Building a Public Domain Voice Database for Odia

Subhashish Panigrahi
subhashish@theofdn.org
O Foundation
Bhubaneswar, Odisha, India

ABSTRACT

Projects like Mozilla Common Voice were born to address the challenges of unavailability of voice data or the high cost of available data for use in speech technology such as Automatic Speech Recognition (ASR) research and application development. The pilot detailed in this paper is about creating a large freely-licensed public repository of transcribed speech in the Odia language as such a repository was not known to be available. The strategy and methodology behind this process are based on the OpenSpeaks project. Licensed under a Public Domain Dedication (CC0 1.0), the repository currently includes audio recordings of pronunciations for more than 55,000 unique words in Odia, including more than 5,600 recordings of words in the northern Odia dialect Baleswari. No known public listing of words in this dialect was found by the author prior to this pilot. This repository is arguably the most extensive transcribed speech corpus in Odia that is also available publicly under any free and open license. This paper details the strategy, approach, and process behind building both the text and the speech corpus using many open source tools such as Lingua Libre, which can be helpful in building text and speech data for different low-medium-resource languages.

CCS CONCEPTS

• **Information systems** → **Multimedia databases**.

KEYWORDS

Odia, Baleswari Odia, speech corpus, lexeme, Wikidata, low-resource languages

ACM Reference Format:
Subhashish Panigrahi. 2022. Building a Public Domain Voice Database for Odia. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3487553.3524931

1 INTRODUCTION

Mozilla Common Voice project sets an important benchmark to explain how a transcribed speech corpus for a language can be created in a decentralized and open manner. This rationale was to ensure that speech technology-related research and development are not hampered because of the high cost of speech data or the unavailability of the same in the first place [9]. In the case of the Odia language of India, despite it being in active use across different audiovisual mediums, any openly-licensed speech data repository beyond Common Voice is not known. The pilot strategy and methodology explained here are based on the OpenSpeaks Project, and address the issue of meager availability of voice data in most low-medium-resource languages. Additionally, the paper also details the process for building such data for speech research and development. Most indigenous, endangered, and other languages are not officially recognized or lack a strong language digital activism. Some officially-recognized languages would fall under the umbrella category of "low-medium-resource languages". This initiative highlights the experiments in the author's native language Odia that between 35–45 million people speak in the Indian state of Odisha. The broader phonological and Natural Language Processing (NLP) scholarship has been minimal in research and functional application creation such as speech synthesis or Automatic Speech Recognition (ASR). The lack of publicly available voice data in general, and more specifically, the lack of diversity in data and lack of availability of open data, are widely seen in the small sample size of theoretical and experimental scholarships. Furthermore, there have not been enough attempts to make voice data available under open licenses in Odia, and no known attempt in its northern dialect "Baleswari Odia". Baleswari Odia did not even have a comprehensive public listing of all words until the creation of this voice data. The need for the availability of openly-licensed word corpus and speech data has been widely felt in the NLP and ASR community [11]. Recording natural voices and capturing the diverse range of speech by including a vast sample size is always a recommended process for building the foundation of voice data. Spoken forms of languages include the complexity of accents and intonations that are affected by regional influences and socioeconomic mobility and even access to education and the resulting fluency and warrant recording of different speakers. While speech technologies are becoming ubiquitous through commercial products and platforms, there are also humanitarian areas where these technologies can play an extremely critical role. The availability of voice data would be vital for such technological innovations. For instance, building applications using text-to-speech and speech-to-text engines for public announcements in case of emergencies or interactive voice response (IVR) for phone-based services, creating assistive web technologies to support people with disabilities, and ensuring speech-based accessibility at public service kiosks in physical spaces are some of the areas for speech tech innovations. These are where multilingual speech technologies are most needed and absent in languages like Odia. In the case of Odia, there is not much progress in any publicly available and open-source tool development process to make active use of natural voice. The same goes for all the dialects and languages under the Odia macrolanguage. There is very little data

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9130-6/22/04…$15.00
https://doi.org/10.1145/3487553.3524931

available at the moment in an established and officially-recognized language such as Odia to create a decent and functional speech synthesis system. Much less or no speech data is available for languages and dialects are spoken by communities with lower access to financial and other resources. This disparity, in turn, hampers any speech-related research and application development.

This paper describes how a speech corpus of more than 55,000 (over 21 hours of voice data) Odia-language words was created using different open-source tools, particularly Lingua Libre (https://lingualibre.org/), which is a web platform for recording pronunciations. These audio files were uploaded under Creative Commons CC0 1.0 Universal Public Domain Dedication (https://creativecommons.org/publicdomain/zero/1.0/), making them perpetually free for NLP research and development. Qualitative and strategic analysis of the process behind this project led to the foundation of a robust workflow for recording low-resource languages and dialects. The "Methodology" section details the process of collecting words, compiling a wordlist, recording the pronunciation of those words, and uploading the speech data to Wikimedia Commons using Lingua Libre. This workflow was tested both for Odia and the Baleswari Odia, a dialect that still does not have a public listing of lemmas (dictionary forms that are also known as headwords or catchwords). 5,600 Baleswari Odia words have already been recorded through this project, and the learning and recommendations are captured as a framework to provide insights for similar initiatives in different low-resource languages. Though initially planned as a voice-only project, this project also benefited the creation of wordlists and Wikidata lexemes which can further the related NLP work as explained in the "Result" section. Apart from elaborating the technical strategy, this paper also discusses the integration of lexemes for different low-resource languages focusing on voice data.

