Towards a Richer IndoWordNet with New Additions for Hindi and Gujarati Languages

Milind Kumar Audichya[1]
SJD International,
Surat, India - 395009

Jatinderkumar R. Saini[2]*
Symbiosis Institute of Computer Studies
and Research, Symbiosis International
(Deemed University), Pune, India - 411016

Jatin C. Modh[3]
Gujarat Technological University,
Ahmedabad, India - 382424

***Abstract*—The authors of this research paper present a mechanism for dealing with loanwords, missing words, and newly developed terms inclusion issues in WordNets. WordNet has evolved as one of the most prominent Natural Language Processing (NLP) toolkits. This mechanism can be used to improve the WordNet of any language. The authors chose to work with the Hindi and Gujarati languages in this research work to achieve a higher quality research aspect because these are the languages with major dialects. The research work used more than 5000 Hindi verse-based data corpus instead of a prose-based data corpus.As a result, nearly 14000 Hindi words were discovered that were not present in the popular Hindi IndoWordNet, accounting for 13.23 percent of the total existing word count of 105000+. Working with idioms was a distinct method for the Gujarati language. Around 3500 idioms data were used, and nearly 900 Gujarati terms were discovered that did not exist in the IndoWordNet, accounting for nearly 1.4 percent of the total of 64000+ Gujarati words in the IndoWordNet. It will also contribute almost 14000 Hindi words and around 900 Gujarati words to the IndoWordNet project.**

***Keywords*—Gujarati; Hindi; Indian language WordNet; IndoWordNet; loanwords; WordNet**

I. Introduction

Languages transmit knowledge from one generation to the next, as well as from one culture to another, through communication. There are multiple modes of communication, including speech, writing, and sign language. No matter what the communication medium is, Communication becomes a systematic process when using any language because each has its vocabulary, grammar, and components. There are 7151 different languages in the world [1], some of which are well-known while others are on the verge of extinction.

Hindi [2], a truly mellifluous Indo-Aryan language, is one of the most popular languages in the world, which is scripted using Devanagari [3] and is currently supported by The Unicode Standard [4]. Although Hindi is widely spoken throughout the world, it is mostly used in India, particularly in the Hindi Belt, which encompasses sections of India's four major zones: eastern, western, central, and northern [2]. The rich literature and long history of Hindi users make up the language's legacy. Hindi-based research and related efforts in relevant research disciplines are currently strongly emphasized.

Gujarati[5], like Hindi, is a sweet-sounding language that belongs to the Indo-Aryan language family. Although Gujarati is used by Gujaratis all over the world, it is especially utilised for communication in Gujarat, located in India's western region, which is considered the origin place of Gujarati. Gujarati is written in the Devanagari script as well and is currently supported by The Unicode Standard [6]. Gujarati is also thriving in terms of research and development in recent years [7], [8].

Natural languages are those that have evolved gradually and are used by humans to communicate. It is a well-known fact that computers don't understand natural languages. To make computers understand different kinds of languages, many research works are going on for different languages. Natural Language Processing (NLP) [9] is used to assist computers in understanding these languages. Computer Science (CS) [10], Computational Linguistics (CL) [11], and Artificial Intelligence (AI) [12] all intertwine in NLP. To grasp a language, computers, like humans, must understand the alphabet, words, meanings of words, pronunciation, vocabulary, sentence structure, context, and all grammatical rules associated with it.

Because computers lack cognitive intelligence, making them grasp any language is difficult. WordNet [13]–[15], a systematically managed correlated lexical database that usually consists of words and semantic relations with the words including synonyms, hyponyms, and meronyms, is commonly used to help computers overcome this limitation. As a practical matter, WordNet can be thought of as a hybrid of a dictionary and a thesaurus. There are various WordNet-based research projects underway [16], [17], some of which are language-specific exclusively and others that are multilingual.

The authors used IndoWordNet [18], [19], a well-known WordNet based on Indian regional languages, for this study. IndoWordNet is a WordNet that has a base of 18 different Indian languages. The Center for Indian Language Technology (CFILT) in the Computer Science and Engineering Department at IIT Bombay created IndoWordNet. Hindi is the default base language of this WordNet. The IndoWordNet is used in this research for both Hindi and Gujarati languages. Even when a well-built WordNet exists for popular languages [20], [21], there is always room for improvement.

