Field to Model: Pairing Community Data Collection with Scalable NLP through the LiFE Suite

R Karthick Narayanan[1], Siddharth Singh[1,2], Saurabh Singh[2], Aryan Mathur[2], Ritesh Kumar[1,2], Shyam Ratan[2], Bornini Lahiri[1,3], Benu Pareek [1,2], Neerav Mathur[2], Amalesh Gope[1,4], Meiraba Takhellambam[1,5], Yogesh Dawer[2],

[1]Council for Diversity and Innovation, [2]Unreal Tece LLP,
[3]Indian Institute of Technology-Kharagpur, [4]Tezpur University, [5]Manipur University
Correspondence: riteshkrjnu@gmail.com

Abstract

We present LiFE Suite as a "Field-to-Model" pipeline, designed to bridge community-centred data collection with scalable language model development. This paper describes the various tools integrated into the LiFE Suite that make this unified pipeline possible. Atekho, a mobile-first data collection platform, is designed to empower communities to assert their rights over their data. MATra-Lab, a web-based data processing and annotation tool, supports the management of field data and the creation of NLP-ready datasets with support from existing state-of-the-art NLP models. LiFE Model Studio, built on top of HuggingFace AutoTrain, offers a no-code solution for building scalable language models using the field data. This end-to-end integration ensures that every dataset collected in the field retains its linguistic, cultural, and metadata context, all the way through to deployable AI models and archive-ready datasets.

1 Introduction

Mobilising language documentation resources to produce language technologies for low-resource and Indigenous languages faces two significant challenges:

1. the lack of accessible, community-friendly data collection tools, and
2. fragmented workflows that separate field linguistics from computational modelling.

Despite advancements in crowdsourcing, linguistic data collection remains predominantly expert-driven. The tools available to Indigenous language speakers are often either complex proprietary systems, such as Karya[1], or basic audio recorder apps that lack essential features such as prompt integration, multilingual support, and metadata capture, all of which are critical for systematic language documentation. Similarly, the tools used by field linguists and computational linguists rarely support direct interoperability. Field linguistics tools seldom leverage the benefits of automation and machine learning that NLP technologies can offer, while NLP tools often struggle to process the rich, multi-layered annotations typical of language documentation corpora. We introduce LiFE Suite, an integrated pipeline that enables communities and researchers to collect, process, and model language data without requiring programming skills or specialised infrastructure. We describe how the suite supports multimodal, multilingual, and metadata-rich workflows that empower both field linguists and NLP practitioners to build language technologies from real-world field linguistic data.

2 Review of Existing Tools

A variety of tools have been developed to support field linguists, community language workers, and NLP practitioners in data collection, management, annotation, and lexicon creation. However, these tools tend to be fragmented, often serving either field linguistics or NLP, but rarely both. Below, we review commonly used tools in these domains and highlight the gaps that motivate the design of LiFE Suite.

2.1 Field Linguistics Tools

Tools primarily used by field linguists or community members for speech and multimodal data collection, management, and lexicon creation include:

1. **Toolbox (formerly Shoebox)**[2]: One of the earliest linguistic tools developed by SIL International, designed for text data entry and dictionary creation (Robinson et al., 2007).

[1] https://www.karya.in

[2] https://software.sil.org/shoebox, https://software.sil.org/toolbox

Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics (FieldMatters), pages 76–84

2. **FieldWorks Language Explorer (FLEx)**[3]: A widely used SIL tool for managing linguistic and cultural data, including lexicon development and interlinear glossing. LiFE Suite is designed to interoperate with FLEx, supporting the import of LIFT XML data produced by FLEx (Butler and Volkinburg, 2007).

3. **LexiquePro**[4]: Software for creating and formatting lexicon databases, mainly focused on dictionary publication and sharing, with limited editing capabilities (Guérin and Lacrampe, 2007).

4. **WeSay**[5]: Designed to help non-linguists and native speakers build dictionaries of their own languages. It is based on SIL's Semantic Domain and Rapid Word Elicitation methods, promoting community-led lexicon development (Perlin, 2012).

