Konkani Wordnet Resources

Hanumant Redkar[1], Mahadev Gawas[2], Anjali Desai[1], Jyoti Pawar[1]
[1]Discipline of Computer Science and Technology, Goa Business School, Goa University, India
[2]State Higher Education Council, Directorate of Higher Education, Govt. of Goa, India
[1]hanumantredkar@unigoa.ac.in, [2]gawas-dhe.goa@gov.in, [1]desai.anjali.ad@gmail.com, [1]jdp@unigoa.ac.in

Abstract

Konkani is a low-resource Indo-Aryan language spoken along the western coast of India, characterized by significant dialectal variation, multi-script usage, and limited standardized computational resources. This paper presents a consolidated and analysis-ready lexical resource derived from the Konkani Wordnet, built under the IndoWordNet framework. The resource comprises **32,370** synsets, 37,719 unique lexical entries, 32,370 glosses, and 33,318 example sentences, enriched with pronunciations, semantic relations, and illustrative examples. We describe the systematic extraction, normalization, and structural integration of wordnet data, resolving identifier inconsistencies and ensuring semantic coherence across distributed lexical files. To demonstrate the practical utility of this resource, we present an API-based bilingual vocabulary exercise generation system that leverages shared synset identifiers to automatically produce semantically aligned Hindi–Konkani word pairs for e-learning applications. The resulting resource enhances accessibility, reproducibility, and computational readiness for NLP tasks, while providing a foundational infrastructure for developing technology-driven teaching and learning tools for Konkani.

Konkani, Language Resources, Konkani Wordnet, Wordnet, IndoWordNet, Low-Resource Language, Synset, Lexical Resources, Corpus, KWN

1. Introduction

The expansion of digital learning platforms has increased the need for scalable tools that support multilingual education, particularly for Indian languages where structured learning resources are limited. Vocabulary acquisition remains central to language learning, and interactive formats such as match-the-pairs exercises are commonly used to reinforce word associations (Redkar et al., 2018).

WordNet provides a structured lexical framework in which words are organized into synsets representing language-independent concepts (Miller et al., 1990; Fellbaum, 1998). IndoWordNet (Bhattacharyya, 2010) extends this model to multiple Indian languages, including Hindi and Konkani, using a shared synset structure (Walawalikar et al., 2012; Kashyap et al., 2016). This shared conceptual layer enables cross-lingual alignment based on semantic equivalence rather than direct translation (Sarma et al., 2012; Dash et al., 2017).

This paper presents a consolidated Konkani Wordnet resource and demonstrates its application in automatically generating bilingual Hindi–Konkani match-the-pairs vocabulary exercises. The approach uses synset identifiers as a pivot: words from both languages belonging to the same synset are paired to form semantically aligned vocabulary questions. The system is implemented as an API-based module that retrieves synset–word mappings, performs cross-lingual synset matching, and generates word pairs for use in e-learning applications (Redkar et al., 2017b,a, 2018).

1.1. Contributions

The primary contributions of this work are as follows:

- **Dataset Extraction and Normalization:** A systematic consolidation of the Konkani Wordnet comprising **32,370** synsets, 37,719 unique lexical entries, 32,370 glosses, and 33,318 example sentences. The resource resolves structural redundancies, harmonizes identifier mappings, and ensures semantic consistency across distributed lexical files, making it analysis-ready for NLP tasks.
- **Cross-lingual Pairing System:** An API-based bilingual exercise generation system that uses shared IndoWordNet synset identifiers to automatically produce semantically aligned Hindi–Konkani word pairs without relying on direct translation dictionaries.
- **Educational Application Demonstration:** A working demonstration of how structured lexical databases can be operationalized into scalable, concept-based vocabulary generation systems suitable for integration with e-learning platforms and language learning applications.

2. Literature Review

Work on building lexical and computational resources for Konkani has moved in steps — first laying down basic word lists, then expanding them using real language data, creating visualization tools,

Proceedings of WILDRE-8, pages 49–54
12 May 2026.

adding new features like audio, and testing these resources in actual applications. Konkani, being a low-resource language with multiple dialects and scripts, presents challenges in standardizing vocabulary, maintaining consistency, and ensuring interoperability. Below we trace the development of Konkani WordNet within the broader context of Indian language wordnet development.

