IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Detection and Correction of Grammatical Errors in Hindi Language Using Hybrid Approach

M. Mittal[1], S K Sharma[2*], A Sethi[3]

[1]Department of Computer Science and Engineering,Guru Kashi University,Talwandi Sabo,Punjab, India
[2]Department of Computer Science and Applications, DAV University, Jalandhar,Punjab, India
[3]Department of Computer Science and Engineering,Guru Kashi University,Talwandi Sabo,Punjab, India

**Corresponding Author: sanju3916@rediffmail.com, Tel.: +91-94650-05780*

DOI: https://doi.org/10.26438/ijcse/v7i5.421426 | Available online at: www.ijcseonline.org

06/May/2019, Published: 31/May/2019

Abstract:-Grammar checking or proof reading is one of the major tool incorporated in almost every word processor software. Almost all the word processor software contains spell checker and grammar checker as an essential component. The function of the grammar checker is to check the grammatical mistakes in the text typed by the user. In this research article, authors have developed a grammar checking system for Hindi language using hybrid approach. All the components (Morphological analyzer, POS tagger, error detection system and error correction system) required for development of grammar checker have been developed from scratch. Some components like morph, POS tagger and error detection systems have been developed using statistical approach and grammar correction system has been developed using rule based approach. Hence overall hybrid approach has been used for development of complete Hindi grammar checker. The system is tested for four different types of errors (Adjective noun agreement errors in terms of number and gender, Noun Verb agreement errors in terms of number and gender) and on testing, the system shows an overall precision of 0.83, Recall as 0.91 and F-measure as 0.87.

Keywords: Hindi Grammar checker, Hindi POS tagger, Hindi morph, HMM, Hybrid approach.

I. INTRODUCTION TO GRAMMAR CHECKER

Language is define as a way of communication between two entities and natural language is used by humans to communicate with each other. Grammar of a language is defined as set of instructions followed while writing the sentences in that language. These instructions may include instruction regarding presence of agreements between different part of speech of the sentence like noun should be grammatically in agreement with verb in terms of number, gender and case, similarly modifier should also be in agreement with the noun to whom it is modifying and many more. Every language has its own syntax and hence its own grammatical rules. Although two language belonging to similar family may have similar structure or similar grammar rules but some differences are always there. Hence the grammar checking system developed for one language cannot be used for other language without any modification. The general syntax of grammar checker is shown in figure 1:

As shown in figure 1, an input sentence is checked by the grammar checker software and if the sentence is grammatically incorrect then the suggestion to rectify the sentence will be provided by the software otherwise if the sentence is correct then it will be displayed as it is.

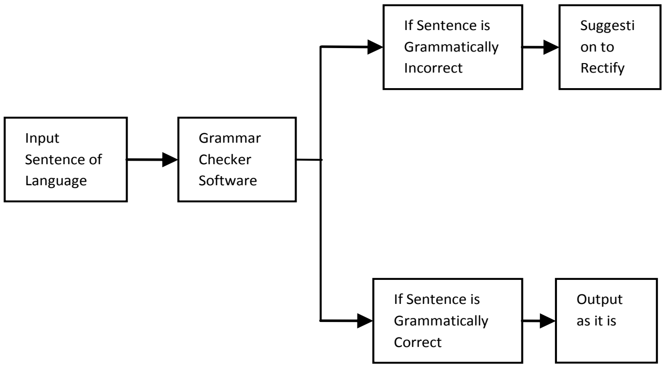
Input Sentence of Language
Grammar Checker Software
If Sentence is Grammatically Incorrect
Suggestion to Rectify
If Sentence is Grammatically Correct
Output as it is

Figure 1: General Architecture of Grammar Checker

II. INTRODUCTION TO HINDI LANGUAGE

Hindi language belongs to Indo-Aryan family of languages.Other members of this family areAsamiya (Assamese, about 13,175,000 speakers), Bangla (Bengali,83,875,000), Gujarati(46,100,000), Kashmiri (5,52 5,000),Konkani(2,500,000), Marathi(71,950,000), Nepali (2, 875,000), Oriya (33,025,000), Punjabi (29,100,000), Sindhi (2,550,000),and Urdu (51,550,000)[39]. Hindi is the national language of India and is widely written, spoken and understand by all the north Indian states of India. Today, the Hindi language has over 480 million speakers worldwide and is understood or spoken in Nepal, Pakistan, Bangladesh, and even Fiji. It is fifth most spoken language in the world. It is written in Devanagari script.Hindi is a descendant of Sanskrit and has been influenced by the Dravidian language, Arabic, Portuguese, English, Persian, and Turkic. The dialects of Hindi include Awadhi, Braj and Khari Boli. Hindi words are divided into five categories, namely Tatsama, Ardhatatsama, Tadbhava, Deshaj, and Videshi. [35]

