IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 12, December 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112037 |

English to Kokborok Bilingual Thesaurus for Natural Language Processing

Pankaj Debbarma[1], Pusparwng Hrangkhawl[2]

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

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

ABSTRACT: The unstructured characteristics of a natural language is a challenge that has to be dealt by computational linguist especially in case of vulnerable and resource-scarce language like Kokborok which is one of the official language of Tripura, a state in the north-east India. Kokborok does not even have any official script till date and use of Bengali as well as Roman script are prevalent. This has contributed partly to the slowing down of overall development of the language as well as enrichment of the literature. This paper emphasizes on developing English to Kokborok bilingual thesaurus so that the existing vocabulary can be preserved and further help in contributing in the field of Natural Language Processing for Kokborok language.

KEYWORDS: Kokborok Language, Information Retrieval, Controlled Vocabulary, Bilingual Thesaurus, Natural Language Processing.

I. INTRODUCTION

Natural languages may consist of many unstructured forms of communication including signs, signals or dialects. This brings along complication for searching for particular information by a person who may have a vague knowledge of the topic he is searching. Humans try to interpret their search according to the collection of vocabulary they might be having. Therefore, having a collection of controlled vocabulary can maximize information retrieval (IR). IR is concerned with finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) [1].

Table 1. Examples of homographs and polysemy in Kokborok language.

Word	Meaning	Example
cha	to eat	Bo mai da **cha**kha? Did he/she **eat** the foot?
cha	to become, to happen	Bo ba tamo kobor se **cha**khana bla? Did he/she **become** mad?
cha	to prosper, to do well	Bórok se **cha**kha bla. They have done **well**.
kha	to tie	Bini yak se **kha**jagwi tongkha. His/her hands are **tied**.
kha	to be, to become	O manwi le ha se **kha**kha. This thing has **become** soiled.

Common types of controlled vocabularies include subject headings lists, authority files, and thesauri [2]. For consistency and improved retrieval, various institutions attempt to restrain the disorder of natural language when it comes to describing the relevance of resources. Subject classification or indexing is more reliable if the vocabulary is controlled [3]. A thesaurus provides a structure to the natural language including the concepts and terminologies found in it. Thesauri provide lists of words or terms, often indicating a relationship between the terms.

The entries in a thesaurus are generally listed alphabetically to facilitate easy searching of entries, with some being set hierarchically. Entries are sometimes classified as broader terms (BT) or narrower terms (NT). In the hierarchy

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112037 |

broader terms, often representing a superclass, such as vertebrate, are above narrower or subclass terms, such as primates or hoofed.

Kokborok (ISO 639-3:trp, Ethnologue) is a resource-scarce and vulnerable language [4]. It belongs to the Tibeto-Burman (TB) group of languages put along with languages Bodo, Kachari and Garo [5]. It was once known as Tippera or Hill Tippera also [6]. A unique characteristic of Kokborok is that verbs are almost always in bound form i.e. the verbs have to be in combined form with a suffix, however, free verbs are found when used in context of questions. Further, it is also observed that pronouns in Kokborok are strictly used with neuter gender e.g. nwng (you), ang (I), bo (he/she), bini (his/her), bono (to him/to her), bórok (they). It is also common to find homographs and polysemy in Kokborok language. Few examples representing these characteristics of Kokborok are shown in Table 1.

Vocabulary learning is a major component of language learning [7], thus, it is of utmost importance to design and enrich a thesaurus for Kokborok as literature and digital resources are very scarce for this language. This research work can highly contribute in preservation of the existing vocabulary and play a vital role in enriching the literature of Kokborok language.

II. RELATED WORK

Avenues of research in Kokborok have been there since long but work started very recently. Only few research papers are available till date, few of them have been detailed briefly in Table 2.

Table 2. List of few research work done in the field of Kokborok NLP.

