Shabdocchar: Konkani WordNet Enrichment with Audio Feature

Sunayana Gawde
Goa Business School,
Goa University

Shrikrishna Parab
Goa Business School,
Goa University

Jayram Gawas
Vidyaapati Project Lab,
Goa University

Shilpa Desai
Fr. Agnel College,
Pilar Goa

Jyoti D. Pawar
Goa Business School,
Goa University

Abstract

Konkani WordNet, also called *Konkani Shabdamalem*, was created as part of the Indradhanush WordNet Project Consortium between August 2010 and October 2013. Currently, the Konkani WordNet includes about 32,370 synsets and 37,719 unique words. There is a need to enhance the Konkani WordNet both quantitatively as well as qualitatively. In this paper we are presenting a Game-Based Crowdsourcing approach adopted by us to add audio feature to the Konkani WordNet which has resulted in an increase in the number of users using and getting exposed to the capabilities of the Konkani WordNet to aid in the Konkani language teaching-learning process as well as for creation of resources to initiate further research. Our work presented here has resulted in the creation of an audio corpus of 37,719 unique words which we have named as ***'Shabdocchar'*** within a short time span of four months covering five dialects of Konkani. We are confident that *Shabdocchar* will prove to be a very useful resource to support future research work on Dialects of Konkani and support voice-based search of words in the wordnet. This approach can be adopted to enhance other wordnets as well.

Keywords— Konkani Wordnet, Corpus Collection Tool , Speech Corpus

1 Introduction

The first WordNet, developed at Princeton University[1], was created for the English language. This was followed by the creation of WordNets for several European languages, such as the EuroWordNet project (Vossen, 1999). Since the year 2000, efforts to build WordNets for various Indian languages have gained momentum, starting with the Hindi WordNet[2], which was developed at the Indian Institute of Technology, Bombay(IITB). These efforts have expanded to include WordNets for several 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 is one of the 22 languages listed in the Eighth Schedule of the Indian Constitution and serves as the official language of Goa. It is an Indo-European (Indo-Aryan) language that 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 for Konkani, it is also written in Roman and Kannada scripts.(Wikipedia, 2024)

According to the 2011 survey by the Census Department of India, only 0.19% of India's population speaks Konkani. From 2001 to 2011, the number of Konkani speakers dropped by 9.34%. Konkani has various spoken dialects, including Antruzi, Bardeskari, Saxxtti, Canconi and Pednekari. These dialects vary based on factors such as region, religion, caste, and the local languages of the area. (Goa365, 2018)

Advances in machine learning have led to the use of deep learning, which depends on resources like Wordnet for understanding meaning. The creation of Konkani WordNet (Walawalikar et al., 2010) has been a major step, along with developing tools like a database for WordNet (Prabhu et al., 2012), a Concept Merging Tool (Nagvenkar et al., 2014), and an Application Programming Interface (API) (Prabhugaonkar et al., 2012).

Konkani Wordnet needs to be enriched both qualitatively and quantitatively. Recently, a quantitative approach on corpus-based enhancement was explored using Crowd sourcing and Konkani Shabdarth corpus was released comprising 71 new synsets and 21 additional unique words (Manerkar et al., 2022).

To the best of our knowledge, no work has been reported which attempts to enhance the Konkani Wordnet qualitatively. In this paper we present *Shabdochhar*, a qualitative enhancement to Konkani WordNet, wherein we add pronunciations of words to the Konkani WordNet. We have also added a visualizer for semantic relationships in Konkani Wordnet, which has been presented separately.

This paper is organized as follows - section 2 briefly introduces the Konkani WordNet and its features, Section 3 describes the proposed Game-Based Approach for audio corpus creation, the implementation details of the audio corpus creation are discussed in Section 4 and Section 5 presents the Results. Section 6 presents the future scope and the conclusion is discussed in section 7.