2 BACKGROUND AND RELATED WORK

2.1 Odia language

Odia (formerly "Oriya") is a South Asian language that originated in the present-day Odisha (formerly "Orissa"), an administrative province on the east coast of India. As per the 2011 Indian census [1], Odia is spoken by about 35 million people, and Ethnologue mentions another 4 million speakers who speak it as their second language (L2) [7]. Odia is the most spoken macrolanguage (81.32% of the population) in Odisha, where the neighboring languages include over 21 other minorized languages [23] which are clubbed as the Adivasi languages. Adivasi is a heterogeneous and socioeconomic group comprising many indigenous peoples. Sixty-two such groups live in Odisha [23]. Their languages have immensely influenced Odia and are also influenced by Odia themselves. Odisha also shares borders with the neighboring states such as West Bengal and Jharkhand to the north, Andhra Pradesh to the south, Telangana to the southwest, and Chhattisgarh to the west [28]. Angika, Bajjika, Bengali, Bhatri, Hindi, Ho, Kui, Sadri, Santali, Urdu, and Telugu are some of the other neighboring languages whose speakers live both in the aforementioned neighboring states and inside Odisha [23].

2.2 Baleswari Odia (Balesoria)

Baleswari Odia (an exonym, the endonym is often shortened as "Balesoria") which is also used as an eponym for the people from the larger Balasore area) is the northern dialect of the Odia language. It is spoken primarily in the Balasore (Baleshwar) district, an administrative region in India, of the Indian state of Odisha and the neighboring districts of Mayurbhanj and Bhadrak, and the bordering regions of West Bengal state such as the Purba Medinipur district [13].

2.3 Voice data in Odia and Baleswari Odia

Recording and transcribing voice data has been limited for Odia. Much less data is published under different public and open licenses [15]. Creating audio descriptions to accompany textual data has also led to creating some amount of speech data. For instance, dictionary entries for the Odia Wiktionary project sometimes carry the pronunciations of words. Similarly, adding the native pronunciation of some of the Wikipedia article titles, such as biographical articles or articles with proper nouns, are helpful for the readers as audio helps avoid any ambiguity in such cases. Such needs have seen responses from the volunteer contributor community contributing in recording audio. Many articles on biographies of living personalities also occasionally contain their recorded voice. The Odia Wikipedia has existed since 2003 and has close to 16,000 articles at present. There have been sporadic contributions to enhance such articles. For instance, 464 audio recordings containing text from Wikipedia articles on medicine were uploaded by some Odia Wikipedia editors [29]. Similarly, several Odia Wikipedia editors made 50 recordings of notable personalities through the "Project Parichay" [22] and under the Voice Intro project [31]. Panda and Nayak [18] have performed an analysis by contrasting their text-to-speech system in Odia with the free software Dhvani TTS [14] that make use of compressed audio and concatenation to produce "intelligible speech". As documented on OpenSpeaks, experimental open-source projects such as Lekatha also demonstrate natural voice data for speech synthesis. [19] Mozilla's Common Voice project emphasizes the use of high-quality and openly-licensed (Public Domain) natural voice recordings as a recommended practice for the future of speech technology as natural speech that inherently carry intonations and accents can be helpful in speech recognition [9]. A dataset containing more than nine hours of the audio recording of sentences in Odia is uploaded by volunteers to Common Voice by the end of 2021 and over one hour of this recording is validated by a community of contributors [12]. The 55,000 recordings that are elaborated about in this paper add nearly 22 hours of voice data to such existing openly-licensed data [20].

2.4 Platforms and applications used

The OpenSpeaks project includes a broad set of resources including Open Educational Resources (OER), strategies, and also templates for specific needs in different linguistic and other demographic environments to conduct audiovisual documentation of languages, especially low and medium-resource languages. It is hosted on the English Wikiversity (https://en.wikiversity.org/wiki/OpenSpeaks). Lingua Libre (https://lingualibre.org/) is a browser-based online platform for recording pronunciations of a list of words in any language or dialect that can be written using a writing system as long as the writing system has a Unicode encoding. Common Voice (https://commonvoice.mozilla.org/en) is a web platform by Mozilla

that encourages contributors to submit sentences available under a Public Domain release in supported languages, record pronunciations of the sentences, and review recordings made by others. The recordings are made are made available under a Public Domain for downloading. Unlike LinguaLibre, where the username is tied to each recording, Common Voice strips the username and completely anonymizes the recordings. The Python script created by T. Shrinivasan [27] for creating a unique list of words from any text file (with .txt or even .xml or .json extensions) is available on GitHub and can be tailor-made to accommodate the needs of different languages typed using varied writing systems.

