Reg. No.: VARA7452 ISSN 0974 7222

PARISHEELAN

International Multidisciplinary Peer Reviewed Quarterly Research Journal

U.G.C. Sr.No.-19365 U.G.C.Journal No.-40790

Vol-XVIII No.-1&2 Jan.-June-2023

Editor in Chief
Dr. Anjani Kumar Mishra

Editors
Dr. Archna Dubey
Dr. Sanjay Kumar

Published by
SURUCHI KALA SAMITI, VARANASI
Approved by U.G.C., New Delhi

A Brief Study Report On Group Of Twenty G-20

Dr Ratnesh Ranjan*

The group of G20 comprises Argentina, Australia, Brazil, Canada, China, France, Germany, India, Indonesia, Italy, Japan, Republic of Korea, Maxico, Russia , Saudi Arabia , South Africa, Turkey, United Kingdom and The United States along with the European Union . G20 members currently account for more than 80% of the Global Trade and 60% of the Global population.

ORIGIN AND EVOLUTION

The G20 was founded in 1999 after the Asian financial crisis as a forum for the finance Ministers and Central Bank Governors to Discuss Global economic and financial issues. The G20 was later upgraded to the level of heads of the state and was designated the "premier forum for International economic cooperation ". Since 2011, the G20 summit is held annually, under the policy, but it has since expanded.

ORGANISATIONAL STRUCTURE

The G20 dose not have a charter or a secretariat. The Presidency, Aided by the Countries holding the two tracks, there are thematically oriented working groups in which representative from the relevant ministries of the members as well as from invited countries and various international organization participate .

PREVIOUS G20 SUMMITS

The first G20 Summit was held in 2008 in Washington DC (USA) it set the Scene for the most dramatic reform of global Finance in over 60 years . At the follow -up summit in 2009 in London (UK) , the G20 agreed to blacklist states that refused crisis , the G20 resolve to imposed stricter controls on hedge funds and rating agencies .

***HOD, Political science Government College, Sundargarh, Odisha**

A Descriptive Study of Explicator and Reverse Compound Verbs in Magahi - The Case of a Less-resourced Language

Shivek Kumar Sicky, Ph.D. Scholar
Centre for Linguistics, SLL & CS, Jawaharlal Nehru University, New Delhi, India
loveshivek89@gmail.com
Pitambar Behera
Assistant Professor, Department of English, Govt. College, Sundargarh, Odisha
Ph.D. Scholar, Centre for Linguistics, SLL & CS, Jawaharlal Nehru University, New Delhi, India

Abstract

The phenomena of Reversed Compound Verbs (RCVs) and Explicator Compound Verbs (ECVs) have attracted a great deal of attention and focus recently by Linguistics and Computational Linguistics, as they are one of the salient features in almost all the South Asian languages. An ECV refers to the sort of constructions wherein the two different verbs i.e., v1 and v2 are sequentially arranged in order to form a meaningful expression. Here v1, which is the main verb of the sentence, is being followed by an auxiliary verb v2, which is later de-lexicalized.

On the other hand, an RCV is a type of compound verb formation in south Asian languages that was first noticed by Hook in 1974. An RCV is the reordered structuring of the same Compound Verb (CV), which is found in almost all south Asian languages as a special feature. The present paper is an attempt to examine the formation of both of these types of verbs and analyze their functions in Magahi. One could also find it interesting in investigating what happens to other linguistic properties of Magahi like syntactic, semantic and morphological properties when the structure is rearranged. It will also be interesting if both of these verb forms namely ECVs and RCVs show any sort of resemblance among them in Magahi once formed and how well they are accepted in the society. The present paper is also an attempt at making an aerial survey of Magahi language, which is one of the regional variations of Bihari languages and to figure out if there are any issues or challenges faced while reshuffling, reordering or restructuring CVs.

Keywords: Magahi, Bihar, South Asian Languages, Reverse Compound Verb, Explicator Compound Verb, Compound Verb.

