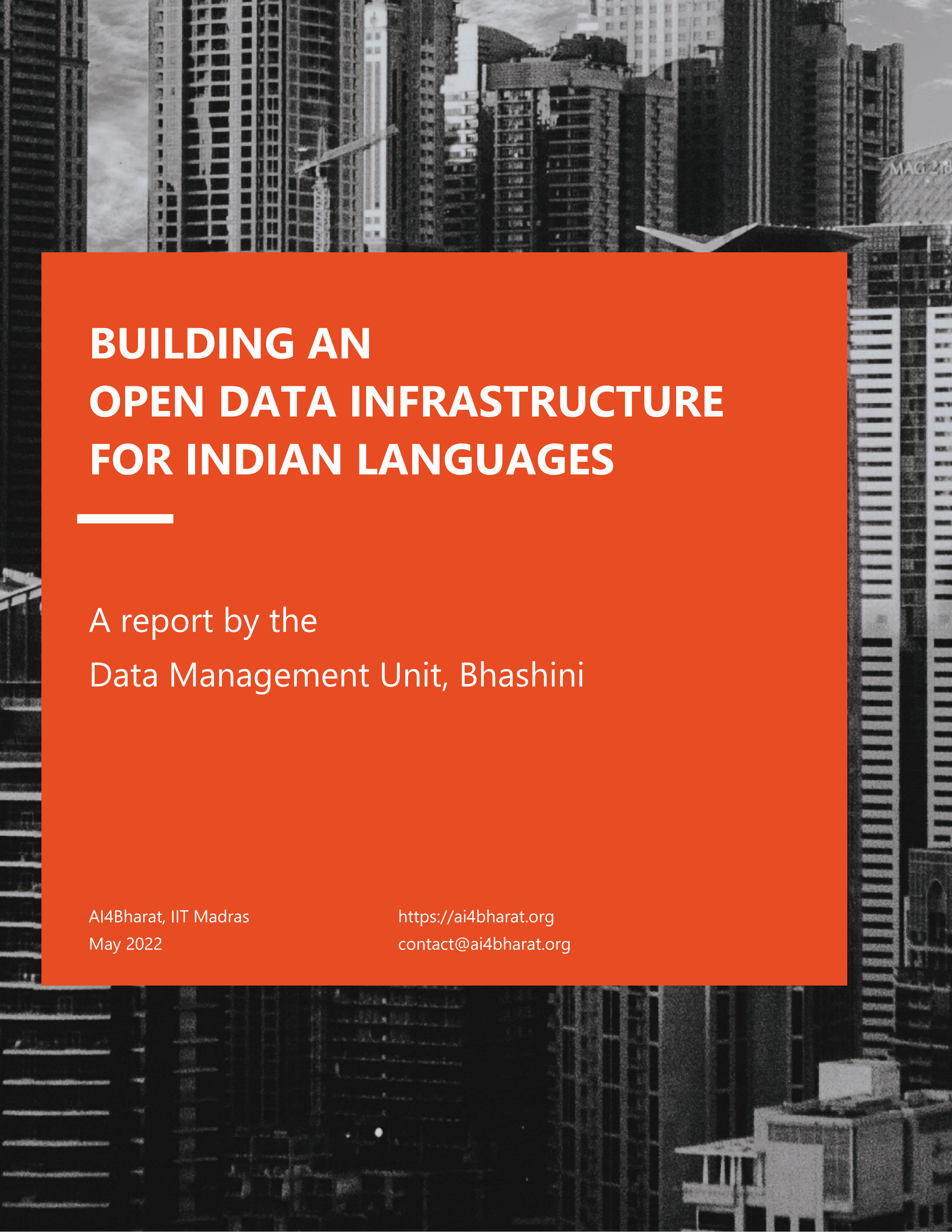


BUILDING AN OPEN DATA INFRASTRUCTURE FOR INDIAN LANGUAGES

A report by the
Data Management Unit, Bhashini

AI4Bharat, IIT Madras
May 2022

https://ai4bharat.org
contact@ai4bharat.org

Executive Summary

The National Language Translation Mission (NLTM), **Bhashini** has been set up under the aegis of the **Ministry of Information and Technology (MeitY)** with the aim of making rapid progress in speech, text, and vision technology for Indian languages. The broader goal is to bring parity in AI technologies for Indian languages with respect to English with open-source contributions in tools, datasets, neural models, and reference applications. This mission is cross-cutting in terms of supporting

(a) all 22 constitutionally recognized languages in India

(b) a wider veriety of tasks including machine translation, speech recognition, speech synthesis, character recognition, and language understanding.

AI4Bharat will be serving as the Data Management Unit (DMU) for Bhashini and will help in **creating open-source datasets across languages and tasks** often improving on existing datasets by orders of magnitude. AI4Bharat will also be **building an open-source tool for collaboration** on language data collection and annotation. In this document, we report on the current state of data across languages and tasks, and the current plan of data collection as part of the DMU's roles.

Table of Contents

Introduction

The National Language Translation Mission (NLTM), Bhashini has been set up under the aegis of the Ministry of Information and Technology (MeitY) as a mission mode project for making rapid progress in speech, text, and vision technology for Indian languages. AI4Bharat is a center at IIT Madras with the mission of bringing parity in AI language technology with respect to English for Indian languages. AI4Bharat has been chosen to serve as the Data Management Unit (DMU) of Bhashini.

As part of this DMU role, AI4Bharat is working to create high quality open-source datasets for all the 22 constitutionally recognised languages across multiple tasks – Machine Translation, Automatic Speech Recognition, Text-to-Speech, and Optical Character Recognition. Specifically, the aim of the DMU is to provide a base layer of data infrastructure across languages and tasks, while other projects in the mission focus on building datasets and models for specific tasks and languages. The DMU will release all created datasets in the open source with permissible licenses and will also upload them to the ULCA repository following open standards. It is expected that this spurs development of AI models and applications for large scale use across the nation.

In this document, we report on the current state of data across languages and tasks, and the current plan of data collection as part of the DMU's roles. In the remainder of this chapter, we discuss the broader context of challenges and opportunities in the langauge space in India and the specific responsibilities of the DMU.

Challenges and Opportunities

Indian languages are an example of the diversity of India and their rich morphological structures. India is home to the fourth highest number of languages. Hindi, Urdu, Bengali, and Punjabi are in the top 20 spoken languages across the world. As per the 2011 census, there are 1,369 rationalised mother tongues (with 10,000+ speakers) and they are grouped into 121 languages including 22 constitutionally recognized languages. 191 of these rationalised mother tongues are classified as vulnerable or endangered. 27 of the non-scheduled languages have more than 1 million speakers and most of these are dialects or variants that are grouped under the Hindi language. Sanskrit, Kannada, Telugu, Malayalam, Tamil, and Odia have been given classical language status for their rich heritage and their independent nature. Indian languages are characterised by their rich morphological structures and have a rich diversity that supports multiple dialects and accents, and their heterogeneity enables diverse cultural ecosystems that are part of India.

This rich, diverse, heterogeneous, and multilingual aspect of our country has its pluses. However, on the flip side, it presents a major challenge in achieving the goals of Digital India and its objective that all the citizens in the country would prefer to access digital content and services in their native language(s). This is mainly because resources needed to enable AI technologies in Indian languages are limited and must be scaled to a significant extent. Comparison between high resource languages like English and Spanish and Indian languages for AI readiness indicates that to achieve parity, **there exists a need for a foundational language infrastructure layer as a public good that addresses many gaps**. These gaps include the need for (a) tools for dataset creation, (b) large datasets for training AI models, (c) reference state-of-the-art AI models, and (d) reference applications demonstrating the use of these models.

This language infrastructure layer needs to be broad-based in two aspects: languages and AI tasks. First, the infrastructure should span many languages and dialects, at least covering the 22 scheduled languages. This is an important principle given that a large set of languages may lie outside the commercial interest of current technology providers. Second, the infrastructure should span a range of AI tasks such as text translation, text input tools, speech recognition, speech synthesis, character recognition, and others. Often applications, such as speech-to-speech translation, require a combination of tools spanning these tasks to work accurately with each other. This broad basing of the infrastructure layer requires a fundamentally different approach in creating networks of contributors, efficient and extensible tools, and standardized processes for data collection.

Another fundamental principle is to ensure that the language infrastructure is standardized, interoperable, and open. Several past efforts, supported by the government, academia, and non-profit foundations have created language resources, but many of them remain locked in custom formats or require complex processes for access. There is a need to thus have a centralised, standardised, and open infrastructure to make available tools, datasets, models, and applications. These must be available to all - government, academia, startups, and technology companies - to build upon with a liberal licence.

Finally, another requirement of the infrastructure layer is to ensure that it is available widely across deployment scenarios. This has implications for both datasets and AI models. Datasets need to be created to cover various domains where language usage is critical, spanning areas such as news, government announcements, legal content, women empowerment, digital payments, entertainment, etc. On the other hand, models across AI verticals need to be available in form factors to run on low-powered devices and even in scenarios without consistent connectivity.

The opportunity exists to create an infrastructure for Indian languages that is open, standardized, interoperable, and widely available.