3 METHODOLOGY

This section describes the end-to-end process followed in this project which has two major steps:

(1) Creating a wordlist of unique words (collecting words, cleaning up by correcting spelling and other inherent mistakes, sorting unique words alphabetically, removing words already recorded)
(2) Recording pronunciations of words using LinguaLibre and uploading to Wikimedia Commons

3.1 Wordlist creation

Scraping through online textual content was the starting point for this process. Periodical raw data dumps of several Wikimedia projects are listed under the Wikimedia Downloads project [35]. Three Odia-language projects for which data dumps are available are Odia Wikipedia [8], Odia Wikisource [32], and Odia Wiktionary [33]. These three projects have a vast expanse of words of a different era: loanwords related to science and technology in the Odia Wikipedia, pre-1941 vocabulary in the Odia Wiktionary, and words from the 14th century and contemporary use in the Odia Wikisource. To space out this process, I downloaded from the Odia Wikipedia dump first and extracted only the unique Odia-language words from all Mediawiki namespaces (titles of about 15,000 articles and the same discussion pages and different non-article content). The Python script that T. Shrinivasan originally developed to work for the Tamil-language Wikimedia projects data dumps [27] was modified slightly to match the requirements for Odia. This script helped process XML files downloaded from the Wikimedia data dump and remove Latin characters and signs, except hyphens/dashes ("-"), as hyphens are occasionally used to create forms in Odia. The script makes a unique wordlist and sorts the words alphabetically. Running this script for the Odia Wikipedia dump resulted in creating 1,559,195 words. While this process helped list words that start with an Odia character, it did not help remove Latin characters attached to Odia words. Copying the words to the LibreOffice Calc Spreadsheet and sorting based on word length helped move all the exceptionally-long words to the very end of the list. Manually searching for Latin characters, replacing them with spaces, and replacing them later to line breaks helped sort unique words. This process helped bring down an entire list of 1,559,195 words to 301,608 unique Odia words. So far, this wordlist has been the essential source for the recording process. I also collected words from other sources beyond Odia Wikipedia for more contemporary words.

The script mentioned above was also used to extract words from data scraped from sources such as Odia-language news sites, science magazine the "Bigyana Diganta" [5]. These sources helped increase contemporary and science and technology-related terms, generally loanwords transliterated into Odia. The current wordlist has over 530,000 unique words [24], and the wordlist is available publicly under a Public Domain Dedication. It is updated periodically when new words are sourced from different sources.

A	B	C	D	E	F	G	H
କ୍ରମ	ଶବ୍ଦ୧	ଶବ୍ଦ୨	-ତର	-ତମ	-ତା	pp.	ପ୍ରତିଶବ୍ଦ
3	ମହାର୍ଘ	ମହାର୍ଘ	ମହାର୍ଘତର	ମହାର୍ଘତମ	ମହାର୍ଘତା	6383	ଦାମିକା; ଅତି ମୂଲ୍ୟବାନ୍
3	ବସ୍ତୁୟ	ବସ୍ତୁୟ	ବସ୍ତୁୟତର	ବସ୍ତୁୟତମ	ବସ୍ତୁୟତା	5788	ସଂଶୟ
3	ଶ୍ରେଷ୍ଠ	ଶ୍ରେଷ୍ଠ	ଶ୍ରେଷ୍ଠତର	ଶ୍ରେଷ୍ଠତମ	ଶ୍ରେଷ୍ଠତା	7865	ପ୍ରଧାନ
3	କଠିନ	କଠିନ	କଠିନତର	କଠିନତମ	କଠିନତା	1246	କଠୋର; ନିଷ୍ଠୁର; ନିର୍ଦ୍ଦୟ
3	କଠଣ	କଠଣ	କଠଣତର	କଠଣତମ	କଠଣତା	1246	କଠନ
3	ଉଷ୍ଣ	ଉଷ୍ଣ	ଉଷ୍ଣତର	ଉଷ୍ଣତମ	ଉଷ୍ଣତା		
3	ପ୍ରବଳ	ପ୍ରବଳ	ପ୍ରବଳତର	ପ୍ରବଳତମ	ପ୍ରବଳତା	5100	
3	କଠୋର	କଠୋର	କଠୋରତର	କଠୋରତମ	କଠୋରତା	1247	
3	ସକ୍ରୟ	ସକ୍ରୟ	ସକ୍ରୟତର	ସକ୍ରୟତମ	ସକ୍ରୟତା		
3	ନିଷ୍ପ୍ରୟ	ନିଷ୍ପ୍ରୟ	ନିଷ୍ପ୍ରୟତର	ନିଷ୍ପ୍ରୟତମ	ନିଷ୍ପ୍ରୟତା	4412	କ୍ରିୟାହୀନ; କର୍ମରହିତ
3	ଅକୁଣ୍ଠ	ଅକୁଣ୍ଠ	ଅକୁଣ୍ଠତର	ଅକୁଣ୍ଠତମ	ଅକୁଣ୍ଠତା	45	ଯାହା ଦଳୁଡ଼ା ହୋଇନାହିଁ; ତୀକ୍ଷ୍ଣ
3	ଆବିଳ	ଆବିଳ	ଆବିଳତର	ଆବିଳତମ	ଆବିଳତା	788	କଳୁଷିତ; ଅସ୍ୱଚ୍ଛ

Figure 1: Creation of forms from lemmas using LibreOffice Spreadsheet. Screenshot by author. [CC-BY-SA 4.0].