5. **Woefzela**[6]: A smartphone-based tool for offline data collection, supporting multiple sessions and metadata capture. It has been successfully deployed in South Africa for quality-controlled data collection (Vries et al., 2014).

6. **SayMore**[7]: A tool for organising multimedia recordings and their metadata. It also supports basic transcription and translation workflows (Moeller, 2014).

7. **Living Dictionaries**[8] help communities build and manage their own word collections. People can add words, meanings, sounds, pictures, and videos. They can search, filter, and organize entries by topics. The tool works offline and allows data to be shared or imported using common file formats like CSV, PDF, and JSON (Daigneault and Anderson, 2023).

8. **Aikuma and LIG-Aikuma**[9] are mobile apps designed to support speech data collection for under-resourced and endangered languages. Originally developed as Aikuma and later extended as LIG-Aikuma, these apps offer features such as audio recording, respeaking for clarity, oral translation, and elicitation using prompts. LIG-Aikuma also supports metadata capture, geolocation tagging, and data export compatible with ELAN. While LIG-Aikuma remains available, its development has slowed since 2018. A more recent adaptation, Williaikuma, offers updated features for sentence-level elicitation and Praat integration, demonstrating continued interest in mobile tools for linguistic fieldwork (Bird et al., 2014; Gauthier et al., 2016).

While these tools have advanced the practice of field linguistics, they suffer from several limitations that hinder their broader adoption and integration into computational workflows. Most of these tools are standalone desktop or mobile applications, often lacking compatibility with Linux operating systems and restricting usage to Windows or Mac environments. Users are typically required to switch between multiple specialised tools for different tasks, such as ELAN for video transcription, Audacity or Praat for audio segmentation, and FLEx for lexicon management, each with its own steep learning curve. Additionally, data produced by these tools is often stored in non-standard or tool-specific formats, making interoperability with NLP systems difficult without additional processing or programming expertise. Finally, data sharing in reusable, open formats remains cumbersome, limiting long-term accessibility and cross-tool usability.

2.2 NLP Annotation Tools

In contrast, NLP practitioners use a different set of tools for data annotation and management, including:

1. Label Studio[10]: An open-source, web-based data labelling platform supporting audio, text, image, video, and time-series annotation. It allows export to multiple ML-ready formats, making it popular for preparing training data (Tkachenko et al., 2020-2022).

2. Shoonya[11]: An open-source platform focused on enhancing digital content for India's underrepresented languages, supporting large-scale annotation for machine translation and other language technologies.

[3] https://software.sil.org/fieldworks
[4] https://software.sil.org/lexiquepro
[5] https://software.sil.org/wesay/
[6] https://sites.google.com/site/woefzela/
[7] https://software.sil.org/saymore, https://github.com/sillsdev/saymore
[8] https://livingdictionaries.app
[9] https://lig-aikuma.imag.fr
[10] https://labelstud.io/; https://github.com/heartexlabs/label-studio
[11] https://ai4bharat.iitm.ac.in/shoonya, https://github.com/AI4Bharat/Shoonya

3. BRAT[12], doccano[13], and INCEpTION[14]: Popular open-source tools for text annotation at the token, span, and document levels. These tools offer features for text classification, sequence labelling, and sequence-to-sequence applications (Stenetorp et al., 2012; Nakayama et al., 2018; Klie et al., 2018).

However, these NLP tools generally do not support field data collection or integrate with linguistic data management workflows. They are designed for annotation and model preparation, often assuming pre-processed, clean data rather than raw, community-collected, multimodal datasets.

2.3 The Need for an Integrated Pipeline

As highlighted, field linguistics tools and NLP tools often operate in isolation, each addressing specific stages of the data lifecycle but failing to offer a cohesive, integrated experience. LiFE Suite seeks to bridge this gap by providing a unified, no-code pipeline that supports the full workflow—from community-led data collection with structured metadata, to linguistic data management and annotation, and ultimately to NLP model training and deployment. Designed to serve both field linguists and NLP practitioners (Figure 1), LiFE Suite reduces the technical barriers that currently separate these communities, enabling them to collaboratively develop language technologies for low-resource and Indigenous languages.

