi being spoken differently in various regions of Bihar also has issues with its orthographical system. This is so because Magahi exhibits the property of non-uniformity while writing. This means that any single expression while writing is written differently, which is dependent on the region and area in which it is spoken. This is also because the words of Magahi adjoin in a disorganized fashion.

For example: - The word /rɑmɑjəɳə/, in standard Hindi is written with /ɳ/ but in Magahi the same word is written as /rəmɑjənə/ having some changes in its writing formats like the use of /rə/ in place of /rɑ/ as well as the use of /nə/ in place of /ɳə/ at the end. This created issues for the tool to identify whether it is the correct convention used for the given word or not. This difference thus creates a hindrance for the tool to identify such terms or words. It is difficult for machines to detect such minute orthographical differences in Magahi which occurred frequently in this language. This is also due to its variety of unorganized usages, from place to place or person to person. One of the most important reasons behind this is because, it does not have any standardized form of writing and separate scripts like Hindi wherein it can use such standard forms. This disorganized usage of Magahi scripts creates huge issues during both human and automatic annotations simultaneously. Due to this huge difference in Magahi writing systems can be seen across different places in Bihar and its

other neighbouring regions like Jharkhand. Therefore, to address all such alterations in writing system of Magahi, separate conventional labels must be designed for its annotation purposes, or else a standard writing script must be developed just like it has Kaithi in past, which should be revived again to analyse such basic differences thoroughly and deeply.

6.4.9. Hindi-Oriented Syntax

Magahi is much similar to that of Hindi in terms of its syntax, grammatical constructions, and usage of words in a sentence. This is due to the reason that it belongs to the same Indo-Aryan language family group as well as it has a strong influence of Hindi as most Magahi speakers have adopted Hindi in majority of their uses. It is also due to the reason that Magahi belongs to such geographical areas which are much closer to prominent Hindi-speaking regions such as states like Jharkhand and Bihar. These states are much more influential towards the use of standard Hindi in most of their day-to-day daily needs. The use of a similar word order pattern by Magahi is strong evidence of this claim.

Therefore, the corpus collected has several similar words in Hindi such as /rɑjɑ-rɑni/, /beʈi-beʈɑ/, /gʰərə-d̪vɑrə/ etc. Such instances occurred in the corpus because, as already mentioned, it is an intermixing of the standard Hindi data along with Magahi. The data produced here is not an actual form of Magahi rather, as translated version from Hindi. It is due to this reason there are instances like mentioned-above can be seen here thus creating an intermixing with the corpus collected This intermixing has thus created confusion for the tool in identifying these words properly, thus creating an issue in its evaluation as most of them have been left unidentified and have received much overlapping to that of Hindi entities.

6.4.10. Ambiguities

This was a major challenge which affected the smooth functioning of the tool. In the corpus collected for Magahi, there exists a lot of ambiguous words or sentences that were not very easy to identify in terms of their structures or POS classes. It is due to this reason they were unable to receive a proper tag, which created issues for the tool in terms of their accuracies. This created a major problem while classifying the sequence of a serial verb. This also posed a challenge in identifying whether it is a compound or serial verb set of two or three or even more constituents. The ambiguity issues were faced because of the conflict in the two-label set or three-level set of words or even more.

This was most prevalent while identifying words like /sonɑ/ meaning ‘gold’ and also ‘sleep’ /hɑrə/ meaning ‘defeat’ and also ‘garland’ /ɖigri/ meaning ‘temperature’ and also ‘qualification’ /ɑmə/ meaning ‘mango’ fruit and also ‘general’, /pʰələ/ meaning ‘fruit’ or else ‘result’ at times whenever or wherever required as per the need and demand of the sentence. All of such categories of words keeps altering their class or categories while identification process by the tool. it is due to this reason they are sometimes tagged wrongly or else have received different tags irrespective of their usage in the data. This overlapping of tagsets is mostly seen with verb phrases, which conveyed two or even more than two different meanings at the same time. This has confused the tool a bit thus affecting its performance which later needs manual clarifications and corrections in order to make them properly functional.

6.4.11. Homonyms

Languages have a special feature of holding multiple meanings of one single word. The same is also true for Magahi. This also has words which exhibit multiple meanings. This is due to the changing situation and nature of the speaker. As this is a corpus collected from a literary blogging website which has the literal Magahi translation of fairy tales in Hindi therefore, it has such homonyms present in them. These are basically borrowed or loaned from Hindi. Some of them are /rɑsə/, /ʈə̪ʈvə/ /ɑnⁿɖ̪ə/, etc. The tool also found difficulties to differentiate between the words like /kʰusɪ/ and /ɑnⁿɖ̪ə/ as to what grammatical category it should go with. Prima-facie the tool tagged such instances as an adjective but, in most cases, it is being tagged as a verb or noun depending on the situation and pragmatic requirement of the sentence. This created a state of confusion for the tool, thus giving errors for all such instances, which have later been corrected manually by comparing the words or tokens with the context and the reference for which it has been used.

Apart from these classified and specific issues there also existed some common issues wherein we have also found instances wherein a single word is further fragmented into several parts. For example- /cɑpɑkələ/ meaning ‘handpump’ is fragmented into two different words such as /cɑpɑ/ and /kələ/, after crawling and sanitizing the corpus. This is because the IL sanitizer and crawling tool consider these both as two different entities while creating separate classes. Later on, when the fragmented data is trained with the machine, then they both receives two different tags, which of course have posed a great challenge in identifying such terms. Here the tool was found it a difficult to classify and identify them as one single

entity. Such issues were handled by making manual corrections in order to increase the efficiency of the tool. Along with this, there were also some English terms, which were written as it is in the data, which the tool found a bit difficult and challenging to identify and diagnose its CP category. Few among those instances were /vɑʈərə ɦɑrvesʈiⁿgə/, /pərəmɑneⁿʈə seʈələmeⁿʈə/, /holɖiⁿgə sʈeɕənə/ etc. The tool wrongly tagged such elements of the data considering those terms from the English language and not from the view of the concerned corpus.

Likewise, there were also some ambiguous terms like /sonɑ/, which means ‘to sleep’ in some cases in the corpus while it is also referred as ‘gold’ in some other instances. Considering such cases, the tool has committed errors while tagging words or tokens like these. Such issues were very common while tagging the corpus at the sentence level. Therefore, it can be seen in the corpus that the ambiguity issues were more frequent while tagging the corpus at the sentence level.

As per the issues mentioned above, it is evident to note here that human languages can be modelled and modified in numerous ways because of which the same meaning can be expressed vividly. Therefore, the approaches of NLP and ML recognise their essential qualities purely based on these directions. Due to this, it is important to recognise that human language does not always adhere to any specific conventions thus resulting in alterations to it depending from one language to another. Most often, the differences occur because of the inference of the meaning of phrases applied in a different context, which might not have any impact on communication but it has a huge effect on words wherein certain words are being omitted or deleted along with the rule of grammar applied which may be possible not up to the mark.

6.5. Suggested Solutions for the Magahi CP Statistical Identifier

This section of the chapter includes various different ways that can be implemented and used as solutions for the future betterment of the statistical CP identifier tool. This includes several ways that can be used later on for enhancing the performance of this statistical tagger in order to make it perform a bit better in terms of quality, reliability, and efficiency. This includes several different developmental approaches, including the data approach, word sense disambiguation method, and hybridising taggers by developing and formulating certain linguistic rules for the language concerned for which the tool has been developed. All these approaches can be taken as suggestive measures for futuristic

researchers in order to make this developed tool a bit better in its performance as this needs a lot more improvements from time to time. These were not been implemented during this research because of two reasons wherein time constraint issue is one and developing the tool as a pilot module for such low-resourced languages is the other. Hence these solutions could be seen as suggestions by the researcher and not an implementation.

6.5.1. Framing Language-specific Linguistic Rules

This can be seen as one of the most specific steps taken towards the improvement of the tool. This will not only increase the efficiency of the tool by reducing its errors but can also be quite helpful for structuring the grammatical patterns of the language in a more well-ordered manner. This can make the language more effective, accurate and globally accepted. This can be done more specifically by closely monitoring the errors made by the developed statistical Magahi CP tagger. These linguistic rules are completely a hybrid approach which can later be encoded inside the tool in order to improve its performance and accuracy as a whole. These encoded linguistic rules make the tool a bit more hybrid.

As we all know that SVM module follows the learned module techniques for its identification and classification purposes, therefore this only annotates the data after imitation therefore, it is important to make necessary syntactical changes in the gold file of the data before we put them to training and evaluation. This will help the SVM-based modules to learn the grammatical patterns and corrections which are induced in the raw data thus resulting in more accurate and appropriate outputs. This will also help in eventually identifying the number of desired occurrences inside the raw data thus imitating the same and allocating the respective required tags to the concerned tokens as it is there in the gold file. This means that it will directly affect the chances and numbers of a single CP class, thus preventing them from being imitated thus allocating the same to a wrong token or word class. This will eventually help in the evaluation process by reducing the number of imitations of CP tags of the output file.