Sl.	Title of Research Paper	Year of Publication	Objective
1	Morphological Analysis of Kokborok for Universal Networking Language Dictionary	2012	This paper emphasizes on bringing works on morphological analysis in the frame, with the goal to produce a Kokborok-dictionary, as well as Machine Translator [8].
2	A Light Weight Stemmer in Kokborok	2012	It describes the design of a simple rule based stemmer for Kokborok using an affix stripping algorithm by stripping the affixes and applying boundary rules where needed [9].
3	Part of Speech (POS) Tagger for Kokborok	2012	It describes the development of a POS tagger using rule based and supervised methods in Kokborok by employing Conditional Random Field (CRF) and Support Vector Machines (SVM) [10].
4	Morphological Analyzer for Kokborok	2012	This paper introduces the design and implementation of a database driven affix stripping algorithm has been used to design the Morphological Analyzer [11].
5	Stemmer for resource scarce language using string similarity measure	2014	This work tries to build a Kokborok stemmer using a statistical approach based on string measure [12].
6	Frequency based named entity recognition system for under resource language	2014	This paper studies the issues and challenges for developing a Named Entity Recognition (NER) system for Kokborok by using the frequency based approach [13].
7	Named Entity Recognizer for less resourced language Kokborok	2015	This paper describes the development of a rule based and a supervised Named Entity Recognizer for the Kokborok language which is less computerized and agglutinative [14].

III. PROPOSED METHODOLOGY AND IMPLEMENTATION

A. *Data Collection:*

The proposed model of thesaurus focuses taking English term as input and output the equivalent synonyms and antonyms in Kokborok language. The model design of the Kokborok bilingual thesaurus will mainly focus on verbs. In order to achieve this, collection of most commonly used verbs in English was obtained from [15] and entered in the thesaurus database. For each verb entered, the respective synonyms and antonyms were also entered. The equivalent synonym and antonym meanings were in turn collected from Koktipra Online Dictionary [16].

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112037 |

B. *Software Design:*

The bilingual thesaurus has been designed by using PHP-MySQL. The Data Flow Diagram (DFD) for a simple search in the Kokborok bilingual thesaurus has been illustrated in Fig. 1.

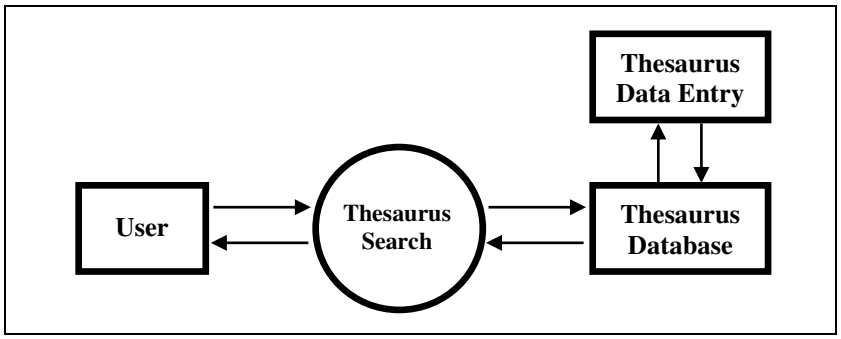


Fig. 1. DFD for Thesaurus Search (Context Diagram)

Firstly, the database for the bilingual thesaurus was created with a table consisting of four fields as given in Table 3. The fields created are 'id', 'english', 'synonyms' and 'antonyms' as shown in Fig. 2. The database is then populated with the English verbs that were collected in the 'english' field. Following this, the respective synonyms and antonyms in Kokborok language is entered in the 'synonyms' and 'antonyms' field, respectively, as shown in Fig. 3.

Table 3. Table structure of the database for Kokborok bilingual thesaurus in MySQL.