By creating Wikidata Lexeme Forms [21], I have also been able to generate forms of headwords/lemmas that are already included in the wordlist. I use a spreadsheet to generate forms containing prefixes, suffixes, and modified forms in the back-end, which generally follow the grammar rules. Such a process is also helping create lexemes on Wikidata containing lemmas and forms. By using "Twofivesixlex" (TFSL) [17], a Python-based framework developed by Wikimedia contributor Mahir Morshed, bulk-creation of lexemes in Odia is now possible in a moderately automated process. Before uploading the words using Lingua Libre, I create categories for the words on Odia Wiktionary (about 15 different categories), including Arabic, Turkish, and Persian loanwords. By manually going through words in each of these categories, it was possible to create a list of lemmas and use a range of different Spreadsheet formulas. It was possible to create multiple forms by adding postpositions and suffixes (see "Figure 1"). This process helped create a basis for using the same treatment for the Baleswari Odia lexemes in the future. Creating lexemes is an additional step that is beyond creating speech data. This step is relevant for any dictionary and machine translation application development. The words created using this process are being added to the wordlist.

3.2 Recording using Lingua Libre and uploading to Wikimedia Commons

Lingua Libre was used for recording the pronunciations of the words once the wordlist was ready. It is generally a four-step process:

(1) **Step 1: One-time speaker profile setup**
RecordWizard, a component of Lingua Libre that is primarily used for the front end of the recording process, encourages the input of some demographic details to capture a speaker's gender and accent. This step is done while creating a speaker profile for a speaker. Once set up, this step can only be skipped (RecordWizard allows to modify any details) for the forthcoming batches.

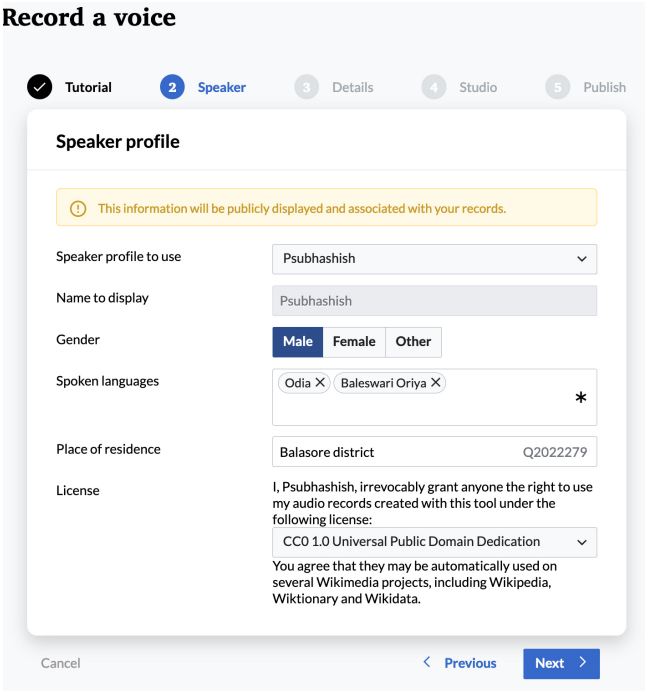

Figure 2: Inputting demographic details into Lingua Libre RecordWizard. Screenshot by author. [CC-BY-SA 4.0].

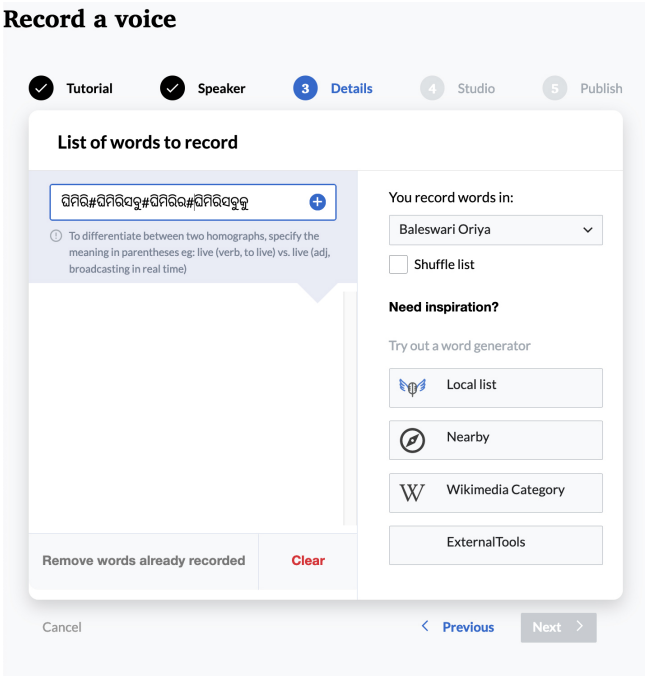

Figure 3: Uploading words into Lingua Libre RecordWizard. Screenshot by author. [CC-BY-SA 4.0].

(2) **Step 2: Loading wordlist**
In a typical batch, words are separated using a "#" (hash) while inputting words locally. Article names in a Wikipedia article category or entries under a Wiktionary category can also be loaded at the moment. The "Remove words already recorded" button is recommended for skipping previously-recorded words.

(3) **Step 3: Recording pronunciations of words**
Once words are ready for recording, each word is pronounced by giving a slight pause between the words to allow Lingua Libre to detect the gap and move to the following word (see "Figure 4" and "Figure 5"). The current workflow makes it possible to record an approximate average of 300-400 words in Odia (of varied character lengths) in an hour.

