ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 6, Issue 3, pp: (142-147), Month: July - September 2018, Available at: www.researchpublish.com

Grammar Checking Using Pattern Matching Approach

[1]Leekha Jindal, [2]Vijay Rana

[1]Research Scholar, Sant Baba Bhag Singh University, Jalandhar
[2]Assistant Professor, Department of Computer Science and Applications, Sant Baba Bhag Singh University, Jalandhar

Abstract: **In this research articles author proposed a pattern matching approach for grammar checking of Punjabi sentences. In the proposed approach author used the pattern of sentences whose length ranges from 4 to 6. An annotated corpus of around 10,000 sentences is used for generating the patterns. Different patterns are generated for sentences of different lengths.**

Keywords: **grammar checker, pattern matching approach.**

1. INTRODUCTION

Grammar checker also called syntactic analyzer is a software that verify the syntax of a specific language against the grammatical rules of that language. If the written text is according to the rule of the language then no error will be produced by the grammar checker otherwise if the written text is not according to the grammar of the language then it will generate an error message and will also display a list of suggestions to rectify these errors. Generally grammar checking systems are of not used independently rather these are part of some word processor like MS WORD. Many grammar checkers have been developed for English and other languages. Very little work has been done for Indian languages. Probably, Bangla grammar checker [19], Urdu [21] and Punjabi grammar checker [16] [26][27][28] are the only systems developed for Indian languages.

2. EXISTING WORK

Various techniques have been used by various authors for different languages [22][23][24][25]. Some of these includes Schmaltz, A. et.al (2017) [1] proposed seq2seq approach for error detection and correction in sentence that achieved a $F_{0.5}$ score as 50.12 for deletion method, 42.51 for insertion method and 50.39 for replacement method as compare to SMT approach that has $F_{0.5}$ score 46.56 for deletion, 31.48 for insertion and 42.21 for replacement methods when tested on Automated Evaluation of Scientific Writing dataset. Sharma, S.K. et.al (2016) [2] put their efforts to improve the existing rule based Punjabi grammar checker and observed 5-6% improvement in morph and 8-9% improvement in POS tagger. Schmaltz, A. et.al (2016) [3] proposed an attention based encoder-decoder model for checking grammar errors present in a sentence and showed precision as 0.5444, recall as 0.7413 and F-score as 0.6278. Lin, C. J. et.al (2015) [4] proposed a system that checked the grammatical accuracy of Chinese sentences generated by deleting, inserting or exchanging characters or words and achieved a precision of 23.4% and a recall of 36.4% in the identification level. Boroş, T et.al (2014) [5] described the development of RACAI's (Research Institute for Artificial Intelligence) hybrid grammatical error detection and correction system. Temesgen, A. et.al (2013) [6] described the statistical grammar checker for Amharic language. This system showed 59.72% precision and 82.69% recall for simple sentences using bi-gram and if trigram is used then system showed 67.14% precision and 90.38% recall. Further for complex sentences system shows 57.82% precision and 65.38% recall for bigram and 63.76% precision and 67.69% recall for trigram. Xing, J. W. et.al (2013)[7] described the hybrid approach based NLP-CT Grammatical Error Detection and Correction system for the CoNLL 2013 shared task. Nazar, R. et.al (2012)[8] explored the possibility of using a large n-gram corpus (Google Books) to derive lexical transition probabilities from the frequency of word n-grams and then use them to check and suggest corrections in a target text without the need for grammar rules and obtained a precision of 64.58, Recall 47.69 and F measure 54.86. Sharma, S.K. et.al (2011) [9] used Hidden Markov Model (HMM) to improve the accuracy of existing rule based Punjabi POS tagger. Jiang, Y. et al. (2011) [10] proposed a rule based grammar checker. Tesfaye, D. (2011) [11] developed a rule based grammar checker for Afan Oromo (language widely spoken and used in Ethiopia). Kasbon, R. (2011) [12] proposed a rule based grammar checking system for Malay language. Deksne, D. et.al (2011) [13] explained the implementation of the Latvian grammar checker based on a parser. Henrich, V et.al (2009) [14] proposed a Language Independent Statistical Grammar (LISG) checking system. They developed this system by using N-gram statistical technique. Singh M et.al (2008)[15] developed a rule based grammar checker for Punjabi that reported precision of 76.79%, recall of 87.08%, and F-measure of 81.61%. Kumar, A. et.al (2007) [16] provided a corpus based approach for grammar checking that is based on the principles of an Artificial Immune System (AIS). Bal, B. K. et.al (2007) [17] proposed an architecture of rule based grammar checking system for Nepali language. Alam, M. Jahangir et.al (2006)[18] used Part of speech (POS) tags and n-gram technique to check the grammatical correctness of a sentence. The system showed an accuracy of 63% for English and 53.70% for Bangla when tested on manually tagged correct sentences. Sjöbergh, J. et.al (2005) [19] describes a method to create an automated grammar checker that did not require any manual work. Kabir, H. et.al (2002) [20] proposed a computational model for Urdu Grammar Checker. Two pass parsing approach was used for analysis of sentence in which first, some base Phrase Structure Grammar (PSG) Rules are used to parse the sentence and in case of failure, Movement Rules are applied and sentence is reparsed. Arppe, A. et.al (1999) [21] developed a commercial grammar checking system for Swedish at Lingsoft Inc. This grammar checker is also a part of the Microsoft Office since 2000.