2.1. Princeton WordNet and the WordNet Model

The foundation of all modern wordnet development was laid by Miller et al. (1990) with the creation of Princeton WordNet for English. WordNet organizes lexical knowledge into synsets — groups of synonymous words representing a single underlying concept — connected through semantic relations such as hypernymy, hyponymy, meronymy, and antonymy (Fellbaum, 1998). This model proved highly influential, inspiring the development of wordnets for dozens of languages worldwide. The core insight — that meaning is best captured through conceptual groupings rather than individual word definitions — remains the guiding principle for all IndoWordNet development.

2.2. IndoWordNet and Indian Language Wordnets

Bhattacharyya (2010) introduced IndoWordNet as a multilingual lexical database linking wordnets of major Indian languages under a shared synset framework. The shared synset structure means that languages do not need direct translation dictionaries — instead, words from different languages are linked through a common concept identifier, enabling cross-lingual alignment based on semantic equivalence (Sarma et al., 2012; Dash et al., 2017).

Hindi WordNet, developed at IIT Bombay, is the most mature and largest Indian language wordnet and serves as the pivot language for IndoWordNet development (Kashyap et al., 2016). It has been used extensively for educational applications, including the Hindi Shabdamitra tool for vocabulary learning (Redkar et al., 2017b,a, 2018). Other Indian languages represented in IndoWordNet include Bengali, Marathi, Telugu, Tamil, Gujarati, Punjabi, and Assamese, each developed using either the expansion approach — translating from Hindi synsets — or the merge approach — building independently then linking. Sanskrit WordNet has also been developed and applied for educational purposes (Kulkarni et al., 2019). The breadth of IndoWordNet demonstrates both the scalability of the shared synset model and the growing importance of structured lexical resources for Indian NLP.

2.3. Building Konkani WordNet

Walawalikar et al. (2010) initiated Konkani WordNet[1] development using the expansion approach pioneered for Hindi WordNet at IIT Bombay. They systematically translated and adapted existing synsets into Konkani, producing around 1,969 core synsets. This process required careful handling of Konkani's dialectal diversity — including Goan and Saraswat varieties — multi-script usage across Devanagari, Roman, and Kannada scripts, and numerous vocabulary items without straightforward equivalents in other languages. Their work established Konkani's place within the broader IndoWordNet network and enabled data sharing with other Indian language wordnets (Desai et al., 2010, 2016). Further, Konkani Wordnet Dictionary, KWN-Dict, has been developed as one of the applications of Konkani Wordnet by (Redkar et al., 2026).

2.4. Surveying Konkani NLP Resources

Rajan et al. (2020) conducted a comprehensive survey of computational resources for Konkani, cataloguing tools such as the ILCI parallel corpus (approximately 50,000 sentences), POS taggers, morphological analyzers, and early speech datasets. Konkani WordNet was identified as a key resource, though the survey highlighted persistent gaps — small corpora, insufficient annotated data, few benchmarks, and limited integration between existing tools. The authors called for dataset standardization and stronger connections between lexical databases and real NLP applications.

2.5. Corpus-Based Enhancement

Manerkar et al. (2022) moved beyond simple synset expansion by using the Shabdarth crowdsourcing platform to identify gaps through real language data analysis. They identified 572 missing words and added 71 new synsets, improving coverage in certain domains by up to 27%. This work marked a methodological shift — from top-down list expansion to bottom-up, data-driven enrichment — and demonstrated the value of community participation in resource building for low-resource languages.

2.6. Visualization and Learning Tools

As the Konkani WordNet database grew beyond 32,000 synsets, concerns around usability and accessibility emerged. Gawde et al. (2024a) addressed this by developing a tree-based visualizer that allows users to explore semantic relationships such as hypernyms and hyponyms interac-

[1] https://konkaniwordnet.unigoa.ac.in/

tively. This tool improved the transparency of the resource, supported concept-based teaching, and helped identify structural inconsistencies within the network.

2.7. Multimodal Enrichment

The Shabdocchar project (Gawde et al., 2024b) extended Konkani WordNet beyond text by adding audio pronunciation recordings through a gamified crowdsourcing mechanism. This multimodal enrichment is particularly significant for Konkani where dialectal pronunciation differences are substantial. The audio data supports development of speech recognition and text-to-speech systems, pushing Konkani WordNet toward becoming a truly multimodal lexical resource.

2.8. Application-Focused Studies

Ghosarwadkar et al. (2024) demonstrated an NLP application using zero-shot transfer learning from Marathi for Konkani sentiment analysis, exploiting the linguistic proximity between the two languages to compensate for limited Konkani training data. Their work underscores the importance of structured lexical resources as a foundation for downstream NLP tasks and highlights the broader potential of cross-lingual transfer for low-resource Indian languages.