(4) **Step 4: Reviewing recording and uploading to Wikimedia Commons**
Lingua Libre allows to review each recording after they are recorded, once after pronouncing, and once more after moving to the next phase of reviewing all recordings, finally before uploading. An average batch of 100 words could contain about five poorly recorded words with either a poor accent or any background noise recorded inadvertently. I have used a USB-based desktop studio microphone connected to a MacOS-based computer for recording and a studio monitor for monitoring the recorded audio. After reviewing and skipping any poorly recorded words, the pronunciations are uploaded to Wikimedia Commons by clicking the "Publish on Wikimedia Commons" button.

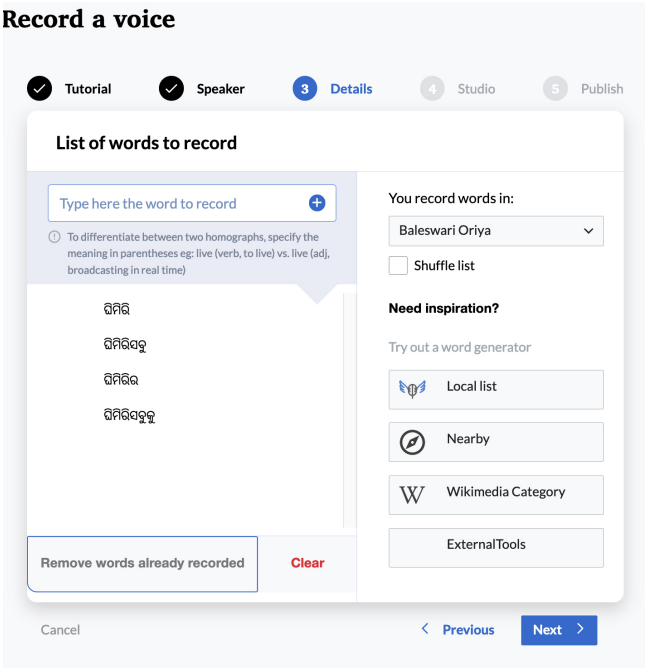

Figure 4: Using the "Remove words already recorded" button on Lingua Libre. Screenshot by author. [CC-BY-SA 4.0].

(5) **Step 5: Categorizing words on Wikimedia Commons**

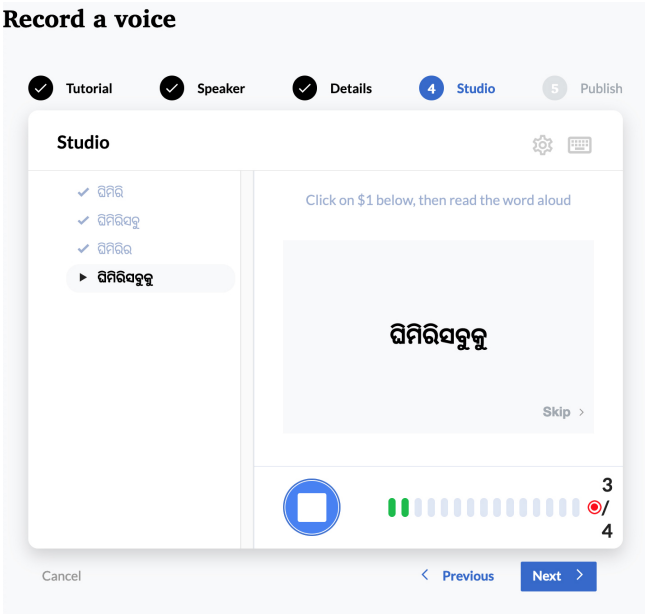

Figure 5: Ongoing recording of pronunciations of words using Lingua Libre RecordWizard. Screenshot by author. [CC-BY-SA 4.0].

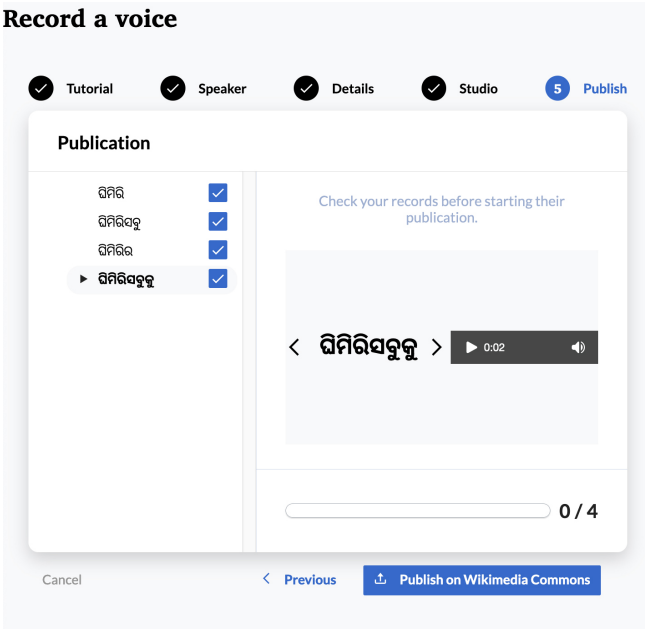

Figure 6: Reviewing recordings inside RecordWizard. Screenshot by author. [CC-BY-SA 4.0].

By default, the data collected in "Step 1" helps add categories to each word on Wikimedia Commons. While that suffices for general purposes, further categorization of all Odia words into the "Odia pronunciation" category using the Wikimedia Commons Gadget Cat-a-lot (https://commons.wikimedia.org/wiki/Help:Gadget-Cat-a-lot) helps with a larger category containing all pronunciation audio files in Odia.