3. PROPOSED MODEL

The proposed architecture for classification of sentences on the basis of the length is shown in figure 1.

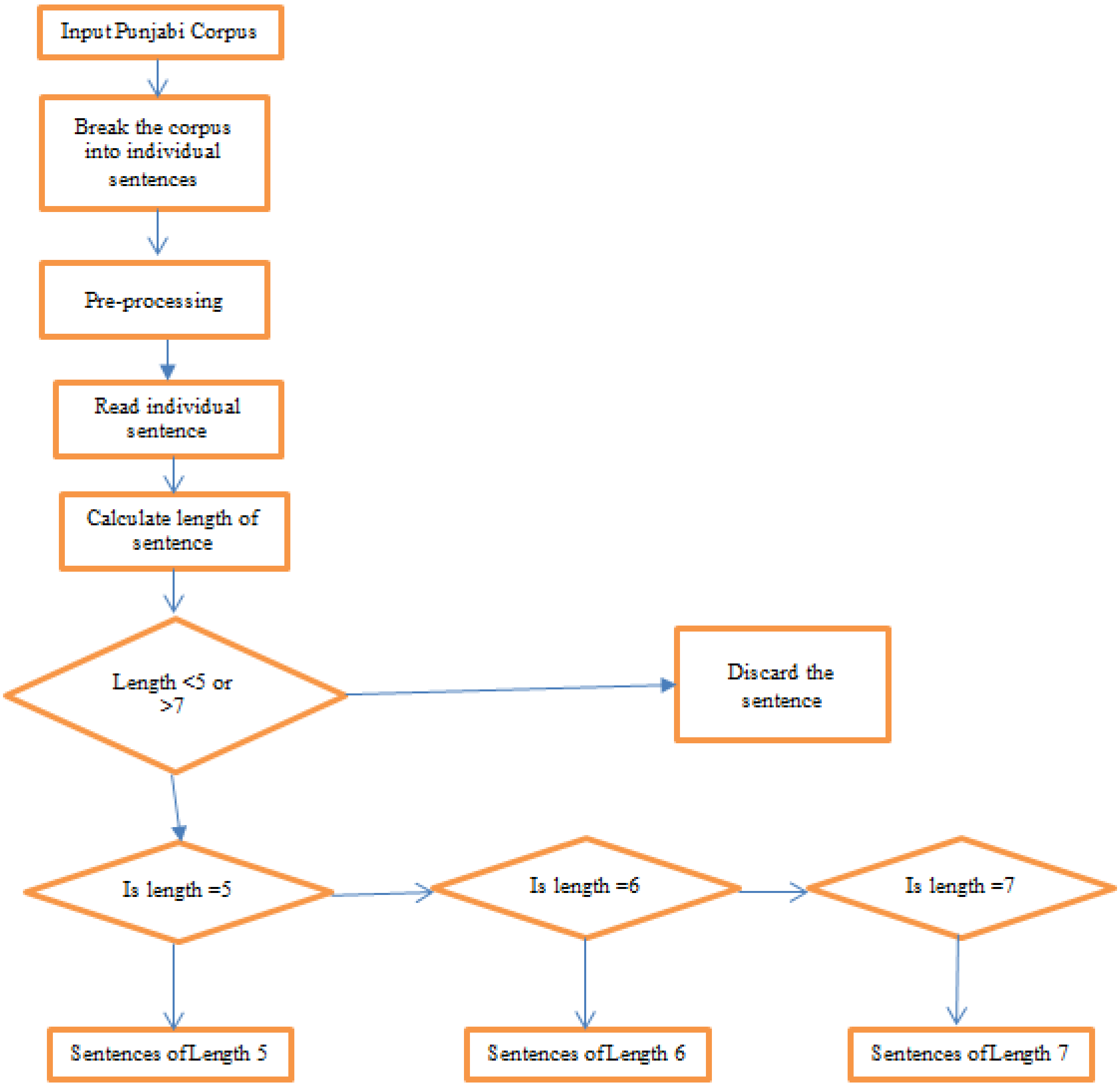

Figure 1: Proposed architecture for classification of sentences

As shown in figure1, after inputting the corpus, the corpus is broken into individual sentences. Each individual sentence undergo pre-processing in which spell checking is performed so as to make the corpus free from spelling errors. Other than spell checking, duplicate entries of sentences as well as incomplete sentences are removed from the corpus. After pre-processing, the corpus is classified into various classes according to size of sentence. For classifying the sentences, each individual sentence is inspected for its length and separate class is formed for each specific length. In this research author experimented with three lengths i.e. from length 5 to length 7. Sentences having length less than 5 and greater than 7 are discarded.