To address the challenge of resource limitation in Indian languages at scale there is the opportunity to create a language infrastructure as a public good that provides tools, datasets, AI models, and applications. This infrastructure should be broad-based, covering a large set of languages and various AI verticals. It must be standardized, interoperable, and open to enable innovation from different sectors. Finally, it must be available widely across deployment scenarios addressing the needs of many Indians in connecting to the benefits of Digital India.

DMU’s Mission

Given the above context, the mission of the DMU is to bring parity with English in AI tech for Indian languages with open-source contributions of tools, data, models, and solutions.

The DMU’s focus would be to achieve the right balance of reuse of best practices adopted for building language technology for high resource languages such as English while ensuring richness, diversity, heterogeneity of Indian languages is retained in digital ecosystems using them for delivery of services in native languages to citizens of India. For example, considering rich morphological structures in Indian languages and diversity in terms of dialects, accents and scripts, one of the key objectives of DMU is to ensure this diversity/heterogeneity is retained to extent feasible in data collection, models and tools developed as part of the platform. The goals of DMU can thus be summarized below:

- Build a cross country team of language experts spanning academia, social sector and data collection agencies which help in collecting high quality diverse data at scale for all the 22 constitutionally recognized languages.
- Build a unified tool for data collection that supports all languages and all tasks of interest. This tool would standardize data collection across tasks and languages and ensure that all data is compatible with the ULCA (Universal Language Collection API) specification prescribed by MeitY.
- Build reference applications for translation, transcription and optical character recognition which act as a starting point for academia, industry and start-ups to build upon.

The remainder of this document is organised as follows. We first discuss the current state of data as captured on the ULCA platform for which of the tasks mentioned above. We then outline our goals for data collection followed by a discussion on the approach that we will use for achieving these goals. We then discuss our timelines and mention the guidelines that we will use for data collection in the Appendix.

Current State of Data

In this section, we will first provide a quick summary of the current state of data for the four important tasks in language technology, viz., machine translation, automatic speech recognition, text-to-speech and optical character recognition. Over the past several months, Bhashini, through its ULCA (Unified Language Contribution API) platform has ingested data from multiple sources and made it available in a central repository. This repository contains the following types of data.

- **Manual**: This category of data is of the highest quality and involves human effort in creation without dependence on any existing models. In our classification, we consider a data to be of Manual type if it is created by language experts with strict adherance to guidelines on data collection. Examples include translating sentences from scratch or transcribing audio files. All benchmark quality data is often manual.
- **Crowsourced**: This category of data is also manual, but the data collection is done through crowdsourcing platforms, including the Bhashadaan platform.
- **Post-editing**: This category of data includes manually labelled data but with inputs from existing models. Examples include translating text by editing the output of a translation model such as IndicTrans. Such data may have biases introduced by the model that is used for generating the translation but comes with the benefit of faster human effort.
- **Machine generated**: This category includes different types of data generated by models entirely with no human effort. The model used could be for different purposes. For instance, transcription models can be used to align chunks of audio data with transcripts available in document form or language encoders can be used to align bitext pairs across large monolingual corpora. Under high thresholds of alignment scores, such machine generated data can be of high quality. Another type of machine generated data is data created directly by models. For instance, a parallel corpus can be created by a translation model such as IndicTrans. Such data is not considerd of high quality given that the data is representative of the model that is used to create it.
- **Unsupervised**: The final category of dataset includes unsupervised data typically useful for pretrainign models and for mining aligned pairs. For instance, for speech transcription, unsupervised audio data is used to pretrain models such as wav2vec. Similarly, for mining bitext pairs large monolingual corpora are used.

All participating institutes have been contributing such data to ULCA. Such data has been collected using five different modes: (i) curated from government sources on the web (e.g., News on AIR) (ii) curated from non-government sources on the web (e.g., Times of India) (iii) curated from sources which are free of any copyright (e.g., books whose copyright period has expired) (iv) collected manually using crowdsourcing platforms with explicit consent of the participants (v) collected manually using in-house or outsourced annotators who are explicitly paid for the content. There are a few other public sources of data which have not been ingested into ULCA yet. We are working with the contributors/creators of these datasets to get them added to ULCA at the earliest.

In the following, we have listed the datasets avaialble in each of the different modalities with language-wise counts on the data items.

TABLE 1: MACHINE TRANSLATION DATASETS REPORTED IN THOUSANDS OF SENTENCES
DATA TYPE LEGEND – M: MANUAL, P: POST-EDITED, MT: MACHINE TRANSLATED, MA: MACHINE ALIGNED

Source	Type	In ULCA	as	bn	brx	doi	gu	hi	kn	ks	gom	mai	ml	mni	mr	ne	or	pa	san	sat	sd	ta	te	ur
ELRC_2922	MA	N	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0
GNOME	M	N	30	41	0	0	39	30	24	0	0	0	23	0	27	0	21	34	0	0	0	31	38	0
GlobalVoices	MA	N	0	138	0	0	0	3	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0
JW300	MA	N	46	269	0	0	306	511	316	0	0	0	371	0	289	0	0	374	0	0	0	718	204	0
KDE4	M	N	7	35	0	0	32	86	14	0	0	0	40	0	12	0	8	79	0	0	0	80	15	0
Mozilla-I10n	M	N	8	22	0	0	0	1	13	0	0	0	13	0	16	0	9	0	0	0	0	17	25	0
OpenSubtitles	M	N	0	373	0	0	0	82	0	0	0	0	358	0	0	0	0	0	0	0	0	29	23	0
TED2020	M	N	1	11	0	0	16	47	3	0	0	0	6	0	23	0	0	1	0	0	0	12	6	0
Tanzil	M	N	0	185	0	0	0	186	0	0	0	0	185	0	0	0	0	0	0	0	0	93	0	0
Tatoeba	M	N	1	6	0	0	1	11	1	0	0	0	1	0	54	0	1	1	0	0	0	1	1	0
Ubuntu	M	N	21	28	0	0	28	26	22	0	0	0	23	0	26	0	20	30	0	0	0	26	25	0
alt	M	N	0	21	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
banglanmt	MA	N	0	2380	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bible-uedin	M	N	0	0	0	0	16	62	62	0	0	0	61	0	61	0	0	0	0	0	0	0	62	0
cvit-pib	MA	N	0	92	0	0	59	267	0	0	0	0	44	0	115	0	95	102	0	0	0	116	45	0
iitb	M	Y	0	0	0	0	0	1604	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mtenglish2odia	M	N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	35	0	0	0	0	0	0	0
nlpc	P	N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	32	0	0
odiencorp	M	N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	91	0	0	0	0	0	0	0

pmi	MA	N	7	24	0	0	42	51	29	0	0	0	27	0	29	0	32	29	0	0	0	33	34	0
sipc	P	N	0	21	0	0	0	38	0	0	0	0	30	0	0	0	0	0	0	0	0	36	44	0
tico19-terminologies	M	N	0	1	0	0	1	1	1	0	0	0	1	0	1	0	0	1	0	0	0	1	1	0
ufal	MA	N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	167	0	0
urst	MA	N	0	0	0	0	65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
wikimatrix opus (threshold 1.04)	MA	N	0	281	0	0	0	232	0	0	0	0	72	0	125	0	0	0	0	0	0	96	92	0
wmt-2019-wikipedia	MA	N	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
wmt2019-govin	MA	N	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Samanantar	N	Y	33	5109	0	0	2457	7309	3622	0	0	0	4688	0	2870	0	770	2350	0	0	0	3809	4354	0
FLORES Benchmark	B	Y	2	2	0	0	2	2	2	0	0	0	2	0	2	2	2	2	0	0	0	2	2	2
WAT 21 Benchmark	B	Y	3	3	0	0	3	3	3	0	0	0	3	0	3	0	3	3	0	0	0	3	3	0
Indic Educational books IndicTrans (Anuvaad)	MT	Y	450	586	0	0	591	607	584	0	0	0	576	0	598	0	571	595	0	0	0	570	582	0
Total			609	9629	0	0	3688	11181	4696	0	0	0	6525	0	4251	2	1659	3603	0	0	0	5873	5557	2
Total in ULCA			1335	3737			3688	7533	3134				3456		7791		1779	3941				3422	4634	165

TABLE 2: ASR DATASETS REPORTED IN HOURS OF AUDIO
DATA TYPE LENGEND – M: MANUAL, C: CROWDSOURCED, P: POST-EDITED, MT: MACHINE TRANSCRIBED, MA: MACHINE ALIGNED, U: UNSUPERVISED