3. Statistical Analysis of the Consolidated Resource

3.1. Core Resource Statistics

The consolidated Konkani WordNet resource comprises **32,370** synsets and 37,719 unique lexical entries. In addition, the dataset contains 32,370 glosses and 33,318 example sentences. These figures reflect the scale of the resource following normalization, identifier harmonization, and structural integration across distributed lexical files.

Component	Count
Synsets	32,370
Unique Lexical Entries	37,719
Glosses	32,370
Example Sentences	33,318

Table 1: Statistics of Konkani Wordnet

3.2. Derived Quantitative Metrics

To assess semantic coverage and structural density, quantitative metrics were computed relative to the total number of synsets.

Gloss Coverage:

$$\frac{32{,}370}{32{,}370} \times 100 = 100\% \tag{1}$$

Every synset contains a definitional description, ensuring complete semantic coverage.

Example Coverage:

$$\frac{33{,}318}{32{,}370} \times 100 = 102.9\% \tag{2}$$

Some synsets contain more than one example sentence, resulting in a coverage ratio exceeding 100%.

Lexical Density:

$$\frac{37{,}719}{32{,}370} = 1.17 \tag{3}$$

This ratio reflects the presence of polysemy and shared lexical realizations across conceptual nodes.

Average Words per Synset:

$$\frac{55{,}530}{32{,}370} \approx 1.72 \tag{4}$$

Each synset contains 1.73 lexical items on average, indicating a semantically rich but compact lexical organization.

3.3. Sample Synset Entries

To illustrate the structure and richness of the consolidated resource, Table 2 presents representative synset entries from the Konkani Wordnet across different parts of speech.

These examples demonstrate several key properties of the resource. First, multiple synonymous words are grouped under a single synset, capturing the synonymy relation. Second, glosses provide definitional clarity in natural language. Third, example sentences ground each concept in authentic Konkani usage, supporting both language learners and NLP researchers. The structured synset representation further enables bilingual applications such as the exercise generation system described in Section 4.

4. Konkani Wordnet Resources

The Konkani WordNet used in this work is part of the larger IndoWordNet initiative, which has taken inspiration from the Princeton WordNet model by organizing lexical knowledge into synsets representing language-independent concepts. Each synset consists of a set of synonymous words and is linked to other synsets through semantic and lexical relations such as hypernymy, hyponymy, and meronymy. In the database schema, lexical

Synset ID	POS	Konkani Words	Gloss	Example Sentence
99	Noun	गिन्यान, ज्ञान (*ginyaana, nyaana*) - knowledge	मनाक वा विचारांक जाता अशी वस्तूंची वा विशयांची माहिती (*manaaka vaa vichaaraanka jaataa ashee vastoonchee vaa vishayaanchee maahitee*)	ताका संस्कृताचें बरें गिन्यान आसा (*taakaa sanskrutaacheM bareM ginyaana aasaa*)
86	Adjective	हजर, उपस्थीत (*hajara, upastheeta*) - present	लागीं बशिल्लें, मुखार वा लागीं आयिल्लें (*laageeM bashilleM, mukhaar vaa laageeM aayilleM*)	आज वर्गांत हजर आशिल्ले विद्यार्थी उणें आसले (*aaja vargaanta hajara aashille vidhyaarthee uNeM aasale*)
2078	Verb	पियेवप, पिवप, घोंटप (*piyevapa, pivapa, ghoMTapa*) - to drink	पातळ द्रव, रोस, उदक, दूद, बी पदार्थ पोटांत घालप (*paataLa drava, raosa, udaka, dooda, bee padaartha poTaanta ghaalapa*)	उदक पियेवप भलायकेक बरें (*udaka piyevapa bhalaayakeka bareM*)
123	Adverb	सारकें, प्रमाणें (*saarakeM, pramaaNeM*) - like that	कोणायच्या मतान वा नदरेन (*koNaayachyaa mataana vaa nadarena*)	तो म्हजे सारकें काम करपाक सोदिना (*to mhaje saarakeM kaama karapaaka sodinaa*)
28106	Adjective	बरें, उत्कृश्ट, उत्तम (*bareM, utkrushTa, uttama*)	जो बरे तरेन परिणामाच्या रुपान आसा वा येता असो (*jo bare tarena pariNaamaachyaa rupaana aasaa vaa yetaa aso*)	हांव थंय नाशिल्लों ही बरी गजाल (*haaMva thaMya naashilloM hee bari gajaala*)

Table 2: Sample synset entries from Konkani Wordnet with Part-of-Speech information

items are stored in the `wn_word` table, while the mapping between words and concepts is maintained in the `wn_synset_words` table through the `synset_id` field (Redkar et al., 2015, 2016).

