IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

Development of Thesaurus for Hindi

Mandeep Kaur

Department of Computer Science, Punjabi university, Patiala, India

*Corresponding Author: meenu03kaur@gmail.com

Available online at: www.ijcseonline.org

Accepted: 25/Dec/2018, Published: 31/Jan/2019

Abstract— NLP is a vast field and thesaurus is its integral part. Thesaurus is a software tool which is inbuilt in few word processors that provides synonyms for selected words. A thesaurus is used on a computer while writing an e-mail, letter, or paper to find an alternative meaning for words. A thesaurus is a reference work that lists the synonyms and sometimes antonyms of words. Synonyms are words with similar meanings, and antonyms are words with opposite meanings. The research work in the paper elaborates the development of thesaurus in relevance with the Hindi language. The paper focuses on the development of framework which may assist the people finding it difficult to write and dealing with Hindi.

Keyword— *Thesaurus, Hindi language, Synonyms, Antonyms*

I. INTRODUCTION

Thesaurus is one of the real research works of NLP. Natural Language Processing (NLP) is a region of research and application that investigates how PCs can be utilized to comprehend and control normal dialect content or discourse to do valuable things [1, 11]. A thesaurus is a kind of scientific classification concentrating particularly on the connections between the terms. It gives an institutionalized phrasing or also controlled vocabulary for a specific territory of learning [2, 12]. A thesaurus is an asset that gatherings words as indicated by likeness. Some thesauruses, generally manual ones, have a progressive structure including various layers. Others, as a rule, the programmed ones, basically contain gatherings of words (so might be seen as one-level progressive systems) [3, 13]. Hindi dialect has a place with an Indo-Aryan dialect that is the part of Indo European; it has created from the Sanskrit dialect. Hindi is additionally talked by an extensive populace in India principally (Northern locale of India). As indicated by 2011 statistics, 528,347,193 individuals communicate in Hindi dialect. This is the roughly 43.63% of the aggregate populace of our nation India. Most valuable content (लिपि) is Devanagari (देवनागरी) to compose Hindi, Nepali and furthermore Marathi dialect. Undertakings which could profit by an astounding thesaurus incorporate parsing, anaphor goals, building up content lucidness and word sense disambiguation [4, 14]. In the Hindi Language, Synonyms are also called समानार्थक शब्द या पर्यायवाची शब्द and antonyms in like manner called by some other different names which called विलोम शब्द या विपरीतार्थक शब्द. Antonyms and Synonyms are a fundamental bit of any vernacular[18, 19].

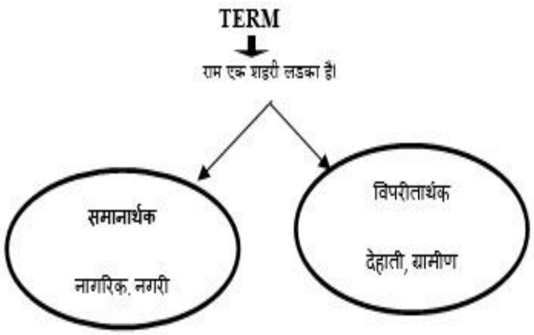


Figure 1. Example of Hindi thesaurus

II. RELATED WORK

Tannin et al. [1] have proposed a cemented methodology for machine explanation in which two unmistakable strategies are used to manage based and outline based are related to making another structure for English to-Thai sentence understanding. Yamamoto et al. [2] have proposed the change of bilingual word reference which is Korean to Japanese by using two extraordinary vocabularies. Likewise, they use one tongue as center lingo between these two-word references. These vocabularies are Japanese to English and Korean to English. The widely appealing vernacular is English that is used for making the bilingual dictionary. Vidyasagar et al. [3] have proposed the fundamental examination of Kannada WordNet. Disregarding the way that the manner in which that the outline has been energized by the extraordinary English WordNet, and to a certain degree, by the Hindi WordNet. Kannada WordNet fills in as an online thesaurus. Sarkar et al. [4] have proposed the issues identifying with the made course of action and change out of a summed up expanded WordNet for Indo Aryan dialects with stand-out reference to Hindi and Bengali. The outline of ABHIDHA depends upon the way that the central relations and lexical affiliations are between the sunsets. Shingo et al. [5] have proposed two figuring's to regularly disconnect an English/Arabic bilingual word reference from parallel messages that exist in the Internet document. Utilizing the made bilingual word reference and parallel corpus we can execute a not that entire terrible English-Arabic machine understanding structure. Tjoa et al. [6] have proposed a novel semantic similitude procedure. This depends on apprehensive counting which joins WordNet and space ontologies written in Web Ontology Language (OWL). Wu et al. [7] have proposed a novel method is to evaluate semantic comparability between words using HowNet. A Chinese thesaurus is used to improve the likeness assessing. Kanzaki et al. [8] have proposed the Japanese WordNet. They incorporate synsets in Japanese dialect. The Japanese word net creates from Princeton word net, Spanish word net and French WordNet. Yu et al. [9] have proposed a novel method to manage to improve the part based Word Sense Disambiguation (WSD). In light of the straight piece, two external data sources are fused. One is the semantic principles to find illustrative words. Kulkarni et al. [10] have proposed the Sanskrit WordNet. They created it in Indian establishment of innovation at Mumbai. They portray the extension approach which is utilized for building up the Sanskrit WorldNet's database [15, 16, 20].