By following the above steps, it is a high chance that the tool selects the most likely and frequently used CP label for generating outputs for any specific input token belonging to that specific proper token or word class. For example, if in case of serial verbs there exists three or four different consecutive labels such as SV1, SV2, SV3 or SV4 then the first two have the most occurrences inside the whole data or the latter has the most occurrences or even either of them has the least ones, then definitely the tool will take the maximum level of

occurrences for its evaluation purposes, which will surely increase the tool proficiency thus reducing much of the human efforts of corrections and manual insertions. This insertion of occurrence and patterning the same in terms of their relevance and order can be seen as the most hybrid method which can not only contextualize the data but also its performance and results which will be completely based on the rules mentioned above for data insertion. Let us see some proposed rules mentioned below which can be used for improving the tool efficiency: -

- Wherever there are nouns coming along with any verbs, we have marked the nouns as nominal conjuncts with the tag CP_CONJ_NN and the verb as the explicator verbs as this explicated the meaning of the entire phrase. These explicators are marked with the tag CP_COMP_EXPL. This is done so because sometimes it bears the entire meaning of the phrase or tokens.
- Sometimes there existed some long sentences which required tagging, therefore, the entire sentence or the phrase was tagged. Here the beginning of the sentence was marked with a capital B whereas the internal part of the sentence was marked as intermediate, with the tag symbol I. this entire phrase then received the required appropriate tag depending on the nature of the sentence. This means the beginning was marked with the notation B_CP whereas the intermediate part was marked as I_CP.
- Compound Verb, in general, is a two-verb construction which comprises v1 and v2, wherein v1 the first verb is polar and v2 acts as the second verb which is vector. But in case of a CV formation, any verb can occur at any place as v1 or v2. Such constructions in Hindi or Magahi are framed only with instances like lenɑ/d̪enɑ or ɑnɑ/jɑnɑ or else uʈʰənɑ baiʈʰənɑ. For all such cases, v1 acts as the main verb stem, while all other instances act as v2 receiving all the inflections for gender, number and tense aspect and mood. Once this happens V1 loses its lexical meaning but adds a more meaningful sense to the main verb.
- For any CP formation comprising compound verbs, lenɑ is used when the effect of the verb v1 goes towards the subject or is self-directed, but d̪enɑ is used when the effect of the v1, denotes an action which is directed in some other ways.
- It is important to note that compound verb construction cannot be used in negative sentences or with adverbs of negative sense.

- In case of Serial verbs, four different categories were made in order to mark the sequencing and patterning of such constructions.
- Whenever any word or token is preceded by an adjective or a noun or any conjunct word among these two, it is marked as a conjunct by default, Nainwani (2012).
- Whenever there exists a single verb in the entire phrase or sentence along with a noun or an adverb, and the verb bears the entire meaning of the sentence then in such cases, it is marked as an explicator.
- The cases of reverse compounds are not much found in the corpus collected however, for all such cases wherein such instances were faced, these were marked either explicator, normal verbs, compound verbs or else reverse compounds depending on the nature and functions of the same, which of course was decided as final once the context or the pragmatic relation among them was set up successfully.

6.5.1.1. Corpus-Related Approaches

This includes several different corpus-related solutions that will eventually help in increasing the accuracy and performance of the tool developed. These can be comprised of correcting and increasing the corpus. in order to make the tool effective in terms of its accuracy one needs to obtain this data-driven approach. This is crucial as this will not only improve the accuracy of the tool but also to develop a well-defined corpus for any future searches concerning NLP applications or general linguistics in Magahi.

This corpus-related solution can be implemented manually also, as we all know that Magahi being very less used as a means of communication in today's scenario does not have much exposure. Therefore, it is our duty and responsibility to promote languages like Magahi more and more into our daily needs and uses of day-to-day life. This will also avoid any further intermixing of the data with its similar and identical languages like Hindi, which was one of the major hindrances of this research. This means that if one has only Magahi-specific data wherein more and more raw and original Magahi words or expressions are used, then it will become quite easier to get it verified by the local speakers and tag them accordingly, which will later be supplied to NLP tools in order to execute any COLING related task smoothly.

Under this approach, the results were evaluated and analysed accordingly after rectifying each of the errors encountered. This involved multiple cycles of evaluations wherein the data was corrected time and again while each evaluation cycle was performed. With each

assessment step, the result and performance of the tool were noted and errors encountered during the cycles have been rectified and sorted accordingly. This was a very trivial task as it involved much time, effort and corpus corrections however, there still existed some issues which is due to the similarities between Hindi and Magahi which hampered the tool's performance. It is due to this reason having a specific Magahi corpus from the core Magahi-specific language has been suggested for future approaches in the concerned field.

6.5.1.2. Assigning appropriate CP Labels to avoid Overlapping of Words or Tokens

While performing the task of annotation it was often found tough to decide as to what annotation label a token or word or even a sentence should receive. Sometimes it was made easier by understanding the pragmatics of the sentence but, there also existed a few situations wherein even the pragmatics were not good enough to decide as with what tag a token should be given. Such issues were faced most often when there existed ambiguous sentences or word tokens. During such cases, the core knowledge of the Magahi speakers was given utmost priority in order to decide the appropriate tag for such tokens. This is so because there were no written grammar or rules concerning Magahi is present for now. This was one of the most complex decisions taken while tagging such a corpus. This made the task a bit easier in annotating the data at POS labels at first and then a bit further for CP annotation. This step was also taken to avoid the issue of collocations, which have been found in most of the words or tokens.

6.5.1.3. Rhetorical Linguistic Replacements

While performing the task of annotation we have come across a lot more words that have been directly taken from English and are written as it is in the corpus. this is so because the corpus has not been taken from any Magahi-specific source rather, from a blogging website which has used a very common language of layman style. It is due to this reason words of these English instances have been translated further to first in standard Hindi and then to local Magahi in order to make it adaptive for the corpus. the terms which were finally translated into Magahi were hence made equal in terms of their meaning in order to make it fit for the corpus undertaken for evaluation.

These are most importantly the terms denoting the natural environments, elements of uses, commonly used articles and items etc. After being translated into the desired target language which is Magahi, such terms, words or tokens or even sentences do not possess much issues while annotation or at final stage of further evaluation. This is so because most

of the terms have received the simple terms of their uses which were common at large in both the languages which are Hindi and Magahi. These were mostly done at the levels of CONJ_NN which is nominal conjunct and CONJ_JJ which is adjectival conjunct. One of the most common examples of such linguistic replacements was of /suⁿɖ̤ərə/, or /kʰubəsurəʈ̤ə/ which is also referred as /suɕilə/ in standard Hindi along with /suⁿɖ̤ərə/, and /kʰubəsurəʈ̤ə/ but both of them were further translated as /besə/, /nimnə/ or /nikə/ in Magahi. In connection with this, there were also some similar translations made for many abstract entities which designate or denote the characteristics, amounts, and degrees of things, events, actions etc.

6.5.1.4. Replacing Pronouns

Since pronominal forms are used in the most semantically identical and consistent manner in both Hindi and Magahi hence, it was believed that there will not be much issues when the pronominal forms are translated between the two languages. However, this phenomenon was not proved to be true in terms of actual translation of instances when performed at the sentence or phrase level. These were only true and successful at tokens or word levels. These were only effective to demonstratives, relatives and interrogatives set of pronominal occurrences. While performing the task of manual tagging and making the data set feasible for evaluating the model, we have translated a bit of personal Hindi occurrences of pronouns to Magahi, so that this can help in increasing the level of accuracy percentage of the model or the tool, however, this has not made any big differences in the task. Some of the changes made in the pronouns were from /ʈ̤umə/ to /ʈ̤ohərə/, meaning ‘you’ plural informal. /ʈ̤uməlogə/ to /ʈ̤ohəni/, meaning ‘you people’ or ‘you all’ /həməlogə/ to /həməni/ meaning ‘We’ etc. This also includes the transitions of some most commonly used Hindi honorifics such as /ve logə/ as /okʰəni/ or else /ohəni səbə/ meaning ‘You’ or ‘You people’ as honorifics for Magahi.

6.5.1.5. Suggestive Structural Closedness

This is concerning the structural closeness and its relation to the respective linguistic entities present in the sentence. As Hindi and Magahi both belong to the same language family, therefore, they both follow the same word order of their structural formations which is SOV. As Magahi and Hindi both belong to the same language family group, therefore, they both are considered as sister languages. However, sometimes certain structural shifting in word orders of Magahi language can be seen while framing sentences. This is done in order to obtain some newly framed sentences with possible changes in their structural patterns. The

same has been seen here in the collected corpora as this is a very literary and stylistic form of corpora. This is so because it was not obtained from any purely existing Magahi source but rather, an online literary blog which mostly has the translated and distorted forms of this language.

This structural closedness was closely observed in order to analyse the data structures at the sentence level. This was so because the data was also examined and evaluated at the sentence level as well. This was to avoid the scarcity of the available corpus in the concerned language. Performing such alterations have most importantly changed the positions of phrases of the sentences, thus showing no effects in changing the meaning of the sentence but, making the task a bit easier to analyse the structures and also to see and prove whether this can allow alterations in the same without change or not. This has helped a lot in training the system as it seems to have been quite effective and resultative in terms of establishing the idea of structural equivalence in the concerned language.

6.5.1.6. Handling Possible Ambiguities

This is more of a knowledge-based solution in order to correct and increase the efficiency of the tool in this regard. We all know that every single word may have several different meanings and these meanings might prove to be ambiguous at times when used within a sentence or else a word or token. Such ambiguities are so trivial that they can only be removed and handled through knowledge-based approaches or solutions. This is so because, in machine learning or any AI-related task, the issues related to ambiguities can only be tackled and eradicated through this only. This is also because it is the human brain only which can contextualize the meaning of each word or token or even a sentence and uses them wisely depending upon its situation and need. Therefore, following this knowledge-based approach the machine could also be made well-trained in the concerned task, by applying this approach. ding this cannot only increase the efficiency of the AI toll rather, the functioning and final result of the same as well. This could help the tool in pattern prediction, which could also be of great help while analysing any structural pattern of the sentence.

NLP system uses this sort of approach to make their systems more effective, evaluative, functional and reliable. In order to make it more efficient, it needs more effective human knowledge which needs to feed in the machine in order to make the machine learn and think like humans. This can also help the machine to contextualize the order of events, thus finally helping in proper functioning, arrangements of sentences and usage of words. This is

most effective when used for sentence predictions. This helps in increasing the optimal knowledge of the machines thus helping them in auto predictions. These auto-predictive solutions can solve the issues of ambiguities quite easily and effectively once the machines become capable of analysing the errors and their patterns while training. This can also solve the issues of creating inadequacies among the sentences and word tokens. Once such issues are handled effectively by implementing this solution the machine will become self-capable of detecting errors and irregularities among the tokens or sentences, identify them and will solve them accordingly by applying automatic solutions for the same.