The Konkani WordNet provides structured lexical coverage across multiple parts of speech, including nouns, verbs, adjectives, and adverbs. Each word entry is associated with a unique `word_id`, and multiple words may be linked to the same `synset_id`, representing synonymy at the conceptual level.

The system operates on two IndoWordNet-compliant lexical databases, `wordnet_hindi` and `wordnet_konkani`, which follow a standardized relational schema. Both databases contain two core tables:

- `wn_word (word_id, word)` — stores lexical entries in the respective language.
- `wn_synset_words (synset_id, word_id)` — maps each word to one or more synsets representing concepts.

Cross-lingual alignment is performed by identifying synset identifiers present in both databases. These shared synsets form the semantic bridge between Hindi and Konkani. For any shared synset s, where H_s represents the set of Hindi words and K_s represents the set of Konkani words linked to s, the number of potential bilingual word pairs derived from that synset is:

$$|H_s \times K_s| \tag{5}$$

Thus, the total number of automatically generable pairs depends on the lexical coverage of the two WordNets and the extent of their synset overlap. The system is designed to operate on any dataset conforming to this schema, making it independent of specific corpus sizes and suitable for future expansion as WordNet resources grow.

5. Discussion

The proposed system highlights the usefulness of lexical knowledge bases for educational technology. By leveraging synset identifiers as language-independent concept markers, the system avoids the limitations of direct translation dictionaries, where one-to-one word mappings may ignore pol-

ysemy and contextual meaning. Instead, words are paired based on shared conceptual membership, making the learning process semantically grounded.

The modular architecture allows the system to function purely as a backend question generator, independent of any specific user interface. This separation of concerns enables integration with different e-learning platforms, learning management systems, or mobile applications.

5.1. Replicability to Other Indian Languages

A significant strength of the proposed system is its replicability across other IndoWordNet language pairs. Since all languages within the IndoWordNet framework share a common synset identifier structure, the bilingual pairing mechanism described in this work can be extended to any two IndoWordNet languages with minimal modification. For example, the same system architecture could generate vocabulary exercises for Hindi–Bengali, Hindi–Marathi, Hindi–Telugu, or Konkani–Marathi pairs simply by substituting the corresponding WordNet databases, provided they conform to the standard IndoWordNet relational schema comprising the `wn_word` and `wn_synset_words` tables.

This cross-language scalability positions the system not merely as a Konkani-specific tool but as a general-purpose vocabulary exercise generator for Indian language e-learning. As IndoWordNet resources continue to grow and improve in coverage and quality, the number of valid bilingual pairs generated by the system will increase proportionally, making the approach increasingly powerful over time. Future work may explore multi-way alignment — generating exercises across three or more languages simultaneously — further extending the educational utility of structured lexical resources in Indian language contexts.

5.2. Limitations

The system's effectiveness depends on the lexical coverage and quality of the underlying WordNet resources. Incomplete synset mappings or uneven vocabulary distribution across languages may limit the number of valid bilingual pairs. Additionally, while synset-level alignment ensures semantic equivalence at a conceptual level, differences in usage frequency, register, or regional variation are not explicitly addressed.

6. Conclusion

This paper presented a consolidated Konkani Wordnet resource and demonstrated its application in automatically generating bilingual Hindi–Konkani match-the-pairs vocabulary exercises. By using shared IndoWordNet synset identifiers as the basis for cross-lingual alignment, the system produces semantically equivalent word pairs suitable for vocabulary learning in e-learning environments. The architecture separates configuration, database access, and service logic, resulting in a modular and extensible API-based solution.

The work demonstrates how structured lexical databases can be transformed into practical educational tools that support concept-based language learning. The approach is replicable across all IndoWordNet language pairs, making it a scalable solution for Indian language vocabulary learning at large. Future extensions may include integration of additional IndoWordNet languages, incorporation of difficulty levels based on word frequency or part of speech, and the use of glosses or example sentences to further enrich learning content.

References

Pushpak Bhattacharyya. 2010. IndoWordNet. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Malta.

Niladri Sekhar Dash, Pushpak Bhattacharyya, and Jyoti D. Pawar, editors. 2017. *The WordNet in Indian Languages*. Springer Singapore.