Field	Data type	Comment
id	int(11)	Sequence of items in the table
english	varchar(255)	English term input by user to search the thesaurus
synonyms	varchar(255)	Kokborok synonyms for the English term
antonyms	varchar(255)	Kokborok antonyms for the English term

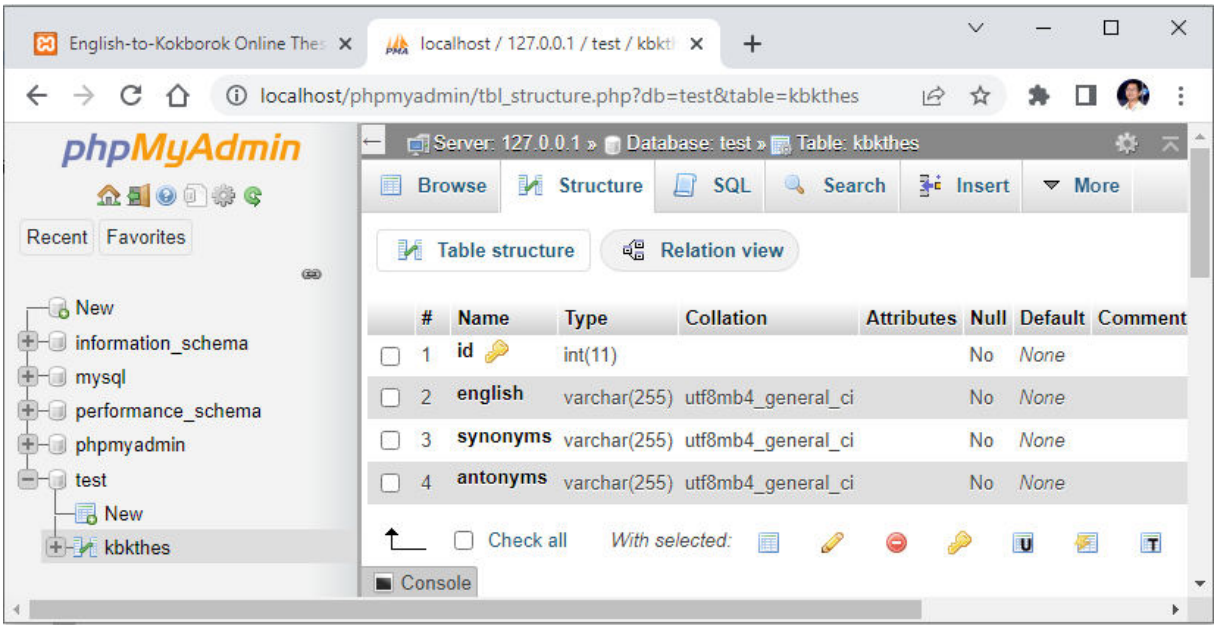
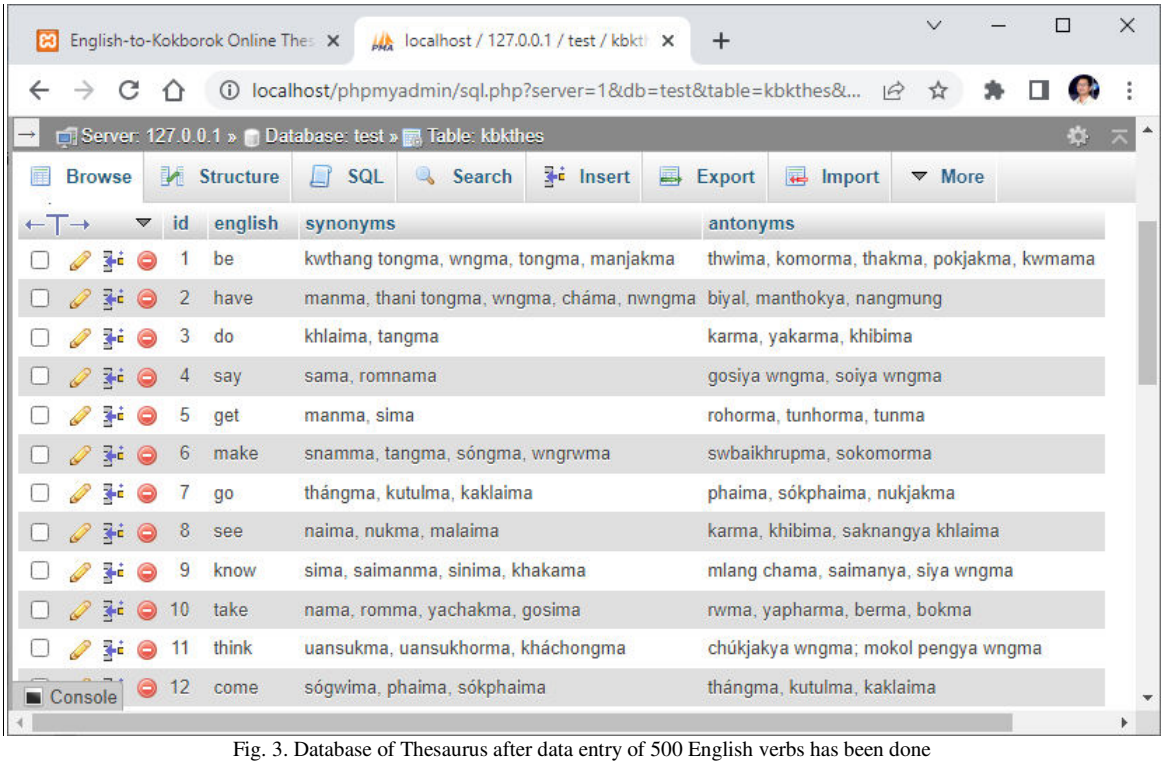


