



IJIRCCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 12, December 2022

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.165



9940 572 462



6381 907 438



ijircce@gmail.com



www.ijircce.com

Online Monolingual Thesaurus for Kokborok Language

Pankaj Debbarma¹, Pusparwng Hrangkhaw²

¹Assistant Professor, Department of Computer Science & Engineering, Tripura Institute of Technology, Narsingarh, India

²Assistant Professor, Department of Information Technology, Ambedkar College, Fatikroy, India

ABSTRACT: Kokborok (ISO 639-3:trp, Ethnologue) being one of the vulnerable languages in India needs special focus in preserving the language in any form possible. A thesaurus can be a significant medium in preserving the vocabulary and its morphology so that it can be used for future purpose in Natural Language Processing (NLP). The paper illustrates the design and development of an efficient system for a monolingual thesaurus in Kokborok language. This system takes as input a word in Kokborok language and gives synonyms and antonyms, also in Kokborok language, as its output. The thesaurus follows the basic grammatical rules of the Kokborok language.

KEYWORDS: Natural Language Processing, Monolingual Thesaurus, Kokborok Language, Synonyms, Antonyms.

I. INTRODUCTION

Kokborok (ISO 639-3:trp, Ethnologue) is spoken by around 1 million people in Tripura [1] and surrounding states of Assam and also in neighbouring regions in Bangladesh. It is one of the official languages of the north-east Indian state of Tripura. The United Nations has identified Kokborok as one of the vulnerable languages of India [2]. It is considered to be under the Tibeto-Burman (TB) language group [3] and is closely related to TB languages like Bodo and Garo [4]. Kokborok was the official language of the erstwhile kingdom of Tripura or Tipperah 1949 AD before being annexed to India [5]. Kokborok is being taught from primary up to the university level [6].

Kokborok has a high agglutinative nature morphologically. It shows a high level of inflection. Very few work has been done till date on Kokborok language in natural language processing (NLP). We can find some ambiguities in Kokborok words also which brings in challenge in the design or development of any NLP system for Kokborok language. Like other languages Kokborok also comes across the effects of semantic bleaching in its vocabulary.

Written form of Kokborok is accepted in Roman script as well as Bengali script as it does not have its own script [7]. The paper describes the designing and development of framework for a monolingual thesaurus for Kokborok language with Roman script. The proposed design will help in compiling a controlled multilingual vocabulary or thesaurus and can also function as the word repository for Kokborok language [8]. This may become the foundation for research work in Word Sense Disambiguation (WSD) of Kokborok language and addresses the gap of existing resources available online which are not able to translate directly from English into Kokborok or vice versa.

II. RELATED WORK

Thesauri are considered important for Information Retrieval (IR) systems as they are utilized for controlled usage of the vocabulary. A thesaurus provides with semantic relationships between words which help to expand or alter search queries (or query expansion) and help in retrieval of more relevant items.

Chansanam et al. [9] constructed a digital thesaurus platform to compile a controlled vocabulary for Thai and English which can enable semantic search. Bayekeyeva et al. [10] demonstrated the use of a controlled multilingual thesaurus capable of being implemented in a Digital Library (DL) that enables the user to search for a efficient translation of industry-specific terms. It also devises the rudiments for subsequent inception of Kazakh language corpus and its multilingual interoperability.

Kaur (2019) elaborates the development of thesaurus in relevance with the Hindi language [11] and framework which may assist the people finding it difficult to write and deal in Hindi. Apart from other conventional methodologies for thesaurus construction Mohsen et al. [12] opted for an automatic approach instead of manual. The paper presents various approach to achieve automatic building of a large-scale thesaurus through Term Frequency-Inverse Document Frequency (TF-IDF), Pointwise Mutual Information (PMI) and Latent Semantic Analysis (LSA).

III. PROPOSED ALGORITHM

The design and development of the Kokborok Monolingual Thesaurus (KMT) can be divided into three sections, namely,

- (a) Collection of Kokborok words with their synonyms and antonyms,
- (b) Creating and Populating the Database and
- (c) Designing of Graphical User Interface (GUI) for Connectivity with Thesaurus Database.

A. Collection of Kokborok Words:

There are only few dictionaries available for Kokborok language which has sufficient stock of words for implementation of this thesaurus [13]. However, an online platform of Kokborok dictionary is also available [14] with a word stock of more than 40,000 words from where vocabulary can be directly collected to populate the database of the proposed Thesaurus.

