International Journal of Innovative Research and Advanced Studies (IJIRAS)
Volume 4 Issue 1, January 2017

ISSN: 2394-4404

Developing A Heterogeneous Contextual Corpus Rater For Odia Language

Rudranarayan Mohapatra

P.G. Department of Odia, Utkal University, Vanivihar,
Bhubaneswar, Odisha, India

Abstract: ***It is a common knowledge that having access to an error free corpus is of great benefit for developers, learners, instructors and users. Each corpus provided a unique perspective. When developing the Corpus Rater for Odia language content having heterogeneous in nature, certain linguistic issues like current theories of language, language learning and good practice in assessment are taken into consideration. The paper describes self-made classification of Odia language errors types using Odia language corpus data. The paper also presents and elaborates the developing a corpus rater in principle of zero quality control using 'Process writing' behavior in order to reach its perfection. The knowledge can be a great use in assessing soundness of our error annotation.***

Keywords: Heterogeneous Corpus Rater (HCR), Intention Counter, Error Mode and Impact Analysis Model, Error Tree Builder, Bootstrapping Algorithm'

I. INTRODUCTION

The cost of ignoring Official Communication has become substantial as the proliferation of electronic information amidst an increasingly regulatory environment has made it prohibitively expensive to take a reactive approach to the issue very specific to Indian Languages. The dramatic growth in the use of electronic information within Official Communication and expectation of standards in business practice requires new technology solutions that can fully address governance in an automated fashion in Indian Languages.

Looking this aim to build a surveillance tool for Official languages domain (Odia language) for automatically editing and rating the official communication documents in Indian languages and would help to try to fix the improvised suggestion not only from morphological or syntax direction but also from the direction of contextual semantic nature looking the whole text subjective mood and expectation in a multilingual or cross lingual environment to cater the client interest in desired languages.

II. CORPUS RATER & ITS FUNCTION

The Corpus Rater system has two major features. Firstly, the texts known as warrant are to be analyzed by our pre-machine learning library to identify the features that make them distinctive. Secondly, machine-learning strategies will be used to analyze the texts to derive other contextual features that may be useful in classification and would try to automatically fixing the probable improvements with an expected rating.

This task will comprise of major technologies like Internet document classification system query based analyzer and Machine Translation system.

The Corpus Rater can:

- ✓ Detect the text languages, domains & domain names
- ✓ Make Automatic Formatting with the given setting parameters.
- ✓ Find those pesky mistakes like (spelling, punctuation and Conjunction markers) and correct them before turning the text.
- ✓ Automatic handle the Syntax, morphology the text and makes possible grammatical agreements.
- ✓ Automatic alerts to opportunities to improve the writing

ISSN: 2394-4404

- ✓ Automatic try to replace the domain specific & contextual vocabulary looking assessing the mode and mood of the written text.
- ✓ The user can personalize the proofreading process by giving probable real-time committed mistakes if necessary.
- ✓ Automatically Grade the Corpus, its style and Word Choice Analysis

The system would combine the power of Natural Language Processing (NLP), Artificial Intelligence (AI), machine learning, Information Retrieval (IR), Computational Linguistics, Data Mining, and Advanced Pattern Matching (APM) process in Indian Languages looking to the future huge data management and reducing the task of large human intervention.

So we wanted to create tools that do not muddy the waters in the process of assisting with the craft of writing in Indian languages but to behave as a surveillance servicing tool for large data management in improving the editing quality of Indian language corpus very specific to here is for Odia language.

III. LITERATURE SURVEY

Generally a Corpus rater behaves as an eco-system for Automatic Language production. In the early 1980s, NLP began to facilitate the development of new writing tools as well: Writer's Workbench was designed to help students edit their writing. This software provided automotive feedback mostly related to writing mechanics and grammar (MacDonald, Frase, Gingrich, & Keenan, 1982). Some suggested resources in this direction are MLA Guide, APA Guide etc. And some suggested products are: Grammarian Pro X, After the Deadline, style-check.rb etc. However, looking to the Odia language prospective, a standard Corpus rater tool is a day dream to language market.