The authors were motivated by the various WordNets they used and the difficulties they encountered while using them in different research studies due to missing or borrowed words. Current research is an attempt to address such concerns. This research will be useful in strengthening the WordNet and adding new words, loan words, and missing words.

Loanwords, missing words and newly developed terms are

*Corresponding authors.

several issues that must be addressed in the building of any language's WordNet. Each language typically has loanwords, which are words taken from other languages with slight or no alterations. Another challenge is dealing with words that are not in WordNet and dealing with the newly developed terms in the recent times. This research aims to efficiently address these issues in order to improve WordNets throughout time.

This research paper will also explain how to use WordNets to make effective use of information created by WordNets in order to strengthen WordNet research. The Section II, Literature Review will discuss various research works in a similar or related field to determine the research gap and the need for current research work. The Section III, Methodology segment describes the actual mechanism that can be used to continuously improve the WordNets. The Section IV, Results part will summarise the outcome of the applied mechanism for the Hindi and Gujarati languages utilising the IndoWordNet. The Section V, Conclusion portion will indicate how the improvements should be implemented and expanded. Finally, Section VI, Future Enhancement include some relevant and helpful remarks for upcoming research works.

II. Literature Review

The authors attempted to delve deeper into the WordNet research area first, as they were working with various WordNets and encountering issues with missing words from WordNets. WordNets are lexical hybrid resources that are an interconnected combination of a dictionary and a thesaurus built using different approaches for different languages. Professor George A. Miller[13], [14] oversaw the creation of the first WordNet for the English language at Princeton University's cognitive science lab. Following the development of Princeton's American English WordNet, a multilingual wordnet EuroWordNet[15] was created on similar principles during March 1996 to June 1999, specifically designed for the various European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). The development of WordNets for various Indian languages began in the early 2000s, spearheaded by the Hindi WordNet [16]. Later, in the direction of Bhattacharyya research project IndoWordNet [17] was expanded to include multiple Indian languages, and it now supports 18 languages (Assamese, Bangla, Bodo, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Tamil, Telugu, Urdu).

Authors attempted to locate some advancement and improvement associated research works related to WordNet after exploring for foundation-related research works. Redkar et al. [18] attempted to create an online multilingual dictionary with 19 languages for researchers and non-researchers to utilise for various purposes. To access the IndoWordNet [17], Panjwani et al. created pyiwn [19], a python-based Application Programming Interface (API), pyiwn is used to access IndoWordNet in this study.

In other recent work, McCrae et al. [20] worked on English WordNet 2020, an open-source project to improve and extend the Princeton WordNet, which hasn't been updated in a long time. In other recent work, McCrae et al. worked on English WordNet 2020, an open-source project to improve and extend the Princeton WordNet, which hasn't been updated in a long time. They made around 15000 modifications since the last version was updated.

Kanojia et al. [21] worked on linking 18 different Indian Language WordNets to Princeton WordNet, which aids in the exchange of knowledge and comprehension of various terminologies and their meanings in a multilingual environment. Fellbaum [22] described the WordNet's latest approaches, which are focused on language mapping concepts. He also discusses Crosslinguistic WordNets, as well as all of the components of how WordNets are managed over time.

In an another introductory work, Bhensdadia et al. [23] reviewed the development of the Gujarati wordnet. That research was also a component of the IndoWordNet, and the source language for Gujarati language development was solely Hindi. As a result, the authors of the current study decided to conduct parallel research in both languages. Zankhana and Sajja [24] also worked on the issue of word sense disambiguation in the Gujarati language using a Knowledge-Based Approach and a Genetic Algorithm. Using the IndoWordNet, Modh and Saini attempted to contextually improve Gujarati machine translation [25]. They used the n-gram model and tests with varying frame sizes to try to enhance the translation of Gujarati idioms.

The authors reviewed numerous study works linked to WordNets of various languages, as well as improvement-related works and a few studies focusing on the trip of the WordNets' development thus far.The authors realised that there is still a need to identify some systematic technique that might aid streamline the process of improving and strengthening WordNets. Inclusion of loanwords, missing words, and newly formed terminology in any WordNet of any language around the world must be done in a systematic and organized way.