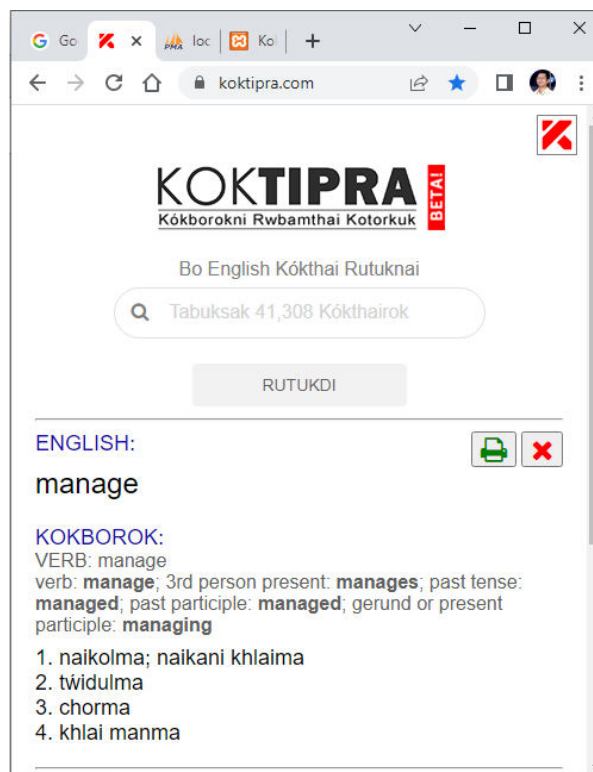


Figure 1: Homepage of Koptipra Kokborok Online Dictionary.

The related synonyms and antonyms have been identified and recorded by approaching various indigenous Kokborok speaking persons and through various textbooks in Kokborok language available for Schools. Roughly 3,000 words with their synonyms and antonyms have been collected manually which will eventually be used for populating the database of the Kokborok Monolingual Thesaurus.

B. Creating and Populating the Database:

The database and repository, namely “kbkthes”, for the Thesaurus has been created with PHP as the interface and PHPMyAdmin interface for MySQL at the backend. The KMT database has two tables called “thes” and “login”. The table “thes” is used as the repository of word with its synonyms and antonyms, whereas, the table “login” contains information of users permitted to do enter data in the Thesaurus word repository.

The table “thes” consists of parameter fields with the name “word”, “synonym” and “antonym” as shown in Figure 2. The primary parameter “word” contains the exact search query of Kokborok words to be done by users through the

search option available in the Homepage of KMT. The second parameter “synonym” contains all the relevant synonym(s) entered in the KMT repository. The third parameter “antonym” consists of appropriate antonym(s) for the word searched by the user.

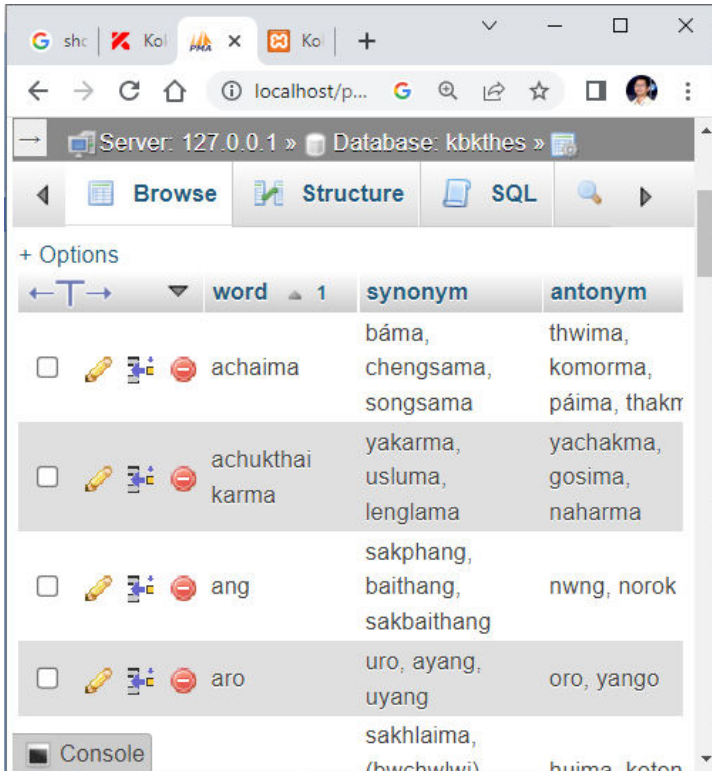


Figure 2: KMT Database with Word and its Synonyms and Antonyms.