Fig. 2. Table created for Thesaurus with four fields.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112037 |



id	english	synonyms	antonyms
1	be	kwthang tongma, wngma, tongma, manjakma	thwima, komorma, thakma, pokjakma, kwmama
2	have	manma, thani tongma, wngma, cháma, nwngma	biyal, manthokya, nangmung
3	do	khlaima, tangma	karma, yakarma, khibima
4	say	sama, romnama	gosiya wngma, soiya wngma
5	get	manma, sima	rohorma, tunhorma, tunma
6	make	snamma, tangma, sóngma, wngrwma	swbaikhrupma, sokomorma
7	go	thángma, kutulma, kaklaima	phaima, sókphaima, nukjakma
8	see	naima, nukma, malaima	karma, khibima, saknangya khlaima
9	know	sima, saimanma, sinima, khakama	mlang chama, saimanya, siya wngma
10	take	nama, romma, yachakma, gosima	rwma, yapharma, berma, bokma
11	think	uansukma, uansukhorma, kháchongma	chúkjakya wngma; mokol pengya wngma
12	come	sógwima, phaima, sókphaima	thángma, kutulma, kaklaima

Fig. 3. Database of Thesaurus after data entry of 500 English verbs has been done

Thereafter, the front-end page for the thesaurus was created so that query can be input by user which will then connect with the database and search for a match-case for the user input query. On getting a hit, the values of the respective fields are populated in the output portion of the thesaurus home page as shown in Fig. 4.