3 LiFE

LiFE Suite [15] is an open-source, AI-powered platform developed by UnReaL-TecE [16], a venture led by linguists to enable seamless language data collection, management, processing, and analysis. All the components of the suite are developed as web apps with HTML, CSS and JavaScript at the frontend, Python and FastAPI (for serving different APIs) at the backend, MongoDB as the backend database and IndexedDB as the frontend database. The apps are served using Flask.

The suite integrates state-of-the-art technologies, including Large Language Models (LLMs), Automatic Speech Recognition (ASR), Optical Character Recognition (OCR), Machine Translation (MT), and advanced text processing models, providing an efficient and scalable solution for linguistic data workflows. LiFE Suite brings together three key components to support this pipeline. Atekho, a mobile-first progressive web application (PWA) designed for community-centred, offline and online multimodal data collection with integrated metadata scaffolding; MATra-Lab, a web-based platform for organizing, segmenting, transcribing, translating, and annotating linguistic datasets; and LiFE Model Studio, a no-code model-building environment built on top of HuggingFace AutoTrain, enabling the training and deployment of speech and multimodal models. Together, these components form an end-to-end system that bridges community-led data collection with scalable NLP model development and long-term archival, making linguistic technologies more accessible and sustainable for low-resource and Indigenous languages.

3.1 Atekho

Named after the Great Andamanese word for “language", Atekho (Figure 2) is a mobile-first, progressive web application for data collection, designed to empower communities and researchers to co-create living linguistic and cultural archives. By supporting multimodal data capture, including audio, video, image, and text, Atekho enables the documentation of linguistic, cultural, environmental, and oral traditions in both spontaneous and staged settings.

[12] http://brat.nlplab.org, https://github.com/nlplab/brat

[13] https://doccano.herokuapp.com, https://github.com/doccano/doccano

[14] https://inception-project.github.io, https://github.com/inception-project/inception

[15] https://github.com/unrealtecellp/life

[16] UnReaL-TecE is an organisation that is founded to develop and maintain this platform. Unlike a large number of other platforms, which could not be maintained because of various practical reasons, we expect this organisation to take care of long-term maintenance of the app and ensure that it remains available in the future.