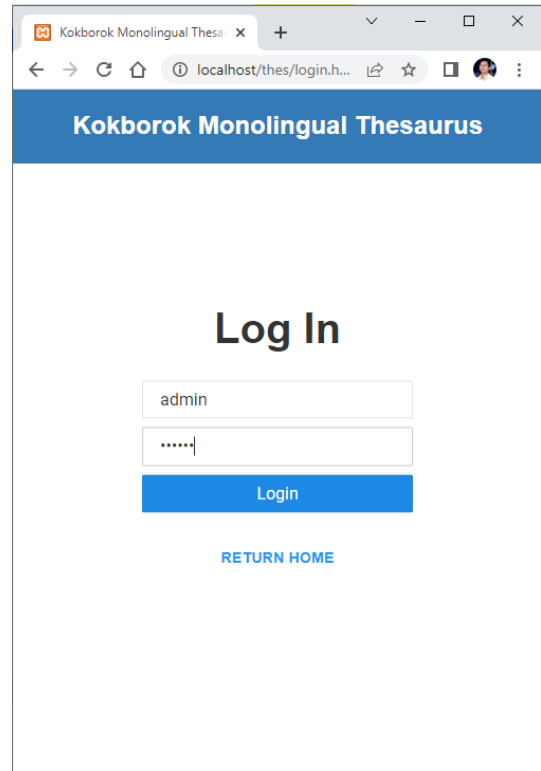


Figure 3: Data Entry Log In page for KMT.

The Thesaurus repository thus created can be populated directly through PHPMyAdmin application or through another web interface, namely, the “Log In” page as shown in Figure 3. The “Log In” page has been designed for users who are not familiar with handling MySQL database through the PHPMyAdmin application. Users will be given access to the page for data entry by entering a valid User ID and Password.

In order to store the valid User ID and Password for the “Log In” page the KMT database has another table called “login” with three parameters, namely, “id”, “username”, and “password”. The parameter “password” is a secret key. Any unauthorized users will not be able to access the next page. For the purpose of this paper only one user named “admin” with password “123456” has been entered in the “login” table.

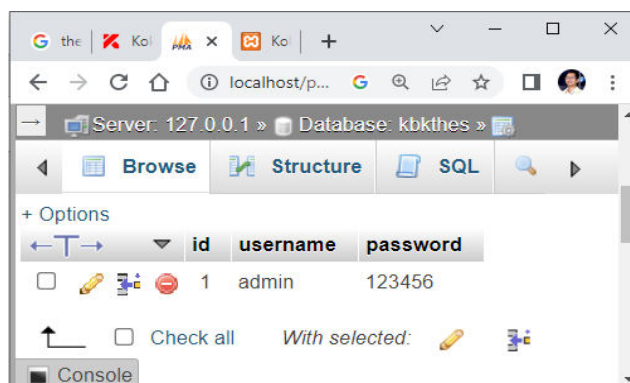


Figure 4: Table for storing User ID and Passwords.

When a user inputs the valid User ID and Password, another page called “Enter Data” is loaded in the browser which will enable the user to enter the word, its synonyms and antonyms in the KMT repository as shown in Figure 5. The “Enter Data” page is a user-friendly GUI which will help users in expanding the repository created for the monolingual thesaurus for Kokborok language.

The “Enter Data” page consists of three input boxes for a user to input a Kokborok word, its synonym(s) and antonym(s), respectively. The “Add Data” submit button in the page will immediately connect with the KMT “kbkthes” database and send the new input data to be updated in its “thes” table. The new data so input will now be available for any user accessing the KMT Homepage to search for that word to find its synonym(s) and antonym(s).