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

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

2 Introduction to Konkani Wordnet

WordNet (Miller, 1995) is essentially a large, graph-like structure of words. It serves as an electronic lexical database and is a valuable resource for researchers in fields such as computational linguistics, text processing, and other related NLP tasks. Konkani wordnet (Walawalikar et al., 2010; Desai et al., 2017) was developed at Goa University as part of the Indradhanush WordNet Consortium Project funded by Technology Development for Indian Languages (TDIL), Department of Electronics & Information Technology (DeitY), Ministry of Electronics & Information Technology (MeitY). In this project seven wordnets were constructed for seven Indian languages namely Bengali, Gujarati, Kashmiri, Konkani, Odia, Punjabi and Urdu using the expansion approach with Hindi WordNet (Jha et al., 2001; Narayan et al., 2002) as the source. These seven wordnets were later linked to Indowordnet (Bhattacharyya, 2010). The table below shows the POS category-wise break-up of Konkani WordNet Synsets

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 Game-Based Approach for audio corpus creation

A simple web-based game was developed to allow Konkani speakers to record the pronunciations of the Konkani words. Based on the number of words recorded for 15 minutes of time; they were awarded with the score. There are around 37,719 unique Konkani words available in the Konkani WordNet with unique IDs which are further linked to the Konkani Synsets.

The recording tool is developed such that the words displayed for pronouncing and the audio recordings are linked to each other. Once the Konkani word is displayed, a user has to pronounce it clearly and accurately. Users can also listen to the recorded audio before saving and proceeding to the next word. The process goes on for 15 minutes and the users are awarded with scores based on the number of recordings they make in the given time slot.

This interesting game is developed to capture and verify user-spoken words from the Konkani WordNet. The application is built with a focus on user accessibility and audio quality control.

The figure 1 shows how an app collects word pronunciations. The Word Allocator assigns words to users based on their dialect, ensuring that a word recorded in one dialect is also recorded by users in other dialects, so each word has pronunciations in all dialects. Users record the words using a Recording Web Interface and check the audio themselves for noise and clarity in a self-validation phase. Once approved, the recording is saved to the Cloud. It then goes through a quality enhancement process, and the improved version is also saved to the cloud.

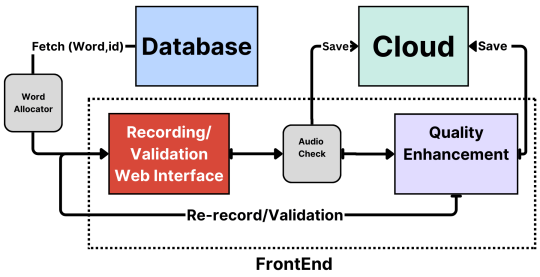

Figure 1: Game-Based Application Block Diagram

It begins with a user login screen, ensuring that only authorized users can access the recording tool. Once logged in, users are presented with a word fetched from a MySQL database, which stores the lexical data for Konkani WordNet. The word is displayed prominently on the screen which also triggers the timer allowing the user to record and listen to the audio playback for self-verification. The application runs a decibel level and noise analysis on the recorded audio to ensure it meets predefined quality thresholds.

4 Speech Collection Tool for Audio Corpus creation

The pronunciation of each word was collected with the help of Speech Collection tool/game specially developed for this task. This game was played as a trial within the team of 5 Konkani speakers and these pronunciations were validated in a later stage. Entire activity went on for two weeks with a team of 5 members recording for 15 minutes a day for a week and cross-checking the recordings for 10 minutes a day during the consecutive week. As a result of this, pronunciations for 3,000 words out of 37,719 unique words of Wordnet were recorded and validated. With this, the working of the game was tested and collected audio files were processed and linked to the Wordnet.

For the remaining 34,719 words, we used the crowd sourcing approach where a variety of Goan Konkani speakers from different regions, varying age groups and both the genders were requested to play the game and record the pronunciations for words in the Konkani Wordnet.

4.1 Audio Recording Phase

A group activity was conducted amongst the team members having Konkani language expertise as a game which they had to play for 15 minutes. All of them were given a score at the end of the game based on the number of words they recorded. This activity was conducted as a mind-refreshing activity in the afternoon time for 15 minutes every day for 1 week. Team comprises 4 female members and 1 male member. On average, 500-600 words were recorded daily by 5 speakers with the Speech Recording App. The web app can be used on Android as well as iOS systems. The recorded audio files are further cleaned to remove any noise and make them sharp and clear.