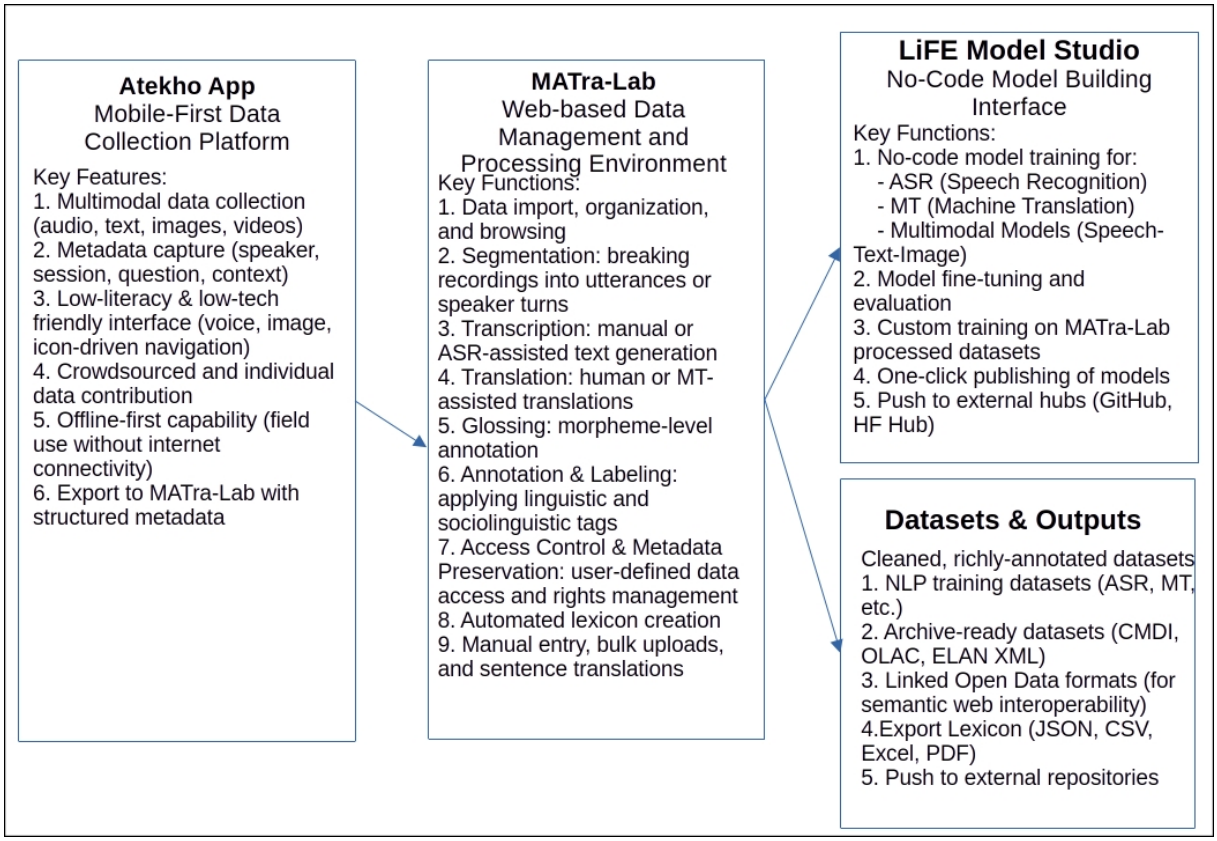

Figure 1: LiFE Pipeline

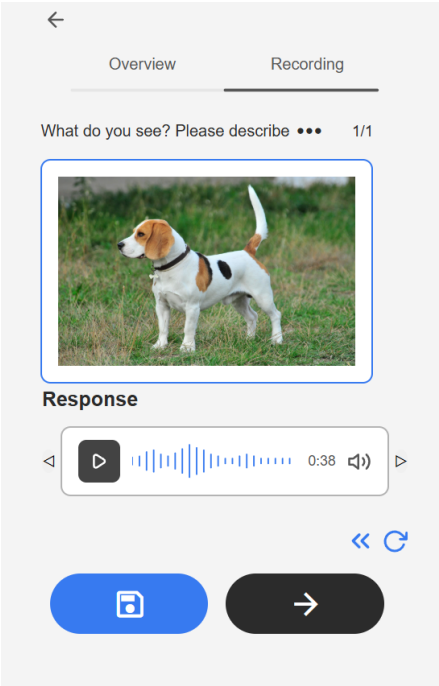

Figure 2: Atekho Interface

Atekho is built with accessibility and inclusion at its core. Its voice and icon-based user interface is designed for users with limited literacy or technology experience, making it particularly suited for community-centered projects in remote or low-resource contexts. The application operates in offline-first mode, ensuring functionality in rural areas with limited or no internet connectivity. In addition to capturing content, Atekho supports collaborative metadata scaffolding, allowing users to tag recordings with speaker details, contextual information, and community-generated annotations. Its workflows are customizable, enabling projects to define their own data structures and inventory formats. This flexibility makes Atekho adaptable to a wide range of documentation initiatives, including sociolinguistic surveys, oral literature preservation, and environmental knowledge documentation. As part of the LiFE Suite, Atekho seamlessly integrates with MATra-Lab, allowing collected data and metadata to flow directly into more advanced processing pipelines. Once synchronized with MATra-Lab, recordings can be segmented, transcribed, translated, glossed, and annotated using an AI-in-the-loop mechanism. This interoperability positions Atekho not just as a data collection tool, but as the starting point of an end-to-end "Field to Model" pipeline, bridging community-driven documentation with scalable NLP model

development.

By placing ownership and control in the hands of the communities whose heritage it seeks to preserve, Atekho supports the creation of living archives that are ethically grounded, accessible, and sustainable.