Chapter- 7 CONCLUSION

7. Conclusion

This is the final chapter of the entire work. This will give a brief conclusion of the task undertaken. It is divided into four main sections and further subsections. The very first section of this chapter will briefly summarize all the research, while others deal with the results shown, which obviously concern the developed model. along with this, the possible defined frontiers of the research and their future implications and the way forward for any future research that has a strong chance of being undertaken in this agreement are discussed.

7.1. Summarizing the Research

Through this work I have tried to summarize the technical and detailed theoretical aspects of CPs in Magahi. The pursued concept was not only treated theoretically, but also discussed with possible technical aspects of COLING using NLP platforms. It also examined the detailed analysis of CPs in general and structural linguistics through a morpho-syntactic perspective. Throughout this research, there have been instances of Magahi CPs resembling several other Indo-Aryan languages, including many South Asian languages such as Hindi, Bangla, Oriya, Bhojpuri, Sanskrit, Maithili, and Marathi. This entire study was also interesting and relevant because, apart from these parallels, Magahi and other Indo-Aryan languages also have some structural differences that have been closely observed and discussed in detail in the appropriate sections of the chapters concerned.

Although Magahi has been one of the ancient languages of India for centuries from both an analytical and theoretical point of view, it has not received much attention for study and analysis by the masses and this is still neglected for research today. For this reason, it is not only resource-poor, but also extinct or threatened with extinction. This work is a detailed analysis of the possible complex predicates and their different types, together with the tool development technique that will be used in the future for CP identification and an analysis that is intended to be used in an automatic mode by the users of this language. It is important to understand the Magahi language, the environment in which it is spoken and the geography before explaining these difficulties, so a separate chapter has been designed and discussed for this. For this reason, a simple and detailed description of all these concepts has been attempted in this thesis.

Through this dissertation we have also found that the word order in Magahi is quite conventional, allowing for subject-object-verb ordering along with the special property of scrambling. This is one of its common traits. In this language, the sequence of nouns, adjectives, main verbs, and light verbs form a whole verb phrase, which together forms the form of a CP. In this dissertation we have witnessed that how beautifully the two different elements which is the first and second elements of a complex predicates which are arranged in an order sometimes changes their positions, without affecting the semantics of the language. CPs are capable of moving the entire verb phrase from initial position to final position or vice-versa or else to any other places without affecting its meaning. This is done with Magahi as well depending on its pragmatic use, however, in case of Magahi the change of meaning can be seen at times. At the same time, we have also witnessed the magical reordering and rearrangements of its places from one to another, when we were discussing reverse compound verbs which is RCVs which was not possible in Magahi, but can be seen in languages like Hindi.

Here we were also able to observe for Magahi that the two different components of a compound verb known as v1 and v2, can be positioned differently. This was seen in case of conjunct verbs wherein nouns and verb sequences may occasionally switch its respective places depending on the pragmatics of the entire sentence. This shows no effect on the meaning of the entire sentence some times. On the contrary, the order of the two parts remains fixed when an adjective and a light verb together makes up a conjunct wherein the two elements cannot switch its respective positions. Therefore, one can say that Magahi sometimes also follows the rigid word order arrangements for its CP constructions or sentence constructions.

Apart from these, there also exists few light verbs in Magahi that plays a significant role in its idiomatic constructions. In idiomatic expression of Magahi, the light verb contributes a bit more towards the meaning the constructed predicate in comparison to that of the major verb. This means that all such constructions bear the concealed meaning of the entire phrase or predicate formed. Some examples of such words or expressions are like /-ɑ/, /jo/, /pɪ/, /uɽə/ etc.

In the next chapter, which is the second chapter of research, any major technical or theoretical work that has been carried out for the idea in question is discussed. At the same time, the next immediately following chapter discussed the possible technical tools and

methods applied and used for the development of the tool and the collection of the desired corpus of the language in question, which was later used for the CP identification and analysis method. This was achieved by applying a machine learning model called SVM. It also discussed in detail the different approaches of NLP and COLING applied not only in Magahi but also in relation to several other Indic languages, some of which were underserved, like Magahi as one while Odia as the other. But with ongoing technological advancements it has also been seen how current NLP research trends have advanced the idea of research and future approaches in these languages as well.

It can also be seen that the architectural design of the tool, developed through a protracted process, has also been presented. Not only that, but it also thoroughly discussed the applied computational model, system modelling, feature extraction techniques along with the detailed analysis of the errors that may have occurred, and suggested some valuable evaluation measures to reduce such problems and challenges.

With the help of various diagrams and figures, research has also tried to outline the practical aspects of the same. Likewise, the SVM architecture for the NLP tool and the result and analysis have been easily represented through various specific bar charts, charts and figures. The different set of tag sets was represented by a modelled table to show how the tagging task was transformed by using BIS tag set. It is the tag set used by all Indic languages for the NLP tasks with minor modifications according to the requirements of the languages concerned. The functionality of the ILCI tool was also discussed to show how the large amount of data is tagged with this technological tool. The entire procedures and functionalities of these tools were discussed in detail with the help of the respective concerned sections of chapter five.

Through the penultimate chapter, i.e., chapter six, the evaluation and analysis part were briefly discussed. This includes discussing possible problems and challenges, suggested actions, failure analysis techniques, etc. This has also suggested several fruitful solutions to further enhance the capabilities of this tool. The sections and subsections of this chapter have not only discussed error analysis and the types of errors we encountered, but also to easily identify and analyse why these errors occurred halfway through.

After carefully analysing the errors, the machine was able to identify errors between the "known" and "unknown" data types. visible vs. invisible along with gold vs. raw data set types. It also encountered issues like ambiguities between words and data types, which

became one of the main reasons why the tool made so many mistakes. The reason for such ambiguous constructions is that most Magahi speakers have switched their spoken language to standard Hindi. This is because both belong to the same language families and therefore shares the most similarities in their linguistic and structural patterns. Another major reason for such ambiguity was the scarcity of data platforms, meaning the availability of concrete Magahi data. This implies that we do not have clear access to the raw and actual Magahi speeches, which bears the original raw structures of the language concerned. As a result of this the collection was done through a popular blogging website which was mentioned in the concerned chapter. This website has the data concerning, an intermixing of both Hindi and Magahi language. Though there existed a lot of Magahi data structures in the same but this also had some limitations within just because of the reason mentioned above. This means that the Magahi structures were not proper and pure rather distorted and have adopted a lot of structures from standard Hindi. Also, it was a data from an online blogging website therefore, it does not support pure constructions of standardized types of Magahi. This means it comprised of the structures, which were influenced mostly from Hindi. These includes structures and datasets mainly from stories, translations of novels into Magahi and several other fields of literature and non-literature backgrounds. It is also due to the lack of knowledge of proper Magahi structures. It is due to this reason why these bloggers who wrote these stories or biographies over the blog failed to translate the given sentences in pure and accurate Magahi which ended up the data as of mix up variety of Hindi and Magahi both.

This research has developed the identification and analysis tool not only by examining individual tokens of the Magahi language, but also at sentence labels as well. For this reason, the nature of ambiguity was also briefly discussed in relation to each and every token decided and designed for this separate category. In addition, one will also find a descriptive error analysis including defined percentages for each sentence label as well as various tokens at the end which involved a deep and thorough analysis of the tool.

In the last chapter of the work, the researcher provides an overall summary with all the necessary details of the work comprising all necessary details and procedure followed and applied. It is a summary of the entire work done. It discusses the necessary developments and related findings. Also, the limitations and future impacts are further discussed with the possible limiting elements. This will help summarize the work and explore the futuristic approaches for the future researchers in this agreement not only regarding Magahi but for all other low-resource language. This is done by providing them with an important platform for

NLP and COLING, which helps in further advancing the development of all such languages in India.

7.2. Results of the research

This section, as the name suggests, deals with the result of the research carried out. This includes several different sub-sections that completes this entire research through a little more detailed explanation of ideas like designing a tagset for Magahi's CP tag along with its little brief synopsis, method of annotating the desired corpora, the developed mechanism for the Magahi CP tag identifier and the extractor tool which is specially designed for Magahi, as the core and the sole idea of this work.

7.2.1. Designing Tagsets

In terms of tagset development, the idea has been shared by the BIS tagset module, which is the one that all Indian languages utilises for corpus tagging and labelling tasks in NLP. The tagset used here has been adopted from (Renu; et. al. 2015), and has undergone some changes and modifications in order to use for Magahi CP identification and classification. This is due to the reason because, Magahi and Hindi shares same common boundaries and also the scripts which is Devanagari.

Corpus tagging by implementing these tagsets has not only reduced the human effort but also greatly simplified the task of testing and evaluation of the tool to achieve a tremendous percentage of accuracy. Therefore, the task of assigning the desired tags to the raw data according to the required linguistic categories helped us a lot in classifying and refining the corpus much more easily. This has not only reduced the complexity of the tool but has also resulted in providing a much more acceptable model for the language undertaken which is Magahi.

7.2.2. Brief Summary of Designed and Assigned Tagsets

For COLING, the task of creation of tagsets for any low-resource languages like Magahi is extremely rare and not easy to find. The creation of such tagsets not only beautifies the language tokens, but also helps to assign a unique linguistic characteristic to each word or tokens of the language. Here in this work, we have tagged all of Magahi's CPs, specifically the verbs with the appropriately styled tagsets. This not only refined the corpus for the next level which is testing, but also helped us to create a COLING digital database for languages which are not exceptionally rich in technological advances and implementations. This CP

labelling of the collected data has not only bleached it computationally but has also made it a bit more advanced linguistically, so that it can be easily implemented by the machine learning tools to get the desired output. Doing this can make more refined data. It simply works on the principle of having more accurate data gives more specified accuracy of the designed tool. Hence, one can understand how well these two tasks are related and directly proportional, which are very important for successful execution of the NLP, COLING or ML related tasks.