III. EXISTING GRAMMAR CHECKING SYSTEMS AND TECHNIQUES USED

There are basically three techniques used for developing grammar checker. These includes rule based approach, statistics based approach and third is syntax based approach. Although combination of any two has also been tried by some of the authors. Some of the grammar checker tools developed for Indian languages includes grammar checker for Bangla [2,32], Urdu [11] and Punjabi[10]. Rule based approach has been used for the development of all these three types of grammar checkers. Besides these three Indian grammar checkers, many other grammar checkers have been developedby using these three techniques. Syntax based technique has been used for Korean language [Young-Soog 1998], Danish language [Bick, E. (2006).][37], French language [Vandeventer 2001][36] and for Urdu language [Kabir et. Al. 2002][11]. Further statistics based approach has been used for English language [Park et al. 1997][39], French language [Tschichold et al., 1997], English language [Powers 1997], Brazilian Portuguese language [Martins et al. 1998], Swedish language [Arppe 1999], Bangla and English language [Alam et al. 2006], Swedish language [Sjöbergh 2006], Persian language [Ehsan and Faili 2010] and Amharic language [Temesgen and Assabie 2012]. Further a Language Independent Statistical Grammar (LISG) checking system was developed by [VerenaHenrich and Timo Reuter 2009]. Further rule based approach has been used for Dutch language [Vosse 1992], Czech and Bulgarian language [Kuboň and Plátek 1994], Swedish language [Hein 1998], French, German, and Spanish languages [Helfrich and Music 2000], Swedish language [Carlberger et al. 2002, 2004], English language [Naber 2003], Brazilian Portuguese language [Kinoshita et al. 2006], Nepali language [Bal and Shrestha 2007][3], Persian language [Ehsan and Faili 2010], Chinese language [Jiang et al. 2011] and for Malay language [Kasbon et al. 2011].

IV. PROPOSED METHODOLOGY

Hindi sentence written in Unicode will be given as input to the system. The developed system checks one sentence at a time. The sentence must be terminated by sentence ender. The developed system is mainly developed for simple sentences of Hindi language, although some compound sentences can also be processed for detection and correction of errors. Proposed architecture is shown in fig. 2.

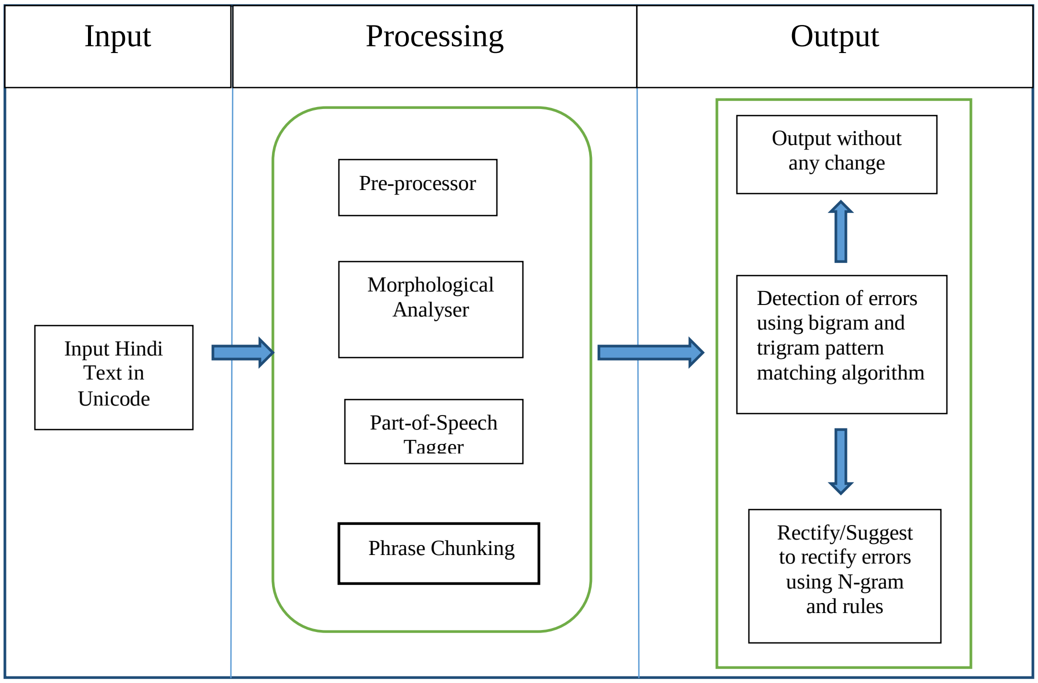
Input
Processing
Output
Input Hindi Text in Unicode
Pre-processor
Morphological Analyser
Part-of-Speech Tagger
Phrase Chunking
Output without any change
Detection of errors using bigram and trigram pattern matching algorithm
Rectify/Suggest to rectify errors using N-gram and rules