Source	Type	In ULCA?	as	bn	brx	doi	gu	hi	kn	ks	gom	mai	ml	mni	mr	ne	or	pa	san	sat	sd	ta	te	ur
AI4Bharat	U	N	843	1035	64	614	1061	1075	1012	436	499	38	857	464	1054	707	1018	863	500	9	107	1012	1052	721
MUCS + MSR	M	Y	0	0	0	0	40	95	0	0	0	0	0	0	94	0	95	0	0	0	0	40	40	0
OpenSLR	M	N	0	71	0	0	0	0	0	0	0	0	0	0	0	70	0	0	0	0	0	0	0	0
IITM	M	Y	0	0	0	0	0	178	0	0	0	0	0	0	0	0	0	0	0	0	0	112	0	0
IITH Telugu	C	N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2000	0
Bhashini	C	N																						
MUCS benchmark	M	Y	0	0	0	0	5	6	0	0	0	0	0	0	1	0	5	0	0	0	0	4	5	0
OpenSLR benchmark	M	N	0	5	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0
MSR benchmark	M	Y	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4	0
IITM benchmark	M	Y	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0
Vakyansh synthetic (labelled ASR)	MT	Y	0	798	0	0	385	2330	460	0	0	151	0	0	1318	442	582	0	125.42	0	0	1071.56	983.11	0
Vakyansh synthetic	U	Y	649.988	841.494	28.992	298.901	352.887	2383.74	457.575	251.297	340.075	130.997	700.936	220.986	1390.05	514.482	565.49	851.94	122.391	5.899	45.403	1157.36	936.65	415.119

(unlabelled ASR)																								
IIITM-K	M	Y	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
SMC	M	Y	0	0	0	0	0	0	0	0	0	0	1.58	0	0	0	0	0	0	0	0	0	0	0
Total labelled			0	873	0	0	424	2607	459	0	0	150	3	0	1411	516	676	0	125	0	0	1227	302	0
Total labelled in ULCA			0	797	0	0	384	248	459	0	0	15	3	0	1317	441	0	581	125	0	0	1195	983	0
Total unlabelled			1492	1876	92	912	1413	3458	1469	687	839	168	1557	684	2444	1221	1583	1714	622	14	152	2169	1988	1136
Total unlabelled in ULCA			649	841	28	298	352	2537	457	251	340	130	702	220	1390	514	565	851	122	5	45	1282	936	415

TABLE 3: TTS DATASETS REPORTED IN HOURS OF AUDIO DATE TYPE LEGEND – M: MANUAL

Source	**Type**	**ULCA submitted**	**as**	**bn**	**Brx**	**doi**	**gu**	**hi**	**kn**	**ks**	**gom**	**mai**	**ml**	**mni**	**mr**	**Ne**	**or**	**pa**	**san**	**sat**	**sd**	**ta**	**te**	**ur**
IITM	M	Y	27.39	20.07	9.78	0	31.69	41.86	19.16	0	0	0	20.89	20.75	16.4	0	9.66	27	0	0	0	53.59	36.71	0
IIITH	M	N	0	0	0	0	0	21.47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total			27	20	10	0	32	63	19	0	0	0	21	21	16	0	10	27	0	0	0	54	37	0
Total in ULCA			27.39	20.07	9.78	0	31.69	41.86	19.16	0	0	0	20.89	20.75	16.4	0	9.66	27	0	0	0	53.59	36.71	0

TABLE 4: OCR (SCENE DETECTION) DATASETS REPORTED IN THOUSANDS OF IMAGES DATA TYPE LEGEND – M: MANUAL, MG: MACHINE GENERATED

Source	Type	In ULCA	as	bn	brx	doi	gu	hi	kn	ks	gom	mai	ml	mni	mr	ne	or	pa	san	sat	sd	ta	te	ur
MLT-19	MG	N	0	49	0	0	0	33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	52
IITM	M	N	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ICDAR-19	M	N	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
IIIT-ILST	M	N	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1
Total			0	495	0	0	406	593	316	0	0	0	495	0	0	0	426	483	0	0	0	483	396	555

TABLE 5: OCR (SCENE RECOGNITION) DATASETS REPORTED IN THOUSANDS OF IMAGES DATA TYPE LEGEND – M: MANUAL, MG: MACHINE GENERATED

Source	Type	In ULCA	as	bn	brx	doi	gu	hi	kn	ks	gom	mai	ml	mni	mr	ne	or	pa	san	sat	sd	ta	te	ur
IIITH	M	N	0	0	0	0	1	4	0	0	0	0	1	0	0	0	0	0	0	0	0	3	2	2
ICDAR	M	N	0	4	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
Kaggle	M	N	0	4	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AI4Bharat	M	N	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total			0	5508	0	0	5501	6015	4500	0	0	0	6001	0	0	0	2400	5800	0	0	0	6003	6002	5307

TABLE 6: OCR (DOCUMENT) DATASETS REPORTED IN THOUSANDS OF IMAGES
DATA TYPE LEGEND – M: MANUAL, MG: MACHINE GENERATED

Source	Type	In ULCA	as	bn	brx	doi	gu	hi	kn	ks	gom	mai	ml	mni	mr	ne	or	pa	san	sat	sd	ta	te	ur
IndicOCR-v2	MG	N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	335	0	0	0	0	0
Multilingual OCR (IIIT-H)	M	N	0	3	0	0	5	5	5	0	0	0	5	4	5	0	5	0	0	0	0	5	5	0
iiit-indic-hw-words	M	N	0	113	0	0	116	95	103	0	0	0	116	0	0	0	101	0	24	0	0	103	120	100
IndicOCR-v2	M	N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	0	0	0	0	0
Indian Language Benchmark Portal-Offline (IIIT-H)	M	N	1	1	0	0	1	0	1	0	0	0	1	1	1	0	1	0	0	0	0	1	1	0
Indic Educational books OCR (anuvaad)	MG	Y	74	135	0	0	341	599	384	0	0	0	633	0	233	0	212	541	0	0	0	732	320	0
Total			75	252	0	0	463	699	493	0	0	0	755	5	239	0	319	541	383	0	0	841	446	100
Total in ULCA			58	146	0	0	153	234	267	0	0	0	333	0	194	0	144	115	0	0	0	314	226	0

DMU’s Data Collection Goals

For each of the 4 tasks, viz., Machine Translation, Automatic Speech Recognition and Optical Character Recognition, we list down the desired characteristics of the data as well as our data goals in the Table below.

Machine Translation

To train and evaluate Machine Translation systems for a given language pair we need parallel sentences for this language pair. Such parallel data should have the following characteristics:

- **Diversity in domains**: The parallel sentences should cover a wide variety of domains such as Legal/Govt, History, Geography, Tourism, STEM, Religion, Business, Sports, Entertainment, Health, Culture and News
- **Diversity in lengths:** For every domain of interest, it is desired that the following ranges of sentence lengths have a good representation: 6-10 words, 11-17 words, 18-25 words, > 25 words.
- **N-way parallel**: The data should have n-way parallel sentences, *i.e.*, a large fraction of the data should contain the same English sentences translated to all the 22 constitutionally recognised languages.
- **Source original**: For each language, the data should contain some sentences which were originally return in that language and then translated to English
- **Discourse level translations**: Instead of collecting translations of isolated sentences, it is preferred to translate entire paragraphs or a collection of contiguous sentences so that the data can also be used for training/evaluating discourse level translation models
- **Downstream applicability**: While collecting training data from a variety of domains would be useful, one should also focus on collecting data for building practical applications, such as translation for everyday usage/conversations

Based on the above wishlist, we will collect 100K parallel sentences between English and each of the 22 languages. The distribution of these 100K sentences would be as follows:

- 50K English sentences taken from Wikipedia and government sources from 12 different domains, viz. Legal, Government, History, Geography, Tourism, STEM, Religion, Business, Sports, Entertainment, Health, Culture, News. These sentences would be translated to all the 22 languages to create n-way parallel data. This will ensure that the parallel data has diversity in domains and contains formally written content.

- 30K English sentences from daily conversations in the Indian context in 20 different domains (e.g., railway stations, Indian tourist spots, etc). These sentences would be translated to all the 22 languages to create n-way parallel data. This will ensure that the parallel data has diversity in domains and contains informally written content with a focus on everyday conversations (a primary use case of speech-to-speech translation systems).

- 5K English sentences corresponding to reviews of 500 popular products which will be translated to all the 22 languages to create n-way parallel data. This will again ensure that the parallel data has some commercial content and diversity in writing style.

- 10K English sentences taken from government acts and policies which will be translated to all the 22 languages to create n-way parallel data. This will ensure representation of content that is typically translated by government bodies.

- 5K regional languages sentences taken from books which were originally written in the regional languages. For each of the 22 languages, such 5K sentences will be translated to English (this will not be n-way parallel).

10% of the above data will be reserved as benchmark data and the rest will be used as training data. Our MT data collection goals are summarised in the table below.

TABLE 7: GOALS FOR COLLECTING PARALLEL SENTENCES BETWEEN ENGLISH AND ALL 22 CONSTITUTIONALLY RECOGNIZED LANGUAGES

Domain	Source	Number of sentences	Direction	n-way parallel
Legal	Government	5,000	En-X	Yes
Government Policies	Government	5,000	En-X	Yes
History	Wikipedia	5,000	En-X	Yes
Geography	Wikipedia	5,000	En-X	Yes
Tourism	Wikipedia	5,000	En-X	Yes
STEM	Wikipedia	5,000	En-X	Yes
Business	Wikipedia	5,000	En-X	Yes
Sports	Wikipedia	5,000	En-X	Yes
Entertainment	Wikipedia	5,000	En-X	Yes
Health	Wikipedia	5,000	En-X	Yes
Culture	Wikipedia	5,000	En-X	Yes
News	Wikipedia	5,000	En-X	Yes
Everyday conversations	AI4Bharat	30,000	En-X	Yes
Product Reviews	AI4Bharat	5,000	En-X	Yes
Literature*	Books	5,000	X-En	No

* Source original content is mainly derived from books written in that language. However, taking content from such books may have copyright issues. We are working to resolve this.