7.2.3. Task of Annotating Corpora

Annotation of corpora comprises basically the task of adding interpretive information to each dataset that makes them recognizable by the system tool, Leech (2005). In order to commute this task total of 45k word tokens have been successfully annotated as per the CP tagging guidelines[37]. Out of all these around 10k tokens were taken for testing and training in first round, 20k tokens in second while a total of 15k tokens have been taken for the final and the third round. For each round of testing a total of 2k gold tokens were taken for the training purpose which is required for commuting the task. Here One may wonder as to why we divided this task into three different phases? This was a necessary procedure and very much required for the task, as the single test set could overlap some functions, which can affect the accuracy of the tool. Therefore, before each round of testing and training, the data introduced into the tool have been thoroughly checked and examined, and then repeatedly induced after making some changes or modifications required if any, to verify the accuracy of the tool, and also to see whether it is overlapping some feature or remain static.

The task of annotation was also a bit challenging as sometimes the data set was not so refined and accurate to determine the exact category of the same, which means that the data set collected and tested was very ambiguous, which of course hampered the task a bit. This is also because it comes from an online blogging website that fully translates data from Hindi to Magahi. Therefore, the ambiguity in their structures and word types was remarkably high, thus finally affecting the annotation procedures.

7.2.4. Designing Magahi CP Identifier and Extractor

This section can be seen as the end result of the entire work. It is important to note here that the development of CP identifier and extractor tool has undergone very deep learning and rigorous method which consists of several basic steps like crawling, sanitizing,

[37] Refer tagset chart in section 4.3.2. of chapter 4

tagging, evaluating and analysing etc. From all these steps, the assessment required a bit more attention as this will help to obtain the final result with maximum accuracy and thus contribute to the development of a successful assessment tool. Missing the evaluation part could lead to system inaccuracies and thus miss the model to be built. Due to the accurate monitoring technique performed throughout the research task, the tool has reported an overall accuracy of 64.57% for both the classification and identification modules. However, the number was somewhat low as it was the very first development of its kind in relation to CP identification and analysis for Magahi language. This can be increased in future by applying the suggestive measures and rules given in this research.

This played a major role out of all the steps followed in this regard. This so because it required a lot more awareness not only in terms of the data collection and tagging but also in terms of its technicalities in order to handle and execute the final result properly. This has also helped us in analysing the issues and challenges concerned with the tool along with corpus handling and collection. This implies that one needs to follow all the necessary methods and approaches required and suggested in this work for the successful execution of the undertaken task.

7.3. Limitations of Current Research and Developed Statistical based Tagger Tool

As stated earlier, it is the research concerning the CP identification and analysis of Magahi language and the development of a possible identifier for the same. This task has been executed by applying SVM techniques using unigram features. For this a large set of Magahi data were taken. Out of this huge dataset of almost more than 5lacs words or tokens from different domains, around 50k word tokens were then tagged as different CPs with its uniquely designed tagsets. These were then set for training the system in three different phases, containing almost 16k raw tokens along with around 2k gold tokens in all the phases in order to check if the tool overlaps during any of the evaluation steps. Once all the steps were finished combining all three stages of training and testing, the maximum accuracy percentage among all three were recorded and considered the best as the tool output.

After getting over with all these steps, now comes the possible limitations. So far as researches concerning NLP, COLING and ML is concerned, Magahi as a language is not exposed much, especially in the field of NLPs. Concerning these facts, we have also decided certain limitations of the research undertaken for the language, so that it may not create any issues for the concerned field. Till date there only exists few automatic taggers in this

language. These are POS taggers for Magahi by Kumar et al. (2012)) and Classification and Identification tagger for MWEs in Magahi, which also is a SVM based module by Kumar and Behera (2017).

This research work was in continuation with the previous researches in order to just check the idea as to how efficient these technologies are for not only Magahi but also for such Indian languages which are not very exposed on to the platforms of NLP. This is also to ensure as to how effective the statistical based SVM module is for such NLP related tasks. In connection with this it is also to investigate whether this given SVM module be more effective for such tasks or else it needs some more advancements in the same. Concerning all these major things, the entire process of this research involved several steps involving from data collection to validation and testing for final result was very challenging as this was done on a two-fold process of data collection method that involve manual and machine learning.

In addition to this, every issue encountered has not only been noted carefully but also has been dealt properly but there still remains some challenges as in the availability of the data in raw form and not an intermixed one. However, each issue and challenges have been tackled wisely, but even after this the machine needs some more refined corpus in order to perform better. This is because, for ML related approaches only data sufficiency could result into maximum accuracies. Here in this case, it was an intermixed data of Magahi along with that of Hindi. Therefore, it is a dire need of creating more and more data sources either online or offline for all such languages that are low-resourced.

The feature selection technique while designing the model through the proposed SVM mechanism was another possible limitation. During the annotation process, attribute like medium verbose (-V2) and sequencing system left-right-left (LRL) mode were employed while the rest of the remaining functions for the tagger have been put to its default settings. Special CP features like RCVs and ECVs were not very much found in Magahi however, getting few in numbers cannot be overlooked. This was possible only because it is similar and identical to Hindi. It is due to this reason that the issue of ambiguity was also very prevalent, hence affecting a bit to the development and accuracy of the system. Therefore, in order to enhance the NLP aids in languages like Magahi, one should must continue their future researches more into this language, which will not only enhance the data collection system but will also help in getting some more tools like Named entity recognition system, Sentence identifier, Sentiment analysis techniques, Morph-analysers etc.

7.4. Implications of this Research and the Way Ahead for Future

As mentioned already, there exists only a few researches concerning Magahi in the field of AI, COLING and NLP, therefore this idea of tool development for CP identification and analysis through automotive technologies will prove out to be quite advantageous. This research has shown us that the constructions like RCVs and ECVs are quite difficult and are rarely found in Magahi. NLP being one of the most recent research trends in the field of data technologies will thus help in further advancements of languages like Magahi through its assistive technological system. It reduces the human efforts by providing a major digitalize platforms for such languages to be safeguarded and protective for upcoming generations. These technological aids are helping in digitally documenting these languages along with their structures which can also be seen as one of major approach to safeguard them from being extinct.

This research will help the future researchers to develop new technological NLP based tools such as text summarizer, parser, chunker etc. This all could be made possible only when these low-resourced languages are exposed towards such supervised learning modules and techniques. The desired exposures towards NLP aids for Magahi will thus result into further advancements which will lead to free access of the electronic data for a long period of time if created digitally. This will also help in safeguarding these ancient languages from getting extinct. The ongoing technical advancements will also eradicate the availability issue of the Magahi data and getting it intermixed with Hindi, which was one of the sole reasons of having ambiguities of Magahi with Hindi and other similar languages in terms of its structures, words, sentences etc.

As this is the first NLP related research of its kind, hence this can also be enhanced in near future by upcoming researchers in order to improve and enhance the performance of the tagger tool for the concerned language. Therefore, deducing from all the above facts, one could conclude that creating such digitalized models for an endangered language like Magahi could help a lot in documenting it and also helps in creating a safety armour for them. These aids could easily help in creating the digital libraries of documents for all such languages wherein the data can be made easily approachable to work and develop such advanced technology-based systems for these languages.

Also, these can be made better if one uses more effective feature selection in the module selected. Further the issue of ambiguity could also get improved along with the

model accuracy by taking into account the proper linguistic characteristics of Magahi as whole and not any similar languages like Hindi. In order to increase the accuracy rate of the taggers, one can also additionally use lexical databases of this language which could only be made possible through NLP aids. Tool development task can further be enhanced by designing case-based system tool wherein both seen and unseen data can easily be supervised and hence trained for results. Further, in order to increase its accuracy one can re-work on it and apply some more concrete and effective data in order to check the viability of the same. This will be a two- way development method wherein data development would be decided as goal one while the advancements of tool with maximal accurate percentage would be the second aim.

Apart from this, there also exists some vital future approaches which will help for the further advancements of this language. Some of the major futuristic approaches are listed here as under: -

(i) Based on this current research one can develop the idea of hence improving the same, a step ahead in future for the development of assistive machine learning tools for text summarization in Magahi like many other languages such as Hindi, Bhojpuri, Maithili etc.

(ii) This could help in constructing advanced lexicons for Magahi and find some better ways in order to make such languages digitally rich.

(iii) This could also help in developing more supervised learning models if the issue of scarcity of data for Magahi is solved and is easily made available as per future needs and requirements. This could also tackle the issue of ambiguity quite well hence, reducing its possibilities in data.

(iv) To import properly designed grammatical rule governing tool for Magahi which could assist in solving the issue of ambiguity hence, reducing its possibilities in data.

(v) Based on this, one could also easily develop the verb analysis system tool concerning Magahi language, which will help the masses in exploring further and future research fields concerning not only general linguistics but also the technical ones. This CP identification tool could also help in developing a verb analyser for the mentioned language.

(vi) This can help in the development for making an online e-dictionary for Magahi in near future.

(vii) The issue of ambiguity can also be handled and tackled wisely as these researches would lead to the availability of relatively better and huge amount of supervised data in order to commute the task.

(viii) Once the tool designed becomes able to identify and classify the CPs in Magahi, this can help in making possible huge digital library of the language concerned which is Magahi.

Last but not the least, CPs in Magahi is a detailed, theoretical, empirical as well as practically supported effort that aids to deeper comprehension of this language. It is an encouraging piece of work for languages like Magahi, that hasn't been studied in detail so far and that too for such technologically enriched researches. It also advances the typological understanding of all the possible nearby Indian languages that are on the verge of extinction.