Figure 2. Architecture for Proposed system

A. Pre-processing

In pre-processing, four basic operations are performed. These includes spell checking, identification of punctuation mark, identification of sentence ender (if not present) and tokenization by splitting all the word into individual tokens. The tokenization is performed by considering the space as delimiter. The spell checking is performed by scanning each word and searching it with the Hindi lexicon. if the word is not present in the lexicon then most nearest word i.e. word matching maximum characters is suggested for the replacement. For identification of punctuation mark and sentence ender, pre-developed databases containing all punctuation marks and all possible sentence enders are used.

B. Morphological Analysis

Morphological analysis is performed by using Morphological analyzer. Morphological analyzer will provide the grammatical information to each word of the sentence in the form of tags called part of speech tags. For this research a tagset having more than 630 tags has been proposed. This is the modified version of tagset used by [Singh and Lehal] for development of Punjabi Grammar checker.To develop the morphological analyzer, author used parallel corpus of Hindi and Punjabi language. Since the morphological analyzer of Punjabi has already been developed with accuracy of 87.64%[10], so author took the advantage of this and developed Hindi morph using Hindi Punjabi parallel corpus. This parallel corpus was taken from Technical Development of Indian languages (TDIL) web site.

C. Part of speech tagger

The morph used in the previous step assigns multiple tags to those words which are used in more than one word class in different context. In order to assign the appropriate tag to these words, POS tagger is used. Author developed the part of speech tagger using Hidden Markov Model Based technique. For training the model author collected the Hindi corpus from the Indian language Corpora Initiative (ILCI) and tagged the corpus using morphological analyzer developed in the previous step. This annotated corpus is then filtered to separate those sentences which do not have any ambiguous tag.From this filtered corpus various HMM parameters i.e. initial probability, transition probability and emmition probabilities were calculated. Maximum likelihood approach was used to calculate these parameters.

D. Detection of errors

For detection of errors in an input sentence, pattern bases bigram and trigrams approaches are used. These bigrams and trigrams probabilities are calculated from tagged Hindi corpus. After applying these pre-calculated bigram and trigram probabilities, the sum of product of all the probabilities of input sentence is calculated. Now if the sum of product of all the bigrams/trigrams of a sentence is positive then the sentence is grammatically correct otherwise it is incorrect.

E. Correction/Possible suggestion to rectify errors

For the correction of sentence, hybrid approach is used. In this approach, first statistical technique is applied to rectify those errors that are within phrase. After resolving the errors within phrases, rules are used to rectify the error related with different phrases. Hand written Hindi grammar rules are used. These rules are developed by linguistic and are implemented in the form of regular expressions.