Figure 2 shows the web app which was developed considering the user login so that we could identify the speaker for a particular pronunciation. Each time only one word from the WordNet is displayed to a user.

Figure 2: User Interface of the Game-Based Application

Figure 3: User Interface of the Audio Validation

Users can read the word and record. Additionally users can listen to the recorded pronunciation to confirm its correctness and clarity. If the recording is found incorrect or noisy, the user can record it again and only when the user is confident of the pronunciation, can submit the recording. Audio files are stored over cloud storage. System is designed in such a way that each audio file is named after the particular word's ID so that later, we could link the audio files to the unique words in the Wordnet database.

4.1.1 Audio Quality Check Module

To enhance the quality of the recorded pronunciations, an additional module was integrated into the system to enforce stringent audio quality standards. This module ensures that users cannot submit their recordings unless the audio meets predefined criteria, including appropriate pitch, clarity, acceptable decibel levels, and minimal external noise. By implementing this mechanism, the system effectively prevents the submission of low-quality recordings caused by user inattention or negligence, thereby maintaining the overall quality of the pronunciation corpus. This mechanism also discourages users from recording in unsuitable environments, particularly in the case of the crowd-sourcing method, by ensuring that recordings made in noisy or inappropriate settings do not meet the required quality standards for submission.

4.2 Audio Validation Phase

The subsequent phase involved validating the pronunciations before integrating the audio files into the WordNet. A user-friendly interface was developed for this purpose, where each unique word from the WordNet was displayed along with its corresponding audio pronunciation, as illustrated in Figure 3. Users were required to listen to the recordings and evaluate them based on correctness and clarity. Any audio files marked as incorrect were automatically added to a database queue for re-recording. These flagged recordings were subsequently re-recorded by the same team, dedicating 10 minutes daily over a week to complete the task. To ensure impartiality, the interface was designed to prevent team members from validating the recordings they had originally contributed. The remaining corpus obtained by the crowd-sourcing approach is also validated by the in-house team by using the same validation interface. This process enabled the creation of a high-quality speech corpus for the WordNet.

4.3 Team formation and roles allocation

Team consists of four females and 1 male, native Konkani speakers having fluency in the language so that the pronunciations sound natural. These speakers are chosen keeping in mind their age, gender, regional origin and religion. Having 2 female members in their 30's and 2 female members in their 20's added the variation in the pronunciations. All five speakers were selected from 5 different regions of Goa specifically known for their completely different dialect of Konkani language, namely Antruzi, Bardeskari, Saxxtti, Canconi and Pednekari. Additionally, there is a huge difference in the dialect of people from two major religions in Goa while speaking Konkani (Hindu Konkani and Kristanv Konkani), so we have chosen the speakers from both the religions to cover all the variations.

4.4 Crowdsourcing Approach for the Remaining Corpus

After successful validation of the game for recording the small section of the word corpus, the actual task was initiated by circulating the developed game among the Konkani speakers for them to play and record the pronunciation in their voices. This enabled us to collect the speech data which could be accurately linked to the Konkani Wordnet's unique concepts.

Users could register with their mobile number or email ID by providing one time generated password and create the account for playing this game. We also make them enter their age, gender and location so as to keep track of their demography. Each speaker's voice varies with their age or gender and their dialect varies with their area of origin.

We floated this game for the period of four months to cover the remaining 34,719 unique words in Konkani Wordnet. The Konkani speakers across the state (User statistics given as follows) have played this game in their free time producing a diverse speech data which we integrated with the newly redesigned Konkani Wordnet.

Shabdocchar game has been developed for enriching the wordnet with pronunciation audios. Initially a team of five native konkani speakers were tasked with testing the game. Suggestions were sought; scores and weekly ranking, noise validation features were added. This score module improved the popularity of the game. With score and weekly ranking we see that user login on average was increased three times a week to play the game. Currently we have reached out to 27 + 5 volunteers who have contributed to 37,719 word pronunciations in at least one of the dialects. We plan to cover each word in all five dialects. As we have collected the entire corpus with one dialect recording, current users are recording in the remaining 4 dialects. Since we will have the pronunciation of the same word in multiple dialects, this audio corpus can be used for automatic feature extraction of dialects of Konkani. The Table shows the current statistics with respect to the gender age group and dialect.