3.2 MATra-Lab: Web-Based Linguistic Data Management and Processing

MATra-Lab (Figure 3) is a web-based platform designed to support the management, processing, and annotation of multilingual and multimodal datasets, with a particular focus on the linguistic diversity of Indian languages. It provides researchers across subfields, such as field linguistics, sociolinguistics, and computational linguistics, with an integrated environment for processing audio, video, text, and image data. By combining multimodal data processing, AI-powered tools, collaborative management, and an intuitive interface, MATra-Lab offers an end-to-end environment to produce scalable, reproducible, and NLP-ready linguistic resources.

Figure 3: Transcription in Matra Lab

3.2.1 Data Ingestion and Management

In addition to providing automated ingestion of data collected using Atekho, MATra-Lab supports the direct upload of field-collected datasets along with rich metadata, including participant information and item-level metadata. This metadata enables users to sort, filter, and organise data for efficient navigation and management of large and heterogeneous collections (Figure 4).

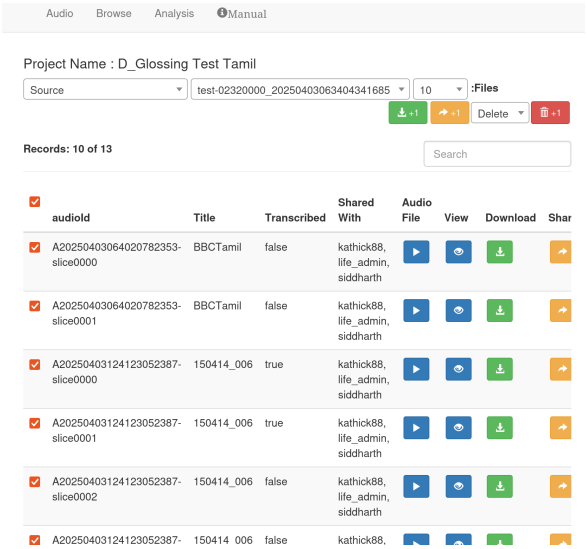

Figure 4: Browse and Filter in Matra Lab

3.2.2 Supported Workflows and Tasks

MATra-Lab supports multilayered annotation and processing workflows that linguists and computational linguists may use. It allows users to apply document-level (for all kinds of multimodal documents) and span-level (for audio, video and text documents) labels for tasks such as morphological analysis, syntactic tagging, discourse annotation, and semantic labeling. It currently allows for the following kinds of tasks -

1. Time-aligned audio and video transcription in multiple scripts at any level from individual phones to complete discourse. It also allows for speaker diarisation, mapping audio files or parts of audio files to prompts and anonymising parts of the audio.

2. Translation of audio, video, text and images.

3. Annotation of audio and video chunks with custom labels.

4. Annotation of text at both document and span level with custom labels.

5. Interlinear glossing of audio, video and text documents.

6. OCR and labelling of images.

3.2.3 Building Lexicons

In addition to providing data processing support, MATra-Lab also includes a Lexicon Module for building and managing multilingual glossaries and dictionaries. It supports:

1. Automated extraction of lexical items from existing annotated data,
2. Manual entry of lexicon items,
3. Bulk uploads from external sources, and
4. Sentence-aligned translations to support context-rich dictionary development.

Users can browse, edit, and export lexicons in multiple formats, including JSON, RDF, CSV, Excel, and PDF, facilitating integration with other linguistic tools or dissemination to wider audiences. The module also supports collaborative dictionary development with fine-grained access control, allowing teams to manage user permissions for viewing, editing, and exporting lexicon data.

This added functionality extends the utility of MATra-Lab beyond corpus management, allowing dictionary making, terminology development, and community-led lexical documentation within the same unified workflow.

3.2.4 AI-in-the-loop

A core feature of MATra-Lab is its graphical user interface (GUI), which allows users to apply pretrained models for a range of natural language processing (NLP) tasks without requiring coding expertise. The tasks where AI models currently provide support are the following:

1. Transcription (both in IPA and native scripts for supported languages), speaker diarisation, translation, and glossing for audio data using Automatic Speech Recognition (ASR), Machine Translation (MT) and other relevant models,
2. Digitisation of scanned documents and images using Optical Character Recognition (OCR) tools,
3. Advanced text processing using Large Language Models (LLMs) and multimodal models.

All of these automation facilities are made available by integrating models and APIs from different sources.