Table 1: Various Possible Errors in Hindi Language

<table>
<tr><th></th><th colspan="2">Example</th><th></th></tr>
<tr><td>Error type</td><td>Incorrect</td><td>Correct</td><td>English Translation</td></tr>
<tr><td>लिंग (Ling) – Gender</td><td>मैंने यह पुस्तक देखा हूँ। (मैंने_PNPBSDF|PNPBSOF यह_PNDBSD|PNDBSO|PNDBPD पुस्तक_NNFSD|NNFSO देखा_VBMAMSXXPTNIA हूँ_NNFSD|NNFSO|IJ ।_Sentence)</td><td>मैंनेयहपुस्तकदेखीहैं। (मैंने_PNPBSDF|PNPBSOF यह_PNDBSD|PNDBSO|PNDBPD पुस्तक_NNFSD|NNFSO देखी_IJ हैं_VBAXBPT1 ।_Sentence)</td><td>I have seen this book.</td></tr>
<tr><td>वचन (Vachan) – Number</td><td>प्राणनिकल गया।(प्राण_NNMSD|NNMSO|NNMPD निकल_NNMSD|NNMSO गया_VBMAMSXXPINIA ।_Sentence)</td><td>प्राणनिकलगए।(प्राण_NNMSD|NNMSO|NNMPD निकल_NNMSD|NNMSO गए_VBOPMPXXPINIA।_Sentence)</td><td>Life is gone.</td></tr>
<tr><td>कारक (Karak) –</td><td>मैंने घरजानाहैं। (मैंने_PNPBSDF|PNPBSOF</td><td>मुझेघरजानाहैं।(मुझे_PNPBSTF</td><td>I have to go</td></tr>
</table>

Case	घर_NNMSD|NNMSO|NNMPD जाना_VBMAMSXXXINNA हैं_VBAXBPT1 ।_Sentence)	घर_NNMSD|NNMSO|NNMPD जाना_VBMAMSXXXINNA हैं_VBAXBPT1 ।_Sentence)	home
विषेय-विशेषण (VishyVishesha d) – Attributive Adjectives	हवागरम चलरहीहै।(हवा_NNFSD|NNFSO गरम_AJU चल_VBMAXSS3XINO रही_VBMAFSXXPINIA है_VBAXBST1 ।_Sentence)	गरमहवाचलरहीहै।(गरम_AJU हवा_NNFSD|NNFSO चल_VBMAXSS3XINO रही_VBMAFSXXPINIA है_VBAXBST1 ।_Sentence)	Warm wind is blowing.
क्रिया (Kriya) – Verb	सीताखाना माँगता है।(सीता_NNFSD खाना _NNFSD माँगता_VBMAMSXXXTNDA है_VBAXBST1 ।_Sentence)	सीताखानामाँगतीहै।(सीता_NNFSD खाना_NNFSD माँगती_VBMAFSXXXTNDA है_VBAXBST1 ।_Sentence)	Sita wants food.

V. TYPES OF ERROR COVERED

This grammar checker cover the following types of errors:

- Error due to mismatch between adjective and noun in terms of number (Adj-NN-N).
- Error due to mismatch between adjective and noun in terms of gender (Adj-NN-G).
- Error due to mismatch between Noun and Verb in term of number(NN-VB-N).
- Error due to mismatch between Noun and Verb in term of gender(NN-VB-G).

Table 2: output of the test data

Error Type	Total Number of sentences in the test data	Number of sentences correctly handled by the system	Number of sentences in-correctly handled by the system	Number of sentence s not handled by the system
Adj-NN-N	150	118	12	20
Adj-NN-G	140	116	8	16
NN-VB-N	160	138	16	6
NN-VB-G	150	133	7	10

Table 3: Precision and Recall

Error type	Precision	Recall	F-Measure
Error due to mismatch between adjective and noun in terms of number.	0.78	0.90	0.83
Error due to mismatch between adjective and noun in terms of gender.	0.82	0.93	0.87
Error due to mismatch between Noun and Verb in term of number.	0.86	0.89	0.87
Error due to mismatch between Noun and Verb in term of gender.	0.88	0.95	0.91
Overall	0.83	0.91	0.87

6. Error due to mismatch between Adverb and Verb in term of gender.

VI. RESULT AND DISCUSSION

The developed system is tested using a test data. This test data was designed in such a way that it contains sentences having errors for which this grammar checker is developed.Table2 shows the observations after testing the grammar checker on test data:

From table 2 it is clear that there are some sentences which are not grammatically checked by the system. This is because the developed system did not find any error in the sentence. Some false alarm are also observed. False alarm means the developed system incorrectly mark the correct sentence as incorrect and vice versa. The main reason behind these false alarms is presence of unknown words i.e. the words which are not present in the dictionary.

Table 3 shows the precision and recall of the developed system for various types of errors. The developed system shows an overall precision of 0.83, Recall as 0.91 and F-measure as 0.87.