III. PROPOSED WORK

The research work utilize different apparatuses and procedures for database outline and execution. We utilize MySQL (organized question dialect) for the database stockpiling and PHP for usage utilizing a WAMP server. There are different advances depicted for the database outline and usage. Initially, we display the structure of the proposed think about.

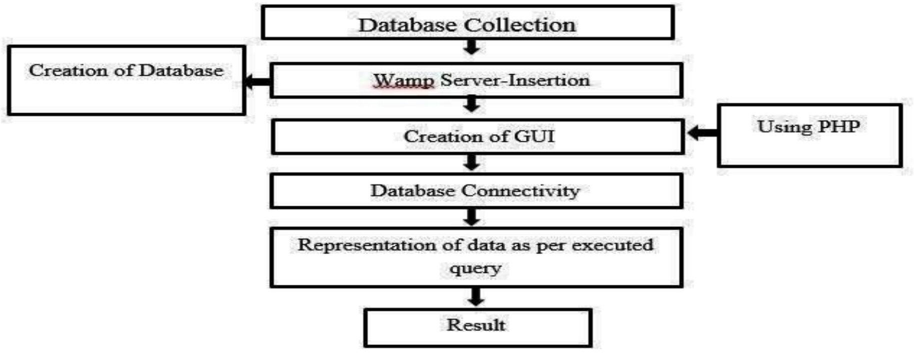


Figure 2: Framework of the proposed work

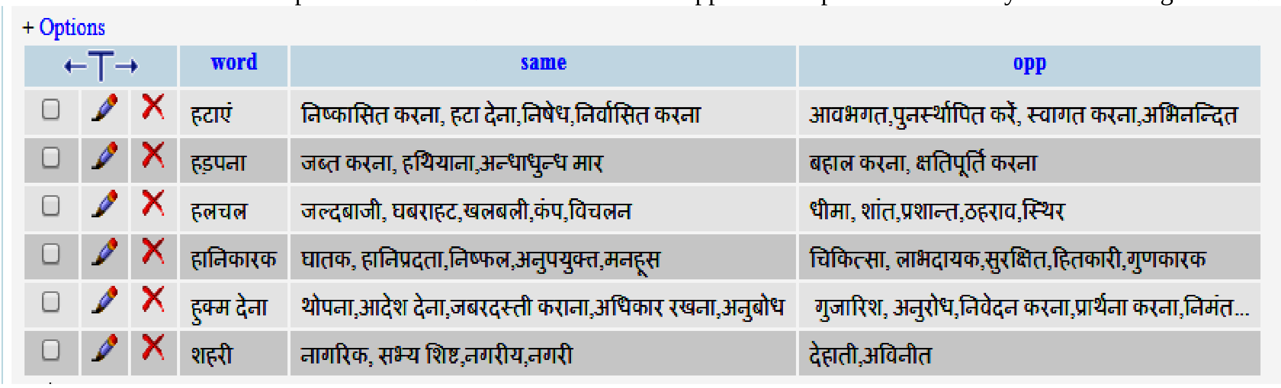
Preparation of Database

The database is depicted as the efficient accumulation of the related information. The database is populated, structure and worked with information for the specific utilize [17]. Our information is gathered from different assets which are Hindi sentence structures, Internet and Hindi lexicons. Database of this task work comprises two tables:

1) data1
2) user1

These are described as:

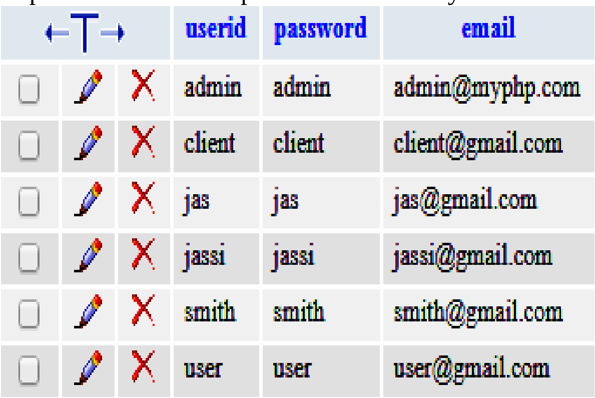
data1: This table contains three parameters appeared underneath in figure 3 is that the primary section is "word" which is the exact word for seeking Antonyms and Synonyms. The second section is "same" which demonstrate the equivalent words of the sought word. The last section is "opp" which speak to the antonyms of the sought word.



←T→			word	same	opp
□	✎	✗	हटाएं	निष्कासित करना, हटा देना,निषेध,निर्वासित करना	आवभगत,पुनर्स्थापित करें, स्वागत करना,अभिनन्दित
□	✎	✗	हड़पना	जब्त करना, हथियाना,अन्धाधुन्ध मार	बहाल करना, क्षतिपूर्ति करना
□	✎	✗	हलचल	जल्दबाजी, घबराहट,खलबली,कंप,विचलन	धीमा, शांत,प्रशान्त,ठहराव,स्थिर
□	✎	✗	हानिकारक	घातक, हानिप्रदता,निष्फल,अनुपयुक्त,मनहूस	चिकित्सा, लाभदायक,सुरक्षित,हितकारी,गुणकारक
□	✎	✗	हुक्म देना	थोपना,आदेश देना,जबरदस्ती कराना,अधिकार रखना,अनुबोध	गुजारिश, अनुरोध,निवेदन करना,प्रार्थना करना,निमंत...
□	✎	✗	शहरी	नागरिक, सभ्य शिष्ट,नगरीय,नगरी	देहाती,अविनीत

Figure 3. Perspectives of Database Table of Synonyms and Antonyms

user1: This table shows us in figure 4 is that all the enrolled clients who can utilize this entrance for performing the different actions. For new enlistment customer need to enter "customer id", "Email Id" and "Mystery word" in "reg.php". No duplicate customer will select in the database. After productive registration customer can log in and incorporate/look for comparable word/antonym.

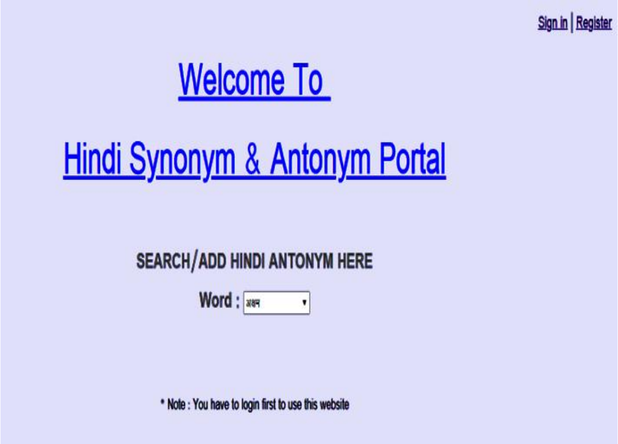


←T→			userid	password	email
☐	✎	✗	admin	admin	admin@myphp.com
☐	✎	✗	client	client	client@gmail.com
☐	✎	✗	jas	jas	jas@gmail.com
☐	✎	✗	jassi	jassi	jassi@gmail.com
☐	✎	✗	smith	smith	smith@gmail.com
☐	✎	✗	user	user	user@gmail.com

Figure 4. Database administration of enlisted client of Hindi Thesaurus gateway

Experimental Result: There are nine-screen captures of Hindi thesaurus entryway that portrayed the execution through a different view.

Hindi thesaurus portal start with "MyGuset.php" page displayed in Figure 5, in this page user will get an option to List the words available in the database. To search or add new Synonym and Antonym user need to login by click on Sign_in link. If the user is new then he/she will register through Register.