1. Overview

This paper is an attempt at tracing the forms of the Reverse Compound Verbs and Explicator Compound Verbs in Magahi. It is an attempt to trace such verbs in a very less

International Journal of Dravidian Linguistics
Vol. 50 No. 1 January 2021

ISSN-0378-2

AN EXPLORATORY STUDY OF LEXICO-SYNTACTIC VARIATIONS IN SAMBALPURI AND ODIA*

PITAMBAR BEHERA, BISWANANDAN DASH &
DEBI PRASANNA PATTANAYAK

Abstract

This paper attempts to identify the changes that are found at the levels of lexis and syntax within Sambalpuri and Odia. This study explores how the structures of syntax and lexicon change from one geographical area to another. In doing so, the study attempts at bringing out the different lexical and syntactic phenomena from the morpho-syntactic perspective. Under the morphology section, it throws light on case endings, pluralisation, classifiers, post-positions, honorifics, emphatic particles and spatio-temporal nouns or frequentative adverbs agglutinating with various categories. Under the section of syntax, the paper gives adequate emphasis on copular verbs, negation and serial verbs.

[Keywords: Odia, Sambalpuri, Lexical variation, Syntactic variation, Morpho-syntactic]

---

* This is the revised version of the paper presented at the *UGC National Seminar on Minority and Minor Languages of India in the Present Context* at CPeDL, Dravidian University, Kuppam held on 21-23 February, 2017. We are indebted to Prof. B. Ramakrishna Reddy (Formerly at PS Telugu University, Hyderabad) and Prof. G. Balasubramanian (Presently the Vice Chancellor, Tamil University, Thanjavur) for giving their valuable comments that greatly enriched the present paper.

International Journal of Advanced Science and Technology
Vol. 29, No. 11s, (2020), pp. 1649-1659

A Comparative Analysis between Paninian and Tesni`ere's Dependency Formalisms with Specific Reference to Annotating Karta

Pitambar Behera

Assistant Professor (OES-I), Dept. of English, Govt. College, Sundargarh, Dept. of Higher Education, Govt. of Odisha, India

Ph.D. Scholar, Centre for Linguistics, School of Language, Literature and Culture Studies, Jawaharlal Nehru University (JNU), New Delhi-110067, India
pitambarbehera2@gmail.com

Abstract

Paninian Dependency (PD) framework takes into consideration both the syntactic and semantic aspects whereas the Tesni`ere's Dependency (TD[1]) formalism is oriented more towards the semantic aspect for annotation. On one hand, the former is modeled upon the Paninian model of Sanskrit grammar. On the other hand, the latter is based on Tesni`ere's initial experimentation in French; later universalized with the advent of Natural Language Processing and inspires Universal Dependency (UD) formalism. Researchers like Kiparsky and Staal in 1969 and Bharati and others in 1995 have emphasized on the point that the PD formalism is accommodative of Indian languages drawing empirical observations from Hindi-Urdu. Therefore, it is hypothesized that PD will be more suitable and accommodative of Odia language as well. For that said purpose, a comparative analysis between the Paninian and Universal Dependency Formalisms has been conducted to figure out which formalism works better for Odia with special reference to the annotation of the karta 'k1' dependency relation.

Keywords: *Panini, Tesni`ere, Dependency Grammar Formalism, Universal Dependency, Odia, Karta, Dependency Linguistics, NLP, Karaka, Vibhakti.*

1. Introduction

One of the most salient objectives of dependency grammar is to create a tree-like structure for connecting to each and every word in a sentence. Dependency grammar is considered to have influenced some linguistic theories; specifically pertaining to semantics, for instance, semantic relations/cases/theta roles and related to the predicate. The former set of theories upholds the view where various arguments do connect to either their head or predicate whereas the latter deals with the relations where arguments directly connect to their predicates. Paninian Grammar is considered to be suitable for accommodating natural languages at the syntactic level as compared to models such as Phrase Structure[2] (PS) grammar [1] and Government & Binding (GB) theory [2]; for it encapsulates both syntactic and semantic information [3]. So far as the word order of Odia is concerned, it is relatively free [4, 5] and is head-final [5, 6].