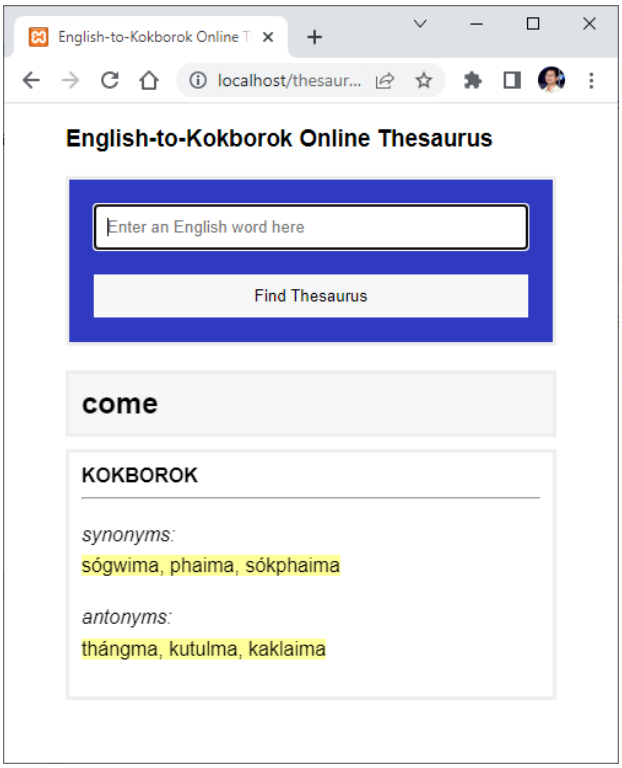


Fig. 3. Home page of the English to Kokborok Bilingual Thesaurus

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112037 |

IV. CONCLUSION AND FUTURE WORK

The English to Kokborok bilingual thesaurus has been designed to help preserve and enrich the vulnerable and resource-scarce Kokborok language. However, as it is known that there are many more languages in India which are not only vulnerable but are identified by the United Nations to be endangered and may face extinction in near future, this software can extensively help in preserving and in turn relating the languages with each other. People involved in publishing activities can take full advantage of this bilingual thesaurus to increase the existing literary standard through the word options that are being available in this thesaurus.

REFERENCES

1. C. D. Manning, P. Raghavan and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, New York, NY, 2008.
2. https://guides.lib.utexas.edu/metadata-basics/controlled-vocabs
3. https://www.librarianshipstudies.com/2020/03/controlled-vocabulary.html
4. C. Moseley, “Atlas of the World's Languages in Danger,” in Memory of Peoples, 3rd edition, UNESCO Publishing, Paris, 2010.
5. A. Hale, Research on Tibeto-Burman Languages, De Gruyter Mouton, Berlin, Boston, 2020.
6. Imperial Gazetteer of India, Vol. 1, Clarendon Press, Oxford, p. 119, 1909.
7. V. Abou-Khalil, S. Helou and B. Flanagan, “Learning isolated polysemous words: identifying the intended meaning of language learners in informal ubiquitous language learning environments,” Smart Learn. Environ. 6, 13, 2019.
8. K. Debbarma, B. G. Patra, S. Debbarma, L. Kumari and B. S. Purkayastha, “Morphological analysis of Kokborok for universal networking language dictionary,” in Recent Advances in Information Technology (RAIT), 2012 1st Intl. Conf., IEEE, p. 474-477, 2012.
9. K. Debbarma, B. G. Patra, S. Debbarma, "A Light Weight Stemmer in Kokborok," in 24th Conf. on Computational Linguistics and Speech Processing, 2012.
10. B. G. Patra, K. Debbarma, D. Das and S. Bandyopadhyay, "Part of Speech Tagger for Kokborok", Intl. Journal of Computational Linguistics and Natural Language Processing, vol. 1, no. 3, p. 85-91, 2012.
11. K. Debbarma, B. G. Patra, D. Das and S. Bandyopadhyay, "Morphological Analyzer for Kokborok," in 24th International Conference on Computational Linguistics, p. 41, 2012.
12. A. Debbarma, B. S. Purkayastha and P. Bhattacharya, "Stemmer for resource scarce language using string similarity measure", Proc. Intl. Conf. on Optimization Reliability and Information Technology (ICROIT), p. 96-98, 2014.
13. A. Debbarma; P. Bhattacharya and B. S. Purkayastha, "Frequency based named entity recognition system for under resource language," in Intl. Conf. on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014.
14. B. G. Patra, N. Debbarma, T. A. Abahai, D. Das and S. Bandyopadhyay, "Named Entity Recognizer for less resourced language Kokborok", in Intl. Conf. on Asian Language Processing (IALP), p.164-168, 2015.
15. https://www.educall.com.tr/blog/post/500-most-common-english-verbs
16. 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: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details