IV. APPROACH TO BE FOLLOWED

The Corpus Rater would follow the 'Process Writing' behavior for proofing the heterogeneous corpus from the very beginning of language recognition to review, edit and the final proof.

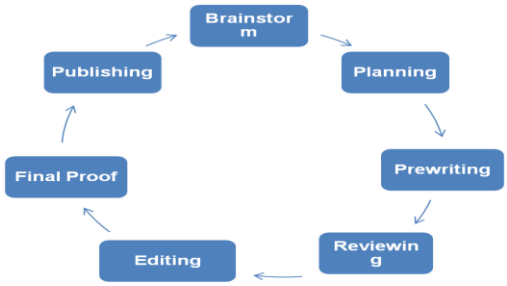


Figure 1: over all process flow of the System

V. ARCHITECTURE OF CORPUS RATER SYSTEM

The Sub-components of the system can be described through (1) LMS Data base Creation, (2) "Error Mode and Impact Analysis Model: EMIA", (3) Intention Counter Module, (4) Error Tree Builder. The details of system Architecture is illustrated at the end of the paper at Fig. No. 2.

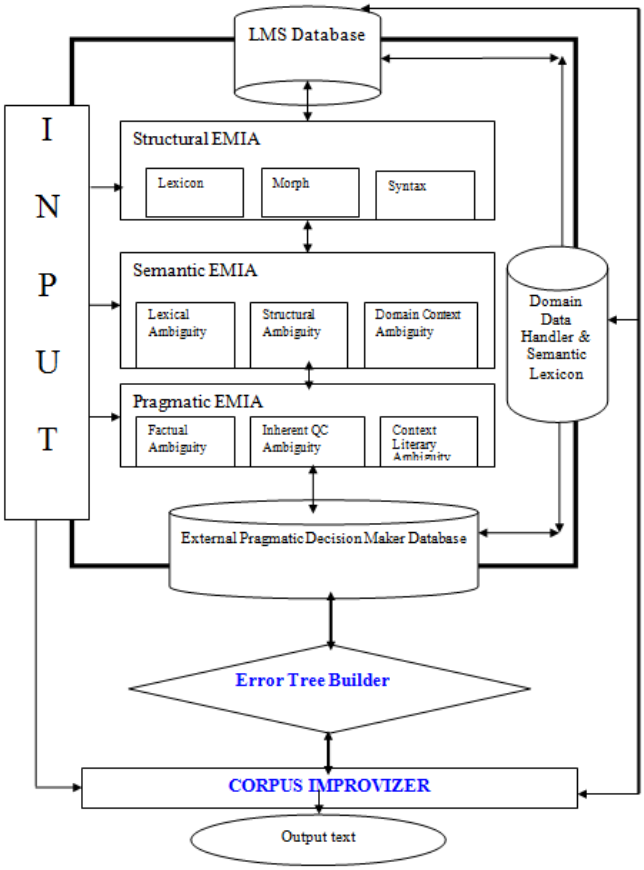


Figure 2: System architecture flow of the System

VI. ERROR MODE AND IMPACT ANALYSIS MODEL (EMIA)

The GUI of EMIA would be would be a fourfold divisional model. It starts from input and followed by improvement, annotation and then evaluation. The life cycle process of EMIA takes its continuation process of initiation to conclusion till it receives perfection.

A. TYPES OF AUTOMOTIVE EMIA

For a Corpus Rater tool, the EMIA starting from Structural level to Semantic and then conceptual level i.e. pragmatic EMIA should be taken into consideration. Where lexical, morphological and syntactic behavior are falls under Structural EMIA, the Pragmatic part is the part of Conceptual EMIA. The semantic part of corpus is the bridge between Structural and conceptual EMIA both.