2. Literature Survey

[1] Universal Dependency and Tesniere's Dependency formalisms are applied complementarily as UD is modeled on Tesniere's Dependency formalism
[2] Phrase Structure grammar is the opposite of dependency grammar so far as their frameworks are concerned

ISSN: 2005-4238 IJAST

ISSN 2319 - 3514

प्रज्ञाभारती

षाण्मासिकी परिवीक्षिता राष्ट्रियशोधपत्रिका

(Half yearly Peer-Reveiwed National Research Journal)

प्रधानसम्पादकः

प्रो.सुकान्तकुमारसेनापतिः

सम्पादकौ

डॉ [illegible]माधवदाशः

[illegible]ेश्वरमिश्रः

प्रज्ञाभारतम्

www.pbharatam.in

Puri, Odisha

ISSN 2319 - 3514

प्रज्ञाभारती

षाण्मासिकी परिवीक्षिता राष्ट्रियशोधपत्रिका
(Half-Yearly Peer-Reviewed National Research Journal)
वर्षम् - ९ जनवरी - २०२० दशमपुष्पम्

प्रज्ञाभारतम्
www.pbharatam.in
Puri, Odisha

प्रकाशकः
प्रज्ञाभारतम् (पञ्जी.संख्या - ७३१३-८०)
पुरी - ओडिशा, M - 9861324413, 8982744588
www.pbharatam.in
अक्षरसंयोजनम्
श्रीमान् राधामाधवदाशः

मुद्रणः -
दाश-कम्प्युटर
Dash Computer
Khadipada, Puri, Odisha
प्रतयः -२००
मूल्यम् - ५००/- पञ्च शतम् रूप्यकाणि मात्रम्

संस्कृते कृषिविज्ञानम्

डॉ. विनोदविहारिमल्लिक
सहायकाचार्य, संस्कृतविभाग
नाशकीयविद्यालय, सुन्दरगढ़

संस्कृत संस्कृतेर्मूल ज्ञानविज्ञानवारिधि ।

वेदतत्त्वार्थसंजुष्ट लोकाऽऽलोककर शिवम्॥ (कपिलसंहिताय)

संस्कृत न केवल ब्राह्मणाना भाषा अपितु समग्रस्यापि विश्वस्य निर्यासभूत विज्ञानमेव । अत संस्कृत न केवल ज्ञानस्य निधि अपितु विज्ञानस्य वारिधि एव । अद्य युग विज्ञानस्य युग । विज्ञान व्यतिरिच्य साम्प्रतिकपरिस्थितौ जीवननिर्वाह दुष्कर प्रतिभाति । पदे पदे ग्रामे नगरे जले स्थले अनले आकाशे च विज्ञानस्य जयगाथा श्रूयते । तर्हि एतत् विज्ञान किमु अर्वाचीनम् ? एतत् अति प्राचीनम् । नहि, निखिलमपि वैदिक लौकिक च तथ्य दर्शनम्, भौतिकविज्ञानम्, रसायनविज्ञानम्, जीवविज्ञानम्, भूविज्ञानम्, खगोलविज्ञानम्, कृषिविज्ञानम् सकल शास्त्रसारम् (आख्यानम् नीतिशास्त्रम् आचारशास्त्रादिक) च संस्कृते (वेदे) निहितम् । अत उच्यते - यदिहास्ति तदन्यत्र यन्नेहास्ति न तत् क्वचित्।

कृषे महत्त्वम् - यथा आत्मान विना शरीरस्य स्थिति नास्ति । तथैव कृषि विना शरीरस्यापि परिकल्पन निरर्थकम् । समग्रे विश्वे सकलजीवानाम् उदरपरिपूरणार्थ कृषे महत्त्वम् सर्वजनविदितमेव । अस्माकं देशे प्रायत अशीतिप्रतिशत जना कृषिद्वारा जीवन्ति । यद्यपि वेदादिशास्त्रेषु कृषिकर्म केवल वैश्यानामेवास्तीति प्रतिपादितम् । तथापि इद कर्म सर्वेभ्य श्रेष्ठम्, कृषिकर्म सर्वोत्तम् यदनेन चराचरस्य कल्याण विदधाति । गीतायां भगवता उक्तम् -

अन्नाद भवन्ति भूतानि । पर्जन्यादन्नसम्भव ॥(श्रीमद्भागवद्गीतायाम्) इति ।

वेदेऽपि उक्तम् - **जीवन्ति स्वधयाऽन्नेन मर्त्या ॥** (अथर्ववे 12.1.22) **इति । अत** मानवजीवन अन्नोपरि निपतति । अन्नस्य प्राप्ति कृषिद्वारा सम्भवति । सृष्टे उत्पत्तिना सह अन्नस्योत्पत्ति ।

कृषे उत्पत्ति - पृथिव्या कृषिविद्याया कदा विकाश संजात इत्यस्मिन् विषये वेदे प्रतिपादितम् । वेदानुसार प्रथमत देवगणा स्वस्वपरशुना वनभूमि परिष्कृत्य कृषे उपयोगिभूमिरूपेण प्रस्तुतवन्त । यथा -