Numerous applied areas of linguistics, including computational linguistics and language training by applying proper NLP modules could lead to such language advancements. This study of "Automatic Identification and Analysis of Complex Predicates in Magahi" can be expanded in a variety of ways to many other Indian languages as well which are low-resourced and less exposed to technologies and further higher approaches of study. With this other major works for further researches could be commuted such as Bhojpuri and Maithili can be compared by applying a comparative study along with its possible similarities and differences between the two. A morphological analyser system or other possible machine translation tool if any can be used to apply and analyse the facts of the difficult predicative constructions in Magahi.

However, this dissertation would not end here as this has just marked the beginning for the development of NLP based applications and its approaches in the language concerned which has been overlooked since ages, despite being one of the oldest form of speeches for people since ancient times.

BIBLIOGRAPHY

- 52nd Report og the commissioner for Linguistic Minorities in India. (2014-2015). https://www.minorityaffairs.gov.in/sites/default/files/2.%2052nd%20Report%20English.pdf
- Abbi, A. (1980). *Semantic Grammar of Hindi: A Study in Reduplication*. New Delhi. India. Bahri Publications.
- Abbi, A. (1990). Reduplication in Tibeto Burman Languages of South-Asia. *Japanese Journal of Southeast Asian Studies*, 28(2), 171-181.
- Abbi, A., & Devi Gopalakrishnan. (1991). Semantics of explicator compound verbs in South Asian languages. *Language Sciences*, 13(2), 161-180.
- Abbi, A., Raghubir Saran Gupta, & Ayesha Kidwai. (Eds.). (2001). *Linguistic structure and language dynamics in South Asia: papers from the proceedings of SALA XVIII Roundtable* (Vol. 15). Motilal Banarsidass Publishing House. New Delhi. India.
- Abney, S. P. (1987). The English noun phrase in its sentential aspect (Doctoral dissertation, Massachusetts Institute of Technology).
- Abujar, S., Mahmudul Hasan, M. S. I. Shahin, and Syed Akhter Hossain. (2017, July). A heuristic approach of text summarization for Bengali documentation. In 2017 8th *International Conference on Computing, Communication and Networking Technologies* (ICCCNT) (pp. 1-8). IEEE.
- Ali, H. (2010). An unsupervised parts-of-speech tagger for the Bangla language. *Department of Computer Science, University of British Columbia*.
- Alok, D. (2016). The syntax of split: The case of Hindi and Magahi. *Formal Approaches to South Asian Languages.*
- Alok, D. (2021). The morphosyntax of Magahi addressee agreement. *Syntax, 24(3),* 263-296.
- Annamalai, E. (1979). Aspects of aspect in Tamil. *International Journal of Dravidian Linguistics (IIDL) Kerala*, *8*(2), 260-267.
- Antony, P. J., & K. P. Soman. (2012). Computational morphology and natural language parsing for Indian languages: a literature survey. *International Journal of Computer Science and Engineering Technology (IJCSET), 3*, 136-146.

- Antony, P. J., S. P. Mohan, & K. P. Soman. (2010, March). SVM Based Part of Speech Tagger for Malayalam. In *Recent Trends in Information, Telecommunication and Computing (ITC), 2010 International Conference on* (pp. 339-341). IEEE.
- Aryani, S. (1965). Magahi Lok Sahitya. *Delhi. Hindi Sahitya Sansar*.
- Asa, A. S., Sumya Akter, Md Palash Uddin, Md Delowar Hossain, Shikhor Kumer Roy, and Masud Ibn Afjal. (2017). A comprehensive survey on extractive text summarization techniques. *Am. J. Eng. Res*, 6(1), 226-239.
- Atreya, L. (2015). Aspects of Magahi Syntax. Doctoral Dissertation. Indian Institute of Technology, Patna.
- Atreya, L., & Rajesh Kumar. (2013). Anaphors in Magahi: A Binding Theoretic Treatment. *International Journal of Linguistics, 5* (4), 109.
- Atreya, L., & Sweta Sinha. (2020). Phonological and functional analysis of diminutive marker-waa in Magahi. *Dialectologia: revista electrònica*, 43-57.
- Baker, M. (1985). The mirror principle and morphosyntactic explanation. *Linguistic inquiry*, *16*(3), 373-415.
- Baker, M. (1988). Theta theory and the syntax of applicatives in Chichewa. *Natural Language & Linguistic Theory*, *6*(3), 353-389.
- Baldwin, J., & Zhengxi Lin. (2002). Impediments to advanced technology adoption for Canadian manufacturers. *Research policy*, *31*(1), 1-18.
- Baldwin, T., & Aline Villavicencio. (2002). Extracting the unextractable: A case study on verb-particles. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.
- Baldwin, T., Colin Bannard, Takaaki Tanaka, and Dominic Widdows. (2003, July). An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment* (pp. 89-96).
- Balusu, R. (2012, December). Complex Predicates in Telugu: A computational perspective. In *Proceedings of COLING 2012: Demonstration Papers* (pp. 1-8).
- Banerjee, S. (1994). *Complex Predicates in Bangla*. M.Phil. Dissertation, Centre for Linguistics and English Studies, J.N.U., New Delhi.
- Bashir, E. (1993). Causal chains and compound verbs. *MK Verma ed.(1993) Complex Predicates in South Asian Languages.*

- Baskaran, S., Kalika Bali, Monojit Choudhury, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Girish Nath Jha, S. Rajendran, K. Saravanan, & L. Sobha (2008). A common parts-of-speech tagset framework for indian languages. In *In Proc. of LREC 2008*.
- Behera, P. (2015). Odia parts of speech tagging corpora: suitability of statistical models. Diss. M. Phil. Dissertation, Jawaharlal Nehru University (JNU), New Delhi, India.
- Behera, P. (2015). Odia Parts of Speech Tagging Corpora: Suitability of Statistical models. M.Phil. Dissertation, Centre for Linguistics, J.N.U., New Delhi.
- Behera, P., & Sharmin Muzaffar. (2018). Developing classification-based named entity recognizers (NER) for Sambalpuri and Odia applying support vector machines (SVM). *Nepalese Linguistics*, 1-7.
- Behera, P., Neha Mourya, & Vandana Pandey. (2016, December). Dealing with linguistic divergences in English-Bhojpuri machine translation. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)* (pp. 103-113).
- Belletti, A., & Luigi Rizzi. (1988). Psych-verbs and θ-theory. *Natural Language & Linguistic Theory*, 291-352.
- Bhatt, R. M. (2008). In other words: Language mixing, identity representations, and third space 1. *Journal of sociolinguistics*, *12*(2), 177-200.
- Bhattacharya, T. (1999). The Structure of the Bangla DP. Doctoral Dissertation, University College, London.
- Birdsong, D., Denis Bouchard, & Kathy Leffel. (1988). On the question of negative evidence in second language acquisition. *Florida occasional contributions to the advancement of linguistics*, *1*(1), 19-41.
- BLLIP 1987-89 WSJ Corpus Release 1. https://catalog.ldc.upenn.edu/LDC2000T43/
- Bouchard, D., & Kathy Leffel. (1988). *Florida Occasional Contributions to the Advancement of Linguistics.* Program in Linguistics, University of Florida.
- Bowern, C. (2006, January). Inter-theoretical Approaches to Complex Verb Constructions: Position Paper; *Department of Linguistics, Rice University*.

- Brants, T. (2000). TnT: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on applied natural language processing* (pp. 224-231). Association for Computational Linguistics.
- Bright, W. (1996). The Devanagari script. In *Peter Daniels and William Bright, editors, The World's Writing Systems. Oxford University Press, New York*, NY, pages 384–390.
- Brill, E. (1992, February). A simple rule-based part of speech tagger. In *Proceedings of the workshop on Speech and Natural Language* (pp. 112-116). Association for Computational Linguistics.
- Brill, E. (1992). *A simple rule-based part of speech tagger*. PENNSYLVANIA UNIV PHILADELPHIA DEPT OF COMPUTER AND INFORMATION SCIENCE.
- British National Corpus. http://www.natcorp.ox.ac.uk/
- Bukhari, N. (2009) Comparative Study of Double Verb Construction in Gojri. *Language in India; 1* (9), 26-51.
- Butt, M. (1995). *The structure of complex predicates in Urdu*. Center for the Study of Language (CSLI).
- Butt, M. J. (1993). The structure of Complex Predicates in Urdu. Ph.D. Dissertation. Department of Cognitive and Linguistic Sciences, Providence, Brown University. department of linguistics and the committee on graduate studies. Stanford University.
- Cattell, R. B., G. Schröder, & A. Wagner. (1969). Verification of the structure of the 16 PF questionnaire in German. *Psychologische Forschung, 32*(4), 369-386.
- Chakrabarti, D., & Pune CDAC. (2011). Layered parts of speech tagging for Bangla. *Language in India, Special Volume: Problems of Parsing in Indian Languages*. *www.languageinindia.com*
- Chakrabarti, D., Hemang Mandalia, Ritwik Priya, Vaijayanthi M. Sarma, & Pushpak Bhattacharyya. (2008, August). Hindi compound verbs and their automatic extraction. In *Coling 2008: Companion volume: Posters* (pp. 27-30).
- Chakrabarti, S., Martin Van den Berg, & Byron Dom. (1999). Focused crawling: a new approach to topic-specific Web resource discovery. *Computer networks, 31*(11-16), 1623-1640.
- Chakraborty, C. (2021). Representing Structural Nuances of the Code-mixed/switched Data: A Case Study of English-Bangla. *Language in India, 21*(9).

