Qualitative Enhancement of Konkani WordNet through Phonetic Transcription and Concept Image Integration

Anonymous ACL submission

Abstract

Konkani WordNet (also known as Konkani Shabdamalem) was developed as part of the Indradhanush WordNet Project Consortium between August 2010 and October 2013. As of now, the Konkani WordNet includes approximately 32,370 synsets and 37,719 unique words. There is a need to enhance the resource both quantitatively and qualitatively. In this paper, we describe the addition of concept images and phonetic representations to each synset in the Konkani WordNet. Our work has produced an enriched version of the WordNet, which facilitates a better understanding of Konkani concepts. This feature is accessible via the project website. The same approach can be applied to enhance other regional language WordNets.

Keywords— Konkani Wordnet , Concept Image , Pronunciation

1 Introduction

The first WordNet, developed at Princeton University known as Princeton Wordnet [1], was created for the English language. This was followed by the development of WordNets for several European languages, notably through the EuroWordNet project (Vossen, 1999). Since 2000, efforts to build WordNets for Indian languages have gained momentum, beginning with the Hindi WordNet [2], developed at the Indian Institute of Technology Bombay (IITB). These efforts have since expanded to include WordNets for many other Indian languages. BabelNet (Navigli and Ponzetto, 2012) is a multilingual lexical database that integrates concepts from WordNet, Wikipedia, and other resources to provide a comprehensive semantic network.

Konkani (Wikipedia, 2024) is one of the 22 languages listed in the Eighth Schedule of the Indian Constitution and is the official language of Goa. It is an Indo-Aryan language evolved from Sanskrit through Prakrit, and has been influenced by several languages, including Marathi, Kannada, Malayalam, Hindi, Portuguese, and English. While Devanagari is the officially recognized script, Konkani is also written in Roman and Kannada scripts.

According to the 2011 Census of India, only 0.19% of the country's population speaks Konkani. Between 2001 and 2011, the number of Konkani speakers declined by 9.34%. The language comprises several dialects such as Antruzi, Bardeskari, Saxxtti, Canconi, and Pednekari shaped by regional, religious, and sociolinguistic factors, as well as local language influences (Goa365, 2018).

With advances in machine learning, especially deep learning, the role of lexical resources like WordNet has become increasingly important in semantic understanding. The creation of the Konkani WordNet (Walawalikar et al., 2010) marked a significant milestone, supported by tools including a dedicated WordNet database (Prabhu et al., 2012), a Concept Merging Tool (Nagvenkar et al., 2014), and an Application Programming Interface (API) (Prabhugaonkar et al., 2012). Despite this progress, the Konkani WordNet still requires both qualitative and quantitative enrichment. A corpus-based enhancement was recently undertaken using crowdsourcing, resulting in the Konkani Shabdarth corpus, which introduced 71 new synsets and 21 additional unique words (Manerkar et al., 2022). Subsequent efforts added pronunciation features to each word leading to Shabhocchar Corpus (Gawde et al., 2024b), and a visual representation of the WordNet through a WordNet Visualizer (Gawde et al., 2024a), aiding in language teaching and learning.

To the best of our knowledge, no prior work has attempted to enhance the Konkani WordNet by incorporating concept images. In this paper, we present Shabda-Chitra, a qualitative enhancement wherein concept images are added to each synset in the Konkani WordNet. Additionally, we extend the phonetics feature to all words in the resource.

This paper is organised as follows - section 2 briefly introduces the Konkani WordNet and its features, section 3 describes the proposed methedology for adding Phonetics Module i.e. explained in section 3.1 and Concept Image Module which is explained in section 3.2. Section 4 presents the future scope and conclusion.

2 Konkani WordNet

WordNet (Miller, 1995) is essentially a large, graph-based structure of words. It functions as an electronic lexical database and serves as a valuable resource for researchers in computational linguistics, text processing, and related natural language processing (NLP) tasks.