देवास आयन् परशुरबिभ्रन् वनावृश्चन्तोअभि विड्भिरायन् ॥ (ऋग् 10.28.8)

तदनुसार कृषिविद्याया प्रथमाविष्कारक भवति राजापृथु , य वेनस्य पुत्र ।

पृथी यद् वा वैन्य ॥(ऋग् 8.9.10)

वेदाध्ययनात् ज्ञायते सरस्वतीनद्या तीरे देवा एका माधुर्ययुक्ता भूमि प्रस्तुतवन्त । यत्र अधिष्ठातारूपेण मरुद्देवा कृषकस्य कार्य कृतवन्त । यथा -

देवा मधुना संयुत यद् सरस्वत्यामधिमणावचर्कृष इन्द्र आसीत् सीरपति शतक्रतु कीशाना आसन् मरुत सुदानव ॥(अथर्व 6.30.1)

शतपथब्राह्मणे कृषिकार्यस्य चत्वारशब्दा प्रयुक्ता यथा कर्षन्त , कर्षणम्, वपन्त (वपनम्), लुनन्त (लवनम्), मृणन्त (मर्दनम्) इति। (शतपथब्रा 1.6.1.3)

कृषिविज्ञानस्य महत्त्व प्रकाशयितु महाभारतशान्तिपर्वणि अपि उक्तम्

तेनेय पृथिवी दुग्धा सस्यानि दश सप्त च ॥ (महाभारत शा.प. 59.124)

कृषिनिमित्तम् उत्तमा भूमि आवश्यकी । उत्तमबीज , वृष्टि , हल , कृषक इत्येतेषामावश्यकता अपरिहार्या । अथर्ववेदस्य नवमन्त्रेषु अस्य चर्चा विहिता । उर्वराखलभूमिभेदेन द्विविधाभूमि । कृषियोग्या भूमि उर्वराभूमि । उर्वराभूमौ उत्पन्न अन्न उर्वरं नाम्ना अभिहित । नम उर्वर्याय इति यजु । यत्र भूमौ बीज न जायते तत् खलभूमीत्युच्यते। तत्रापि कृषियोग्यार्थं प्रक्रियाकरण क्रियते ।

ऋग्वेदानुसारम् उत्कृष्टान्ननिमित्त उत्तमकोटिबीजस्यावश्यकता विद्यते। यजुर्वेद तथा अथर्ववेदानुसार बीजवपनात् पूर्व कृषियोग्यभूमि घृतमधुदुग्धै शोधयेत्। अनेन भूमे उर्वराशक्ति वर्द्धते । यथोक्तम्

घृतेन सीता मधुना समज्यताम्। ऊर्जस्वती पयसा पिन्वमाना ॥ (यजुर्वे 12.70, अथर्व 3.17.9)

गोमहिषादिपालनेन दुग्ध दधि घृत तक्र च प्राप्यते । तेषा मलमूत्रादिसिञ्चनेन कृषिभूमि उर्वरा भवति ।

कृषिनिमित्त यज्ञस्य महत्त्वम् - वेदे कृषिनिमित्त यज्ञस्य महत्त्वमुक्तम् । यज्ञो वै विष्णु , यज्ञो वै भुवनस्य नाभि , वसन्तोऽस्यासीदाज्य ग्रीष्म इध्म शरद हवि (यजु.) (शतपथ ब्रा 1.1.2.13)इत्यादि वेदोक्तवचनानुसार यज्ञो वै कृषे प्राणा । यज्ञद्वारा अन्नस्य (उत्तमफलस्य) विवर्द्धन भवति । अक्षयान्न धान्यादिशस्याना च प्राप्ति भविष्यति । यज्ञद्वारा वृष्टि तथा वृष्टिद्वारा सुमिष्टमन्नम्, अन्नेन प्राणा एतत्स्वीकुर्यात्।

कृषिनिमित्तम् आवश्यकद्रव्याणि - उर्वराभूमि , उत्तमबीज , सूर्यालोक , वायु , वृष्टि , जलम्, सुरक्षित क्षेत्र, कृमिनाशकं औषध, गोमूत्र मलञ्च नितान्त प्रयोजनम् । अस्मिन् विषयेऽपि वेदेऽपि बहुमहनीय तथ्य प्रतिपादित विद्यते ।