Shilpa N. Desai, Ramdas N. Karmali, Shantaram W. Walawalikar, and Damodar Ghanekar. 2010. Tools for IndoWordNet development. In *Proceedings of the International Conference on Natural Language Processing*.

Shilpa N. Desai, Shantaram W. Walawalikar, Ramdas N. Karmali, and Jyoti D. Pawar. 2016. Insights on the Konkani WordNet development process. In *The WordNet in Indian Languages*, pages 101–117. Springer Singapore.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Sunayana R. Gawde et al. 2024a. Konkani WordNet visualizer as a concept teaching-learning tool. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*.

Sunayana R. Gawde et al. 2024b. Shabdocchar: Konkani WordNet enrichment with audio feature. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*.

Rohit M. Ghosarwadkar et al. 2024. Sentiment analysis for Konkani using zero-shot Marathi trained neural network model. In *Proceedings*

of the 21st International Conference on Natural Language Processing (ICON).

Laxmi Kashyap, Salil Rajeev Joshi, and Pushpak Bhattacharyya. 2016. Insights on Hindi WordNet coming from the IndoWordNet. In *The WordNet in Indian Languages*, pages 19–44. Springer Singapore.

Malhar Kulkarni, Nilesh Joshi, Sayali Khare, Hanumant Redkar, and Pushpak Bhattacharyya. 2019. Introduction to sanskrit shabdamitra: An educational application of sanskrit wordnet. In *Proceedings of the 6th International Sanskrit Computational Linguistics Symposium*, pages 117–133.

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, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. volume 3, pages 235–244.

Annie Rajan, Ambuja Salgaonkar, and Ramprasad Joshi. 2020. A survey of Konkani NLP resources. *Computer Science Review*, 38:100299.

Hanumant Redkar, Sudha Bhingardive, Kevin Patel, Pushpak Bhattacharyya, Neha Prabhugaonkar, Apurva Nagvenkar, and Ramdas Karmali. 2016. WWDS APIs: Application programming interfaces for efficient manipulation of world wordnet database structure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Hanumant Redkar, Mahadev Gawas, and Sherwin Pereira. 2026. Kwn-dict: An online konkani dictionary based on konkani wordnet. In *Proceedings of Research Libraries: A Source to Adapt to Digital Form Using the Indian Knowledge Systems (IKS) in Promoting Social Science Research*, pages 110–116.

Hanumant Redkar, Nilesh Joshi, Sayali Khare, Lata Popale, Malhar Kulkarni, and Pushpak Bhattacharyya. 2017a. Hindi shabdamitra: A wordnet based tool for enhancing teaching-learning process. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON 2017)*.

Hanumant Redkar, Rajita Shukla, Sandhya Singh, Jaya Saraswati, Laxmi Kashyap, Diptesh Kanojia, Preethi Jyothi, Malhar Kulkarni, and Pushpak Bhattacharyya. 2018. Hindi WordNet for language teaching: Experiences and lessons learnt. In *Proceedings of the 9th Global Wordnet Conference*, pages 314–323.

Hanumant Redkar, Sandhya Singh, Nilesh Joshi, Anupam Ghosh, and Pushpak Bhattacharyya. 2015. IndoWordNet dictionary: An online multilingual dictionary using IndoWordNet. In *Proceedings of the 12th International Conference on Natural Language Processing*, pages 71–78.

Hanumant Redkar, Sandhya Singh, Meenakshi Somasundaram, Dhara Gorasia, Malhar Kulkarni, and Pushpak Bhattacharyya. 2017b. 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.

Shikhar Kr Sarma, Dibyajyoti Sarmah, Biswajit Brahma, Himadri Bharali, Mayashree Mahanta, and Utpal Saikia. 2012. Building multilingual lexical resources using wordnets: Structure, design and implementation. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, pages 161–170.

Shantaram Walawalikar, Shilpa Desai, Ramdas Karmali, Sushant Naik, Damodar Ghanekar, Chandralekha D'Souza, and Jyoti Pawar. 2010. Experiences in building the Konkani WordNet using the expansion approach. In *Proceedings of the 5th Global WordNet Conference*, Mumbai, India. Narosa Publishing House.

Shantaram Walawalikar, Shilpa Desai, Ramdas Karmali, Sushant Naik, Damodar Ghanekar, Chandralekha D'Souza, and Jyoti Pawar. 2012. Experiences in building the Konkani WordNet using the expansion approach. In *Proceedings of the 5th Global WordNet Conference*. Narosa Publishing House.