Figure 7: Categorizing of recordings on Wikimedia Commons using Gadget Cat-a-lot. Screenshot by author. [CC-BY-SA 4.0].

A detailed step-by-step process for recording pronunciations of entries from any Odia Wiktionary category is explained in Odia through both a video tutorial and screenshot-assisted tutorial [25]. The "Help:RecordWizard Manual" page on Lingua Libre also has a tutorial in English [2] and a video walkthrough containing French subtitles and the French interface of Lingua Libre is also available [3].

3.3 Wordlist creation and recording of Baleswari Odia

Being a spoken variant of Odia, Baleswari Odia has forms common with Odia and forms absent in Odia. The 1931-1040 lexicon "Purnnachandra Ordiya Bhashakosha" lists 790 lemmas.

In 2015 I started an oral history of 93-year-old Musamoni in a slightly archaic version of Baleswari, which will be available in the 2022 documentary film "Nani Ma". The recordings helped collect many Baleswari Odia words, both outdated and current. From 2016, I started documenting more words in everyday use in a private spreadsheet. The third source was a Twitter handle "@balesoria" (https://twitter.com/balesoria), including tweets and nested comments-tweets. Scraping those tweets helped add words from different variants of Baleswaria Odia because of the geographic diversity of the Twitter users. By August 2021, the wordlist contained

1,801 words. By merging all the 790 words from "Purnnachandra Ordiya Bhashakosha" and removing duplicate words, the final list had 2,060 words by January 2022. However, this list neither contained all the lemmas (the dictionary forms) nor the forms and required further development. The postpositions and suffixes in the case of Baleswari Odia often vary from the standard written Odia, and some forms are identical to the latter. It is important to note that the written spelling for many forms for Baleswari Odia and standard Odia might be identical. Still, the intonations could vary widely between these two variations. Like other spoken forms, various stresses and intonations are attributed to the exact words based on the mood in Baleswari Odia. The "Recommendations based on observations" section discusses some of these variations. A wordlist containing 15,748 words in Baleswari Odia is also available publicly [26]. Pronunciation audio files for some of the words are also available for downloading in a periodic manner both as a wordlist [30] and as audio files [6, 34].

4 RESULTS

This section explains the outcomes of the process explained in the "Methodology" section. I generated a wordlist of more than 530,000 unique words, including more than 200,000 lemmas and remaining forms. As the words were collected from different sources, some forms without the related lemmas also exist. This process helped me record and upload pronunciations of over 55,000 unique words between June 2018 and March 2022. The current workflow is based on a working model established by December 2021. Nearly 21,000 words in this list are contemporary (including many loanwords). 5,605 of the 55,000 words are in Baleswari Odia.

The lexemes in Odia currently being created on Wikidata are expected to eventually help create forms for many existing lemmas using the created wordlist. Lemmas and forms can also be accompanied by the relevant pronunciations from the voice data repository. Wikidata also has options for adding multilingual translations to lexemes to develop parallel corpus. While the current text and voice data can be used for the spell-check, dictionary with word pronunciations, and Automatic Speech Recognition (ASR) related work, parallel corpus development can lead to further work on machine translation.

A list of all the recorded Odia words including the Baleswari Odia words, a separate list containing only the recorded Baleswari Odia words are available along with this paper.

5 OBSERVATIONS

Some of main issues and features of the current wordlist:

- There are words in the wordlist with pre-existing spelling mistakes as many Wikipedia articles (or content from other sources) are not always edited for spelling. Odia currently lacks a spell-check with a comprehensive list of all words to check and correct such mistakes. While some mistakes are found while recording and fixed, the remaining words need clean up.
- Many spelling mistakes found in the wordlist were also caused while using either because of the use of script encoding converters (legacy to Unicode encoding standard) or using input tools that were made to work for typefaces of legacy encoding for creating Wikipedia articles. Manually searching and replacing such mistyped words or removing them from the entire document helped clean the list. As explained in the previous point, this still does not guarantee complete accuracy and some spelling mistakes persist.
- In this wordlist, Odia words are also attached with characters in other writing systems.
- Some words have insertions of Zero-width joiner (ZWJ) and Zero-width non-joiner (ZWNJ) characters that make the words look visually the same but are counted as separate words sorting. Both ZWNJ and ZWJ are used in Odia to avoid a character being joined to the previous character when the "halanta" (or "virama"; 0B4D in Unicode block [4]) sign is suffixed to the latter. Such entries result in the exact words being recorded twice when entered into Lingua Libre.
- In terms of depth of topics in this wordlist, Odia Wikipedia articles cover a range of topics, but the content depth is not always uniform [16]. For instance, articles related to specific topics are very well covered, and other areas can be poorly covered. Such instances are because of multiple reasons, primarily the lack of a large, active, and diverse editor community that includes a balanced representation of people of different age groups, genders, and educational/work backgrounds.
- The wordlist contains a significant portion of loanwords, including a long list of medicine-related terms primarily transliterated in the Odia alphabet.