The authors also discovered that while most WordNet-based initiatives focus on prose-based content, verse-based content can also provide some noteworthy contributions. Using verse for such research purposes is still a challenge in and of itself. Languages with a greater number of dialects make such study even more difficult. Taking into account all of these factors, the authors decided to focus on developing a system for continual WordNet improvement and strengthening. The authors opted to work with Hindi and Gujarati languages since they have diverse dialects, and they chose to evaluate the mechanism's quality and efficacy with Hindi's verse-based literary content and Gujarati's idiom-based data.

Following an extensive literature review and identifying a pressing need of the hour while conducting research on metadata generation for Hindi poetry [26], [27] and a few other research works related to Hindi poetry [28] and Gujarati idioms [29] using WordNet, the findings indicate the need for a systematic approach to dealing with loanwords, missing words, and recently developed terminologies. Since there are so many dialects in Hindi and Gujarati, resolving such difficulties is more complicated; Intriguingly, the authors acknowledged this research as a challenging one.

III. Methodology

This proposed research revolves around the use of WordNet, specifically the use of WordNet to find words that aren't currently in the WordNet. Several phases of this research as represented in Fig. 1, Core Methodology includes:

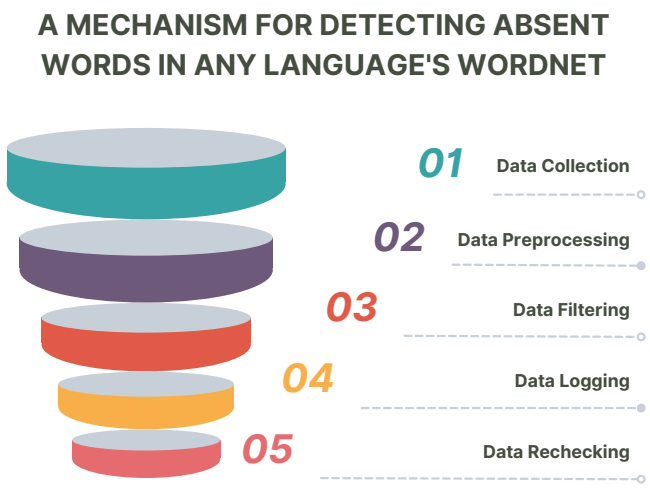

Fig. 1. Core Methodology

- III-A Data Collection
- III-B Data Preprocessing
- III-C Data Filtering
- III-D Data Logging
- III-E Data Rechecking

Authors incorporated transliteration and translation wherever non-English text is included, keeping in mind the international readership of the research work.

A. Data Collection

Special efforts were expended during the data collection phase for both Hindi and Gujarati. After analysing and brainstorming numerous types of data, the Hindi language poetic data and the Gujarati language idiom data were chosen with the goal of finding the most missing terms for a qualitative and quantitative contribution to the improvement of the respective languages' WordNets. Online websites and portals, as well as offline literature, were thoroughly examined for data collection. This study makes use of poetic data from Hindi literature. In Hindi poetry, verses play an important role. To write any type of verse, there are some specific rules that must be followed. Audichya and Saini's Hindi Verse dataset [30], as well as some additional data, were used. A total of 5011 poetic pieces of information were collected.

Example Hindi Data:

लग रहा है आज अम्बर, ज्यों उमड़ता सा समन्दर।
उठ चली चंचल हिलोरे, आ रहे जैसे बवंडर।।

Transliteration of Example Hindi Data:

Lag rahā hai āj ambara, jyoan umadatā sā samandara.
uṭh chalī chanchal hilore, ā rahe jaise bavaṇḍara..

Translation of Example Hindi Data:

Looks like the sky today, as a rising sea.
Agitated tremors got up, like a tornado coming.

Gujarati data was obtained in the same way that Hindi data was, from various books and portals, along with some data which was directly acquired from the authors' previous research studies [31]. All of the standard Unicode-based data was collected, organised, and verified by specialists between December 2017 and April 2022.