Welcome To

Hindi Synonym & Antonym Portal

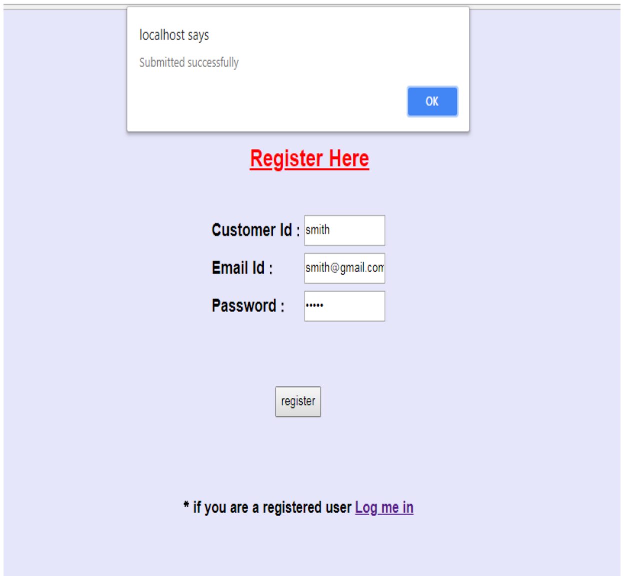
SEARCH/ADD HINDI ANTONYM HERE

Word : अक्षम

* Note : You have to login first to use this website

Figure 5. "Sign in" page expression of Hindi equivalent word and antonym gateway

Figure 6 appeared as beneath for new enrollment client need to enter "client id", "Email Id" and "Secret key" in "reg.php". No copy client will enroll in the database. After an effective registration client can log in and include/seek equivalent word/antonym.



localhost says

Submitted successfully

Register Here

Customer Id : smith

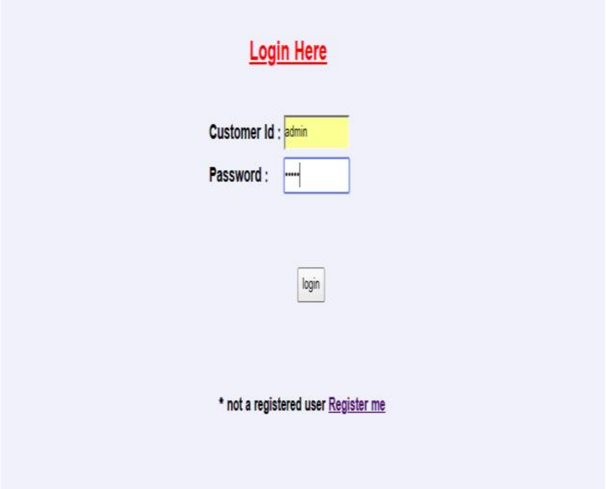
Email Id : smith@gmail.com

Password : •••••

* if you are a registered user Log me in

Figure 6. "Registration page" of Hindi equivalent word and antonym gateway

Figure 7 demonstrates the login page of Hindi portal in which just Registered client can get to the data by login with legitimate "client id" and "secret key". On the off chance that the client isn't enrolled then the Unauthorized client can not get to the data of this entrance.



Login Here

Customer Id : admin

Password : •••••

* not a registered user Register me

Figure 7. "Login page" of Hindi equivalent word and antonym entrance

Figure 8 demonstrates that login effectively by enlist client. Presently client can undoubtedly look through the antonyms and equivalent words of the word by a tap on Search here catch. The client can include new antonyms and equivalent words of words by a tap on interface AddHere.