The Konkani WordNet (Walawalikar et al., 2010; Desai et al., 2017) was developed at Goa University as part of the Indradhanush WordNet Consortium

[1] http://www.wordnet.princeton.edu/

[2] https://www.cfilt.iitb.ac.in/wordnet/webhwn/

Project, funded by the Technology Development for Indian Languages (TDIL) program under the Department of Electronics and Information Technology (DeitY), Ministry of Electronics and Information Technology (MeitY). In this initiative, WordNets for seven Indian languages Bengali, Gujarati, Kashmiri, Konkani, Odia, Punjabi, and Urdu were constructed using the expansion approach, with the Hindi WordNet (Jha et al., 2001; Narayan et al., 2002) serving as the source. These WordNets were later integrated into IndoWordNet (Bhattacharyya, 2010).

Konkani WordNet is organized as a collection of concepts, known as synsets, which are linked through lexical and semantic relationships such as synonymy, antonymy, hypernymy-hyponymy, meronymy-holonymy, and ontological links. Each synset includes a concept definition and example sentences that illustrate the use of the associated words. Synsets are also annotated with their corresponding part-of-speech (POS) categories.

Further enhancements include the addition of pronunciation audio files for synset words, allowing users to hear the correct phonetic rendering of the synset words. Additionally, a WordNet Visualizer was developed to present synset relationships in a more intuitive and visually engaging format. These features collectively support improved accessibility and usability of the Konkani WordNet for both researchers and language learners. The table 1 shows the distribution of synsets over the POS category.

POS Category	**Synset Count**
Noun	23144
Verbs	3000
Adjectives	5744
Adverbs	482
Total synsets	**32370**

Table 1: POS Category-wise break-up of Konkani WordNet Synsets

3 Methodology

For the phonetic component of our work, we utilized the Konkani Raw Speech Corpus released by CIIL (2019). From this dataset, 906 text-audio samples were selected, specifically from the Contemporary Text and Creative Text categories. Each sample included an audio recording (spoken Konkani), transcribed text in Devanagari script, a Romanized transliteration (in ITRANS-style), and associated metadata such as speaker age and gender.

We extracted word-level alignments and created two parallel text files one containing Devanagari words and the other their Romanized equivalents. A custom sequence-to-sequence (Seq2Seq) model was developed using PyTorch, featuring an encoder-decoder architecture based on Gated Recurrent Units (GRUs). Post-processing was performed using a rule-based converter to transform the ITRANS-style Roman output into a simplified phonetic transcription format. In addition to the phonetic enhancement, we integrated concept images into the Konkani WordNet interface to improve visual representation. This was facilitated by leveraging the structural alignment between Hindi and Konkani WordNets, wherein each Konkani synset is mapped to a corresponding Hindi synset through shared synset IDs in the IndoWordNet framework. Using the publicly available Shabdamitra platform originally developed for Hindi we retrieved concept-linked images. The integration was achieved programmatically by querying the Shabdamitra API with Hindi synset IDs, extracting the image URLs, and rendering them within the Konkani WordNet interface.

3.1 Phonetics Module

The Konkani Raw Speech Corpus (Ramamoorthy et al., 2019; Choudhary et al., 2019), released by the Central Institute of Indian Languages (CIIL), Mysore in 2019, is a high-quality resource created to support speech processing and phonetic research in the Konkani language. This corpus contains approximately 156.6 hours of audio data from 504 native Konkani speakers (267 female and 237 male), spanning regions such as North Goa, South Goa, Karwar, and Sindhudurg. The recordings were captured at 48 kHz/16-bit resolution using a 24-bit Linear PCM recorder in stereo format. The data spans various speech domains, including news reading, creative texts, conversations, commands, and isolated words. It serves as a robust foundation for tasks such as Automatic Speech Recognition (ASR), Text-to-Speech (TTS), and phonetic analysis.

For this study, we focused on the phonetic modeling of Konkani using a subset of the corpus containing recordings from the Contemporary Text and Creative Text categories. Each recording was accompanied by a transcription in Devanagari script, its Romanized transliteration (in ITRANS-style), and metadata such as speaker age group, gender, education level, dialect, and recording conditions. From this subset, we extracted 906 audio-text pairs and processed them into two aligned text files: one containing Devanagari script words, and the other containing their corresponding Roman transliterations, aligned line-by-line at the word level.