VII. CONCLUSION AND FUTURE SCOPE

In this research paper author proposed a hybrid approach for developing grammar checker for detection and correction of grammatical mistakes in Hindi language.This is the basic prototype for Hindi Grammar checker. The developed system identify and rectify the grammatical errors related to mismatch of agreement in noun and verb in terms of number and gender. The developed system works only for simple sentences of Hindi language. Further this system checks the errors related to mismatch of agreement between noun and verb, noun and adjective and thus this research work can be further extended to large sentences i.e. compound and complex sentences of Hindi language. Further algorithms for detection and correction of various other types of errors like style error, error due to postposition, use of unnecessary word etc. can also be implemented.

REFERENCES

[1]. Mallikarjun, B, Yoonus, M. Sinha, Samar & A. Vadivel,"Indian Languages and Part-ofSpeech Annotation. Mysore", Linguistic Data Consortium for Indian Language, pp. 22-25. ISBN-81-7342-197-8, 2010.

[2]. Alam, M. J., Uzzaman, N., & Khan, M. (2006). N-gram based Statistical Grammar Checker for Bangla and English. Ninth International Conference on Computer and Information Technology (ICCIT 2006), 3–6.

[3]. B. K. Bal, B. Pandey, L. Khatiwada, P. Rupakheti, and M. P. Pustakalaya, "Nepali grammar checker," PAN/L10n/PhaseII/Reports by Madan PuraskarPustakalaya, Lalitpur, Nepal, pp. 1–5, 2008.

[4]. LataBopche, GauriDhopavkar, and ManaliKshirsagar, "Grammar Checking System Using Rule Based Morphological Process for an Indian Language", Global Trends in Information Systems and Software Applications, 4th International Conference, ObCom 2011 Vellore, TN, India, December 9-11, 2011.

[5]. F. R. Bustamante, "GramCheck: a grammar and style checker," COLING "96 Proc. 16th Conf. Comput. Linguist., vol. 1, pp. 175–181, 1996.

[6]. D. Tesfaye, "A rule-based Afan Oromo Grammar Checker," IJACSA - Int. J. Adv. Comput. Sci. Appl., vol. 2, no. 8, pp. 126–130, 2011.

[7]. J. Carlberger, R. Domeij, V. Kann, and O. Knutsson, "A Swedish grammar checker," Citeseer, 2000.

[8]. N. Ehsan and H. Faili, "Statistical Machine Translation as a Grammar Checker for Persian Language," Sixth Int. Multi-Conference Comput. Glob. Inf. Technol., no. c, pp. 20–26, 2011.

[9]. Y. Jiang, T. Wang, T. Lin, F. Wang, W. Cheng, X. Liu, C. Wang, and W. Zhang, "A rule based Chinese spelling and grammar detection system utility," Syst. Sci. Eng. (ICSSE), 2012 Int. Conf., no. 1, pp. 437–440, 2012.

[10]. M. Gill, G. Lehal, and S. Joshi, "A Punjabi Grammar Checker," Acl.Eldoc.Ub.Rug.Nl, pp. 940–944.

[11]. H.Kabir,S. Nayyer,J. Zaman and S.Hussain. Two pass parsing implementation for an Urdu grammar checker. In Proceedings of IEEE international multi topic conference,pp. 1_8,Dec 2002.

[12]. J. Kaur, "Hybrid Approach for Spell Checker and Grammar Checker for Punjabi," vol. 4, no. 6, pp. 62–67, 2014.

[13]. N. Ehsan and H. Faili, "Towards grammar checker development for Persian language," Proc. 6th Int. Conf. Nat. Lang. Process. Knowl. Eng. NLP-KE 2010, 2010.

[14]. K. F. Shaalan, "Arabic GramCheck: A grammar checker for Arabic," Softw. - Pract. Exp., vol. 35, no. 7, pp. 643– 665, 2005.

[15]. J. Kinoshita, C. Eduardo, and D. De Menezes, "CoGrOO: a Brazilian-Portuguese Grammar Checker based on the CETENFOLHA Corpus," October, pp. 2190–2193, 2003.

[16]. D. Deksne and R. Skadiņš□, "CFG Based Grammar Checker for Latvian," Proc. 18th Nord. Conf. Comput. Linguist. NODALIDA 2011, p. 275 278, 2011.

[17]. A.Schmaltz,Y. Kim,A.M. Rush and S.M.Shieber. Adapting sequence models for sentence correction. arXiv preprint arXiv:1707.09067,2017.

[18]. S.K.Sharma and G.S.Lehal. Improving Existing Punjabi Grammar Checker. In Computational Techniques in Information and Communication Technologies (ICCTICT), 2016 International Conference on,pp. 445_449, IEEE,march 2016.