1. **HuggingFace Hub**[17]: integration with publicly available models on HuggingFace Hub is available out-of-the-box.
2. **Bhashini API**: Bhashini APIs are provided by the Ministry of Electronics and Information Technology, Government of India. These provide access to state-of-the-art open-source models supporting different kinds of tasks in Indian languages viz ASR, speaker diarisation, transliteration, language identification, etc.
3. **LiFE Model Studio**: The models that are trained by the users using the LiFE Model Studio can be used in MATra Lab for automating the tasks.
4. **Stanza**[18]: Stanza models are integrated to provide automatic interlinear glossing and morphosyntactic information including part-of-speech categories, morphological features and dependency relations for the supported languages (Qi et al., 2020).
5. **Language agnostic models**: Some models for language-agnostic tasks such as voice activity detection (viz Silero VAD (Team, 2024) and PyAnnote (Bredin, 2023; Plaquet and Bredin, 2023)), speaker diarisation, (viz. PyAnnote) and universal phoemiser (viz. Allosaurus (Li et al., 2020)) are also integrated into the app.
6. **In-house Models**: Limited support for tasks such as interlinear glossing, sentiment analysis, aggression level, etc in some languages are provided through our in-house rule-based and machine learning-based models.

3.2.5 Data Export and Sharing

The platform allows users to download their dataset in multiple structured and semi-structured formats viz. JSON, CSV, XLSX, Markdown, TextGrid, CHAT, etc, for further processing and use with other apps and libraries.

The platform also offers collaborative workspace functionality, enabling file-level sharing with fine-grained access control. Users can define permissions for:

1. Full access (edit and download),
2. Restricted access (online viewing and annotation without download), or
3. Partial access (limited operational permissions).

[17] https://huggingface.co/docs/hub

[18] https://stanfordnlp.github.io/stanza

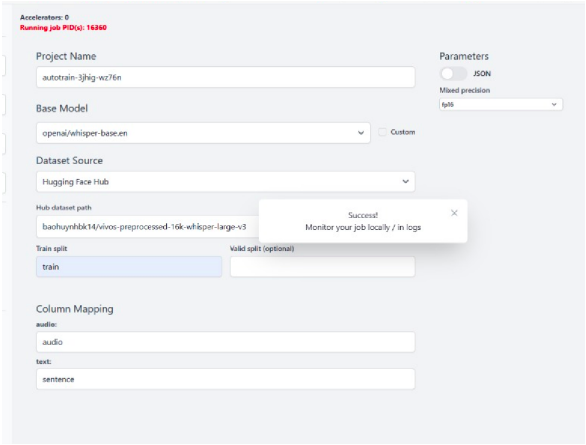

Figure 5: LiFE Model Studio

This collaborative architecture promotes team-based data management while ensuring data security and control.

3.3 LiFE Model Studio

LiFE Model Studio (Figure 5) is a no-code model training interface built on HuggingFace AutoTrain. It allows users to train new models or fine-tune existing ones for a variety of NLP tasks, using the datasets they have created and processed in MATra-Lab. While model building is optional, users who choose to extend their workflow within the LiFE Suite can derive datasets directly from their MATra-Lab projects and use them in LiFE Model Studio to build task-specific models, such as speech recognition, machine translation, or multimodal processing models.

Trained models are immediately available within the platform for further use in data processing workflows, creating a feedback loop that improves annotation, transcription, and translation over time. Additionally, these models can be exported to external repositories, such as GitHub or HuggingFace Hub, for wider public access and reuse. We have added the following functionalities to the existing AutoTrain interface -

1. **MATra Lab Integration**: We have added support for directly importing specific kinds of data from MATra Lab dataset into the interface and use that for fine-tuning the required models.

2. **Audio-based Tasks**: We have added support for audio-based tasks such as Automatic Speech Recognition, speaker diarisation and voice activity detection.

3. **Additional Tasks**: We are in the process of integrating support for additional libraries (such as scikit-learn) and tasks to enable no-code training for additional tasks (such as interlinear glossing).

LiFE Model Studio thus completes the Field-to-Model pipeline, providing users with a scalable, no-code solution for bringing community-collected data all the way to deployable language technologies.