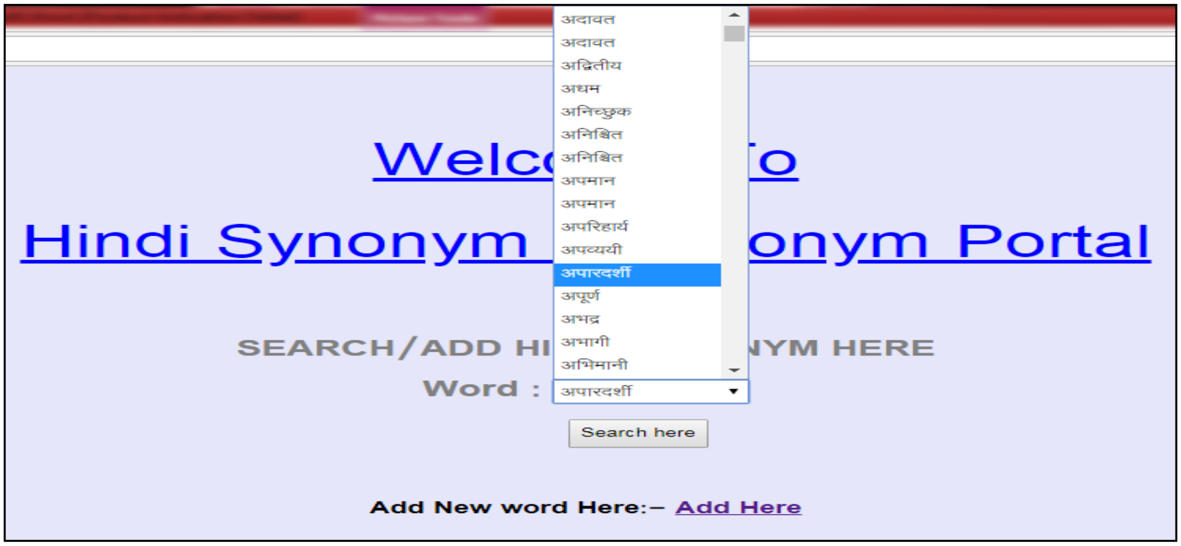
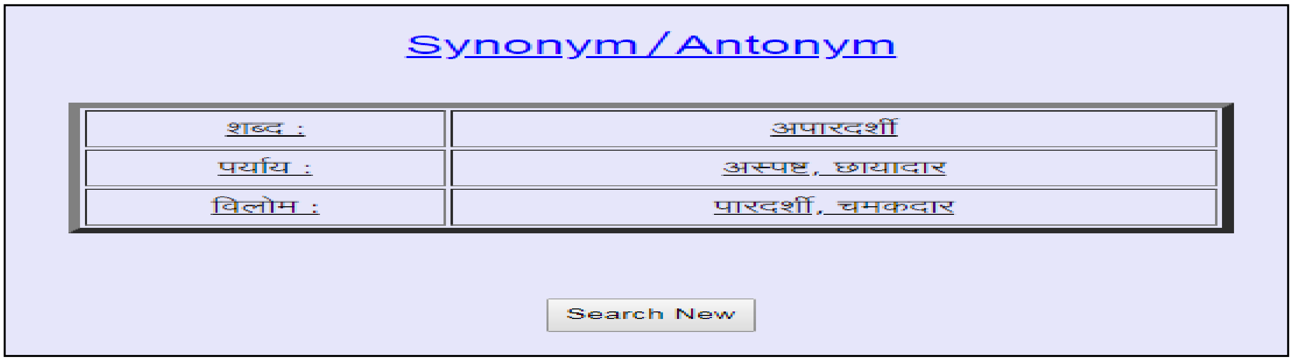


Figure 8. "look page" for word by a registered clientof Hindi equivalent word and antonym entryway

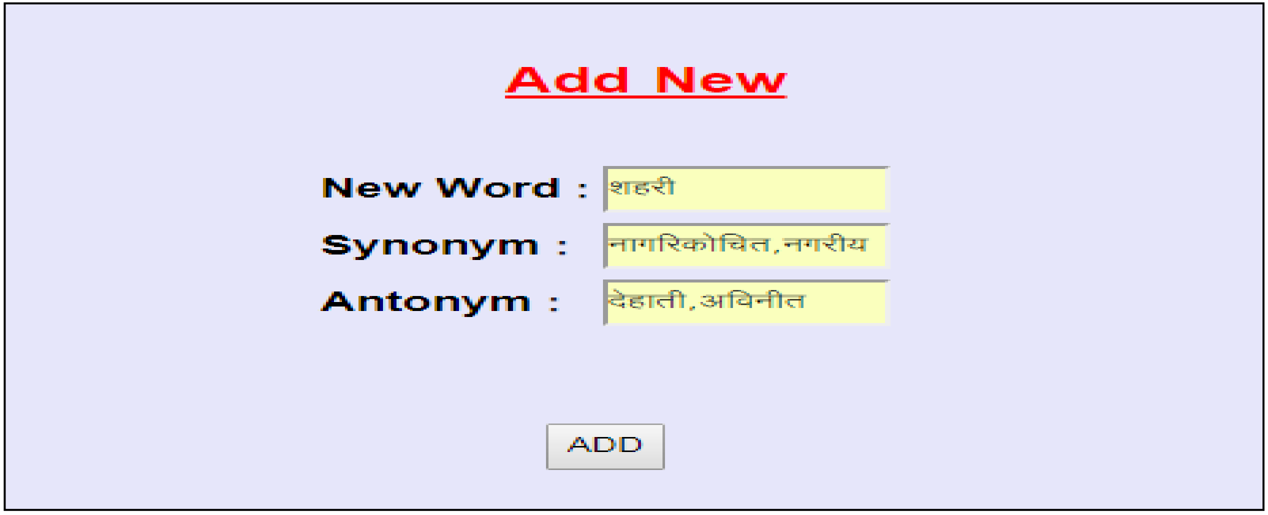
Figure 9 demonstrates that the chose word's equivalent words and antonyms on the screen by tapping on Search here catch. Furthermore, if the client needs to move back on the looking and including word entrance then the client can be a tap on Search New catch.