[19]. A.Schmaltz, Y.Kim,A.M. Rush and S.M.Shieber. Sentence-level grammatical error identification as sequence-to-sequence correction. arXiv preprint arXiv:1604.04677,2016.

[20]. C.J.Lin and S.H.Chen. NTOU Chinese Grammar Checker for CGED Shared Task. In Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications,pp. 15_19,july 2015.

[21]. T.Boroş,S.D. Dumitrescu,A. Zafiu,D. Tufiş,V.M. Barbu and P.I.Văduva. RACAI GEC–A hybrid approach to Grammatical Error Correction. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, pp. 43_48. 2014.

[22]. A.Temesgen and Y.Assabie. Development of Amharic Grammar Checker Using Morphological Features of Words and N-Gram Based Probabilistic Methods. IWPT-2013,pp. 106,2013.

[23]. J.W.Xing,L.Y. Wang,F. Wong,S. Chao and X.D.Zeng. UM-Checker: A hybrid system for English grammatical error correction. In The seventeenth conference on computational natural language learning: shared task,pp. 34_42,Aug 2013.

[24]. R.Nazar and I.Renau. Google books n-gram corpus used as a grammar checker. In Proceedings of the Second Workshop on Computational Linguistics and Writing (CLW 2012): Linguistic and Cognitive Aspects of Document Creation and Document Engineering,pp.27_34,Association for Computational Linguistics,April 2012.

[25]. S.K.Sharma and G.S.Lehal. Using Hidden Markov Model to improve the accuracy of Punjabi pos tagger. In Computer Science and Automation Engineering (CASE), 2011 IEEE International Conference on,Vol. 2, pp. 697_701,IEEE,june 2011.

[26]. Y.Jiang,T. Wang,T. Lin,F. Wang,W. Cheng,X. Liu and W.Zhang. A rule based Chinese spelling and grammar detection system utility. In System Science and Engineering (ICSSE), 2012 International Conference on,pp. 437_440,IEEE,june 2012.

[27]. R.Kasbon,N.A. Amran,E.M. Mazlan and S.Mahamad. Malay language sentence checker,World Appl. Sci. J.(Special Issue on Computer Applications and Knowledge Management),Vol. 12, pp.19_25,2011.

[28]. D.Deksne and R.Skadiņš. CFG Based Grammar Checker for Latvian. NODALIDA 2011 Conference Proceedings, pp. 275_278,2011.

[29]. V.Henrich and T.Reute. LISGrammarChecker: Language Independent Statistical Grammar Checking. Hochschule Darmstadt & Reykjavík University,2009.

[30]. M.S.Gill and G.S.Lehal. A grammar checking system for Punjabi. In 22nd International Conference on Computational Linguistics: Demonstration Papers ,pp. 149_152,Association for Computational Linguistics,Aug 2008.

[31]. A.Kumar and S.Nair. An artificial immune system based approach for English grammar checking. Artificial immune systems,pp. 348_357,2007.

[32]. M.J.Alam, N.UzZaman, and M. Khan. N-gram based Statistical Grammar Checker for Bangla and English. In Proc. of ninth International Conference on Computer and Information Technology (ICCIT 2006),2006.

[33]. J.Sjöbergh and O.Knutsson. Faking errors to avoid making errors: Very weakly supervised learning for error detection in writing. In Proceedings of RANLP,pp. 506_512,Sep 2005.

[34]. A.Arppe. Developing a grammar checker for Swedish. In The 12th Nordic Conference of Computational Linguistics ,pp. 13_27,2000.

[35]. https://www.worldatlas.com/articles/the-most-widely-spoken-languages-in-india.html

[36]. Vandeventer, A. (2001). Creating a grammar checker for CALL by constraint relaxation: a feasibility study. ReCALL, 13(1), 110-120.

[37]. Bick, E. (2006). A constraint grammar based spellchecker for danish with a special focus on dyslexics. A Man of Measure: Festschrift in Honour of Fred Karlsson on his 60th Birthday. Special Supplement to SKY Jounal of Linguistics, 19, 387-396.

[38]. Park, J. C., Palmer, M. S., & Washburn, C. (1997, March). An English Grammar Checker as a Writing Aid for Students of English as a Second Language. In ANLP (p. 24).

[39]. https://www.britannica.com/topic/Indo-Aryan-languages#ref284555