- Chandra, N., Sudhakar Kumawat. and Vinayak Srivastava. (2014, March). Various tagsets for Indian languages and their performance in part of speech tagging. In *Proceedings of 5th IRF International Conference*, Chennai.
- Chandrashekar, R. (2007). Part-of-Speech Tagging for Sanskrit**.** Ph.D. Thesis, Special Centre for Sanskrit Studies, J.N.U., New Delhi.
- Chatterji, S. K. (1926). The origin and development of the Bengali language (Vol. 2). Calcutta University Press.
- Chatterji, S., Tanaya Mukherjee Sarkar, Pragati Dhang, Samhita Deb, Sudeshna Sarkar, Jayshree Chakraborty, & Anupam Basu. (2014). A dependency annotation scheme for Bangla treebank. *Language resources and evaluation*, 48(3), 443-477.
- Chaudhary, S., Zakir Laliwala, & Vikram Sorathia. (2006). Building Semantic Business Services. In *Semantic Web Services, Processes and Applications* (pp. 323-350). Springer, Boston, MA.
- Choudhary, N. (2006). *Developing a Computational Framework for the Verb Morphology of Great Andamanese.* M.Phil. Dissertation, Centre for Linguistics, J.N.U., New Delhi.
- Choudhary, N., & Girish Nath Jha. (2014). Creating Multilingual Parallel Corpora in Indian Languages. In *Human Language Technology Challenges for Computer Science and Linguistics* (pp. 527-537). Springer International Publishing.
- Church, K., William Gale, Patrick Hanks, & Donald Hindle. (1989). Word associations and typical predicate-argument relations. In *Proceedings of the International Workshop on Parsing Technologies*.
- Comrie, B. (2001). *Typology and the history of language* (Vol. 1, p. 21). Oldenbourg Verlag.
- Corbett, G. (2000). *Number*. UK: Cambridge University Press.
- Corpus of Contemporary American English. http://corpus.byu.edu/coca/
- Cortes C, Vladimir Vapnik. (1995). Support-vector networks. *Machine learning. Sep 1; 20* (3):273-97
- Dandapat, S., Sudeshna Sarkar., & Anupam Basu. (2004, December). A Hybrid Model for Part-of-Speech Tagging and its Application to Bengali. In *International conference on computational intelligence* (pp. 169-172).

- Das, B. R., & Srikanta Patnaik. (2014, January). A Novel Approach for Odia Part of Speech Tagging Using Artificial Neural Network. In *Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2013* (pp. 147-154). Springer International Publishing.
- Das, B. R., Smrutirekha Sahoo, Chandra Sekhar Panda, & Srikanta Patnaik. (2015). Part of speech tagging in odia using support vector machine. *Procedia Computer Science, 48*, 507-512.
- Das, P. K. (2009). The form and function of Conjunct verb construction in Hindi. *Journal of South Asian Studies, HUFS, South Korea,* 15(1), 191-208.
- Das, P. K. (2012). The Computation of Basic Verbal-Strings in Hindi^. *Interdisciplinary Journal of Linguistics, Kashmir University, 5*, l3l-148.
- Das, P. K. (2013). Ergativity in Khorṭhɑ: An Enigmatic Phenomenon. *Journal of South Asian Studies, HUFS, South Korea, 18*(3), 321-350.
- Das, P. K. (2015). The linguistic prerequisites and grammaticalization of ‘compound verb’in Hindi. *Journal of South Asian Studies, HUFS, South Korea, 21*(2), 51-76.
- Das, P. K. (2016). Reversed Compound Verbs in Hindi: An Inquiry of its Form and Function. *Journal of South Asian Studies. 22*. 153-192. 10.21587/jsas.2016.22.1.006.
- Das, P. K. (2017). Process of Causativization and De-Causativization in Hindi. *Interdisciplinary Journal of Linguistics, Kashmir University, 10*, 40-54.
- Dasgupta, P. (1989). *Projective Syntax: Theory and Application.* Deccan College P.G. and Research Institute: Pune, India.
- Davies, W. D. (1988). The case against functional Control. *Lingua, 76*(1), 1-20.
- Davison, A. (1991, August). From Parlog to Polka in two easy steps. In *International Symposium on Programming Language Implementation and Logic Programming* (pp. 171-182). Berlin, Springer.
- Deoskar, T. (2006). Marathi Light Verbs. *Proceedings of the 36th Annual Meeting of the Chicago Linguistics Society* Vol. 42 (2), pp183-198. Chicago Linguistics Society, USA.
- DeRose, S. J. (1990). Stochastic Methods for Resolution of Grammatical Category Ambiguity in Inflected and Unified Languages. Ph.D. Dissertation. Department of Cognitive and Linguistic Sciences, Providence, Brown University.
- Edgar, G. J., & Neville S. Barrett. (1997). Short term monitoring of biotic change in Tasmanian marine reserves. *Journal of Experimental Marine Biology and Ecology*, *213*(2), 261-279.

- Efat, M. I. A., Mohammad Ibrahim, & Humayun Kayesh. (2013, May). Automated Bangla text summarization by sentence scoring and ranking. In *International Conference on Informatics, Electronics and Vision (ICIEV)* (pp. 1-5). IEEE.
- Gaikwad, D. K. (2018). Rule Based Text Summarization for Marathi Text. *Journal of Global Research in Computer Science, 9*(5), 19-21.
- Gambhir, V. (1993). Complex Verb Phrase: A Diachronic and Synchronic View. In M. K Verma (ed.), *Complex Predicates in South Asian Languages*. New Delhi: Manohar Publishers and Distributors.
- Garg, N., Shivek Kumar, & Sharma Vaibhav. (2014). Concepts of Difference and Difference- A Comparative Study of Saussure and Derrida. *Language in India*. (pp. 141-147). ISSN 1930-2940. Vol. 14:11.
- Garg, N., Shivek Kumar, & Vaibhav Sharma. (2014). Concepts of difference and differance: A comparative study of Saussure and Derrida. *Language in India.* http://www.languageinindia.com/nov2014/negadifferenceanddifferance1.pdf
- Ghosh, S. (2008). A Generative Lexicon Account of A-V Complex Predicates of Bangla. *Proceeding of ICON*. India: Macmillan Publishers.
- Gill, M. S., Gurpreet Singh Lehal., & Shiv Sharma Josh. (2009). Part of speech tagging for grammar checking of Punjabi. *The Linguistic Journal*, *4*(1), 6-21.
- Gimenez J., & Lluis Marquez. (2006). SVMTool Technical Manual v1. 3.
- Gopal, M., & Girish Nath Jha. (2011, March). Tagging Sanskrit corpus using bis pos tagset. In *International Conference on Information Systems for Indian Languages* (pp. 191-194). Berlin: Springer, Heidelberg.
- Grierson, G. A. (1926). Peasant and Pundit in India. *Nature*. https://doi.org/10.1038/118472a0
- Grierson, G. A. (1927). Linguistic survey of India (Vol. 1, Part 1: Introductory). Calcutta: Government of India.
- Grierson, G. A. (Ed.). (1968). Linguistic survey of India: Specimens of languages of the Eranian family (Vol. 10). Motilal Banarsidass.
- Gupta, V. (2013). Hybrid algorithm for multilingual summarization of Hindi and Punjabi documents. In *Mining Intelligence and Knowledge Exploration* (pp. 717-727). Springer, Cham.
- Gupta, V. (2014). Automatic stemming of words for Punjabi language. In *Advances in signal processing and intelligent recognition systems* (pp. 73-84). Springer, Cham.

- Gupta, V., & Gurpreet Singh Lehal. (2012, December). Automatic Punjabi text extractive summarization system. In *Proceedings of COLING 2012: Demonstration Papers* (pp. 191-198).
- Hardie, A. (2003). Developing a tagset for automated part-of-speech tagging in Urdu. In *Corpus Linguistics 2003*.
- Hardie, A. (2003). The Computational Analysis of Morpho-syntactic Categories in Urdu. Ph.D. thesis submitted to the Dept. of Linguistics and Modern English, Lancaster University, (revised soft copy 2004), p. 40.
- Hardie, A. (2005). Automated part-of-speech analysis of Urdu: conceptual and technical issues (pp49-72).
- Hautli, A., & Sebastian Sulger. (2011). Extracting and classifying Urdu multiword expressions. In *Association for Computational Linguistics* (pp. 24-29).
- He, X., Zemel, Richard S. Zemel, & Miguel A. Carreira-Perpinán. (2004). Multiscale conditional random fields for image labelling. In *Computer vision and pattern recognition, 2004. CVPR, 2004. Proceedings of the 2004 IEEE computer society conference on* (Vol. 2, pp. II-695). IEEE.
- Hellwig, O. (2009). Sanskrit tagger: A stochastic lexical and POS tagger for Sanskrit. In *Sanskrit Computational Linguistics* (pp. 266-277). Springer Berlin Heidelberg.
- Higgins, F. R. (1974). The pseudo-cleft Construction in English, Doctoral dissertation MIT. Cambridge, Massachusetts.
- Hook, P. E. (1974). *The Compound Verb in Hindi*. Ann Arbor, Michigan: Center for South and Southeast Asian Studies.
- Hook, P. E. (1991). The Emergence of Perfective Aspect in Indo-Aryan Languages. In Elizabeth C. Traugott and B. Heine (eds.) *Approaches to Grammaticalization,* Vol. II. John Benjamins: Amsterdam.
- Hook, P. E. (1993). Aspectogenesis and the Compound Verb in Indo-Aryan. In M. K Verma (ed.), *Complex Predicates in South Asian Languages*. Manohar Publishers and Distributors: New Delhi.
- Humayoun, M. (2006). Urdu Morphology, Orthography and Lexicon Extraction. *rXiv preprint arXiv:2204.03071*.
- Indian Language Transliterator. http://sanskrit.jnu.ac.in/ile/index.jsp