The guidelines that will be used for collecting the above data are mentioned in Appendix A.1.

Automatic Speech Recognition

To train and evaluate ASR systems we need audio files and their captions. Such parallel data should have the following characteristics:

- **Diversity in speakers**: For every language, the audio data should be collected from a wider variety of speakers having different accents (e.g., Surati v/s Vadadara), different ages (18-30, 30-45, 45-60, >60), different educational backgrounds (school level, graduate, post-graduate) and different genders.

- **Diversity in collection method**: The data should contain a mix of read speech, extempore conversations, and broadcast content such as news, educational videos, entertainment video.
- **Diversity in vocabulary**: The audio should contain words from a wide variety of domains.
- **Diversity in genres**: The audio sourced from broadcast content should come from a wide variety of genres (e.g., news debates, on-field news reports, comedy shows, reality shows, how-to videos, STEM videos, etc)
- **Downstream applicability**: While collecting training data from a variety of domains, genres and speakers would be useful, one should also focus on collecting data for building practical applications, such as, voice commands for everyday usage in digital payments, e-commerce, etc.

For collecting data for training ASR models, we will adopt two methods: (i) collect data from the field to ensure speaker diversity and coverage of specific content which is hard to obtain elsewhere (e.g., voice commands) (ii) label existing data from news, entertainment, and educational content.

COLLECTING DATA FROM THE FIELD

We will collect data from 600 speakers spread across districts wherein each speaker will:

- read 100 sentences (~10 minutes)
- speak 200 voice commands (~10 minutes)
- participate in an extempore get-to-know-me interview (~10 minutes)
- read 100 English sentences (only from a few speakers who also speak English. This will ensure that we also collect Indian accent English data on the field).

This will ensure that we collect data which has (i) high speaker diversity (number and variety), high content diversity (the 100 sentences will come from a larger pool of 50000 diverse sentences from different domains) and (iii) high downstream applicability (voice commands catering to a variety of use cases).

LABELLING EXISTING AUDIO/VIDEO DATA

We will label existing data from youtube and the content/media industry. This data can be further split into the following types:

- **News**: This will be primarily sourced from news channels and can be further categorised into the following types:
- **Headlines**: This is content of the type "Top 20 headlines of the hour" which does not have high speaker diversity but has peculiar characteristics like jarring background music.
- **On-field reporting**: This is content of the type "cameraman Prakash ke saath…." which is extempore, has background noise and involves common people on the ground.
- **Debates**: Such content would have diversity in content (government policies, banning an outfit, etc.) and will also have peculiar characteristics like emotional outbursts, overlapping chatter, etc
- **Interviews**: Such content would involve a news anchor and 1-2 experts and caters to a variety of topics. The experts do not follow a script, so the content has the flavour of natural speech.
- **Special reports**: Such content involves people on the ground and has good vocabulary spanning multiple domains.
- **Entertainment**: This will be primarily sourced from entertainment channels and would include content from different genres: family shows, comedy shows, crime shows, reality shows, cooking shows, travel shows, songs
- **Education**: This will be primarily sourced from education channels and would contain content from STEM, Health and How-to videos.
- **Call centre**: This will be primarily sourced from call centres catering to one or more of the following domains: agriculture, legal, banking, insurance, health

Our ASR data collection goals are summarised in the Table below.

TABLE 8: GOALS FOR COLLECTING TRANSCRIBED AUDIO DATA FOR ALL 22 CONSTITUTIONALLY RECOGNIZED LANGUAGES

Type of Data	Hours MR/LR	Potential Challenges

On-field data collection	300/300	
Read speech	100/100	The clean sentences to be read by the speakers may either come from native language books or from the translations of English sentences collected by AI4Bharat. If it's the former, then we will have copyright issues. If it's the latter, then we will have to delay this activity till we have enough sentences translated.
Voice commands	100/100	
Extempore conversation	100/100	
Labelling existing audio data	700/200	For low resource languages such as Assamese, Bodo, Dogri, Kashmiri, Konkani, Maithili, Manipuri, Nepali, Odia, Sanskrit and Santali we will label only 200 hours of existing audio data.
News	200/100	We will depend on content providers such as Prasar Bharati to get the raw audio content. As of now it is not clear whether such content will be shared with us and if it will have enough diversity. We may have to revise these goals based on the availability of such data.
Entertainment	200/50	We will depend on content providers such as Prasar Bharati to get the raw audio content. As of now it is not clear whether such content will be shared with us and if it will have enough diversity. We may have to revise these goals based on the availability of such data.
Education	200/50	We will depend on content providers such as Prasar Bharati to get the raw audio content. As of now it is not clear whether such content will be shared with us and if it will have enough diversity. We may have to revise these goals based on the availability of such data.
Call centre data	100/0	We will depend on government/private call centres to share the raw audio data with us. Given the general privacy concerns around such data, we are not sure if we will get access to it. We may have to revise these goals based on the availability of such data.

Text-to-Speech

To train and evaluate TTS systems we need high quality audio recordings from professional voice artists along with textual scripts/prompts. Such data should have the following characteristics:

- **High quality recording**: The data should be collected in a studio setup.
- **High quality voice**: The data should be collected from a professional voice artist.
- **Diversity in domains**: The spoken content should contain words from a wide variety of domains such as Legal/Govt, History, Geography, Tourism, STEM, Religion, Business, Sports, Entertainment, Health, Culture and News
- **Diversity in content**: The scripts used for recording should contain a mix of short statements, long statements, questions, exclamations and short phrases.

Based on the above wishlist, the Text-to-speech data will be collected with the help of professional voice artists hired through a production studio. For each language, we will collect 20 hours of data from a male artist and 20 hours of data from a female artist. The artists will be given prompts from multiple domains. The prompts will be derived from the following two sources:

- **Source original contents from books**: As mentioned earlier (section 4.4.1), we will be sourcing around 5,000 sentences from books which were originally written in the native language. In addition to translating these sentences to English, we will also use them as prompts for the voice artists. Such content taken from books is typically rich in sentence structure, vocabulary, and emotions. Hence, it would be ideal for recording by professional artists.
- **Translations from multiple domains**: As mentioned earlier (section 4.4.1), we will be translating English sentences taken from multiple domains into regional languages. These translated sentences will contain diverse content from multiple domains and will be provided as prompts for the voice artists. This will ensure that the recorded content has a good representation of domains and broader coverage of vocabulary.

The above sentences will contain a mix of statements (70%), questions (10%), exclamations (10%) and short phrases/commands (10%).

TABLE 9: GOALS FOR STUDIO QUALITY TTS DATA FOR ALL THE 22 CONSTITUTIONALLY RECOGNIZED LANGUAGES

Prompts	# Sentences	Characteristics	Hours	Potential challenges
Source original content	5,000	rich in native language structure, vocabulary, and emotions	5	Source original content is mainly derived from books written in that language. However, taking content from such books may have copyright issues.
Translations from multiple domains	15,000	good representation of domains and broader coverage of vocabulary	15	Will depend on the translations collected for MT data and hence there might be some delay before this activity can be started.

Document Optical Character Recognition

Here, we will focus on the task of detecting the layout of a document. For this, we need pdf pages where the different parts of the page are clearly highlighted by a bounding box and the order of the boxes is also specified. The pdf pages that we select should have the following characteristics:

- **Diversity in font sizes**: The page should contain text written in a wide variety of font sizes.
- **Diversity in font types**: The page should contain text written in a wide variety of font types.
- **Diversity in layouts**: The page should have a variety of layouts (one column, two column, magazine style, newspaper style, etc.)
- **Diversity in artefacts**: The page should contain a variety of artefacts such as tables, figures, indentations, sections, sub-sections, bullet lists, etc
- **Diversity in background effects**: The scanned pages should have a variety of background effects such as crumbling, lighting effects, scan marks, page fold marks, etc.

Based on this wishlist, we will create pdf pages where different layout templates will be created, and regional language content will be inserted in these layouts. We will be focusing only on machine-generated PDFs and not scanned PDFs.

TABLE 10: GOALS FOR DOCUMENT LAYOUT DETECTION DATA FOR ALL THE 22 CONSTITUTIONALLY RECOGNIZED LANGUAGES

Image source	Images	Configurations
Books (Synthetic)	2,000	Zoom - [0.5x,1x,2x] Image Quality - [Clean, 2D Degraded, Botched] Lighting conditions - [Normal, Modified] Text orientation – Horizontal
Question papers (synthetic)	2,000	
Application forms (Syntetic)	2,000	
Reciepts and invoices (synthetic)	2,000	
Letters/Orders/Notices/Prescriptions (synthetic)	2,000	
Legal documents and court judgements (synthetic)	2,000	

Scene Optical Character Recognition

Here, we will focus on the task of detecting text in images of natural scenes (such as photos taken by a user's camera). For training and evaluating such a system we need natural images containing text where the area containing the text is highlighted by a rectangular bounding box and the text inside is typed out in unicode. Such data should have the following characteristics.