कृषिसाधनानि - हल, सीता, फाल , ईषा, युग , वरत्रा, अष्ट्रा , बैलाश्च इति वेदेऽपि कृष्युपयोगिसाधनाना विषये विस्तृत वर्णन दरीदृश्यते । तथा तैत्तरीयसंहितासु कृषिसेचनसाधनाना पर्याप्त विवरण मिलति । ऋग्वेदेऽपि दिव्या , खनित्रिया , स्वयजा समुद्रार्था सेचनयोग्या साधनानि कथ्यन्ते ।

अन्नस्य प्रकारम् - वेदे अन्नस्य प्रकार द्विधा विभज्यते । यथा - कृष्टपत्यान्नम् यत् कृषिद्वारा उत्पद्यन्ते गोधूमादय । अपरम् अकृष्टपत्यान्नम् यत् कृषि विना उत्पद्यन्ते निवारादय वन्यधान्यानि ।

उपसंहार - वस्तुगत्या तदानीं कृषिविज्ञानस्य महतीप्रतिष्ठा संजाता । वेदोक्तविधिमालोच्य कृषका कृषिकार्य कृतवन्त वैदिककाले । येन पृथिवी सुजला सुफला शस्यश्यामला चासीत्। तत्र धन्या ते कृषका ये खलु स्वात्मान सुख परित्यज्य जगत्कल्याणाय सूर्योदयात् सूर्यास्त यावत् कृषिक्षेत्रे निजजीवन नियोजितवन्त । अत यथा जीवनस्य महत्त्वम् तथा कृषेरपि । इय भूमि अस्माक माता पुत्रोऽहं पृथिव्या । अत

वदात्कावाधमनुसृत्य यदि वयं प्राचीनकृषिविज्ञानम् अनुसरामः तर्हि क्षणात् [illegible] अवश्यममृतफलम् अस्मान् प्रदास्यति।

सहायकग्रन्थसूची

1. संस्कृतसाहित्येतिहास , बलदेवउपाध्याय , शारदानिकेतन, वाराणसी- 2018
2. वैदिकसाहतऔरसंस्कृति , बलदेवउपाध्याय , शारदानिकेतन, वाराणसी- 2018
3. ऋग्वेदसंहिता, सायणभाष्यम्, रामगोविन्दत्रिवेदी, चौखम्बासंस्कृतप्रकाशन, वाराणसी
4. यजुर्वेदसंहिता, सायणभाष्यम्, रामगोविन्दत्रिवेदी, चौखम्बासंस्कृतप्रकाशन, वाराणसी
5. सामवेदसंहिता, सायणभाष्यम्, रामगोविन्दत्रिवेदी, चौखम्बासंस्कृतप्रकाशन, वाराणसी
6. अथर्ववेदसंहिता, सायणभाष्यम्, रामगोविन्दत्रिवेदी, चौखम्बासंस्कृतप्रकाशन, वाराणसी
7. महाभारतम्, हिन्दी अनुवाद, रामनारायण, गीताप्रेस, गोरखपुर
8. वेद में विज्ञान, के.डि.द्विवेदी, विश्वभारतीपरिषद्
9. श्रीमद्भागवद्गीता, रामसुखदास, गीताप्रेस, गोरखपुर

Language in India www.languageinindia.com ISSN 1930-2940 Vol. 18:10 October 2018
India's Higher Education Authority UGC Approved List of Journals Serial Number 49042

A Qualitative Evaluation of Google's Translate: A Comparative Analysis of English-Urdu Phrase-based Statistical Machine Translation (PBSMT) and Neural Machine Translation (NMT) Systems

Sharmin Muzaffar, M. A., Ph.D. (Linguistics)
Pitambar Behera, M.A., B.Ed., M.Phil., Ph.D. (Linguistics)

Abstract

The paper demonstrates the qualitative evaluation of the English to Urdu Machine Translation Systems, namely PBSMT and NMT hosted on Google's Translate. This system is popularly known as Rosetta, formerly governed by Phrase-based approach and is presently governed by the neural module of source and target languages. In this study, a model corpus set of 100 English sentences has been applied out of 1k cross-domain data considering various types of verbs as input text to evaluate the output of the online systems in Urdu.