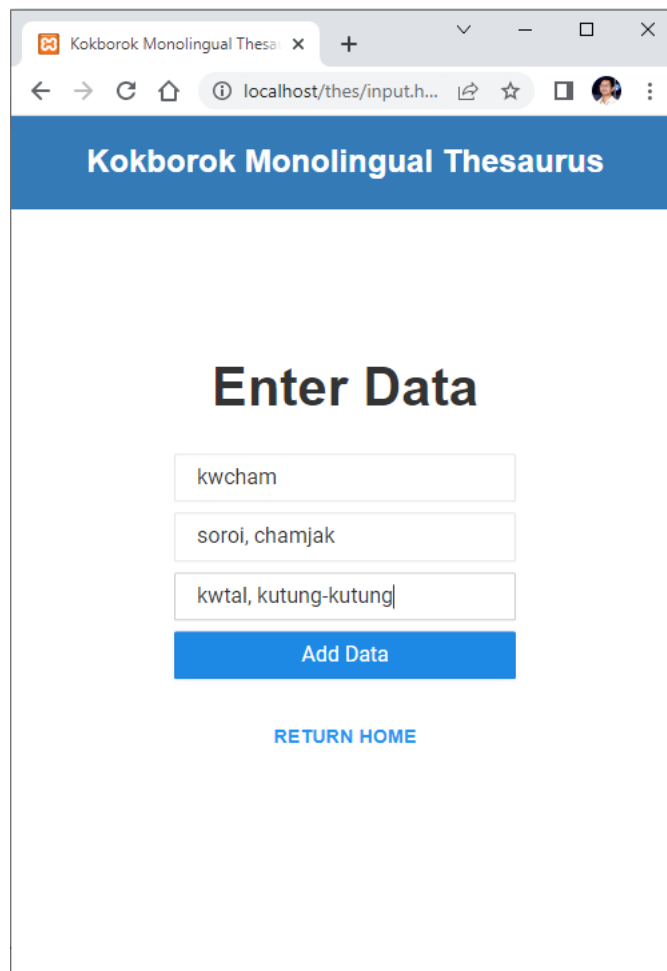


Figure 5: Data Entry page of the KMT repository.

C. GUI for Connectivity with Thesaurus Database:

The Homepage of KMT as shown in Figure 6 is a simplified and user-friendly search page provided for users to input any search query to access the desired information i.e. synonym(s) and antonym(s) of a word, from the Thesaurus repository so created.

When there is a match of user input word in search query and a term in the “word” parameter of the “thes” table, the page returns a response to the Homepage with query results in the form of contents in the “synonym” and “antonym” parameters of the “thes” table as shown in Figure 7, otherwise, it will give the message “No data found” in response.

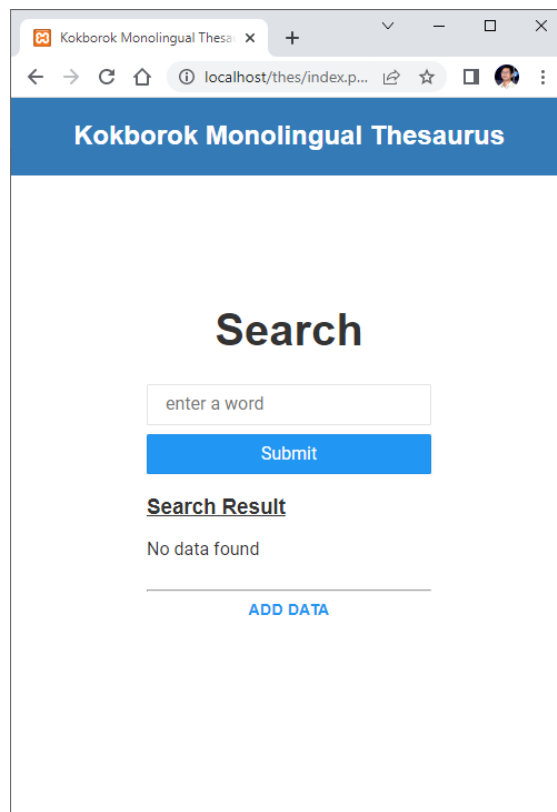


Figure 6: Homepage of KMT with options for searching word in the Thesaurus and also to add data in the KMT repository.

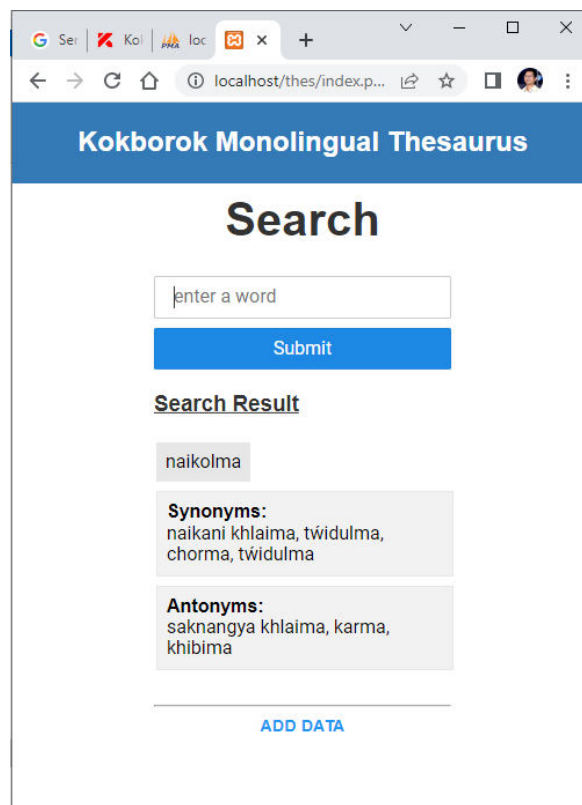


Figure 7: Result of search in the KMT Homepage when there is a match of query.