- **Diversity in background**: The images containing text should have very diverse background such as buildings, sky, trees, etc.
- **Diversity in angles**: The images should be taken from different angles (left, right, top, bottom, etc).
- **Diversity in font types**: The images should contain text written in a wide variety of fonts.
- **Diversity in font sizes**: The images should contain text written in a wide variety of sizes.

- **Diversity in orientation**: The images should contain text with different orientations (horizontal, slanting, circular, etc).
- **Diversity in ambient light**: The images should be taken under different lighting conditions.

Based on the above wishlist, we will create a benchmark of 10000 images for each language as summarised in the table below.

TABLE 11: GOALS FOR SCENE TEXT RECOGNITION FOR ALL THE 22 CONSTITUTIONALLY RECOGNIZED LANGUAGES

Image source	Images	Configurations
Signboards	1,250	Zoom – Zoomed in, zoomed out, normal Image Quality – high, low Lighting condition – Daylight, night, dim light, artificial light Viewpoint of photographer – Top, below, normal Distance of photographer – Far, close, normal
Billboards	1,250	
Movie posters	1,250	
Shop banners	1,250	
Political banners	1,250	
Railway station boards	1,250	
Advertisements	1,250	
Highway milestones / Mile markers	1,250	

DMU's Approach

To collect data at this scale, we need to set up cross country teams comprising of people having deep expertise/experience in creating datasets for specific languages. Second, we need a unified data collection interface which uses best practices from UI/UX design as well as software engineering to ensure that high quality data can be collected at scale. We elaborate on these two ideas in the following subsections. Third, we should mine data from existing sources wherever possible and have a clear policy on how to use data from various sources.

Set up a cross-country team of language experts

To set up a diverse team of language experts across the 22 languages, we will take a 4-pronged approach.

Recruiting in-house translators: For all the 22 languages, we are hiring a team of 5 junior language experts (translators/annotators/transcribers) and 2 senior language experts. These experts will be directly on the payroll of AI4Bharat.

Partnering with universities: For Kashmiri, Urdu and Konkani we will partner with specific academic institutes (Kashmiri University, Goa University).

Partnering with the social sector: For 8 languages (Assamese, Bodo, Dogri, Maithili, Manipuri, Nepali, Sanskrit, Santali) we have partnered with entities or individuals working in the social sector.

Outsourcing to data collection agencies: For TTS, where we need to collect studio quality data from professional voice artists, we will be outsourcing the activity to a professional (this was recommended by our colleagues at IITM who have done a fair amount of data collection with them in the past). Similarly, for voice collection for 11 languages (Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, Telugu, Urdu) we will be partnering with an external data collection agency. Their help will be needed in collecting voice samples from many districts in the country. In addition, for some languages where it is difficult to find language experts, we may partner with some data collection agencies to collect translations also.

The table below summarises our plan for setting up such a cross-country team of language experts.

TABLE 12: SUMMARY OF THE APPROACH USED FOR SETTING UP A CROSS-COUNTRY TEAM OF LANGUAGE EXPERTS

Language	Translation	ASR	TTS	OCR
Assamese	AI4Bharat translators	Pragyam Foundation		Pragyam Foundation

Bengali		Outsourced	Professional studio	Outsourced
Bodo		Pragyam Foundation		Pragyam Foundation
Dogri		J&K Higher Education		J&K Higher Education
Gujarati		Outsourced		Outsourced
Hindi				
Kannada				
Kashmiri		Kashmir University		Kashmir University
Konkani		Goa University		Goa University
Maithili		Aripana Foundation		Aripana Foundation
Malayalam		Outsourced		Outsourced
Manipuri		Koru Foundation		Koru Foundation
Marathi		Mumbai University		Mumbai University
Nepali		Pragyam Foundation		Pragyam Foundation
Odia		TBD		TBD
Punjabi	AI4Bharat translators + Outsourced	Outsourced		Outsourced
Sanskrit	AI4Bharat translators	Aripana Foundation		Aripana Foundation
Santali	AI4Bharat translators	Suchana Uttor Chandipur Community Society		Suchana Uttor Chandipur Community Society
Sindhi	AI4Bharat translators + Outsourced	TBD		TBD
Tamil	AI4Bharat translators	Outsourced		Outsourced
Telugu				Outsourced
Urdu				Kashmir University

While the above clarifies our plans to build a cross-country team of langauge experts, we want to flag the potential risk of not finding expertise (such as in translation from English to Dogri) and representation (such as voice diversity in Santali). We are looking forward to support from academics and the government in this regard.

Build a unified tool for data collection

The goal of the DMU is to collect human annotated data across multiple languages (22 constitutionally recognized languages) and multiple tasks (monolingual text curation, translation, transliteration, speech curation, speech transcription, speech recording, character recognition, natural language understanding). Across both these axes, in total the DMU aims to collect human inputs totalling over 250,000 hours of effort. Collecting data at this scale requires a tool with a few features that are described below.

GOALS AND FEATURES IN THE DESIGN OF SHOONYA

- First, the tool needs to support custom user-interfaces for each of the languages and tasks. These user interfaces must have two properties - (a) they must be efficient for annotators who will repetitively use them to collect data, and (b) it must be easy to create a new interface (eg. an interface to mark sentences as having hate speech) without requiring any major code changes.
- Second, given that most tasks are serially chained (eg. collecting monolingual corpus, then labelling it for offensive content, then translating, then speech recording), it is essential to maintain the links between the data items across all these tasks. Thus, the tool must provide a common ground view of the data that is enriched by different annotation tasks.
- Third, the tool must allow for automated functions and filtering between serially chained tasks as per various criteria. For example, when creating a task for translation from the output of a monolingual text curation task, we may only want to filter sentences that have a length of larger than 10 words. This would require a function to first process the dataset and create a column on the sentence length and then a filter when exporting data based on this column.
- Fourth, the tool must allow deep integration with machine learning tools and processes. These integrations are of two kinds - (a) the human annotation effort should be provided with assistance and automated verification from existing ML models, e.g. an audio transcription task has prepopulated noisy outputs from an ML model which are edited, and (b) the data created by the annotation tasks must be available for building ML models with full access to metadata, e.g. train an NMT model on all sentence pairs marked by annotators to have a rating of 4 and above on a scale of 5.

- Finally, the tool must be able to export data in standardised formats such as the ULCA format and formats commonly used for training deep learning models. Further the data exports need to be version controlled with a global naming scheme across all tasks.

Shoonya is being developed at AI4Bharat with the above properties and will be released in the open source. It will enable the DMU to meet the following goals -

- The efficiency of data collection at the DMU will be enhanced and it will be easier to follow common guidelines of data collection across languages.
- All datasets created with the tool will have rich meta-data linking data items across various annotation efforts. This will provide detailed accountability on how the data was curated/created and annotated.
- It would be easier to coordinate multiple annotation tasks across languages (eg. creating n-way parallel data) and across tasks (eg. speech-to-speech translation).
- All machine learning models that are trained with data created in the tool will be reproducible by logging the data version and the associated metadata.
- Deploying a common tool on the cloud can enable it to accommodate multiple organisations, language teams, and projects to be concurrently serviced.

SOFTWARE ARCHITECTURE AND TECHNICAL TERMS IN SHOONYA

Below is an early draft of how Shoonya will be architectured.

- Shoonya will support multiple organisations which are groups of non-overlapping language contributors working on separate projects that are all hosted on the same deployed instance.
- Each organisation has one or more workspaces which naturally map to groups of people working on related projects.
- Under a workspace there are datasets which are structured databases with defined schema from a set of dataset types. For instance, one dataset type could be for translation related annotations. These dataset instances store the ground-truth view of data that is uploaded, curated, created, or annotated.

- Each annotation effort is called a project from a fixed set of project types. Example of project types that are planned for support in Shoonya include: monolingual text collection, sentence splitting, content domain classification, inappropriate content flagging, translation, marking parts of a text (for example answer phrase in a QA task), transcribing audio both by editing the subtitle and marking the timeline, OCR annotation both by marking text regions and by entering the text, recording audio, etc. Many of these will include versions where automated tools can speed up the annotation process.
- Each project requires input data which is extracted from a dataset and generated annotated data which are also exported back to, often the same, dataset. A project definition includes these ingestion and egestion rules. Further, creation of a project can also be with input data that is obtained by applying specified sampling strategies and content filters to a dataset. For instance, a translation project may be created by randomly sampling one half of rows from a dataset which are from the legal domain.
- Often multiple projects update the data in a dataset. For instance, a dataset can be created by a monolingual text collection project followed by another project on translation on the same dataset type. We group together such projects as project domains. There is often a one-to-one correspondence between project domain and a dataset type: Projects belonging to the same project domain often update the same dataset that is of the corresponding dataset type.
- Each entry in a dataset has additional meta-data on the source of that data-item which includes the project and the annotator that created it.
- The datasets allow versioned exports of the data in different formats. Shoonya will also support an export in ULCA compatible formats.