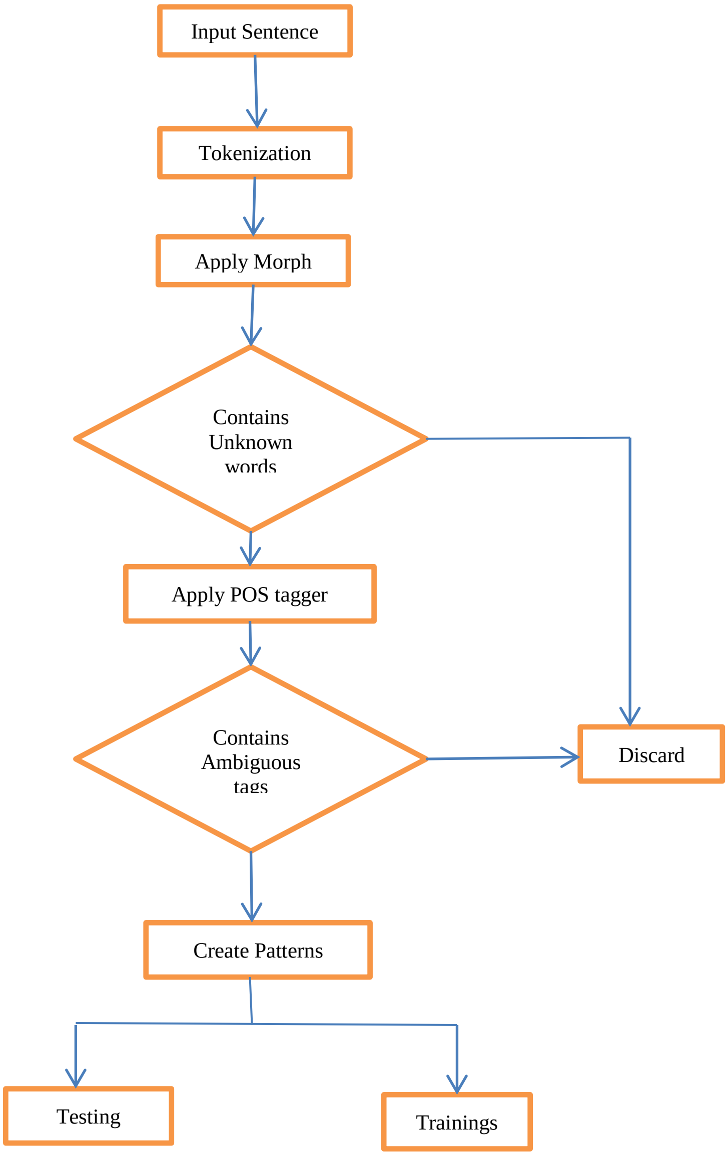

Figure 2: Proposed algorithm for pattern generation

Now as shown in figure2, after classification, each word in every sentence present in each class is assigned its grammatical information in the form of part of speech (POS) tags. Morphological analyzer is used for this task. After applying morphological analyzer (MA), sentences containing unknown tags i.e. those sentences in which one or more words are assigned an unknown tag are removed from the list. After that HMM based POS tagger is applied to keep only appropriate tag with each word out of multiple tags assigned by MA. Now we left with sentences having word tag combinations. After this only tag pattern is extracted from word tag pairs. These tag patterns are divided in to two parts one part is used for training and second part is used for testing.