Synonym/Antonym

शब्द :	अपारदर्शी
पर्याय :	अस्पष्ट, छायादार
विलोम :	पारदर्शी, चमकदार

Figure 9. "equivalent word and antonym" of looked word

Figure 10 demonstrates that the client can include a new word with its equivalent word and antonym by a tap on ADD catch. After this, the new word's antonym and equivalent word put away in a made database.


Add New

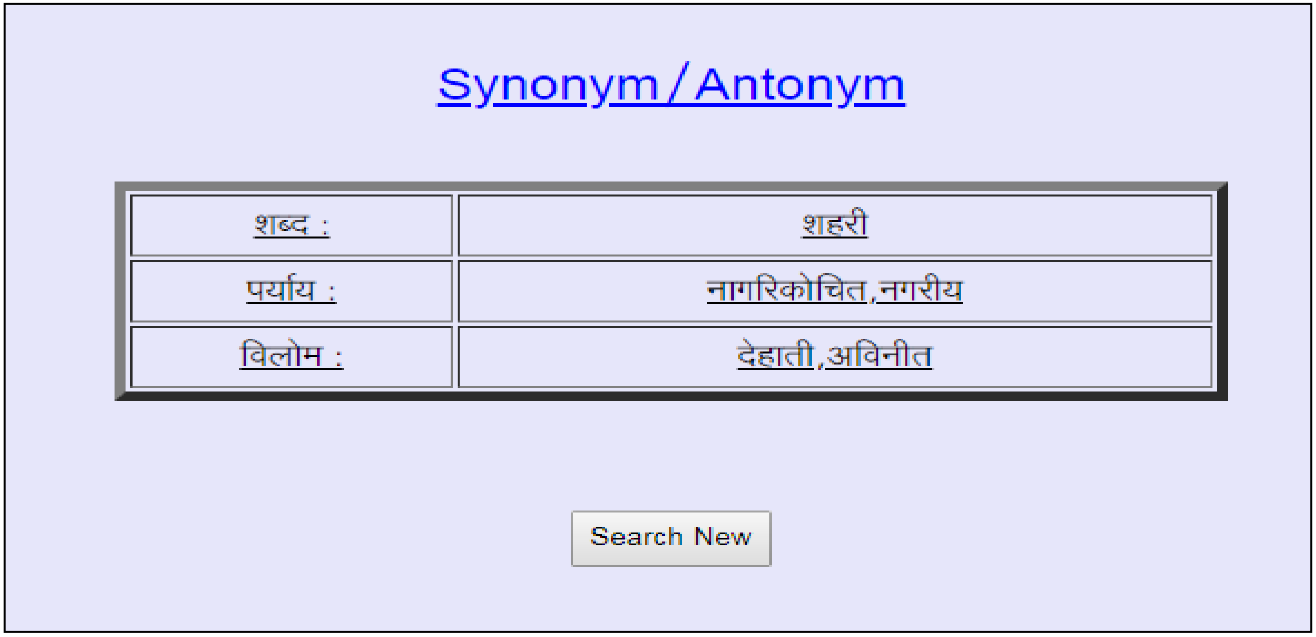
New Word : शहरी

Synonym : नागरिकोचित,नगरीय

Antonym : देहाती,अविनीत

Figure 10. "Expansion of new word" with its equivalent word and antonym

Figure 11 demonstrates that the capacity of a new word, User can easily get to new word's antonym and equivalent word by playing out the searching activity by a tap on Search here catch. At that point, the new included word is appears with its equivalent word and antonym.