Issues and features found in the recording process:

- Lingua Libre allows a user to record a word again even if they had recorded the same word using a different platform/process other than that of Lingua Libre. This is a known issue. However, it also allows a user to record multiple variations of pronunciations of the same word and is useful for ASR training.
- Lingua Libre lets a user record an already recorded word for many consecutive times. However, the consecutive entries rewrites the file on Wikimedia Commons. This is not useful for recording intonations and accents that are spoken by the same user. This can be avoided by creating a separate user profile on Lingua Libre.
- One of the demographic fields of Lingua Libre include "Place of residence". However, the "place of language learning" could be more relevant from a linguistic standpoint. "Place of residence" might not have any impact at all if one such place has a different language environment.
- Wikimedia Commons currently does not detect characters of range of writing systems (including Chakma, Sharada, Grantha, Warang Citi among others) as it was identified during a workshop that I was conducting for the Ho language of India. [10] While this issue does not affect the Odia alphabet, it affects several other writing system and creates a barrier for contributors of such languages.

6 RECOMMENDATIONS

This section includes a set of broad recommendations that would require further consideration while framing strategies and plans

for similar speech corpus development in different low-medium-resource languages and dialects. While these recommendations are neither definitive nor exhaustive, they can be used as a guide or template to help create a workflow for low-limited-resource languages. The native speakers can create a more specific strategy relevant to their constraints and advantages.

(1) The core layer of any speech corpus development is creating a wordlist of unique words. Collecting and compiling words in a target language or dialect is paramount. Also, the lack of such wordlists could hinder NLP research and development. It is highly recommended that the creator of such data publicly release their corpus under open licenses as long as they can compensate for the labor, keeping in mind that volunteer labor is not universal in all parts of the world.
(2) Wordlists created from a monolithic source such as literary works of a single author or literature of a similar genre might lack phonetic diversity. Hence, growing the wordlist gradually and incorporating contemporary words (including known loanwords) are highly recommended. Contemporary vocabulary could also ease any entry-level barrier of everyday users who might not have the fluency that users familiar with the historical use of language through the study of ancient literature might have.
(3) Social media can be a great starting point for community engagement, and such interactions, when done in the target language, can help generate more content. This content can, in turn, be used to generate a wordlist for future speech projects.
(4) Speech corpus development is a slow and gradual process. Hence, incremental and iterative growth in data quality is a known feature of speech data development. Contributors often need to have a long-term strategy for such a project to address the issues of not having a contributor community with diversity in gender, geographical and socioeconomic background.
(5) While it is essential to have the diversity of speech while building voice data, the initial process might be driven by a handful of people, and the initial data would lack diversity. Setting up an initial workflow can follow a contributor community-building exercise.
(6) The socioeconomic mobility of individual contributors (including systemic oppression of certain marginalized groups such as the Dalits and Adivasis in India) and the resulting inequitable access to technical know-how can hinder the NLP research in many languages. It is vital to put early effort into generating financial compensation for paid labor (as opposed to expecting free labor from those who lack affordability) while developing speech corpus.
(7) While developing a speech corpus, it is recommended to evaluate if the target language has multiple dialects or distinct regional accents. Knowing the nuances within a language could help identify individuals within the speaker community and widen the speech data diversity.

7 CONCLUSION

This paper elucidates an end-to-end process of creating speech data for both Odia, a formal and standardized form of a language, and Baleswari Odia, a dialect containing words with different intonations and accents. The process detailed here relies primarily on Lingua Libre, an open-source web platform that allows recording pronunciation of words, phrases, and sentences in a language or dialect and is integrated into the Wikimedia projects. Over 21 hours of voice data of a male contributor is published under a Creative Commons CC0 1.0 Universal Public Domain Dedication. The wordlist used for the pronunciation recording is also published under Public Domain. This license makes the data perpetually free of all copyright restrictions for Natural Language Processing (NLP) research, such as Automatic Speech Recognition, speech-to-text, and other possible applications. The contributions from only a male speaker with higher social and educational access, introducing monotony into this speech data is a known issue. However, the process and the strategy shared here can be replicated for different low-medium-resource languages. There is a vast potential to grow the database by incorporating data from diverse speakers of different genders, socioeconomic mobility, and educational access. The strategy and process summarized in the "Recommendations" section can also be replicated into other languages and dialects with a writing system with Unicode encoding.