IV. CONCLUSION AND FUTURE WORK

The paper discussed the design and development of framework for a monolingual thesaurus for the Kokborok language. The work done in this paper fills the gap in the NLP field of Kokborok language efficiently and effectively. The repository so created contains more than 500 words.

The monolingual thesaurus can be extended further to create bilingual, trilingual or multilingual thesauri for Kokborok language which will not only help in enriching the Kokborok vocabulary but also in preserving the language along with its morphology. The work can also be used for integrating the thesaurus with other web or publishing applications for wider use.

REFERENCES

1. C-16: Population by mother tongue, Tripura – 2011, Population Census of Government of India, 2011.
2. Christopher Moseley, "Atlas of the World's Languages in Danger," in Memory of Peoples, 3rd edition, UNESCO Publishing, Paris, 2010, ISBN 978-92-3-104096-2.
3. Austin Hale, Research on Tibeto-Burman Languages, De Gruyter Mouton, Berlin, Boston, 2020.
4. François Jacquesson, "Discovering Boro-Garo: History of an analytical and descriptive linguistic Category," in European Bulletin of Himalayan Research, vol. 32, pp. 14-49, 2008.
5. Tijana Mamula and Lisa Patti, "The Multilingual Screen: New Reflections on Cinema and Linguistic Difference", Bloomsbury Publishing Plc. pp. 341, 2016
6. <https://tripurauniv.ac.in/Page/departmentsDetailsHome/39-DepartmentsHome>
7. <https://indianexpress.com/article/political-pulse/amit-shah-hindi-remarks-tripura-tribal-body-roman-script-kokborok-7894171/>
8. Kaewchai Chanchaen, Nisanad Tannin and Booncharoen Sirinaovakul, "Sentence-based machine translation for English-Thai," Proceedings of 1998 IEEE Asia-Pacific Conference on Circuits and Systems. Microelectronics and Integrating Systems (IEEE. APCCAS 1998) (Cat. No.98EX242), pp. 141-144, 1998.
9. Wirapong Chansanam, Kanyarat Kwicien, Marut Buranarach and Kulthida Tuamsuk, "A Digital Thesaurus of Ethnic Groups in the Mekong River Basin," Informatics 2021, 8, 50.
10. Ainur T. Bayekeyeva, Saule Zh. Tazhibayeva, Aigul A. Shaheen, Zhainagul S. Beisenova and Gulnar Beysenkyzy Mamayeva, "Developing an Online Kazakh-English-Russian Thesaurus of Industry-Specific Terminology," International Journal of Society, Culture & Language, 2022.
11. Mandeep Kaur, "Development of Thesaurus for Hindi," International Journal of Computer Sciences and Engineering Vol. 7, pp. 67-72, 2019.
12. Ghassan Mohsen, Mahmoud Al-Ayyoub, Ismail Hmeidi and Ahmad Al-Aiad, "On the automatic construction of an Arabic thesaurus," in the 9th International Conference on Information and Communication Systems (ICICS), IEEE, pp. 243-247, 2018.
13. Binoy Debbarma, Anglo-Kokborok Dictionary, KOHM Publication, Agartala, 1996.
14. <https://koktipra.com>

BIOGRAPHY

Pankaj Debbarma is an Assistant Professor in the Department of Computer Science & Engineering, Tripura Institute of Technology, Narsingarh, Tripura (India). He received his M.Tech. in Computer Science & Engineering degree in 2017 from National Institute of Technology, Agartala, Tripura, India. His research interests are Natural Language Processing, Computational Linguistics, Artificial Intelligence and Cryptography.

Pusparwng Hrangkhawl is an Assistant Professor in the Department of Information Technology, Ambedkar College, Fatikroy, Tripura (India). She received her M.Tech. in Computer Science & Engineering degree in 2018 from Tripura (Central) University, Suryamaninagar, Tripura, India. Her research interests are Natural Language Processing, Machine Learning and Computational Linguistics.



INNO  SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

 **doi**[®]
CROSS **ref**

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details