Example Gujarati Data:

મોઢા પર મારવું

Transliteration of Example Gujarati Data:

Moḍhā par māravun

Translation of Example Gujarati Data:

Slap on face

B. Data Preprocessing

Following data collection, data processing takes place, which includes data preprocessing and cleaning. With the exception of the specific language's unicode range, which eventually removes unnecessary special symbols, emojis, junk characters, and so on, all unnecessary data is discarded first. Punctuation marks are also removed (Full Stop, Comma, Question Mark, Colan, Semicolon, Brackets, Exclamation Mark, Quotation Mark, Slash, and all irrelevant marks). The data is chunked into a list of words after the marks are removed and the preprocessing cleaning operation is completed. This procedure is known as Word Tokenization in the domain of NLP.

Preprocessed Hindi Data:

लग | रहा | है | आज | अम्बर | ज्यों | उमड़ता | सा | समन्दर | उठ | चली | चंचल | हिलोरे | आ | रहे | जैसे | बवंडर

Transliteration of Preprocessed Hindi Data:

Lag | rahā | hai | āj | ambara | jyoan | umadatā | sā | samandara | uṭh | chalī | chanchal | hilore | ā | rahe | jaise | bavaṇḍara

Translation of Preprocessed Hindi Data:

Looks | dwell | is | today | sky | like | surge | as | sea | rise | walk | fickle | tremors | come | stay | like | tornado

Preprocessed Gujarati Data:

મોઢા | પર | મારવું

Transliteration of Preprocessed Gujarati Data:

Moḍhā | par | māravun

Translation of Preprocessed Gujarati Data:

Slap | on | face

C. Data Filtering

The data preprocessing stage's list of words is now ready to be filtered. Let us first define filtering and its purpose. Every language has its own StopWords. Stopwords are commonly used words that have a little or no impact on the meaning of a sentence. Because this experiment uses Hindi language data, the authors used Hindi StopWords from a hybrid StopWords-based research of Hindi [32] and Gujarati[33] languages. StopWords from a specific language must be used in this stage to filter based on the language chosen for the implementation of this mechanism. StopWords were filtered while applying filtering to the list of words produced by the data processing stage.

Found Hindi Stop Words:

रहा | है | आज | रहे | जैसे

Transliteration of Hindi Stop Words:

rahā | hai | āj | rahe | jaise

Translation of Hindi Stop Words:

dwell | is | today | stay | like

Filtered Hindi Data:

लग | अम्बर | ज्यों | उमड़ता | सा | समन्दर | उठ | चली | चंचल | हिलोरे | आ | बवंडर

Transliteration of Filtered Hindi Data:

Lag | ambara | jyoan | umadatā | sā | samandara | uṭh | chalī | chanchal | hilore | ā | bavaṇḍara

Translation of Filtered Hindi Data:

Looks | sky | like | surge | as | sea | rise | walk | fickle | tremors | come | tornado

For the given Hindi example, five Hindi StopWords were filtered.

Found Gujarati Stop Word:

પર

Transliteration of Found Gujarati Stop Word:

par

Translation of Found Gujarati Stop Word:

on

Filtered Gujarati Data:

મોઢા | મારવું

Transliteration of Filtered Gujarati Data:

Moḍhā | māravun

Translation of Filtered Gujarati Data:

Slap | face

For the given Gujarati example, a Gujarati StopWord was filtered. Data filtering is critical because it will assist us in reducing computational processing in subsequent stages. The remaining words will now go through the logging process.

D. Data Logging

The act of keeping a record of something is referred to as logging. Words from the filtered data list are checked in WordNet one by one at this stage. If a word is already in the WordNet, one can access all of the relevant properties of that word, such as Synsets, Synonyms, POS tags, Gloss, Example statements, and much more relevant information provided by the specific WordNet.

Now comes the crucial part of this research work: if any of the words are not found while checking the word's existence in the WordNet, some WordNets may produce an information message, while others may generate an error through exceptions. Such cases must be handled properly, and any words that are not found in WordNet must be logged and kept in a list of not found words.

The following words were not found in WordNet for the filter data from the data filtering stage.