Mining Data at Scale

While collecting data with the help of experts and manual supervision is important, it is also expensive and time consuming. Hence, it is important to complement such manual data collection efforts with parallel efforts on mining data from the internet. For example, in the context of Machine Translation, recent works have shown that it is possible to mine parallel sentences from web scale corpora between the source and target languages. While the DMU will continue mining such data from public sources, it should be noted that many translations/transcriptions created by various government agencies are often not easily accessible. To unlock such data the Data Management Unit would seek assistance from MeitY to strike partnerships with various government agencies. A few of these are listed below:

- **Parliamentary proceedings**: Some of the parliamentary proceedings get translated to a few Indian languages. These would be a great source of India specific parallel content if they can be made available by the government.
- **Central Acts/Policies**: Many central acts and policies get translated to multiple Indian languages (e.g., National Education policy). However, they are often not available in machine readable format. It is desirable that respective departments make such data available in an easy to consume digital format.
- **State Legislative Assemblies**: Many bills, speeches, proceedings of the state legislative assemblies are translated into the regional language of the state as well as English. DMU will seek MeitY's assistance in getting such data from state governments.
- **Court orders/judgements**: Many Supreme Court judgements get translated from English to multiple Indian languages. However, these are not very easily accessible. If made accessible, they could be a great source of hard-to-get domain specific parallel data.
- **Government advisories**: Many government advisories (such as COVID protocols) get translated into multiple Indian languages. These are again hard to obtain and appropriate government departments should be contacted to get digital copies of such advisories
- **National Book Trust**: NBT contains a large number of books in all the 22 constitutionally recognised languages. Some of these books are parallel and have been exploited in the past to create the Gyan Nidhi corpus. It is desirable to work closely with the National Book Trust to obtain all such parallel content.
- **National Translation Mission**: NTM is a government body which undertakes various translation activities for the government. An important ongoing activity is translation of higher education books to multiple Indian languages. Such content is valuable and is desirable for training robust Machine Translation systems.
- **Government websites**: Many government websites have parallel content in English and Hindi. Such websites need to be systematically curated to get parallel content.
- **Audio/video content from Prasar Bharati**: Prasar Bharati is India's public broadcaster which includes the Doordarshan Television Network (DD) as well as All India Radio (AIR).

Both DD and AIR broadcast audio/video content in multiple Indian languages. Such raw audio would be very valuable for pre-training speech models and can also be quickly labelled for tasks such as speaker recognition. Further, some shows, such as news broadcasts also have a script which can be used to create parallel speech to text data.

- **Swayam Lectures**: Swayam is India's public MOOCs platform hosting educational content in English. Many videos have English subtitles of which some have been translated to multiple Indian languages. This is a great source of parallel data from technical domain as well as of speech data for training English ASR models for Indian accents.

The DMU believes that successfully building an open data infrastructure for Indian languages requires (a) deep collaboration and partnership of the DMU with government bodies with access to Indian language content listed above, and (b) open acceptance of practices such as web-scale mining, frequently being adopted in industry, amongst researchers.

Policy on sources to collect data from

In the following we discuss the DMU's policy in collecting and curating "source" data required for data collection for different tasks. This is a draft policy and is subject to alignment with mission's data policy.

To create n-way parallel data we will take sentences from Wikipedia and translate them to multiple Indian languages. Similarly, while collecting voice samples in the form of "read speech" we will ask users to read sentences from books, novels, etc. Lastly, for creating transcribed data, we will take audio/video content from multiple sources. It is thus important to have a clear policy on the licensing terms that will be used while curating such data from online/offline sources. DMU's policy will be based on the following principles:

- Any content which comes under CCBY 4.0 license will be taken without any restrictions (e.g., Wikipedia). The transformations done by DMU on such data will be released under permissible license. For example, our policy would be to take sentences from Wikipedia without any restrictions. The translations of these sentences would be released under permissible license. Similarly, if these sentences are spoken out by speakers (after taking appropriate consent), then the voice data along with the original sentences will also be released under CCBY 0 license.

- While dealing with copyrighted content (e.g., books, audio/video, images or any other proprietary content) we will follow the de minimis principle. Specifically,
    - we will not take more than 5% of the content from a single source (e.g., we will not take more than 10 pages from a 200-page book or more than 3 minutes from a 30-minute video)
    - we will not take more than 10 contiguous sentences from a single source or equivalently more than 5 mins continuous audio from a single source
    - we will ensure that the original content cannot be reconstructed from the data that we release
    - we will give due credit to the author and the publisher of the content by including their names in the meta-data

The content here could refer to a book, a website, an article, a magazine, a blog, an image, an audio file, a video file, etc. Any transformations done by DMU on such data will be released under permissible license.

DMU's Timelines

Y1-Q1 (APRIL-JUNE 2022)

- Develop Shoonya v1, as an open-source tool for collecting MT, ASR and NLU datasets for all the 22 languages.
- Set up teams of language experts (annotators, translators, transcribers) for all the 22 languages.
- Run pilot for on-field 100 hours of voice data collection for Tamil.
- Collect a total of 50K English sentences from diverse domains which will subsequently be translated to 22 Indian languages.
- Collect a total of 50K sentences of everyday conversational content in English which will subsequently be translated to 22 Indian languages
- Release 1M mined English-X parallel sentences for 11 languages: Bengali, Gujarati, Hindi, Kannada, Malayalam Marathi, Nepali, Punjabi, Tamil, Telugu, Urdu
- Release 500 hours of mined ASR data for 11 languages: Bengali, Gujarati, Hindi, Kannada, Malayalam Marathi, Odia, Punjabi, Tamil, Telugu, Urdu

Y1-Q2

12 Phase 1 languages (P1): Assamese, Bengali, Gujarati, Hindi, Kannada, Maithili, Malayalam, Manipuri, Marathi, Sanskrit, Tamil, Urdu.

10 Phase 2 languages (P2): Bodo, Dogri, Kashmiri, Konkani, Nepali, Odia, Punjabi, Santali, Sindhi, Telugu.

- Develop Shoonya v2, as an open-source tool for collecting MT, ASR and NLU datasets for all 22 languages.
- Create a MT benchmark containing 10K En-X parallel sentences for P1 languages.
- Create an ASR benchmark of 25 hours for P1 languages containing (a) read speech (b) voice commands (c) transcribed extempore conversations (d) transcribed news content (e) transcribed education content (f) transcribed entertainment content.

- Create 10 hours of TTS data for P1 languages.
- Release synthetic training data containing 100K images each for document layout detection, document text recognition and scene text recognition for all 22 languages

Y1-Q3

- Develop Shoonya v3, as an open-source tool for collecting MT, ASR and NLU datasets for all 22 languages.
- Create a MT benchmark containing 10K En-X parallel sentences for P2 languages.
- Create an ASR benchmark of 50 hours for P2 languages containing (a) read speech (b) voice commands (c) transcribed extempore conversations (d) transcribed news content (e) transcribed education content (f) transcribed entertainment content.
- Create 10 hours of TTS data for P2 languages.

Y1-Q4

- Create 30K En-X parallel sentences (fine-tuning data) for all 22 languages
- Create 100 hours of ASR data for all 22 languages
- Create 10 hours of TTS data for all 22 languages
- Create a benchmark for Scene Text Recognition containing 500 images for all 22 languages (13 scripts)
- Create a benchmark for document OCR containing 500 scanned pages for all 22 languages (13 scripts)

Y2-Q1

- Create 30K En-X parallel sentences (fine-tuning data) for all 22 languages
- Create 100 hours of ASR data for all 22 languages
- Create 10 hours of TTS data for all 22 languages
- Create a benchmark for Scene Text Recognition containing 500 images for all 22 languages (13 scripts)

- Create a benchmark for document OCR containing 500 scanned pages for all 22 languages (13 scripts)

Y2-Q2

- Create 40K En-X parallel sentences (fine-tuning data) for all 22 languages
- Create 100 hours of ASR data for all 22 languages
- Create 10 hours of TTS data for all 22 languages

Y2-Q3

- Create 100 hours of ASR data for all 22 languages
- Create 5K QA pairs for all 22 languages
- Create 5K NER tagged sentences for all 22 languages
- Create 5K sentiment labelled sentences for all 22 languages

Y2-Q4

- Create 100 hours of ASR data for all 22 languages
- Create 5K QA pairs for all 22 languages
- Create 5K NER tagged sentences for all 22 languages
- Create 5K sentiment labeled sentences for all 22 languages
- Create 100K translated QA pairs (noisy training data) for all 22 languages
- Create 100K noisy NER sentences (translation + projection) for all 22 languages
- Create 100K translated SA sentences for all 22 languages

Y3

Having finished the data collection activities in the first two years, we will dedicate the third year to building models across different tasks and languages.

APPENDIX

A1. Guidelines for translating source sentences

Below we describe the guidelines to be used while translating sentences from source data. These guidelines are partly inspired from similar guidelines prepared by LDC for the BOLT Chinese-English translation task.