These aligned pairs were used as input to train a sequence-to-sequence (Seq2Seq) model aimed at automatic phonetic transliteration. The objective was to model the mapping between Devanagari script and its phonetic equivalent in Romanized form, facilitating a deeper analysis of sound-symbol relationships in Konkani.

We initially experimented with the OpenNMT toolkit, which offered a strong baseline for neural sequence modeling. However, to enable more fine-grained control over the architecture and hyperparameters, we implemented a custom Seq2Seq model in PyTorch. The model follows an encoder-decoder design, with both components based on Gated Recurrent Units (GRUs). The encoder transforms the Devanagari input sequence into a context vector, which the decoder then uses to generate the corresponding Romanized output in ITRANS-style. Fig 1 explains the system architechture of our methedology.

A final post-processing step converts the ITRANS-style Romanized text into a Simplified Phonemic Transcription using a rule-based converter (e.g., hello → heh · loh). This module thus enables efficient phonetic modeling and transliteration of Konkani script, supporting further applications in language processing and pedagogy. One of the primary challenges in training the transliteration model was the lack of a large-scale parallel Devanagari–ITRANS dataset. This limited the model's exposure to diverse phonetic variations, reducing its generalization capability. Additionally, some text files in the corpus contained incomplete or corrupt

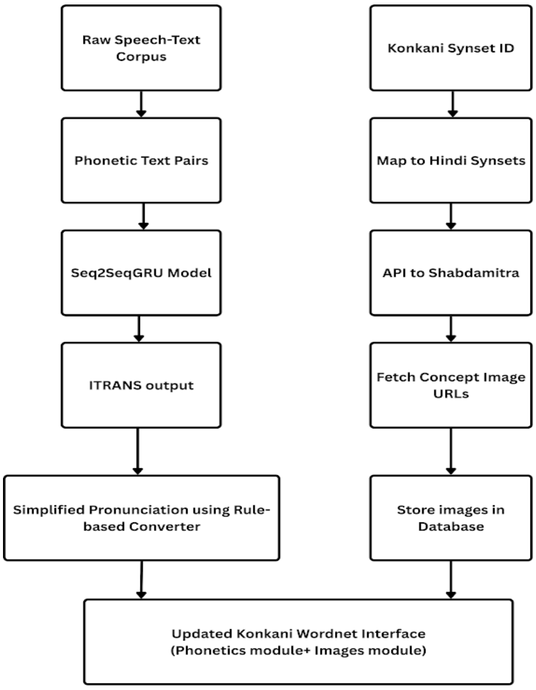


Figure 1: System Architecture

entries, which required manual cleaning and filtering before use.

Another significant issue was the inconsistency in spelling, particularly in handling nasalized sounds and conjunct consonants. These inconsistencies led to parsing errors and introduced noise into the training process.

Given the relatively small size of the dataset, the model initially exhibited signs of overfitting. This was addressed by applying dropout regularization during training to improve generalization. Furthermore, since no pretrained transliteration models existed for Konkani or the ITRANS format, the model had to be trained entirely from scratch, increasing both training time and the need for careful architecture tuning.

ITRANS is case sensitive which requires careful tokenization and post processing to avoid false predictions due to classification errors. Table 2 presents the output of the transliteration phase.

Konkani(Devanagari)	Transliteration
एका	EkA
गांवांत	gAMvAMta
रामू	rAmU

Table 2: Konkani Words with Roman Transliteration

To handle different user groups, we provided two types of phonetic notation for each Konkani word. For linguists and researchers, we offer the ITRANS notation which maintains high phonetic accuracy and also follows a standardized transliteration for computational and linguistic analysis. For general public, we generate a simplified pronunciation format, which is derived from ITRANS using our custom rule-based converter that maps complex phonetic symbols into more intuitive, readable forms.This dual representation ensures that both expert users and everyday speakers can benefit; researchers gain precise phonological structure , while native users can easily understand and pronounce the words. The mapping reference was taken from ITRANS[3]. Table 3 shows some examples of results after simplified transliteration phase.