Logged Hindi Data:

लग | ज्यों | उमड़ता | सा | उठ | चली | चंचल | हिलोरे

Transliteration of Logged Hindi Data:

Lag | jyoan | umadatā | sā | uṭh | chalī | chanchal | hilore

Translation of Logged Hindi Data:

Looks | like | surge | as | rise | walk | fickle | tremors

Logged Gujarati Data:

મોઢા

Transliteration of Logged Gujarati Data:

Moḍhā

Translation of Logged Gujarati Data:

face

This is a continuous process that can be repeated whenever a word is not found in WordNet. This entire mechanism can be embedded in any system that uses WordNet. All that remains is to integrate these various stages and keep track of words that are not found in WordNet while searching for various types of uses.

E. Data Rechecking

This is an optional step in reprocessing the logged data by rechecking it against WordNet. If this mechanism is integrated with any other system that uses WordNet, the logging stage will produce a large list of words that were not found over time.

Now, that list may contain some words that have already been added to WordNet in a later release, so to avoid those words from the logged data, the entire data can be rechecked with the most recent updated WordNet, and any duplicate words that have already been added to WordNet will be removed. Similarly, logged data can be filtered again with the updated StopWords if necessary.

The logged Hindi and Gujarati data example has been rechecked, and one word has been added in Hindi WordNet in the most recent update, so it has been removed from the data. Gujarati data remains as it is as the word is still not added in the Gujarati WordNet.

Rechecked Hindi Data:

लग	ज्यों	उमड़ता	सा	उठ	चली	हिलोरे

Transliteration of Logged Hindi Data:

Lag	jyoan	umadatā	sā	uṭh	chalī	hilore

Translation of Logged Hindi Data:

Looks	like	surge	as	rise	walk	tremors

Rechecked Gujarati Data:

મોઢા

Transliteration of Rechecked Gujarati Data:

Moḍhā

Translation of Rechecked Gujarati Data:

face

As a result, the data rechecking stage is for final checks before producing a clean list of data that are not available in WordNet. Let's take a look at the overall results of this study.

F. Discussion

If this mechanism is applied and followed in a systematic manner, this research will undoubtedly help to strengthen the various WordNet-based research works. This mechanism is useful for almost all WordNet-based projects, regardless of language. It was purposefully tested with Hindi language based poetic data and Gujarati idiom based data to determine its effectiveness because processing poems and idioms differs slightly from processing prose. That is the only explanation for such extraordinary results. There could be several reasons for the large number of words that were not found. It is possible to mention a few borrowed words, missing words, newly developed terminology, and combined or misspelt words. Because Hindi and Gujarati have so many dialects, there are more chances of borrowing words from neighbouring and sister languages.

IV. Results

This research was carried out over a long period of time, between December 2017 and April 2022. During this time, 5011 Unicode Hindi Verse-based poetic literature data and 3472 Gujarati idioms data were processed using this research methodology through all of the stages III-A. Data Collection, III-B. Data Preprocessing, III-C. Data Filtering, III-D. Data Logging, III-E. Data Rechecking as described in the III. Methodology section. Table I. Overall Results representing the different stats of current research work. As a result, the authors were able to populate a list [34] of 13,593 Hindi words which are not available in the Hindi WordNet section of Indian Languages WordNet (IndoWordNet). The total number of Hindi words in the IndoWordNet is 1,05,458, but through this research work, 13,953 new potential words were discovered, accounting for nearly 13.23 percent of the total number of words. In addition, 887 Gujarati words[35] were discovered while analysing 3,472 Gujarati idioms, accounting for 1.38 percent of the existing 64,300 Gujarati words in WordNet. This many words are more than enough to demonstrate the utility of this mechanism. Because there haven't been any similar research and datasets based on the Hindi and Gujarati languages, benchmarking isn't possible at this moment. Due to the limitation that there are currently no other comparable datasets available for both Hindi and Gujarati, future results may vary depending on the availability. If used, the current research methodology can yield significant results for various languages as well. This mechanism will undoubtedly aid in the improvement of WordNets in various languages around the world. While using this mechanism with other languages, the results may vary, but it will significantly improve any WordNet.