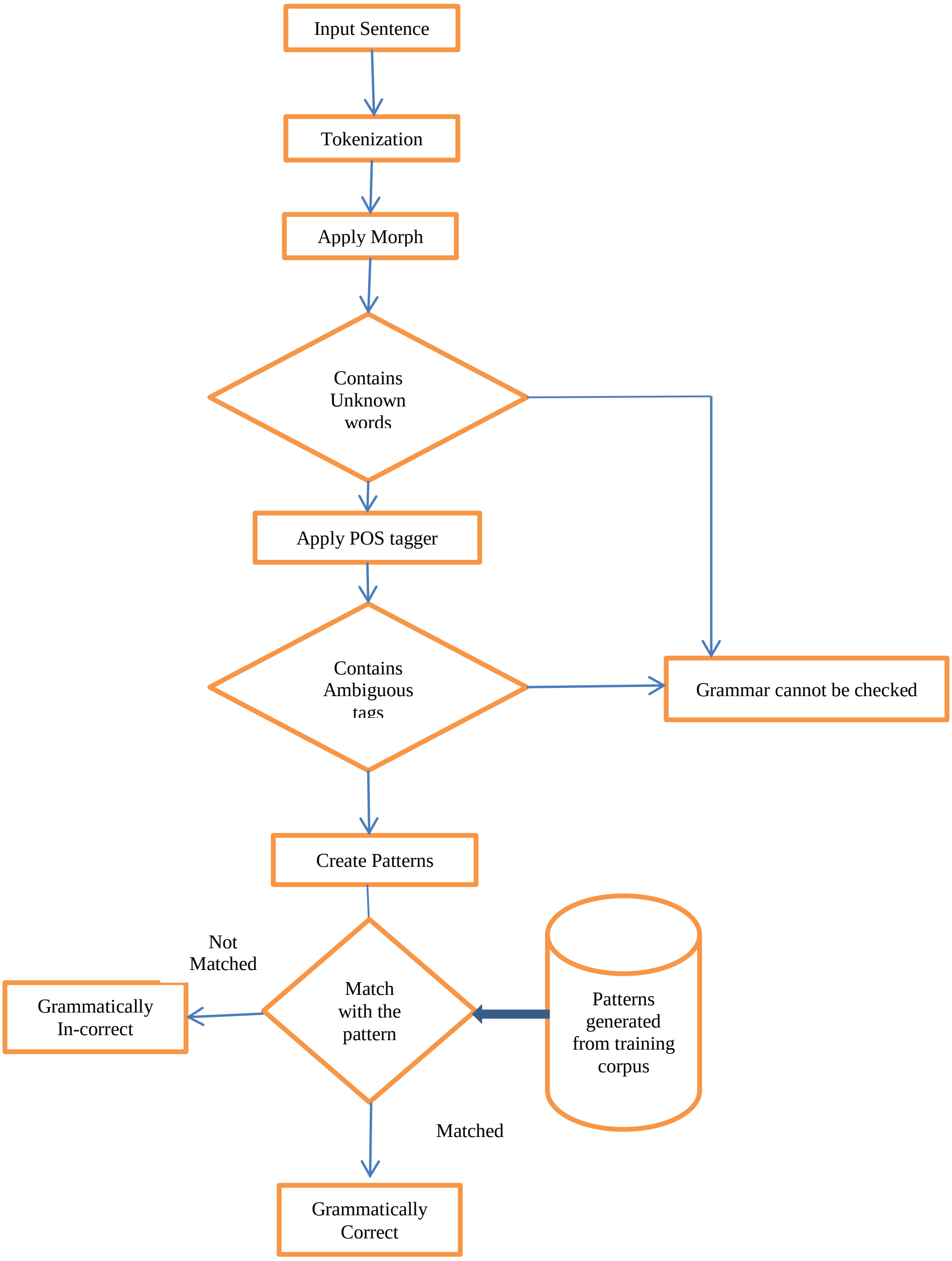

Figure 3: Grammar checking using pattern matching

The grammar checking process is shown in figure 3. Input sentence is first passed through morph and if any of the word in the sentence is marked unknown then the grammar of that sentence cannot be checked. If all the words of the sentence are tagged then POS tagger is applied to remove the ambiguity of the tags if any present. If even after applying the POS tagger some word still contains ambiguous tags then also the grammar of this sentence cannot be checked. And if the sentence does not contain unknown word as well as ambiguous tag then pattern of the tags is extracted from the sentence and this pattern is searched in the existing database of patterns generated during training. If the database contains the pattern then the sentence is grammatically correct otherwise sentence is not grammatically correct.

4. RESULTS AND DISCUSSION

Total Number of sentences in corpus	Number of sentences having length 5	Number of sentences having length 6	Number of sentences having length 7	Number of sentences having length more than 7 (Discarded)
10220	2152	4317	2511	1240

Number of sentences having length 5		Number of sentences having length 6		Number of sentences having length 7	
2152		4317		2511	