Word	ITRANS	Simplified
एका	EkA	ay · kaa
गांवांत	gAMvAMta	gaaM · vaant
रामू	rAmU	raa · moo
शाळा	shAla	shaah · laa

Table 3: Sample Konkani Words with ITRANS and Simplified Transliteration

3.2 Concept Image Module

In addition to the phonetic enhancement, we developed a visual enrichment module that adds concept images into the Konkani WordNet. The primary objective of this module is to associate lexical concepts with relevant images to enhance user interaction, support intuitive concept understanding, and aid language learning.

To implement this feature, we leveraged the existing Hindi Shabdamitra (Redkar et al., 2017) platform, which hosts a curated set of concept images linked to synsets in the Indradhanush Hindi WordNet. Since Konkani synsets in IndoWordNet are already mapped to their corresponding Hindi synsets via shared synset IDs, this existing linkage was utilized as a bridge for image retrieval.

The integration process involved programmatically calling the Shabdamitra Image API. For each Konkani synset, its corresponding Hindi synset ID was identified, and an API request was made using this ID to fetch the associated image. The image URLs returned by the API were then dynamically rendered on the Konkani WordNet interface. This process was scaled across the entire dataset, involving iteration over approximately 32,000 Konkani synsets with parallel API requests to ensure efficient image retrieval and rendering.

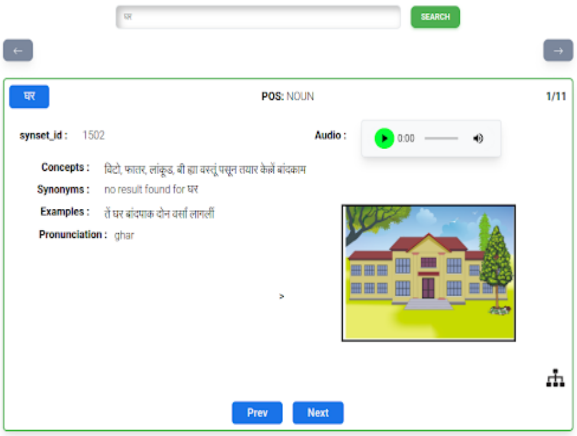
3.3 Enriched Wordnet Website

We redesigned the Konkani WordNet website by incorporating two new functional modules: the pronunciation module and the concept image module. The updated interface is more visually appealing and user-friendly, with an emphasis on improving accessibility and engagement. This redesign aims to enhance the overall user experience, making the resource more effective for language learners, educators, and researchers. Fig 2 shows the screenshot of the Konkani WordNet in which Pronunciation and concept image is added. The word in the example is घर.

4 Conclusion and Future Scope

In this work, we presented qualitative enhancements to the Konkani WordNet through the integration of two key modules: phonetic transliteration and concept image association. The phonetics module utilized a

[3] https://www.aczoom.com/itrans



POS: NOUN

synset_id : 1502

Audio :

Concepts : चिटो, फातर, लांकूड, बी ह्या वस्तूं पसून तयार केल्लें बांदकाम

Synonyms : no result found for घर

Examples : तें घर बांदपाक दोन वर्सां लागलीं

Pronunciation : ghar

Figure 2: Konkani WordNet User Interface with Concept Image and Pronunciation

custom sequence-to-sequence model trained on parallel Devanagari Romanized data, enabling accurate phonetic representations in simplified transcription. The image module leveraged the IndoWordNet linkage with Hindi synsets and the Shabdamitra platform to enrich synsets with concept-level visualizations. These additions not only improve the usability and accessibility of the Konkani WordNet but also support language learning and cognitive understanding of lexical concepts.

The redesigned Konkani WordNet interface now offers a more interactive and learner-friendly experience, aimed at a broader user base including educators, students, and NLP researchers.

For future work, the phonetic module can be extended by incorporating dialectal variations and prosodic features to better reflect spoken diversity. Similarly, the image module can be expanded to support multilingual concept alignment and crowdsourced image validation to improve coverage and accuracy. Finally, the integration of these features into downstream NLP applications such as speech synthesis, transliteration tools, and language teaching platforms remains a promising direction for further development.