TABLE I. Overall Results

Sr. No.	**1**	**2**
Language	Hindi	Gujarati
Existing Words in IndoWordNet	105458	64300
Absent Words in IndoWordNet	13953	887
Absent Words %	13.23%	1.38%

V. Conclusion

To summarise, continual efforts are always required to strengthen the WordNets in order to accommodate freshly produced terms, loanwords, and missing words. Dealing with WordNets of languages with several dialects is difficult.Maintaining a log of words not found in WordNet while using any WordNet is usually beneficial. Later, processing those words with all of the WordNet's accessible words will yield a potential list of words. Such lists may be evaluated for inclusion in WordNet. On such words, the regular techniques for adding words to any language's WordNet can be used.

If these words are still not related to the specific language, they may have been borrowed from another language. In that scenario, these words can be cross-checked with WordNets from neighbouring languages from which they may have borrowed. For the continuing enhancement and strengthening of WordNet research projects, such methods and mechanisms can be included with practically every language's WordNet.

VI. Future Enhancement

The list of populated words after the data rechecking phase can be checked with the nearby and sister languages WordNets to identify the words in case some of the borrowed words

belong to some nearby languages for future enhancement. Other studies, such as joint words or language detection related research, studies can also be conducted to check for missspelt and incorrect words.

References

[1] Ethnologue.com How many languages are there in the world?. *Ethnologue.com*. (2022), https://www.ethnologue.com/guides/how-many-languages

[2] Contributors, W. Hindi - Wikipedia. *Wikipedia.org*. (2021), https://en.wikipedia.org/wiki/Hindi

[3] Contributors, W. Devanagari - Wikipedia. *Wikipedia.org*. (2021), https://en.wikipedia.org/wiki/Devanagari

[4] Unicode, I. Devanagari - The Unicode Standard. *The Unicode Standard,*. **14** (2021), https://unicode.org/charts/PDF/U0900.pdf

[5] Contributors, W. Gujarati - Wikipedia. *Wikipedia.org*. (2021), https://en.wikipedia.org/wiki/Gujarati_language

[6] Unicode, I. Gujarati - The Unicode Standard. *The Unicode Standard,*. **14** (2021), https://unicode.org/charts/PDF/U0A80.pdf

[7] Audichya, M.K. & Saini, J.R. A Study to Recognize Printed Gujarati Characters Using Tesseract OCR. *Engineering, Technology And Applied Science Research*. **5** pp. 1505-1510 (2017,9)

[8] Modh, J.C. & Saini, J.R. Context Based MTS for Translating Gujarati Trigram and Bigram Idioms to English. *2020 International Conference For Emerging Technology (INCET)*. pp. 1-6 (2020), https://doi.org/10.1109/INCET49848.2020.9154112

[9] Contributors, W. Natural Language Processing - Wikipedia. *Wikipedia.org*. (2022), https://en.wikipedia.org/wiki/Natural_language_processing

[10] Contributors, W. Computer Science - Wikipedia. *Wikipedia.org*. (2022), https://en.wikipedia.org/wiki/Computer_science

[11] Contributors, W. Computational Linguistics - Wikipedia. *Wikipedia.org*. (2022), https://en.wikipedia.org/wiki/Computational_linguistics

[12] Contributors, W. Artificial Intelligence - Wikipedia. *Wikipedia.org*. (2022), https://en.wikipedia.org/wiki/Artificial_intelligence

[13] Miller, G. WordNet: A Lexical Database for English. *Commun. ACM*. **38**, 39-41 (1995,11), https://doi.org/10.1145/219717.219748

[14] Miller, G. WordNet: An electronic lexical database. (MIT press,1998)

[15] Vossen, P. EuroWordNet: A multilingual database with lexical semantic networks. *Dordrecht: Kluwer Academic Publishers.*. **10** (1998), https://link.springer.com/book/10.1007/978-94-017-1491-4

[16] Narayan, D., Chakrabarti, D., Pande, P. & Bhattacharyya, P. An experience in building the indo wordnet-a wordnet for hindi. *First International Conference On Global WordNet, Mysore, India*. **24** (2002), https://www.academia.edu/314054/

[17] Bhattacharyya, P. IndoWordNet. *Proceedings Of The Seventh International Conference On Language Resources And Evaluation (LREC'10)*. (2010,5), http://www.lrec-conf.org/proceedings/lrec2010/pdf/939_Paper.pdf

[18] 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 (2015,12), https://aclanthology.org/W15-5910