ISSN: 2394-4404

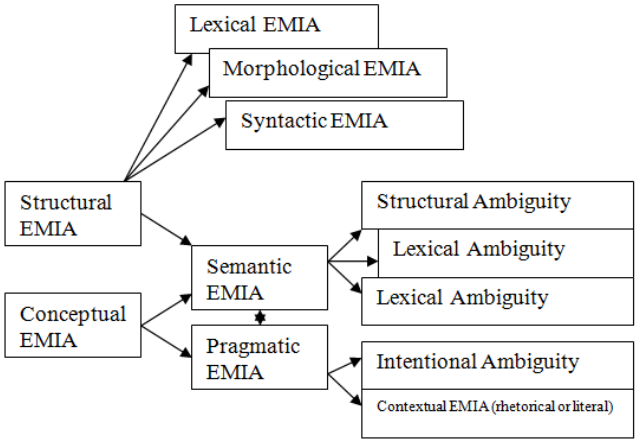


Figure 3: Types of automotive EMIA

B. PRINCIPLES OF AUTOMOTIVE EMIA

The principles of Automotive EMIA are Zero Quality Control. The cause of Errors in a text are Forgetfulness, Errors due to misunderstanding, Errors in identification, Errors made by amateurs, Willful errors, Inadvertent errors, Errors due to slowness, Errors due to lack of standards, Surprise errors, Intentional error.

C. RELATIONSHIP OF AUTOMOTIVE EMIA

There is a great cohesive relationship among the modules of Structural EMIA, Semantic EMIA and Pragmatic EMIA in the foundation of both linguistic theories and technical practice. However we shouldn't Mix up the Semantic Errors and causes with Pragmatic Errors.

Error Mode	Effect	Cause
	Structural EMIA	
Consequence	The Problem	The Cause of Error
	Semantic EMIA	
The Cause of Error of structural EMIA	The error of structural EMIA with better explanation	New Root cause for Semantic EMIA model
	Pragmatic EMIA	
The Cause of Error of Semantic EMIA	The effect of Same error as the Semantic EMIA	Specific Root cause for Pragmatic EMIA

VII. CREATING STRUCTURAL (GRAMMATICAL) EMIA FLOW CHART

A. LEXICAL EMIA

After standard tokenization, each token is passed through the lexical verification process and then inserted into the Lexicon Management System (LMS) which supports automated and manual resolution of unknown tokens. The LMS is a system developed to store the accumulated lexical knowledge and contains categorizations of spelling errors, abbreviations, acronyms and a variety of non-word tokens.

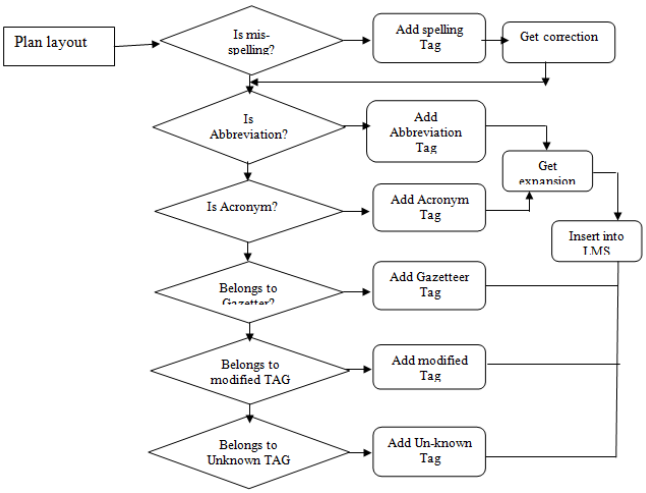


Figure 3: Lexical EMIA Flowchart

B. MORPHOLOGICAL EMIA AND ITS FUNCTIONALITY

In the field of language, morphology means the structure of words, how words are formed, and how the parts fit together. If you get the wrong morpheme (i.e., word part) in the wrong place at the wrong time, you've committed a morphological error. It is the branch of linguistics that deals with the study of the internal structures of words and how new words are created from the existing ones through the use of various morphological processes namely affixation, compounding, conversion, blending, chipping, reduplication etc. (cf. O'Grady and Guzman 1996; Quirk and Greenbaum, 1973).