Dialects	**20's**	**30's**	**40's**	**50+**	**Total**
Antruzi	2	5	1	0	9828
Bardeskari	4	1	0	1	6685
Saxxtti	1	2	0	0	5315
Canconi	2	1	2	2	8719
Pednekari	1	1	1	0	4172
Total Users	**10**	**10**	**4**	**3**	**-**
Total Words	**13403**	**11597**	**5425**	**4294**	**34719**

Table 2: Dialects and User Age-group Distribution of Audio Corpus

Dialects	**Male**	**Female**
Antruzi	3	5
Bardeskari	2	4
Saxxtti	0	3
Canconi	2	5
Pednekari	1	2
Total Users	**8**	**19**
Total Words	**12944**	**21775**

Table 3: Dialects and Gender Statistics of Audio Corpus

4.5 Challenges faced while Audio Recording

Since the audio recording is happening at the user's space, the varied recording environment can create challenges that affect the quality and consistency of the collected audio data, necessitating post-processing and quality control to ensure accurate results.

4.5.1 Background Noise

One of the most common challenges in audio recording is dealing with unwanted background noise. External sounds like traffic, people talking, or electronic interference can degrade the quality of the recording, making it harder to capture clear pronunciations. We consider advantages as well as disadvantages of this noisy data. We use the audio cleaning techniques to save and display the cleaned version for Konkani Wordnet users and the noisy data is also stored to give better accuracy during voice-enabled search. Additionally, noisy data can be used as raw, real, unprocessed audio corpus for research purposes.

4.5.2 Pronunciation Variations

Users from different dialects or regions may pronounce words differently, which can be a challenge for ensuring uniformity and accuracy across recordings, particularly in multilingual projects. But this dialect-wise variation of the pronunciations can be collectively used for the research purposes where one can focus on features specific to dialects.

4.5.3 Handling of Heteronym

There are a few words in Konkani which are spelt identically but pronounced differently. We faced the major challenge while recording such words as the speaker was confused between the multiple possible pronunciations he could record. In such cases we received feedback suggesting adding the gloss definition in the speech recording game, so that one can understand the meaning of the word and pronounce it accordingly. A few examples of such words in Konkani are given in Table: 4 and Table: 5. Each of these words have multiple synsets in Konkani Wordnet having different gloss definitions/meanings, eventually leading to the different pronunciation.

Word	**पाडो**
Gloss 1	दुश्ट वा वायट सभावाचें (dushTa vA vAyaTa sabAvAcEM) English: A person of a wicked or evil nature
Gloss 2	जाका भुरगीं म्हणटात अशी खंयच्या अंकाच्या एका सावन धा मेरेन गुणाकार केल्ली क्रमीक सुची (jAkA bhuragIM mhaNaTAta ashIM khaMyacyA aMkAcyA EkA sAvana dhA mErEna gUnAkAra kellI kramIka sUcI) English: Tables - is a sequential list of numbers obtained when any digit is multiplied from one to ten, usually called by children
Gloss 3	गायेचें नर भुरगें (gAyecEM nara bhUragEM) English: The male child of a cow

Table 4: Examples of Heteronyms along with their Gloss definition

In addition to this, there are technical challenges arising due to Microphone Quality, storage require-

Word	वेळ
Gloss 1	भूतकाळ, वर्तमानकाळ, बी हांचो बोध जाता अशें मिणटां, वरां, वर्सां, बी हांणी मेजतात अशें अंतर वा गती (bhUtakALa, vartamAnaKALa, bI hAMcO bOdha jAtA ashEM mINaTAM, varAM, varsAM, bI hAMNI mEjatAta ashEM aMtara vA gatI) English: an amount of time; a time period of 30 years;
Gloss 2	जीब नितळ करतात असो धातूचो धनुशाकार वा प्लास्टिकाचो लांब पटो (jIba nItaLa karatAta asO dhAtUcO dhanUshAkAra vA plAsTIkAcO lAMba paTO) English: a thin and long plastic strip or a bow-shaped metal strip used to clean the tongue; tongue cleaner
Gloss 3	दर्याची सपाट रेंवताळ देग (daryAcI sapATa rEMvatALa dEga) English: an area of sand sloping down to the water of a sea or lake

Table 5: Examples of Heteronyms along with their Gloss definition

ments, unexpected app-crash, etc which caused the delay in recording process.