GENERAL PRINCIPLES

- The translation in the target language must be faithful to the text in the source language in terms of both meaning and style. The translation should mirror the original meaning as much as possible while preserving grammaticality, fluency, and naturalness.
- To the extent possible, the translation should have the same speaking style, tone or register as the source. For example, if the source is polite, the translation should maintain the same level of politeness. If the source is rude, excited, or angry, the translation should convey the same tone.
- The translation should contain the exact meaning conveyed in the source text and should neither add nor delete information. For instance, if the original text uses Modi to refer to Honourable Prime Minister Narendra Modi, the translation should not be rendered as Prime Minister Modi, Narendra Modi, etc. No bracketed words, phrases or other annotation should be added to the translation as an explanation or aid to understanding.
- All sentences should be spell checked and reviewed for typographical errors before submission.

NAMED ENTITIES

- Named entities in English which have a well accepted conventional translation in the regional language should be translated using this conventional translation. For example, Indian Institute of Technology would be translated as “भारतीय प्रौद्योगिकी संस्थान” in Hindi.

- If a well accepted conventional translation of the English named entity does not exist in the target language, then the named entity should be transliterated. For example, “Pope Francis” should be translated as “पोप फ्रान्सिस” in Hindi.
- In all cases, avoid inventing translations of named entities in the target language if they do not exist already. Use transliteration instead.
- The above rules are language specific, and it is possible that an English named entity gets translated in one Indian language and transliterated in another. The key deciding factor is the presence or absence of a well accepted conventional translation of that named entity in that regional language.

ERRORS IN THE SOURCE SENTENCE

- Factual errors in the source sentence should be retained as it is. For example, if the source sentence says “Ranveer Singh and Alia Bhatt starrer Brahmastra will release in theatres today” then the translation should also contain this factual error and not correct it to Ranbir Kapoor.
- Spelling mistakes and grammatical errors in the source sentence should be corrected.

NUMBERS AND UNITS

- Numbers in the translation should either be spelled out in full or written as digits, according to how they appear in the source text.
- It is acceptable to use English numerals instead of their equivalents in the regional language. However, we leave this choice to the language experts with the understanding that this choice should be consistent across sentences (i.e., either use English digits in all sentences or regional digits in all sentences).
- Big numbers should be translated using the conventions of the target language. For example, 700 million should be translated as 70 करोड़ as opposed to 700 मिलियन.
- For units of measurement that may differ between English and Indian languages (for example "miles" v/s "kilometres" or “gallons” v/s “litres”), the translators should produce a translation which uses the units of measurement familiar and accepted in the target language but adjusts the number so that it reflects the true measurement. For example,

"3 miles" could be translated as "4.8 किलोमीटर" (using a unit which is more familiar in Hindi but change the number accordingly).

DATES

- Dates in the translation should either be spelled out or written as digits, according to how they appear in the source text. For example, 17 January 2022, would be translated as "17 जनवरी 2022" and not "17-01-2022".

- If there is ambiguity in the date format (mm-dd-yyyy v/s dd-mm-yyyy) then the date should be translated as it is in the source sentence. For example, the English date "01-09-2021" could mean 1st September 2022 or 9th January 2022. If the format is not clear from the context, then the date should simply be translated as "01-09-2021" in Hindi and not changed to "09-01-2022".

- The year should be translated using 4 digits or 2 digits depending on how it appears in the source sentence. For example, "01-09-21" should be translated as "01-09-21" in Hindi and not as "01-09-2021" (even though the latter translation has no ambiguity).

TECHNICAL TERMS

- For translating technical terms, the translators should refer to the translation dictionaries prepared by the commission for scientific and technical terminology for different domains (Science, Engineering/Technology, Medical Science, Humanities, Social Sciences, Agricultural Science, Veterinary Science).

- If a technical term does not appear in the above dictionary, then it should be transliterated into the target language.

A2. Guidelines for collecting voice data from native speakers

We now list down the guidelines for collecting voice data from native speakers at a temporary location. These guidelines are taken (almost verbatim) from the guidelines prepared by NIST for the task of speaker recognition.

COLLECTION ENVIRONMENT

The collection environment should:

- Be an indoor space as free as possible from background noises such as air conditioners, generators, fans, or other motorised or electrical devices. Avoid locations that have music, white noise, or other audio playing in the background at any audio level. It may be necessary to turn the interference source off to fully mitigate it.
- Be a location that is not near outside traffic noise (human, animal, vehicular, or aircraft).
- Have a minimum of large, flat, hard sound-reflective surfaces which can cause reverberation and echoes. The effects of a reverberant room can be mitigated by hanging fabric (curtains, blankets, etc.) or other sound deadening materials on the walls or as dividers in the room.
- Allow the subject to be as comfortable as possible, preferably sitting, to lower cognitive/voice stress levels and to facilitate natural conversation.

COLLECTION EQUIPMENT

Although high-quality digital audio recording equipment is preferred, recordings made with equipment meeting the minimum requirements detailed below can be used . If there are multiple recording sites, it would be ideal to use different recording systems at different (or even at the same site) so that there is enough variety in the input devices that are used (e.g., on-device mics of different phones, headsets of different brands, etc). Nothing in these requirements precludes the concurrent use of multiple recording devices if that is required for the intended application. An example of such a requirement would be to create a pair of recordings in which one is a high-quality reference while the other is condition-matched to a specific use case. Recording devices should fulfil the following requirements:

- The speech must be recorded digitally and saved as uncompressed PCM data with at least 16-bit samples at a minimum rate of 16,000 sps. The audio can be mono or stereo. Recording at a higher sample rate and bit depth is greatly preferred if that is possible. Many current devices support the recording of 16 or 24-bit samples at rates up to 48 kHz and storage of the recorded data in PCM-WAV format.

- The audio should be saved in a standard lossless file format such as PCM-WAV, or be in a file which can be converted to a standard format without loss of fidelity. The audio should not be saved in a file format such as MP3 or WMA which use a lossy codec to compress the audio data. Any type of automatic gain control (AGC) on the microphone or recorder should be turned off/disabled during the recording session.

- For recordings made using laptop or other computers, it is preferred to use an external USB condenser microphone with an on-board analog-to-digital (A/D) converter. This is because the internal microphone or external microphones plugged into a "mic" port can pick up noise from internal circuitry.

- The subject's microphone should ideally be a headset mic since:

  it fixes the location of the mic with respect to the mouth.

  it reduces interference from the interviewer's speech and any background sounds.

  the speaker more quickly forgets about its presence.

- Otherwise, a microphone on a stable stand or tripod which places it at an appropriate distance from the subject for the type of microphone being used is acceptable. If the recording device/microphone is directional, it should be situated to best pick up the subject's speech. If not chosen to mimic an operational scenario, the microphone should ideally provide a flat frequency response.

- The interviewer should have some indicator available on the recording device that shows that the audio is being recorded at an appropriate amplitude level and not too low (resulting in a noisy recording due to quantization effects) or too high (which causes clipping and thereby introduces nonlinear distortion into the audio stream).

- There should be some method available to back-up the collected data, such as writing it to optical media, external hard drives, or USB thumb drives.

SPEECH COLLECTION

After the collection environment and equipment have been arranged, the interviewer should record and audibly review an initial sample of test speech in the same recording environment using the same equipment as for the collection to confirm that the equipment is working properly, and the audio quality meets the parameters discussed above. This may also expose other sources of noise not originally noted, such as the buzz of fluorescent lights or sounds from air handlers, which can be addressed. Once the setup is verified, it should be documented, ideally including the model identification and serial numbers of the equipment and a diagram or photographs of how it was connected and arranged. Once recording begins, either the interviewer or the subject must provide a preamble with some subject identifying information along with the date, time and location of the recording session.

During the recording, the interviewer should strive to elicit periods of conversational speech from the subject. Conversational speech could be elicited in multiple ways, such as:

- Asking open-ended questions or prompts. A list of possible questions is given in the table below. This list is not exhaustive, and the interviewer should tailor any questions to be appropriate for the circumstances, the subject's culture, etc.
- Asking the subject to describe or interpret an image. These could be simple drawings of an object or scene, photographs of a general nature, or other images that have content understandable by the subject and which will elicit a conversational response. Giving the subject a choice of several images gives them the freedom to choose one for themselves. The interviewer could ask multiple questions about an image to elicit additional speech, such as "are there any dangerous things in the image?" Alternatively, the interviewer could circle some items in the image, and ask the subject to describe them.
- Ask the subject to discuss an article from a local newspaper, news website, or social media outlet. Note – in the last two methods, the use of paper copies of the image, drawings or newspaper articles should be avoided since their movement can add undesired noises to the recording.

Sample questions to be asked to elicit a natural speech conversation

Who are the members living in your family?

Describe your favourite place in your city/town/village?

Tell us your favourite children's story?

Tell us about your favourite childhood memory?

Can you tell us about your favourite dish and how you make it?

Can you describe a train?

Try to describe your best friend as vividly as possible. What do you like and dislike about him/ her?

Imagine that you have become extremely rich one day. What would you do with all the money you have?

It can be expected that the longer the subject speaks conversationally (presuming that fatigue does not occur), the greater chance that they will become comfortable with the collection situation, resulting in a more "natural" speech sample. The interviewer should avoid interjections while the subject is speaking (e.g., nodding to acknowledge the subject instead of saying "uh-huh"). The subject's portion of the recording (including the identification segment, any answers to questions, and conversational speech) should contain a minimum of ten minutes of speech and preferably up to thirty minutes. This is to be measured after the removal of speech from the interviewer, any noisy segments, and extended pauses. After the completion of the recording session, the interviewer should document any comments on the collection (via the Shoonya tool) prior to beginning the next session.