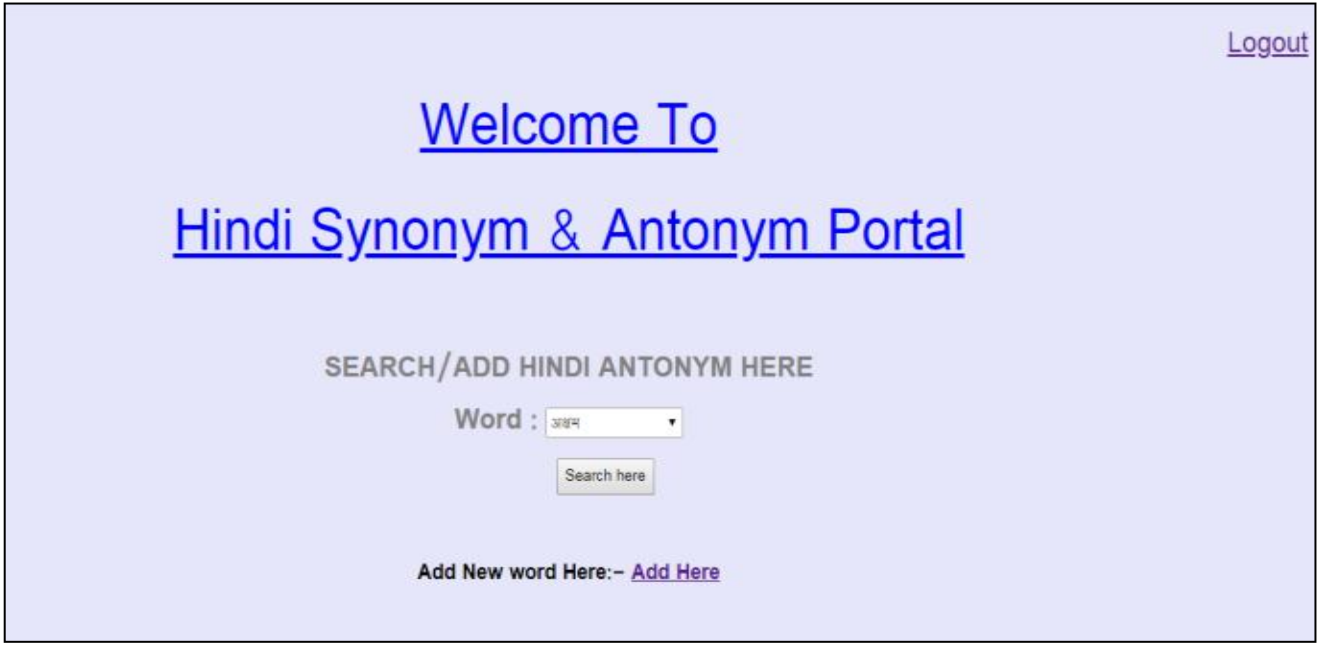


Synonym / Antonym

शब्द :	शहरी
पर्याय :	नागरिकोचित,नगरीय
विलोम :	देहाती,अविनीत

Figure 11. "stockpiling of new word" with its Synonym and Antonym

Figure 12 After performing different activities, for example, looking and expansion of antonyms and equivalent words client can end login session by tapping on logout catch



Welcome To

Hindi Synonym & Antonym Portal

SEARCH/ADD HINDI ANTONYM HERE

Word : अधम

Add New word Here:- Add Here

Figure 12. "Logout page" of Hindi Synonym and Antonyms entryway

Conclusion and future work

The research paper discussed the construction of framework dedicated to Hindi thesaurus. The research work fills the gap in the Hindi thesaurus effectively and efficiently. The research work contains in excess of 400 words. The future research would further extend this database of words to cover almost healthy percentage of Hindi words so that people dealing or working with Hindi words in the computer science may find it comfortable and easy to find out available alternatives.

REFERENCES

[1] Chancharoen, K., Tannin, N., &Sirinaovakul, B., Sentence based machine translation for English-Thai. "*Circuits and Systems", 1998. IEEE APCCAS 1998. The 1998 IEEE Asia-Pacific Conference on* (pp. 141-144). IEEE, 1998.