Limitations

While the proposed enhancements to the Konkani WordNet significantly improve its usability and accessibility, the work is subject to certain limitations. First, the concept images integrated via the Shabdamitra platform are based on mappings from Hindi synsets and have not been manually validated for semantic or cultural relevance in the Konkani context. This may lead to occasional mismatches between the image and the intended concept in Konkani. Second, the phonetic transcriptions generated through the custom Seq2Seq model have not undergone manual or expert linguistic validation. As a result, there may be errors in pronunciation representations, particularly for infrequent words or those influenced by regional dialects.

References

Pushpak Bhattacharyya. 2010. IndoWordNet. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Narayan Choudhary, N. Rajesha, G. Manasa, and L. Ramamoorthy. 2019. Ldc-il raw speech corpora: An overview. In *Linguistic Resources for AI/NLP in Indian Languages*, pages 160–174. Central Institute of Indian Languages, Mysore.

Shilpa N Desai, Shantaram W Walawalikar, Ramdas N Karmali, and Jyoti D Pawar. 2017. Insights on the konkani wordnet development process. *The WordNet in Indian Languages*, pages 101–117.

Sunayana Gawde, Jayram Gawas, Shrikrishna Parab, Shilpa Desai, and Jyoti D Pawar. 2024a. Konkani wordnet visualizer as a concept teaching-learning tool. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 59–67.

Sunayana Gawde, Shrikrishna Parab, Jayram Gawas, Shilpa Desai, and Jyoti D Pawar. 2024b. Shabdocchar: Konkani wordnet enrichment with audio feature. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 531–536.

Goa365. 2018. Are konkani speakers declining? Available at: `https://tinyurl.com/ywkytpp8`.

S. Jha, D. Narayan, P. Pande, and P.A. Bhattacharyya. 2001. Wordnet for hindi. In *Proceedings of the International Workshop on Lexical Resources in Natural Language Processing*, Hyderabad.

Sanjana Manerkar, Kavita Asnani, Preeti Ravindranath Khorjuvenkar, Shilpa Desai, and Jyoti D Pawar. 2022. Konkani wordnet: Corpus-based enhancement using crowdsourcing. *Transactions on Asian and Low-Resource Language information Processing*, 21(4):1–18.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41.

Apurva Nagvenkar, Neha Prabhugaonkar, Venkatesh Prabhu, Ramdas Karmali, and Jyoti Pawar. 2014. Concept space synset manager tool. In *Proceedings of the Seventh Global Wordnet Conference*, pages 86–94.

Dipak Narayan, Debasri Chakrabarti, Prabhakar Pande, and Pushpak Bhattacharyya. 2002. An experience in building the indo wordnet-a wordnet for hindi. In *First international conference on global WordNet, Mysore, India*, volume 24.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnetxplorer: a platform for multilingual lexical knowledge base access and exploration. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, page 393–396, New York, NY, USA. Association for Computing Machinery.

Venkatesh Prabhu, Shilpa Desai, Hanumant Redkar, Neha Prabhugaonkar, Apurva Nagvenkar, and Ramdas Karmali. 2012. An efficient database design for indowordnet development using hybrid approach. In *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing*, pages 229–236.

Neha Prabhugaonkar, Apurva Nagvenkar, and Ramdas Karmali. 2012. Indowordnet application programming interfaces. In *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing*, pages 237–244.

L. Ramamoorthy, Narayan Choudhary, Saurabh Varik, and Rashmi Shet Tanawade. 2019. Konkani raw speech corpus.

Hanumant Redkar, Sandhya Singh, Meenakshi Somasundaram, Dhara Gorasia, Malhar Kulkarni, and Pushpak Bhattacharyya. 2017. Hindi shabdamitra: A Wordnet based E-learning tool for language learning and teaching. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, pages 23–28, Taipei, Taiwan. Asian Federation of Natural Language Processing.

PJTM Vossen. 1999. Eurowordnet.

Shantaram Walawalikar, Shilpa Desai, Ramdas Karmali, Sushant Naik, Damodar Ghanekar, Chandralekha D'Souza, and JD Pawar. 2010. Experiences in building the konkani wordnet using the expansion approach. *5th Global WordNet Conference on Principles, Construction and Application of Multilingual WordNets*.

Wikipedia. 2024. Konkani language. `https://en.wikipedia.org/wiki/Konkani_language`. Retrieved from `https://en.wikipedia.org/wiki/Konkani_language`.