Used for Training	Used for testing	Used for Training	Used for testing	Used for Training	Used for testing
1152	1000	3317	1000	1511	1000

5. CONCLUSION AND FUTURE SCOPE

This grammar checker can be helpful in development of other natural language processing systems like question answering, dialogue generation, paraphrasing and machine translation systems etc. Further the statistical techniques used proposed in this research article can also be implemented in other Indian languages having similar features as that of Punjabi. Also after implementing the statistical approach, hybrid approach can also be used for further enhancement of the system.

REFERENCES

[1] Schmaltz, A., Kim, Y., Rush, A. M., & Shieber, S. M. (2017). Adapting sequence models for sentence correction. arXiv preprint arXiv:1707.09067.

[2] Sharma, S. K., & Lehal, G. S. (2016, March). Improving Existing Punjabi Grammar Checker. In Computational Techniques in Information and Communication Technologies (ICCTICT), 2016 International Conference on (pp. 445-449). IEEE.

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

[4] Lin, C. J., & Chen, S. H. (2015, July). 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).

[5] Boroş, T., Dumitrescu, S. D., Zafiu, A., Tufiş, D., Barbu, V. M., & Văduva, P. I. (2014). RACAI GEC–A hybrid approach to Grammatical Error Correction. CoNLL-2014, 43.