[2] Shirai, S., & Yamamoto, K, "Linking English words in two bilingual dictionaries to generate another language pair dictionary", *Proceedings of ICCPOL*, pp. 174-179, 2001.

[3] Sahoo, K.,Vidyasagar, V. E., "Kannada WordNet-A lexical database", *TENCON 2003.Conference on Convergent Technologies for the Asia-Pacific Region* (Vol. 4, pp. 1352-1356).IEEE, 2003.

[4] Annam, S. R., Choudhury, M., Sarkar, S., &Basu, A., "ABHIDHA: an extended WordNet for Indo Aryan languages", *Research Issues in Data Engineering: Multi-lingual Information Management,RIDE-MLIM Proceedings. 13th International Workshop* (pp. 1-8), 2003.

[5] Fattah, M. A., Ren, F., & Shingo, K., "Internet archive as a source of a bilingual dictionary", In *Information Technology: Coding and Computing, Proceedings ITCC 2004.International Conference on* (Vol. 2, pp. 298-302).IEEE, 2004.

[6] Banek, M., Vrdoljak, B., &Tjoa, A. M. (2007, June). Using ontologies for measuring semantic similarity in data warehouse schema matching process. In *Telecommunications, ConTel 2007. 9th International Conference on* (pp. 227-234). IEEE, 2007.

[7] Dai, L., Liu, B., Xia, Y., & Wu, S. (2008, August). Measuring semantic similarity between words using HowNet.In *Computer Science and Information Technology, ICCSIT'08. International Conference on* (pp. 601-605). IEEE, 2008.

[8] Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., & Kanzaki, K. Development of the Japanese WordNet, 2008.

[9] Jin, P., Li, F., Zhu, D., Wu, Y., & Yu, S. (2008, October). Exploiting external knowledge sources to improve kernel-based word sense disambiguation. In *Natural Language Processing and Knowledge Engineering,. NLP-KE'08. International Conference on* (pp. 1-8). IEEE, 2008.

[10] Kulkarni, M., Dangarikar, C., Kulkarni, I., Nanda, A., & Bhattacharyya, P. (2010, January). Introducing Sanskrit wordnet.In *Proceedings of the 5th Global Wordnet Conference (GWC 2010), Narosa, Mumbai,* (pp. 287-294), 2010.

[11] Chowdhury, G. G., "Natural language processing", *Annual review of information science and technology*, *37*(1), 51-89, 2003.

[12] Slawsky, D., "Building a keyword library for a description of visual assets: Thesaurus basics", *Journal of Digital Asset Management*, *3*(3), pp. 130-138, 2007.

[13] Kilgarriff, A., "Thesauruses for natural language processing.", In *Natural Language Processing and Knowledge Engineering, 2003.Proceedings. 2003 International Conference on* (pp. 5-13), 2003.

[14] Bradeško, L., Dali, L., Fortuna, B., Grobelnik, M., Mladenić, D., Novalija, I., &Pajntar, B., "Contextualized question answering", *Journal of computing and information technology*, *18*(4), pp. 325-332, 2010.

[15] Tayal, A., "THESAURUS FOR INDIAN LANGUAGES AND CONVERSION RULES DURING DESIGN OF PUNJABI THESAURUS", *Journal of Global Research in Computer Science*, *2*(7), pp. 38-41, 2011.

[16] Ramírez, J., Asahara, M., & Matsumoto, Y., Japanese-Spanish thesaurus construction using English as a pivot. *arXiv preprint arXiv:*1303.1232, 2013.

[17] Mohd, M., Zakr, H., Abidin, N. Z., Tiun, S., &Hisham, A. I. I. (2013, December).Word sense disambiguation for English Quranic IR system. NOORIC 2013: Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences, 2013.

[18] Panchal, P., Panchal, N., &Samani, H., Development of Gujarati WordNet for Family of Words. *Development*, 1(4), 2014.

[19] Kanakaraj, M., &Kamath, S. S. (2014, December). NLP based intelligent news search engine using information extraction from e-newspapers. In *Computational Intelligence and Computing Research (ICCIC), IEEE International Conference on* (pp. 1-5), 2014.

[20] Redkar, H., Singh, S., Joshi, N., Ghosh, A., & Bhattacharyya, P., "Indowordnet Dictionary: An Online Multilingual Dictionary using Indowordnet", *Proceedings of the 12th International Conference on Natural Language Processing* (pp. 71-78), 2015.