4.6 Speech Processing for Recorded Pronunciations

4.6.1 Audio decoding and preprocessing

The raw audio is decoded and converted into an audio buffer, and further divided into 100 blocks of equal length. For each block, we calculate the average amplitude by summing the absolute values of the sample points within the block and dividing by the number of samples. This approach helps in reducing the impact of transient noise and allows increased audio quality.

4.6.2 Noise reduction and audio amplification

A high-pass filter was used to remove low-frequency noise, like background sounds, with the cut-off frequency determined by analyzing the input audio. Since each person has a different voice volume, a treble filter with a cut-off frequency of 5000 Hz was applied to soft voices, while a standard volume level was set for louder voices to keep the sound consistent and avoid distortion. The audio was sampled at 48,000 Hz to ensure high-quality sound for clearer listening.

5 Results and Discussion

The contributions of this work can be broadly placed under the following three heads:

5.1 Shabdocchar Corpus

The pronunciations for the unique words from Konkani WordNet create a speech corpus of 37,719 unique concepts in Konkani language. This corpus can be used to develop any learning app for the entry level students to learn Konkani. Corpus will be available for future work by requesting from the Konkani WordNet website[3].

5.2 Enriched Wordnet Website

We have re-designed the Konkani Wordnet website[4] by adding two new modules namely, pronunciations and visualiser module and providing a fresh, updated interface. This version is more visually appealing and user-friendly, aimed at attracting a broader audience and maximizing its usability and benefits for learners and researchers. The Pronunciations module is discussed here and the Visualizer module is discussed in another paper.

5.3 Speech Collection Tool using Crowdsourcing

The speech collection tool[5] developed for this task can be used for recording speech pronunciations for any dataset in future. Also, other language Wordnets can be enriched by collecting pronunciations using this tool. This crowd sourcing tool can also be modified and used to collect image/video content for the unique concept words from Wordnets or any other set of text corpora.

6 Future Scope

Future enrichment of WordNets can be achieved by integrating multimodal resources, such as linking concepts to relevant images and real-world contexts, thereby enhancing their value in language learning and educational applications. Gamified crowdsourcing can play a crucial role in accelerating this process by engaging a diverse user base through interactive tasks, leaderboards, and rewards while ensuring the collection of high-quality data. The use of generative AI to automatically create new glosses for concepts, coupled with gamified validation mechanisms, can further streamline the enhancement process. Additionally, contemporary vocabulary, including regional variations and slang, can be systematically incorporated into WordNets to keep them relevant. The development of domain-specific WordNets tailored to fields like administration, medicine, law, and technology will also enable more precise and impactful applications in both research and industry.

7 Conclusion

In this paper we have presented the creation process of *Shabdocchar*, a speech corpus with 37,719 audio

[3] https://konkaniwordnet.unigoa.ac.in/corpus/english

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

[5] https://github.com/VidyaapatiGU/speech_collection/

files, capturing the pronunciations of words in Konkani WordNet. *Shabdocchar* enriches Konkani WordNet with audio data paving the way to make WordNet accessible to visually impaired and also enable voice based search of WordNet concepts in the future. We capture the pronunciations of the same word in different dialects of Konkani increasing the inclusivity of Konkani WordNet. The game developed to capture the pronunciations is a generic tool that can be adopted by any language to capture audio data with a few data set changes and can act as a source tool by other WordNets to capture audio pronunciations for the other WordNets as well.

Acknowledgements

We would like to thank the team at Vidyaapati Project Lab at Goa University, a project funded by Ministry of Electronics & Information Technology, Govt. of India for contributing to our speech collection activity. We would also like to thank all the Konkani speaking resources who recorded with their voices and covered an entire set of Wordnet words and helped us creating the Shabdocchar corpus.

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).

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.

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.

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.