In this section we present the actual morphological errors in our subjects' scripts and spoken Odia. The errors are subdivided into: (a) Affixation-related errors; (b) Compound-related errors; and, (c) Conversion-related errors

Sl. No.	Error Types	Examples
1. (a)	**Affixation-related errors**	
	Errors arising from the wrong use of prefixes	ଆପଣ ଜଣେ ବେସାଧୁ (ଅସାଧୁ) ବ୍ୟକ୍ତି। *Aapana jane besadhu (asaadhu) byakti.*
	Errors arising from making uncountable nouns countable	ସମୁଦ୍ରର ଜଳଗୁଡିକ ଲୁଣିଆ। (ଜଳ) *Samudrara jalagudika luniaa. (jala)*
	Errors arising from omission of suffixes	ବିକାଶ ଓଡିଆ ବ୍ୟାକରଣ (ବିକଶିତ ଓଡିଆ ବ୍ୟାକରଣ) *bikaasa odia byaakarana (bikasita odia byaakarana)*
2. (b)	**Compound-related errors:**	Compounding is a morphological process which consists in the combination of at least

		two free morphemes such as book + shop (bookshop).
	Compound errors in the rewriting exercise	'ଡାକ ଘର' (Post Office) instead of 'ଡାକଘର' (Post-Office)
3. (d)	**Conversion-related Errors**	
	Few conversion-related errors in the written responses of the subjects.	ମୁଁ ଯିବାଠୁଁ ପ୍ରସ୍ତୁତ (instead of "ମୁଁ ଯିବାକୁ ପ୍ରସ୍ତୁତ") *mu jibaathun prastuta (mu jibaaku prastuta)*

C. SYNTACTIC EMIA AND ITS FUNCTIONALITY

A syntax error is a violation of the syntax, or grammatical rules, of a natural language**.** However the Syntactic EMIA module will handle both the Syntactic & Non- Syntactic error types.

Sl.	Error Types	Examples	Examples
1.	Definiteness form of noun	ତାଙ୍କର କିଛି ବହିଗୁଡିକ ଅଛି। *Taankara kichhi bahigudika achhi.*	ତାଙ୍କର କିଛି ବହି ଅଛି। *Taankara kichhi bahi achhi.*
2.	Subject & verb agreement	ସେମାନେ ଘରକୁ ଯାଉଛି। *semaane gharaku jaauchhi.*	ସେମାନେ ଘରକୁ ଯାଉଛନ୍ତି। *semaane gharaku jaauchhanti.*
3.	Wrong form of vocabulary	ସେ ଏହାକୁ ସୁନିଶ୍ଚିତତା କଲେ। *Se ehaaku sunishitata kale.*	ସେ ଏହାକୁ ସୁନିଶ୍ଚିତ କଲେ। *Se ehaaku sunishita kale.*
4.	Position of adverb	ମଧ୍ୟ ମୋର ଏ ବହିଟି ଅଛି। *madhya mora e bahiti achhi.*	ମୋର ଏ ବହିଟି ମଧ୍ୟ ଅଛି।/ ମୋର ମଧ୍ୟ ଏ ବହିଟି ଅଛି। *mora e bahiti madhya achhi./ mora madhya e bahiti achhi.*
5.	Voice	ବର୍ତ୍ତମାନ ଆପଣ ଆପଣଙ୍କ ଚିଠିକୁ ପଠଉଛି। *bartamaana aapana aapananka chithiku pathauchhi.*	ବର୍ତ୍ତମାନ ଆପଣ ଆପଣଙ୍କ ଚିଠିକୁ ପଠାନ୍ତୁ। *bartamaana aapana aapananka chithiku pathaantu.*
6.	Double negation	ସେ ନଯାଇପୋରିଲେନାହିଁ। *Se najaaipaarilenaahim.*	ସେ ଯାଇପୋରିଲେନାହିଁ। *Se jaaipaarilenaahim.*