- Indian Languages Corpora Initiative (phase 2). http://sanskrit.jnu.ac.in/projects/ilci.jsp?proj=ilci
- Indradev, K. (2007). *Magahi Ki Sanyukt Kriyaon Ka Bhasha Vaignanik Adhyayan*. Patna: Janki Prakashan.
- Jackendoff, R. (1974). A deep structure projection rule. *Linguistic Inquiry*, 5(4), 481-505.
- Jackendoff, R. S. (1972). *Semantic interpretation in generative grammar*. Cambridge: The MIT Press.
- Jayaseelan, K. A. (1988). Complex predicates and θ-theory. In *Thematic Relations* (pp. 91-111). Brill.
- Jayaseelan, K. A. (1988). Emphatic reflexive x-self. *ms. Central Institute of English and Foreign Languages, Hyderabad, Inde.*
- Jayaseelan, K. A. (2004). The possessor-experiencer dative in Malayalam. *Typological Studies in Language*, 60, 227-244.
- Jena, I., Sriram Chaudhury, Himani Chaudhry, & Dipti M. Sharma, D. M. (2011). Developing Oriya Morphological Analyzer Using Lt-toolbox. In *International Conference on Information Systems for Indian Languages* (pp. 124-129). Springer Berlin Heidelberg.
- Jespersen, Otto. (1924). *The Philosophy of Grammar*. London: George Allen & Unwin LTD.
- Jha, G. N. (2010). Sanskrit Computational Linguistics (Vol. 6465). *4th International Symposium, New Delhi, India, December 10-12, Proceedings.*
- Jha, G. N. (2010). The TDIL program and the Indian Language Corpora Initiative (ILCI). In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010). European Language Resources Association (ELRA).*
- Jha, G. N., Lars Hellan, Dorothee Beermann, Srishti Singh, Pitambar Behera, & Esha Banerjee. (2014). Indian languages on the TypeCraft platform–the case of Hindi and Odia. *WILDRE-2, LREC.*
- Jha, G. N., Madhav Gopal, & Diwakar Mishra. (2009, November). Annotating Sanskrit corpus: adapting IL-POSTS. In *Language and Technology Conference* (pp. 371-379). Springer, Berlin, Heidelberg.
- Jha, G. N., Muktanand Agrawal, Sudhir K. Mishra, Diwakar Mani, Diwakar Mishra, Manji Bhadra, & Surjit K. Singh. (2009). Inflectional morphology analyzer for Sanskrit.

In *International Sanskrit Computational Linguistics Symposium, International Sanskrit Computational Linguistics Symposium* (pp. 219-238). Springer, Berlin, Heidelberg.

- Jha, G. N., Sudhir Kumar Mishra, & R. Chandrashekar. (2005). Developing a sanskrit analysis system for machine translation. In *Proc. National Seminar on Translation Today: state and issues, Dept. of Linguistics, University of Kerala, Trivandrum* (pp. 23-25).
- Joachims, T. (1999, June). Transductive inference for text classification using support vector machines. In *Icml* (Vol. 99, pp. 200-209).
- Katz, G., & Eugenie Giesbrecht. (2006, July). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties* (pp. 12-19).
- Kaur, J., & Jatinderkumar Saini. (2015). A Natural Language Processing Approach for Identification of Stop Words in Punjabi Language. 10.5958/2249-3220.2015.00015.4
- Kidwai, A. (2022). Unlabeled Structures and Scrambling Asymmetries: Hindi-Urdu style. *Formal Approaches to South Asian Languages, 1*(1).
- Kingsbury, P. R., & Martha Palmer. (2002). From treebank to propbank. In *LREC* (pp. 1989-1993).
- Knublauch, H., Ray W. Fergerson, Natalya F. Noy, & Mark A. Musen. (2004). The Protégé OWL plugin: An open development environment for semantic web applications. In *International semantic web conference* (pp. 229-243). Springer, Berlin, Heidelberg.
- Kulkarni, A., Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, Jayashree Jagdale, & Raviraj Joshi. (2022). Experimental evaluation of deep learning models for marathi text classification. In *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications* (pp. 605-613). Springer, Singapore.
- Kumar, A., Rajesh Kumar Mundotiya, & Anil Kumar Singh. (2020, December). Unsupervised approach for zero-shot experiments: Bhojpuri–Hindi and Magahi–Hindi@ LoResMT 2020. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages* (pp. 43-46).
- Kumar, C. (2016). Particle ‘wa’and its various linguistics and socio-linguistics implication in Magahi. *Interdisciplinary Journal of Linguistics, 9*, 150-162.

- Kumar, C. (2018). Plurality in Magahi Language and Reference to count/mass Noun. *Language in India*, *18*(3).
- Kumar, K. V., Divakar Yadav, & Arun Sharma. (2015). Graph based technique for Hindi text summarization. In *Information systems design and intelligent applications* (pp. 301-310). Springer, New Delhi.
- Kumar, R. (2006). *Negation and Licensing of Negative Polarity Items in Hindi Syntax*. New York: Routledge.
- Kumar, R. (2015). *Politeness in Hindi Online texts: Pragmatic and Computational Aspects*. Centre for Linguistics, J.N.U., New Delhi.
- Kumar, R., Bornini Lahiri, & Deepak Alok. (2012). Developing a POS tagger for Magahi: a comparative study. In *Proceedings of the 10th Workshop on Asian Language Resources* (pp. 105-114).
- Kumar, R., Bornini Lahiri, & Deepak Alok. (2021). Descriptive Study of Eastern Hindi: A mixed language. https://doi.org/10.31235/osf.io/aw49z
- Kumar, R., Shiv Kaushik, Pinkey Nainwani, Esha Banerjee, Sumedh Hadke, & Girish Nath Jha. (2012, March). Using the ILCI annotation tool for POS annotation: a case of Hindi. In 13th *International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2012*).
- Kumar, S., Pitambar Behera, & Girish Nath Jha. (2017). A classification-based approach to the identification of Multiword Expressions (MWEs) in Magahi Applying SVM. *Procedia computer science*, 112, 594-603.
- Kumar, V., & Divakar Yadav. (2015). An improvised extractive approach to hindi text summarization. *Information Systems Design and Intelligent Applications*. Springer, New Delhi, 291–300.
- Kunchukuttan, A., & Om Prakash Damani. (2008, December). A system for compound noun multiword expression extraction for hindi. In *6th International. Conference on Natural Language Processing* (pp. 20-29).
- Leacock, C., Martin Chodorow, & George A. Miller. (1998). Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, *24*(1), 147-165.
- Leech, G., & A. Wilson. (1996). EAGLES: Recommendations for the Morphosyntactic Annotation of Corpora (EAGLES Document EAG–TCWG– MAC/R). *Pisa, Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale*.

- Linguistic Data Consortium for Indian Languages. http://www.ldcil.org/resourcesTextCorp.aspx
- Lohar, G. T. (2020). A Grammar of Bhojpuri. PhD Thesis. Tribhuvan University, Kathmandu, Nepal.
- Mahajan, A. (1992). The specificity condition and the CED. *Linguistic Inquiry*, 510-516.
- Mahajan, A. K. (1990). The A/A-bar distinction and movement theory. Doctoral dissertation, Massachusetts Institute of Technology.
- Mahapatra, B. P. (1996). Oriya writing. In Peter Daniels and William Bright, editors, *The World's Writing Systems*. Oxford University Press, New York, NY, pages 404–407.
- Majhi, T. D. (2007). Descriptive Oriya Morphology in the Paninian Model. Ph.D. Thesis, Centre for Linguistics, J.N.U., New Delhi.
- Martin, A., D. Maladhy, & V. Prasanna Venkatesan. (2011). A framework for business intelligence application using ontological classification. *arXiv preprint arXiv:1109.1088*.
- Masica, C. P. (1993). *The indo-aryan languages*. Cambridge University Press.
- Megyesi, B. (1998). Brill's rule-based part of speech tagger for Hungarian. Master's thesis, University of Stockholm.
- Menczer, F. (1997, July). ARACHNID: Adaptive retrieval agents choosing heuristic neighborhoods for information discovery. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE*- (pp. 227-235). MORGAN KAUFMANN PUBLISHERS, INC.
- Mhaske, N. T., & A. S. Patil. (2021). Resource creation for opinion mining: a case study with Marathi movie reviews. *International Journal of Information Technology*, *13*(4), 1521-1529.
- Mihalcea, R., & Paul Tarau. (2004, July). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404-411).
- Mitkov, R. (2003). Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing* (pp. 17-22).
- Mohanan, T. (1993). Verb Agreement in Complex Predicates in Hindi. In M. K Verma (ed.) *Complex Predicates in South Asian Languages*. Manohar Publishers and Distributors: New Delhi, 1993.

- Mohanan, T. (1994). *Argument structure in Hindi*. Center for the Study of Language (CSLI).
- Montrul, S. A., Rakesh M. Bhatt, & Archna Bhatia. (2012). Erosion of case and agreement in Hindi heritage speakers. *Linguistic Approaches to Bilingualism, 2*(2), 141-176.
- Muaz, A., Aasim Ali, & Sarmad Hussain. (2009, August). Analysis and development of Urdu POS tagged corpus. In *Proceedings of the 7th Workshop on Asian Language Resources* (pp. 24-29). Association for Computational Linguistics.
- Mukherjee, S., & Shyamal Kumar Das Mandal. (2013, December). Bengali parts-of-speech tagging using global linear model. In *2013 Annual IEEE India Conference (INDICON)* (pp. 1-4). IEEE.
- Müller, S. (2010). Persian complex predicates and the limits of inheritance-based analyses1. *Journal of Linguistics, 46*(3), 601-655.
- Mundotiya, R. K., Shantanu Kumar, Umesh Chandra Chaudhary, Supriya Chauhan, Swasti Mishra, Praveen Gatla, & Anil Kumar Singh. (2020). Development of a Dataset and a Deep Learning Baseline Named Entity Recognizer for Three Low Resource Languages: Bhojpuri, Maithili and Magahi. *arXiv preprint arXiv:2009.06451*.
- Mundotiya, R. K., Shantanu Kumar, Umesh Chandra Chaudhary, Supriya Chauhan, Swasti Mishra, Praveen Gatla, & Anil Kumar Singh. (2020). Basic linguistic resources and baselines for Bhojpuri, Magahi and Maithili for natural language processing. *arXiv preprint arXiv:2004.13945*.
- Nainwani, P., Esha Banerjee, S. Kaushik, & Girish Nath Jha. (2012). Issues in annotating less-resourced languages- the case of Hindi from Indian Languages Corpora Initiative (ILCI).
- Nidhi, R., & Tanya Singh. (2018, August). English-maithili machine translation and divergence. In *2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)* (pp. 775-778). IEEE.
- Nilu. (2013). COMPLEX PREDICATE IN MAGAHI. Doctoral Theses, Indian Institute of Technology, Patna.
- Nongmeikapam, K., Tontang Shangkhunem, Ngariyanbam Mayekleima Chanu, Laisuhram Newton Singh, Bishworjit Salam, & Sivaji Bandyopadhyay. (2011, March). Crf based name entity recognition (ner) in manipuri: A highly agglutinative indian

language. In 2011 *2nd National Conference on Emerging Trends and Applications in Computer Science* (pp. 1-6). IEEE.