In order to evaluate the output text in a qualitative manner, the Inter-translator Agreement (IA) of three human translators has been considered with their scores on a five-point scale. The scores are calculated by the Fleiss' Kappa statistical measure with regard to comprehensibility and grammaticality on the basis of which error analysis and suggestions have been provided for improvement. The Kappa scores of PBSMT for comprehensibility and grammaticality are 0.24 and 0.22 respectively which is indicative of the fact that on both counts the scores are not up to the mark. Furthermore, the system has also been quantitatively evaluated on the basis of word error rate (21.11%) and sentence error rate (72.39%). On the contrary, NMT module has Kappa scores of 0.61 and 1 on comprehensibility and grammaticality respectively. So far as WER and SER are concerned, NMT has 32.58% and 28% respectively. In addition, all the erroneous entities have been analyzed through computational typology. The strategy for evaluation is to evaluate the Urdu output text based on the five-point scale with scores that range from 0-4 where 0 refers to incomprehensible or ungrammatical, 1 = little meaning or disfluent, 2 = neutrality, 3 = comprehensible or grammatical and 4 suggests flawless in both cases.

Keywords: PBSMT, NMT, Google's Translate, MT, Urdu, Indo-Aryan, NLP, Fleiss' Kappa

Overview

As discussed in Castilho et al. (2018), since the advent of the Machine Translation (MT) or automated translation, new methods, approaches and techniques have really created high expectations among researchers. On one hand, the qualitative approaches have paved the way for a graded or incremental improvements in contrary to the significant improvements exhibited by the statistical approaches. Among the statistical techniques, Neural Machine Translation (NMT) has recently emerged as an innovative and robust technique as it has generated a lot of attention because of its high qualitative outputs in comparison to its counterparts.

A Peer-reviewed Journal of Linguistic Society of Nepal

Nepalese Linguistics

Volume 33 (1) November 2018

Nepalese Linguistics is a **peer-reviewed** journal published by Linguistic Society of Nepal (LSN). LSN publishes articles related to the scientific study of languages, especially from Nepal. The authors are solely responsible for the views expressed in their articles.

Published by: Linguistic Society of Nepal
Kirtipur, Kathmandu
Nepal

Copies: 300

ISSN 0259-1006

Price: NC 400/- (Nepal)
IC 350/- (India)
US$ 10/-

The publication of this volume was supported by Nepal Academy.

DEVELOPING CLASSIFICATION-BASED NAMED ENTITY RECOGNIZERS (NER) FOR SAMBALPURI AND ODIA APPLYING SUPPORT VECTOR MACHINES (SVM)

Pitambar Behera and Sharmin Muzaffar

This paper demonstrates the development of named Entity Recognizers (NER) applying Support Vector Machines (SVM) for Sambalpuri and Odia. The Sambalpuri corpus amounts to 112k word tokens out of which 5,887 are named entities. On the contrary, 250k ILCI corpus has been applied for Odia out of which 18,447 tokens are named entities. The former accurately recognizes 96.72% whereas the latter provides 98.10% accuracy.

Keywords: *NER, Sambalpuri, NLP, Odia, SVM, Machine Learning, Indo-Aryan languages, Information Retrieval, Natural Language Processing.*

1 Overview

Named entity recognition (NER) is one of the applications of Natural Language Processing and it is considered as the subtask of information retrieval. NER is the process of detecting Named Entities (NEs) in a document and to categorize them into certain named entity classes such as the names of organization, person, location, sport, river, city, country, quantity etc. In English, we have accomplished a lot of work pertaining to recognizing named entities. On the contrary, we have not achieved remarkable accomplishment with regard to detecting NER in Indian languages. India is an abode for 22 official languages along with endangered, lesser-known and less-studied languages. NER is still considered to be an emerging area of research in the field of NLP in the context of Indian languages.

There are various applications of NER such as Information Extraction, Question Answering, Information Retrieval, Automatic Summarization, Machine Translation, etc. The Named Entities can be made known to us by performing computation on a given natural language through rule-based or statistical approaches. The task of identification, extraction and retrieving necessary information can be made faster, if we are already acquainted with the nature, type and functions of named entities. Therefore, NER is the process of detecting, classifying and extracting Named Entities in a document into their corresponding classes with the application of any of the NER based approaches.