Table 1: Some frequent Syntactic error types

Sl. No.	Error Type	Incorrect Examples	Correct from
1.	Quotation marks	ରାଧାନାଥଙ୍କ ଦୃଷ୍ଟିରେ, "ଚିଲିକାର ପ୍ରାକୃତିକ ରୂପଶୋଭା ଅନନ୍ୟ। (Without closing quotation mark)	ରାଧାନାଥଙ୍କ ଦୃଷ୍ଟିରେ, "ଚିଲିକାର ପ୍ରାକୃତିକ ରୂପଶୋଭା ଅନନ୍ୟ।" (With closing quotation mark)
2.	Date expressions	ତା 2016, 3, ମଇ ରିଖ (in year, date, month pattern)	ତା 3, ମଇ 2016 ରିଖ (in date, month, year pattern)
3.	Several spaces in a row	ଭୁବନେଶ୍ୱର ଓଡିଶା ର ରାଜ୍ଧାନୀ। (Without post position agglutination)	ଭୁବନେଶ୍ୱର ଓଡିଶାର ରାଜ୍ଧାନୀ। (With post-position agglutination)
4.	Abbreviation & Acronyms	ଓ.ଏ. (ଓଡିଶା ସାହିତ୍ୟ ଏକାଡେମୀ)	ଓ.ସା.ଏ. (ଓଡିଶା ସାହିତ୍ୟ ଏକାଡେମୀ)
9.	Parentheses	(କେନ୍ଦ୍ର ସାହିତ୍ୟ ଏକାଡେମୀ} (with different open & close bracket)	(କେନ୍ଦ୍ର ସାହିତ୍ୟ ଏକାଡେମୀ) (with same open & close bracket)
10.	Measurement Units	ଡେସିଲିଟର (ଡେସି. ଲି)	ଡେସିଲିଟର (ଡେ.ଲି.)
11.	Punctuation marks	ଉଦାହରଣ; ଲାବଣ୍ୟବତୀ ଉପେନ୍ଦ୍ର ଭଞ୍ଜଙ୍କ ଦ୍ୱାରା ଲିଖିତ. (with semicolon)	ଉଦାହରଣ: ଲାବଣ୍ୟବତୀ ଉପେନ୍ଦ୍ର ଭଞ୍ଜଙ୍କ ଦ୍ୱାରା ଲିଖିତ। (with colon)

Table 2: Some non-syntactic error type

VIII. SEMANTIC EMIA

To make the semantic EMIA, there is a Semantic lexicon would be prepared by using 'Bootstrapping Algorithm' which discovers words with semantic properties similar to a small set of labeled seed examples. The semantic EMIA would be primarily responsible for:

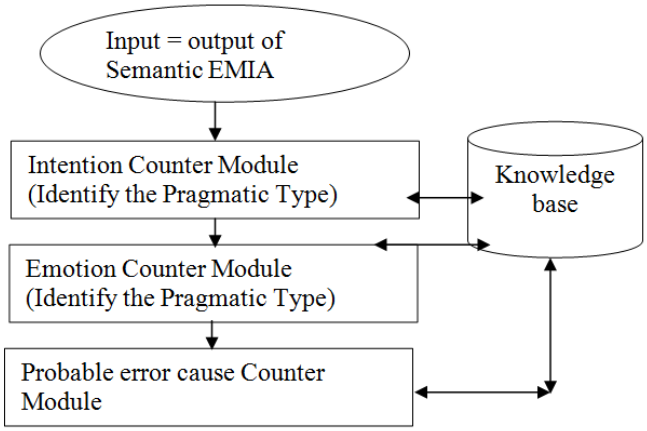
- ✓ Lexical Ambiguity: Would count domain specific probable contextual lexicon based upon thesaurus and Indian language WordNet & FrameNet. Etc.
- ✓ Structural Ambiguity: Would calculate based upon previous and post text sentences about best and Appropriate Syntactic structure out of more than one Structural output.
- ✓ Domain Context Ambiguity: Would calculate the contextual stylistic errors and would suggest probable replacements.

IX. CREATING PRAGMATIC EMIA FLOW CHART

Creating a Pragmatic EMIA flowchart is still in our conceptual level and needs more study for its better form. However some major steps of the same are mentioned here for future study and analysis.