- O'Malley, L. S. S. (1913). Census of India, 1911. Vol. V: Bengal, Bihar and Orissa and Sikkim. Part I: Report.
- Oehrle, R. T. (1976). The grammatical status of the English dative alternation. Doctoral dissertation, Mass. Cambridge.
- Ojha, A. K. (2019). English-Bhojpuri SMT System: Insights from the Karaka Model. *arXiv preprint arXiv:1905.02239.*
- Ojha, A. K., Pitambar Behera, Srishti Singh, & Girish N. Jha. (2015). Training & evaluation of POS taggers in Indo-Aryan languages: a case of Hindi, Odia and Bhojpuri. In *the proceedings of 7th language & technology conference: human language technologies as a challenge for computer science and linguistics* (pp. 524-529).
- Palmer, M., & K. Loper Kipper. (2004). Verbnet. *The Oxford Handbook of Cognitive Science*.
- Pandey, P. (2007). Phonology–orthography interface in Devanāgarī for Hindi. *Written Language & Literacy*, *10*(2), 139-156.
- Pandharipande, R. (1990). *Serial Verb Construction in Marathi*. In A. Zwicky and B. Joseph (eds.), *When Verbs Collide: Papers from the 1990 Ohio State Mini-conference on Serial Verbs*, pp 178-199.
- Pathak, P., Pinal Patel, Vishal Panchal, Narayan Choudhary, Amrish Patel, & Gautam Joshi. (2014). ezDI: A Hybrid CRF and SVM based Model for Detecting and Encoding Disorder Mentions in Clinical Notes. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 278-283.
- Pattanayak, D. P. & S. K. Prushty. (2013). *Classical Odia Language*. KIS Foundation, Bhubaneswar.
- Paul, S. (2004). An HPSG Account of Bangla Compound Verbs with LKB Implementation; Ph. D Dissertation, University of Hyderabad, Hyderabad.
- Poornima, S. (2012). Hindi Complex Predicates at The Syntax-Semantics Interface. Doctoral Dissertation, Faculty of Graduate School the University at Buffalo, State University of New York.
- Priya, Saloni (2020). Morphophonology of Magahi. *International Journal of Science and Research (IJSR).*

- Priyadarshi, A., & Sujan Kumar Saha. (2020). Towards the first Maithili part of speech tagger: Resource creation and system development. *Computer Speech & Language, 62*, 101054.
- Prusty, S. K. (2013). *Classical Odia*. KIS Foundation. http://www.orissalinks.com/odia/classical1.pdf
- Raj, M., Shyam Ratan, Deepak Alok, Ritesh Kumar, & Atul Kr Ojha (2022). Developing Universal Dependency Treebanks for Magahi and Braj. *arXiv preprint arXiv:2204.12633*.
- Rakesh, N., & Rajesh Kumar. (2013). Agreement in Magahi complex predicate. *International Journal of Linguistics*, *5*(1), 176.
- Ramchand, G., & Miriam Butt. (2002). Complex aspectual structure in Hindi/Urdu. *MS, UMIST and Oxford.*
- Ranjan, R., & Rajesh Kumar Dubey. (2016, December). Isolated word recognition using HMM for Maithili dialect. In *2016 international conference on signal processing and communication* (ICSC) (pp. 323-327). IEEE.
- Rathod, Y. V. (2018). Extractive text summarization of Marathi news articles. *Int. Res. J. Eng. Technology 5,* 1204-1210.
- REPORT OF THE COMMISSIONER FOR LINGUISTIC MINORITIES IN INDIA, Govt. of INDIA (July 2014 to June 2015)
- Ritu, Z. S., Nafisa Nowshin, Md Mahadi Hasan Nahid, & Sabir Ismail. (2018, September). Performance analysis of different word embedding models on bangla language. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)* (pp. 1-5). IEEE.
- Ruppenhofer, J., Michael Ellsworth, Miriam Petruck, Christopher Johnson, & Jan Scheffczyk. (2006). Framenet II: Theory and practice. *On-line publication at http://framenet. icsi. berkeley. edu*.
- Sahani, A., Kaustubh Sarang, Sushmita Umredkar, & Mihir Patil. (2016). Automatic text categorization of Marathi language documents. *Int J Comput Sci Inf Technol*, *7*(5), 2297-2301.
- Sarkar, K. (2012). Bengali text summarization by sentence extraction. *arXiv preprint arXiv:1201.2240.*
- Saxena, S., & Raghav Agrawal. (2013). Sanskrit as a programming language and natural language processing. *Global Journal of Management and Business Studies*, *3*(10), 1135-1142.

- Schachter, P. (1985). Lexical functional grammar as a model of linguistic competence. *Linguistics and Philosophy*, 449-503.
- Seiss, M. (2009). On the Difference between Auxiliaries, Serial Verbs and Light Verbs. In M. Butt and T. H. King (eds.), *Proceedings of the LFG09 Conference; CSLI* Publications: Stanford.
- Sicky, S. K. (2017). *Classification and Identification of Multi-Word Expressions in Magahi*. M.Phil. Dissertation, Centre for Linguistics, J.N.U., New Delhi.
- Sicky, S. K., & Pitambar Behera. (2022). A Descriptive Study of Explicator and Reverse Compound Verbs in Magahi – A Case Of Less-Resourced Language. Language in India. http://www.languageinindia.com/dec2022/shivekmaghagicompoundverbsfinal.html
- Sicky, Shivek Kumar & Pitambar Behera. (2022). A Descriptive Study of Explicator and Reverse Compound Verbs in Magahi- The Case of a Less-resourced Language. . *Language in India*. (pp. 1-10). ISSN 1930-2940. Vol. 22:12.
- Singh, D. P. (2011). *A Comparative Study of Hindi Parts of Speech Tagsets*. M.Phil Dissertation, Centre for Linguistics, J.N.U., New Delhi.
- Singh, D., Sudha Bhingardive, Kevin Patel, & Pushpak Bhattacharyya. (2015, December). Detection of multiword expressions for hindi language using word embeddings and wordnet-based features. In *Proceedings of the 12th international conference on natural language processing* (pp. 295-302).
- Singh, R., Atul Kumar Ojha, & Girish Nath Jha. (2016). Classification and Identification of Reduplicated Multi-Word Expressions in Hindi. *Classification and identification of reduplicated multiword expressions in Hindi, WILDRE*, 18-22.
- Singh, S., Rajesh Kumar, & Lata Atreya. (2014). Politeness in language of Bihar: a case study of Bhojpuri, Magahi, and Maithili. *International Journal of Linguistics and Communication, 2(*1), 97-117.
- Sinha, R. M. K. (2009). Mining complex predicates in Hindi using a parallel Hindi-English corpus. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE 2*009) (pp. 40-46).
- Sinha, S., & Girish Nath Jha. (2022). An Overview of Indian Language Datasets Used for Text Summarization. *ICT with Intelligent Applications*, 693-703.
- Soman, K. P. (2011). Parts of speech tagging for Indian languages: a literature survey. *International Journal of Computer Applications (0975-8887), 34*(8).

- The Stanford Log-linear part-of-speech Tagger. http://nlp.stanford.edu/software/tagger.shtml
- TntT – Statistical Part-of-speech Tagging. http://www.coli.uni-saarland.de/~thorsten/tnt/
- Toutanova, K., Dan Klein, Christopher D. Manning, & Yoram Singer. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 252-259).
- TreeTagger – a part-of-speech agger for any languages. http://www.cis.unimuenchen.de/~schmid/tools/TreeTagger/
- Typecraft: The Multilingual Interlinear Glossed Text (IGT) Bank. http://typecraft.org/tc2wiki/Main_Page
- Uniyal, A. (2011). *Issues and Challenges in Hindi Shallow Parsing*. M.Phil. Dissertation, Centre for Linguistics, J.N.U., New Delhi.
- Venkatapathy, S., & Aravind Joshi. (2005). Measuring the relative compositionality of verb-noun (VN) collocations by integrating features. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (pp. 899-906).
- Verma, M. K. (1993). *Complex predicates and light verbs in Hindi*. In M. K. Verma (ed.), Complex Predicates in South Asian Languages; Manohar Publishers and Distributors: New Delhi.
- Verma, M. K., (2007). *'Bhojpuri',* The Indo-Aryan Languages. (pp. 515-537). Routledge.
- Verma, M. K., & Karuvannur P. Mohanan. (Eds.). (1990). Experiencer subjects in South Asian languages. *Center for the Study of Language (CSLI).*
- Verma, S. (2007). 'Magahi', The Indo-Aryan Languages. (pp. 498-514). Routledge.
- Zanuttini, Raffaella. (2001). "Sentential negation", in: Mark Baltin and Chris Collins (eds.), *The Handbook of Contemporary Syntactic Theory*. Oxford: Blackwell, 511-535.