A3. Guidelines for transcribing audio data

We now describe the guidelines for transcribing audio data. These guidelines are inspired by similar guidelines created by a commercial transcription agency and by NIST.

GENERAL PRINCIPLES

- Transcribe a word only if you can hear and understand it properly. If the spoken word/text cannot be understood due to the speaker's manner of speech then mark it as [unintelligible]. On the other hand, if the spoken text cannot be heard due to poor recording, volume or noise then mark it as [inaudible].
- Split longer speeches into smaller transcribed segments. As a rule of thumb a transcribed segment should not be longer than 100 characters (20-30 words)
- Do not paraphrase the speech.
- Do not correct grammatical errors made by the speakers.
- Always use the correct spelling for misspoken words. Example: If a speaker pronounces "remuneration" as "renumeration" then it should still be transcribed as "remuneration".
- Capitalise the beginning of every sentence.
- Do not expand spoken short forms (e.g., ain't, don't, can't, it should be retained as it is)
- Retain colloquial slang as it is (e.g., gotcha, gonna, wanna, etc).

VERBATIM TRANSCRIPTION

The speech should be transcribed verbatim. However, the following rules should be used for transcribing errors made by the speakers.

Errors that should be transcribed as it is:

- Speech errors: "I was in my office, no sorry, home" should be transcribed as it is.

- Slang words: kinda, gonna, wanna, etc should be transcribed as it is.
- Repetitions: “I have I have got the book” should be transcribed as it is.

Errors that should not be transcribed:

- False starts: “I, um, er, I was going to the mall” should be transcribed a “I was going to the mall”
- Filler sounds: um, uh, er, hmm, etc should not be transcribed
- Stutters: “I w-w-went t-t-to the mall” should be transcribed as “I went to the mall”.

Non-speech (acoustic) events

- Background noise such as “fan whirring”, “dog barking”, “engine running”, “water flowing”, etc should not be transcribed.
- Foreground sounds made by the speaker should not be transcribed. These include lip smacks, tongue clicks, inhalation and exhalation between words, yawning coughing, throat clearing, sneezing, laughing, chuckling, etc.
- In the case of transcribing telephone calls, foreground sounds like machine or phone click, telephone ring, noise made by pressing telephone keypad, any other intermittent foreground noise should not be transcribed.

Names, titles, acronyms, punctuations, and numbers

- Proper names should be transcribed in a case-sensitive manner in applicable languages. Initials should be in capital letters with no period following. For example: “M K Stalin would be sworn in as the Chief Minister”.
- Titles and abbreviations are transcribed as words. For example: Dr. → Doctor except if the abbreviated form is pronounced as it is. For example, if the speaker says “Apple Inc” (instead of “Apple Incorporate”), the word 'Inc' should be transcribed.
- Punctuation marks should not be used in transcription unless they are an essential part of the word. For example: "don't"
- Acronyms should be transcribed as words if spoken as words, and as letters if spoken as letters. When transcribing sequences of letters an underscore is inserted between each letter. For example: NASA; I_B_M

- Numbers should be transcribed as full words. For example: 16 → sixteen, 112 → one hundred and twelve.
- Times of the day and dates: always capitalise AM and PM. When using o'clock, spell out the numbers: eleven o'clock.

Speaker Labels

- Mark every utterance with a speaker label
- If the speaker's name has been mentioned in an earlier utterance, then use this as the speaker label
- If the speaker's name has not been mentioned earlier then simply use generic labels such as Speaker 1, Speaker 2, ..., and so on while ensuring that the same label is consistently used for the same speaker.

Incomplete utterances

These are utterances which are incomplete because the speaker forgot what he wanted to say or was stopped mid-way and corrected an error or was interrupted by someone. Indicate such utterances by putting a '--' at the end of the utterance as opposed to a full-stop or question mark.

A.4 Guidelines for collecting voice data from professional artists (for training text-to-speech models)

We now describe the guidelines for collecting voice data from professional artists for training text-to-speech systems. These guidelines are inspired by similar guidelines released by Microsoft.

CHOOSING VOICE ARTIST

- The voice artist should be a professional with proven experience in voiceover or voice character work.
- The natural voice of the artist should be good (as opposed to an “assumed” voice which would be hard to sustain over a long period of time).
- The voice artist must have clear diction and must be able to speak with consistent rate, volume level, pitch, and tone.
- The talent also needs to be able to strictly control their pitch variation, emotional affect, and speech mannerisms.
- Work with your voice talent to develop a "persona" that defines the overall sound and emotional tone of the custom neural voice.A persona might have, for example, a naturally upbeat personality. So "their" voice might carry a note of optimism even when they speak neutrally. However, such a personality trait should be subtle and consistent.

CHOOSING A RECORDING SETUP

- Record your script at a professional recording studio that specialises in voice work and has a recording booth, the right equipment, and the right people to operate it.
- The recording should have little or no dynamic range compression (maximum of 4:1).
- It's critical that the audio has consistent volume and a high signal-to-noise ratio, while being free of unwanted sounds.

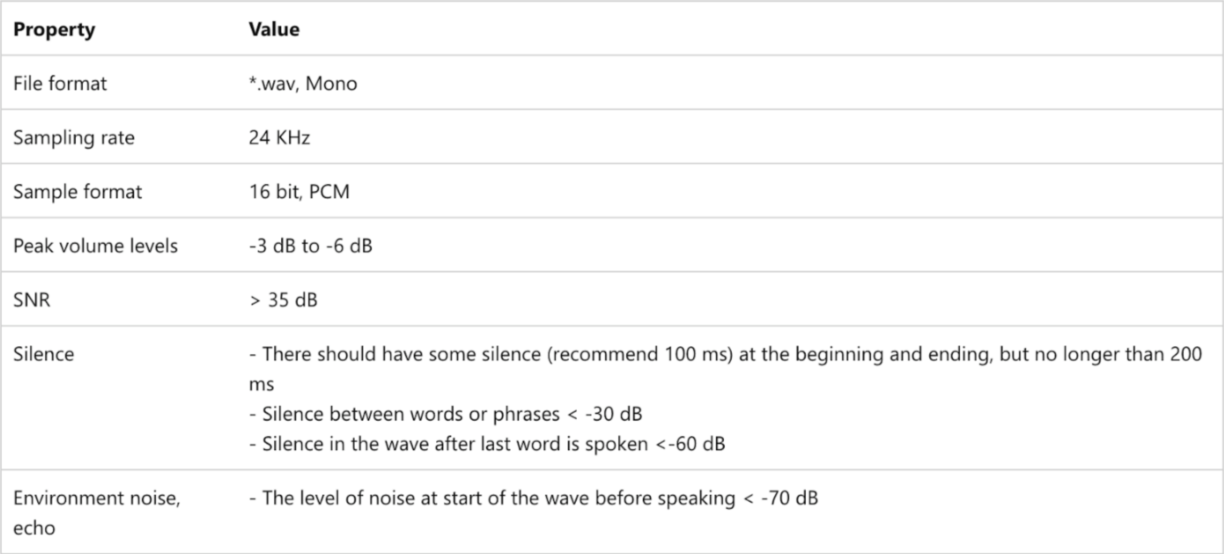
- Your recordings for the same voice style should all sound like they were made on the same day in the same room. You can approach this ideal through good recording practice and engineering.

RECORDING REQUIREMENTS

To achieve high-quality training results, follow the following requirements during recording or data preparation:

- Clear and well pronounced.
- Natural speed: not too slow or too fast between audio files.
- Appropriate volume, prosody, and break: stable within the same sentence or between sentences, correct break for punctuation.
- No noise during recording.
- No wrong accent.
- No wrong pronunciation.

You can refer to the specification below to prepare for the audio samples as best practice.



Property	Value
File format	*.wav, Mono
Sampling rate	24 KHz
Sample format	16 bit, PCM
Peak volume levels	-3 dB to -6 dB
SNR	> 35 dB
Silence	- There should have some silence (recommend 100 ms) at the beginning and ending, but no longer than 200 ms - Silence between words or phrases < -30 dB - Silence in the wave after last word is spoken <-60 dB
Environment noise, echo	- The level of noise at start of the wave before speaking < -70 dB

CREATING A SCRIPT

The script contains the utterances to be spoken by your voice talent. The term "utterances" encompasses both full sentences and shorter phrases. The script should cover different sentence types in your domain including statements (70%), questions (10%), exclamations (10%), short phrases (10%). Each utterance should be between 3 to 40 words.

A5. Guidelines for collecting image data for optical character recognition

We now describe the guidelines for collecting images for scene text recognition.

General principles

- The text in the image should be clearly readable by a human.
- The text content should largely be in the target language.
- Avoid taking a picture of the same scene in different conditions (zoom in/out, natural/artificial lighting, etc).
- Ensure diversity in content and conditions as mentioned in Table 10.
- Ensure there is diversity in the cameras using which the images are clicked (different megapixels, different brands, etc).
- The images should only contain printed text as opposed to handwritten text.