- ✓ Identify the Pragmatic Type
- ✓ Look the Manual Feedback Model
- ✓ Agree on Starting and Ending point of the Context
- ✓ Agree on level of Details
- ✓ Look for Text Areas that to be Improved
- ✓ Construct Pragmatic Flowchart
- ✓ Analyze the Results

Diagram of Pragmatic EMIA flowchart



A. PRAGMATIC ERROR CAUSES

- ✓ Omitted Pragmatics
- ✓ Pragmatization errors
- ✓ Fault in pragmatization texts
- ✓ Missing parts
- ✓ Wrong parts
- ✓ Adjustment error
- ✓ Peripheral text not fully agreed to the text
- ✓ Tools and/or modules improperly prepared

Pragmatic EMIA would follow Error Tree construction steps to count down the probable errors and its necessary resolving possibilities.

X. ERROR TREE FUNDAMENTALS

- ✓ Defining Types of Errors
- ✓ Comparison of Error Occurrence and Error Existence
- ✓ Comparison of Error Causes and Error Effects

A. ERROR TREE CONSTRUCTION STEPS SUMMARY

- ✓ Determine the level to which the error identification and Analysis should be constructed
- ✓ Begin with the Structural Level Errors
- ✓ Fully describe all events which immediately cause this the Error
- ✓ With each lower-level errors, continue describing its immediate causes until a module level error or human error can be attributed to the same error
- ✓ Fully define each branch of the tree before beginning another branch

During the construction of the tree, it is advisable to use a block diagram of the system to simplify determining the main branches

XI. INTENTION COUNTER MODULE

It would analyze the features like: Forgetfulness, Errors due to misunderstanding, Errors in identification, Errors made by amateurs , Willful errors , Inadvertent errors , Errors due to slowness , Errors due to lack of standards, Surprise errors , Intentional error.

- ✓ They are the mental causes of actions, that is, they are what together with some bodily movements constitute an action, as distinct from a mere event.
- ✓ They have conditions of consistency. You can desire p and desire not-p at the same time, but you cannot intend p and intend not-p at the same time.
- ✓ Their object is presupposed to be attainable by the agent. You can desire to go to the moon this afternoon, but you cannot intend to go to the moon this afternoon (unless you are a multimillionaire who has made an arrangement with some spatial agency).
- ✓ Their object represents their conditions of satisfaction.

XII. CONCLUSION

In this paper, we have attempted to identify and classify some of the frequent occurring errors in the Odia language and their probable solution for to develop well managed Odia corpora in an automation process. This paper among other things reveals that the causes of these errors are numerous, and they range from the inconsistency inherent in Odia language itself, overgeneralization of rules, and misapplication of rules interference. Looking to the constraint of paper and our analysis report, we take semantic and pragmatics error rectification part for further study.

REFERENCES

[1] Elizaveta Kuzmenko, Andrey Kutuzov, "Russian Error-Annotated Learner English Corpus: a Tool for Computer-Assisted Language Learning", National Research University Higher School of Economics, http://www.aclweb.org/anthology/W14-3507

[2] Maria Belen Díez-Bedmar, Universidad de Jaén, Marcus Callies & Ekaterina Zaytseva, Johannes-Gutenberg Universität Mainz, "Using Learner Corpora for Testing and Assessing L2 Proficiency", L2 Proficiency Assessment Workshop Montpellier, 24-25 February 2012

[3] Raymond Hickey, University of Munich, "Corpus Data Processing with Lexa", Page 11-17, https://www.uni-due.de/~lan300/06_Corpus_Data_Processing_with_Lexa_(Hickey).pdf

ISSN: 2394-4404

[4] Vít Suchomel, "New features in Corpus Architect (corpus management)", Lexical Computing, 4th Sketch Engine Workshop, Tallinn, October 16, 2013
[5] Wernard Schmit & Sander Wubben,"Predicting Ratings for New Movie Releases from Twitter Content", Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015), pages 122–126, Lisboa, Portugal, 17 September, 2015. © 2015 Association for Computational Linguistics.
[6] Gayatree Ganu & et. al., "Beyond the Stars: Improving Rating Predictions using Review Text Content", Twelfth International Workshop on the Web and Databases (WebDB 2009), June 28, 2009, Providence, Rhode Island, USA.