1.1 Approaches to named entity recognition

There are basically two broad approaches that are employed in the recognition of named entities (Nayan et al., 2008; Sasidhar et al., 2011; Saha, 2008). These include: Rule-based approach and Machine learning based approach (Kaur and Gupta, 2012; Kaur and Gupta, 2010; Srivastava et al., 2011).

1.1.1 Rule-based approach

Under this section, there are list lookup approach and linguistic approach. So far as the former is concerned, gazetteers are exploited that comprise of different lists of named entity classes and a simple look up or search operation is conducted in order to detect whether a word belongs to a named entity class or not. If a particular word belongs to a named entity class, a named entity label, as specified in the annotation schema, is allotted to that word on the basis of the named entity class which it originally belongs to. On the other hand, in linguistic approach, a linguist is entrusted with the work of formulating heuristic linguistic rules, so that the named entities can be identified as well as classified and extracted easily (Ekbal and Bandyopadyay, 2010; Gupta and Lehal, 2011). The formulated rules are language dependent and cannot be applied in order to identify named entities in any other given language (Kaur and Gupta, 2012). Therefore, data-driven statistical approach became indispensable.

1.1.2 Statistical approach

This approach is motivated by the machine learning theories and algorithms, for instance, Hidden Markov Models (HMM), Maximum Entropy Markov Model (MEMM), Conditional Random Field (CRF), Support Vector Machines (SVMs), Decision Tree and so on.

Issues and challenges in corpus collection and annotation of Sambalpuri: The case of a lesser-known language

Pitambar Behera, Research Scholar, Centre for Linguistics, Jawaharlal Nehru University, New Delhi, India. pitambarbehera2@gmail.com (corresponding author)

Abstract

The present paper brings out the issues and challenges underlying the collection and annotation of a voluminous corpus collected for Sambalpuri, a lesser-known language spoken in the eastern region of India (Kushal, 2015). It is an Indo-Aryan (IA) language which is otherwise known as Dom, Kosali, Koshal, Koshali, Western Odia[1]. It is spoken in the ten districts of western and south-western Odisha. The total corpus collected for it amounts to 121k tokens covering five domains: literature, sports, tourism, entertainment and miscellaneous.

Two issues such as corpus collection and annotation have been taken up in this paper. The unavailability of a corpus for a low-density language proves to bear adverse impacts on its future NLP development. If the data is available, it is either in the image form or non-Unicode encoding or not in machine readable format. Since these languages are not technologically empowered, there are not enough number of tools to convert the texts into Unicode. Owing to the facts that Sambalpuri being a lesser-known language and guidelines developed for Indian languages have been devised for only the scheduled languages, there are significant issues and challenges one faces with regard to annotation of a voluminous corpus. These underlying issues mostly pertain to the non-incorporation of unnoticed unique linguistic features of these languages into the uniform guideline devised for the scheduled languages.

Keywords: Corpus Linguistics, Sambalpuri, Lesser-known language, POS tagging, Indo-Aryan language.

1. Overview

Sambalpuri-Kosli (ISO 639-3 spv) owes it's origin to the Indo-Aryan Language family spoken in not only ten districts (Sambalpur, Bargarh, Bolangir, Sonepur, Kalandi, Sundargarh, Boud, Deogarh, Nuapada and Jharsuguda) of western and south-western districts of Odisha with approximately 18 million (Census Report, 2001) population but some parts of Jharkhand and Chattisgarh also.

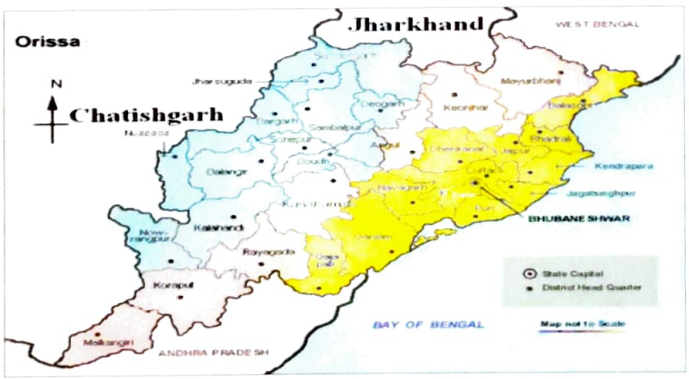

Figure 1: The Sambalpuri-Kosli speaking region (sky blue demarcated)

[1] http://www.ethnologue.com/language/spv

Journal- Language forum (2018)
ISSN- 0253-9071