[19] Panjwani, R., Kanojia, D. & Bhattacharyya, P. pyiwn: A Python based API to access Indian Language WordNets. *Proceedings Of The 9th Global Wordnet Conference*. pp. 378-383 (2018,1), https://aclanthology.org/2018.gwc-1.47

[20] McCrae, J., Rademaker, A., Rudnicka, E. & Bond, F. English WordNet 2020: Improving and Extending a WordNet for English using an Open-Source Methodology. *Proceedings Of The LREC 2020 Workshop On Multimodal Wordnets (MMW2020)*. pp. 14-19 (2020,5), https://aclanthology.org/2020.mmw-1.3

[21] Kanojia, D., Patel, K. & Bhattacharyya, P. Indian Language Wordnets and their Linkages with Princeton WordNet. (arXiv,2022), https://doi.org/10.48550/arxiv.2201.02977

[22] Fellbaum, C. WordNet. *Theory And Applications Of Ontology: Computer Applications*. pp. 231-243 (2010), https://doi.org/10.1007/978-90-481-8847-5_10

[23] Bhensdadia, C., Bhatt, B. & Bhattacharyya, P. Introduction to Gujarati wordnet. *Third National Workshop On Indowordnet Proceedings*. **494** (2010)

[24] Vaishnav, Z. & Sajja, P. Knowledge-based approach for word sense disambiguation using genetic algorithm for Gujarati. *Smart Innovation, Systems And Technologies*. **106** pp. 485-494 (2019), http://dx.doi.org/10.1007/978-981-13-1742-2_48

[25] Modh, J. & Saini, J. Using IndoWordNet for Contextually Improved Machine Translation of Gujarati Idioms. *International Journal Of Advanced Computer Science And Applications (IJACSA)*. **12**, pp. 225-232 (2021), http://dx.doi.org/10.14569/IJACSA.2021.0120128

[26] Audichya, M.K. & Saini, J.R. Computational linguistic prosody rule-based unified technique for automatic metadata generation for Hindi poetry. *2019 1st International Conference On Advances In Information Technology (ICAIT)*. pp. 436-442 (2019,7), https://ieeexplore.ieee.org/document/8987239/

[27] Audichya, M.K. & Saini, J.R. Stanza Type Identification using Systematization of Versification System of Hindi Poetry. *International Journal Of Advanced Computer Science And Applications*. **12**, pp. 142-153 (2021), https://dx.doi.org/10.14569/IJACSA.2021.0120117

[28] Audichya, M.K. & Saini, J.R. Towards Natural Language Processing with Figures of Speech in Hindi Poetry. *International Journal Of Advanced Computer Science And Applications*. **12**, pp. 128-133 (2021), https://dx.doi.org/10.14569/IJACSA.2021.0120316

[29] Modh, J.C. & Saini, J.R. Using IndoWordNet for Contextually Improved Machine Translation of Gujarati Idioms. *International Journal Of Advanced Computer Science And Applications*. **12**, pp. 225-232 (2021), http://dx.doi.org/10.14569/IJACSA.2021.0120128

[30] Audichya, M.K. & Saini, J. R. Hindi Verse Dataset. *Mendeley Data*. (2022), https://data.mendeley.com/datasets/cp6htsbbpp/1

[31] Saini, J.R. & Modh, J.C. GIdTra: A dictionary-based MTS for translating Gujarati bigram idioms to English. *2016 Fourth International Conference On Parallel, Distributed And Grid Computing (PDGC)*. pp. 192-196 (2016)

[32] Jha, V., Manjunath, N., Deepa Shenoy, P. & R, V. Hindi Language Stop Words List. *Mendeley Data*. **V1** (2018), https://data.mendeley.com/datasets/bsr3frvvjc/1

[33] Quanteda Initiative. A List 210 Gujarati Stop Words. *Github.com*. (2022), https://github.com/quanteda/stopwords/issues/11

[34] Audichya, M.K. & Saini, J.R. Additional Hindi Words for IndoWordNet. *Mendeley Data*. (2022), https://data.mendeley.com/datasets/db8sh8js67/1

[35] Audichya, M.K., Saini, J.R. & Modh J.C. Additional Gujarati Words for IndoWordNet. *Mendeley Data*. (2022), https://data.mendeley.com/datasets/3jtm7htsyt/1