*Abstract No: 900* *IT, Mathematics and Statistics*

# HYBRID APPROACH FOR SPELL CHECKING OF TAMIL LANGUAGE

**S. Jananie* and K. Sarveswaran**

*Department of Computer Science, University of Jaffna, Sri Lanka*
**janani.segar@gmail.com*

The spell checkers are specialised application programs that flags words in a document that may be misspelled. Though there are several spell checkers available for languages like English, no fully functional application is available for the Tamil language. The existing systems either find the misspelled words from an existing list of words stored in those systems or *Canti* mistakes. Omission of a required letter or inclusion of an inappropriate letter between two adjoined words is called *Canti* mistake. Further, several issues have been also identified in these systems.

A new approach for Tamil spell checker has been proposed in this research by integrating existing approaches and new approaches such as rule-base, crowd sourcing and suggestions generation using character level *n-gram.* According to the proposed approach, each word is checked whether it exists in the dictionary using a *Levenshtein* distance finding algorithm. If it does not exist, then the *n-gram* based technique is used to generate possible suggestions for the given word. And required rules are written to get the appropriate suggestions by considering *Canti* check as well to identify the appropriate joining letter of two adjoined words. A list of 250,000 unique and error-free words are included in the dictionary. These words have been collected from various sources, including websites. It is very difficult to gather all the words in Tamil language. Therefore, add to dictionary option has been introduced to collect new words from users and add to the existing dictionary after the moderation.

To reduce the search space, the dictionary has been divided into different files based on the first letter of the word. Due to the complex nature of Tamil script compared to English, stacks and lists have been used during the processing of words. These rules have been written in such a way that it can be extended further in future. All these processing is being done without Romanising the Tamil text, while in most of the other approaches Tamil language is processed in Romanised form.

The proposed system gives better accuracy than the existing systems; 85.77% accuracy was noted when considering the suggestions generation. This result had been calculated by analysing the suggestions generated by the system for the words that are not in the dictionary. Hence the proposed approach, which has dictionary check with *Levenshtein* algorithm, suggestions generation with *n-grams*, *Canti* check with a rule-base and crowd sourcing, is a complete solution for Tamil spell checking.