REFERENCES

[1] 2011. *Abstract of Speakers' Strength of Languages and Mother Tongues - 2011.* Technical Report. Registrar General and Census Commissioner of India, New Delhi. 6 pages. https://www.censusindia.gov.in/2011Census/Language-2011/Statement-1.pdf
[2] 2021. Help:RecordWizard Manual. https://lingualibre.org/wiki/Help:RecordWizard_manual
[3] 2021. Tutoriel: Comment contribuer à Lingua Libre? https://commons.wikimedia.org/wiki/File:Tutoriel_Lingua_Libre.webm
[4] 2021. *The Unicode Standard, Version 14.0: Oriya.* Technical Report. Unicode, Inc. 4 pages. https://www.unicode.org/charts/PDF/U0B00.pdf
[5] 2022. Bigyan Diganta. http://odishabigyanacademy.in/bigyan-diganta/
[6] 2022. Index of /datasets: Q322719-mis-Baleswari Oriya.zip. https://lingualibre.org/datasets/Q322719-mis-Baleswari%20Oriya.zip
[7] 2022. Odia: Ethnologue. https://www.ethnologue.com/language/ory
[8] 2022. orwiki dump progress on 20220301. https://dumps.wikimedia.org/orwiki/20220301/
[9] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. *arXiv:1912.06670 [cs]* (March 2020). http://arxiv.org/abs/1912.06670 arXiv: 1912.06670.
[10] Biswajeet3. 2021. T297351 Warang Citi (Ho-language writing system) characters not detected on Wikimedia Commons. https://phabricator.wikimedia.org/T297351
[11] Britone Mwasaru. 2022. Why Voice is Important. https://foundation.mozilla.org/en/blog/why-voice-is-important/ Section: Common Voice.
[12] Common Voice Contributors. 2022. Common Voice by Mozilla. https://commonvoice.mozilla.org/
[13] Artatrana Gochhayat. 2016. Odisha as a multicultural state: from multiculturalism to politics of sub-regionalism. *Afro Asian Journal of Social Sciences* 7, 2 (2016), 28. http://mail.onlineresearchjournals.com/aajoss/art/197.pdf
[14] Ramesh Hariharan, Ravi Masalthi, Rileen Sinha, and Santhosh Thottingal. 2021. Dhvani TTS. https://github.com/dhvani-tts/dhvani-tts original-date: 2013-12-17T05:16:31Z.
[15] Josh Meyer. 2022. Open Speech Corpora. https://github.com/coqui-ai/open-speech-corpora original-date: 2019-01-31T14:57:39Z.
[16] Marc Miquel-Ribé and David Laniado. 2018. Wikipedia Culture Gap: Quantifying Content Imbalances Across 40 Language Editions. *Frontiers in Physics* 6 (2018), 54. https://doi.org/10.3389/fphy.2018.00054 12 citations (Semantic Scholar/DOI) [2022-03-08] Publisher: Frontiers.
[17] Mahir Morshed. 2022. tool-twofivesixlex. https://phabricator.wikimedia.org/source/tool-twofivesixlex/

[18] Soumya Priyadarsini Panda and Ajit Kumar Nayak. 2015. An efficient model for text-to-speech synthesis in Indian languages. *International Journal of Speech Technology* 18, 3 (Sept. 2015), 305–315. https://doi.org/10.1007/s10772-015-9271-y 13 citations (Semantic Scholar/DOI) [2022-03-08].

[19] Subhashish Panigrahi. 2015. OpenSpeaks/toolkit/Lekatha - Wikimedia Commons. https://commons.wikimedia.org/w/index.php?title=OpenSpeaks/toolkit/Lekatha&oldid=246217807

[20] Subhashish Panigrahi. 2022. Lingua Libre pronunciation by Psubhashish. https://petscan.wmflabs.org/?psid=19878687

[21] Subhashish Panigrahi, Mahir Morshed, and Lucas Werkmeister. 2022. Wikidata Lexeme Forms/Odia. https://www.wikidata.org/wiki/Wikidata:Wikidata_Lexeme_Forms/Odia

[22] Project Parichay Contributors. 2022. Project Parichay. https://commons.wikimedia.org/wiki/Category:Audio_files_uploaded_under_project_Parichaya

[23] Minati Singha. 2021. Odisha to impart primary education in 21 tribal languages in schools run by SC/ST dept. *The Times of India* (Sept. 2021). https://timesofindia.indiatimes.com/city/bhubaneswar/state-to-impart-primary-edu-in-21-tribal-languages-in-schools-run-by-sc/st-dept/articleshow/86463363.cms

[24] Subhashish Panigrahi. 2021. Before AI*. https://github.com/ofdn/Before-AI/blob/91b8d20fa3c44305cbbfe0da6ce8579a2ba0380e/data/odia-or-wordlist.txt original-date: 2021-10-20T23:57:09Z.

[25] Subhashish Panigrahi. 2021. Help: Recording pronunciations. https://or.wiktionary.org/s/16h9 Page Version ID: 177311.

[26] Subhashish Panigrahi. 2022. Before AI: nort2660-wordlist.txt. https://github.com/ofdn/Before-AI/blob/45b5dc31c7b2fb376a68767573b472f7cf7861ca/data/nort2660-wordlist.txt original-date: 2021-10-20T23:57:09Z.

[27] T. Shrinivasan. 2021. Odia Wordlist from Wikimedia Dump. https://github.com/ofdn/odia-wordlist-from-wikimedia-dump original-date: 2021-11-29T15:00:15Z.

[28] The Editors of Encyclopaedia. 2020. Odia language: Region, History, & Basics. https://www.britannica.com/topic/Odia-language

[29] Videowiki Contributors. 2022. Videowiki project - or - Wikimedia Commons. https://commons.wikimedia.org/wiki/Category:Videowiki_project_-_or

[30] Wikimedia Contributors. 2022. Category:Lingua Libre pronunciation by Psubhashish. https://petscan.wmflabs.org/?psid=21604903

[31] Wikimedia Contributors. 2022. Category:Voice intro project - Wikimedia Commons. https://commons.wikimedia.org/wiki/Category:Voice_intro_project

[32] Wikimedia Contributors. 2022. orwikisource dump progress on 20220301. https://dumps.wikimedia.org/orwikisource/20220301/

[33] Wikimedia Contributors. 2022. orwiktionary dump progress on 20220301. https://dumps.wikimedia.org/orwiktionary/20220301/

[34] Wikimedia Contrubutors. 2022. Index of /datasets: Q336-ori-Odia.zip. https://lingualibre.org/datasets/Q336-ori-Odia.zip

[35] Wikimedia Foundation. 2022. Wikimedia Downloads. https://dumps.wikimedia.org/