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

[7] Xing, J. W., Wang, L. Y., Wong, F., Chao, S., & Zeng, X. D. (2013, August). UM-Checker: A hybrid system for English grammatical error correction. In The seventeenth conference on computational natural language learning: shared task, Aug. 2013, p. 34-42.

[8] Nazar, R., & Renau, I. (2012, April). 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.

[9] Sharma, S. K., & Lehal, G. S. (2011, June). Using hidden markov model to improve the accuracy of punjabi pos tagger. In Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on (Vol. 2, pp. 697-701). IEEE.

[10] Jiang, Y., Wang, T., Lin, T., Wang, F., Cheng, W., Liu, X., & Zhang, W. (2012, June). A rule based Chinese spelling and grammar detection system utility. In System Science and Engineering (ICSSE), 2012 International Conference on (pp. 437-440). IEEE.

[11] Tesfaye, D. (2011). A rule-based Afan Oromo Grammar Checker. IJACSA Editorial.

[12] Kasbon, R., Amran, N. A., Mazlan, E. M., & Mahamad, S. (2011). Malay language sentence checker.

[13] Deksne, D., & Skadiņš, R. (2011). CFG Based Grammar Checker for Latvian. NODALIDA 2011 Conference Proceedings, pp. 275–278.

[14] Henrich, V., & Reuter, T. (2009). LISGrammarChecker: Language Independent Statistical Grammar Checking. Hochschule Darmstadt & Reykjavík University.

[15] Gill, M. S., & Lehal, G. S. (2008, August). A grammar checking system for Punjabi. In 22nd International Conference on Computational Linguistics: Demonstration Papers (pp. 149-152). Association for Computational Linguistics.

[16] Kumar, A., & Nair, S. (2007). An artificial immune system based approach for English grammar checking. Artificial immune systems, 348-357.

[17] Bal, B. K., Shrestha, P., Pustakalaya, M. P., & PatanDhoka, N. (2007). Architectural and System Design of the Nepali Grammar Checker. PAN Localization Working Paper.

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

[19] Sjöbergh, J., & Knutsson, O. (2005, September). Faking errors to avoid making errors: Very weakly supervised learning for error detection in writing. In Proceedings of RANLP (pp. 506-512).

[20] Kabir, H., Nayyer, S., Zaman, J., & Hussain, S. (2002, December). Two pass parsing implementation for an Urdu grammar checker. In Proceedings of IEEE international multi topic conference (pp. 1-8).

[21] Arppe, A. 2000. Developing a grammar checker for Swedish. In The 12th Nordic Conference of Computational Linguistics pp. 13-27.

[22] Blossom Manchanda, Vijay Anant Athavale and S K Sharma, Various Techniques Used for Grammar Checking. International Journal of Computer Applications & Information Technology, 177-181, Special Issue on NLP, June 2016.

[23] Misha Mittal, Dinesh Kumar, S K Sharma, Grammar Checker for Asian Languages: A Survey. International Journal of Computer Applications & Information Technology, 163-167, Special Issue on NLP, June 2016

[24] Sharma, S. K. Role of Statistical Approaches for Error Detection in English and Other European Languages. AGU International Journal of Research in Social Sciences & Humanities. 234-239, Vol. 5, July Dec-2017.

[25] Sharma, S. K Rule Based Grammar Checking Systems (A Survey). International Journal of Computer Applications and Information Technology.217-220, 10(1) July 2016.

[26] Sanjeev Kumar Sharma, Effect of Statistical POS tagger on Syntactic Analysis of Punjabi Sentences. Indian Journal of Science and Technology, Vol9 (32), August 2016

[27] S.K. Sharma, G.S. Lehal, Improving Existing Punjabi Grammar Checker. IEEE International Conference on Computation Techniques in Information and Communication Technologies held at Indraprastha University, New Delhi, IEEE Xplore, 2016

[28] Sharma, S. K. Detection and Correction of Style Errors Present in Punjabi Sentences. International Journal of Computer Science and Engineering, 196-199, 5(8), August 2017