4 Case Studies: Demonstrating the Field-to-Model Pipeline

To demonstrate the effectiveness of the Life Field-to-Model pipeline, we present two ongoing case studies that apply the pipeline in real-world, community-centred language technology projects. These case studies illustrate how the pipeline enables end-to-end data collection, management, processing, and model development in two distinct linguistic contexts.

4.1 SpeeD-TB

The first case study, Speed-TB (Kumar et al., 2023), focuses on six Tibeto-Burman languages spoken in Northeast India—Bodo, Meetei, Chokri, Kokborok, Nyishi, and Toto—and is funded by the Bhashini initiative of the Government of India. Using the Life Suite, data is collected through structured questionnaires, stimulus-based narration, role-play, and spontaneous speech, with community members actively participating in data contribution and validation. The collected data is processed in MATra-Lab and used in LiFE Model Studio to build baseline speech recognition models. The project explores fine-tuning/training models such as conformer-multilingual-asr by AI4Bharat[19], Whisper[20](Radford et al., 2022), wav2vec 2.0(Baevski et al., 2020), and NVIDIA NeMo[21](NVIDIA, 2025).

4.2 Irula Language

The second case study focuses on Irula, a Dravidian language spoken in Tamil Nadu, India. In collaboration with the Keystone Foundation, a community-based organisation, the project builds on existing

[19] https://dibd-bhashini.gitbook.io/bhashini-apis/available-models-for-usage

[20] Whisper: https://github.com/openai/whisper; Paper: https://arxiv.org/abs/2212.04356

[21] NVIDIA NeMo: https://developer.nvidia.com/nvidia-nemo

resources from a community radio station to collect and process Irula speech data. The team experiments with fine-tuning the conformer-multilingual-dravidian model by AI4Bharat and other multilingual models to develop a dedicated Irula ASR system.

5 Conclusion

In this paper, we have presented a new workflow for building language technologies for underresourced languages using primary data collected from the field. This workflow is enabled through the LiFE Suite, an open-source AI-powered platform. We give details of the suite and how it enables the operationalisation of the Field-to-Model workflow. We also present two case studies where we use the workflow and the suite for building language technologies.

In both the case studies, the community remains at the center—not only as data contributors but as co-creators and validators of the resulting language technologies. These case studies serve as proof-of-concept implementations, demonstrating that the Field-to-Model pipeline is viable, scalable, and capable of supporting community-driven speech technology development for underrepresented languages.

While these case studies demonstrate the practical value and scalability of the Field-to-Model pipeline, our experience also highlights several challenges and constraints that must be addressed to make the workflow more inclusive and widely adoptable.

Limitations

While the LiFE Suite offers a comprehensive, no-code pipeline for community-centered language documentation and NLP model development, several limitations persist. Firstly, although Atekho is designed for offline use, both MATra-Lab and LiFE Model Studio require stable internet connectivity and access to web-based interfaces, which may pose challenges in remote or resource-constrained environments. Secondly, MATra-Lab's reliance on pre-trained models from platforms like HuggingFace means that its performance is contingent on the availability and quality of existing models, which may not adequately represent all low-resource or Indigenous languages. Regarding LiFE Model Studio, while it provides an accessible interface for model fine-tuning, it currently does not support training models from scratch; users can only fine-tune existing pre-trained models . Additionally, despite its no-code design, users may still require a foundational understanding of NLP concepts to effectively navigate model selection and fine-tuning processes. Lastly, the computational demands of model training and inference necessitate access to GPUs, which may not be readily available to all users, potentially limiting the suite's accessibility and scalability.

Acknowledgments

We would like to thank Mission Bhashini, Ministry of Electronics and Information Technology (MEITY), Govt of India, for supporting the SpeeD-TB project and the development of the LiFE suite. We would also like to express our heartfelt thanks to all the community members of Irula, Toto, Chokri, Kok Borok, Nyishi, Bodo and Meitei, who contributed immensely in the two case studies and by providing valuable feedback on the LiFE suite.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*.

Steven Bird, Florian R. Hanke, Oliver Adams, and Haejoong Lee. 2014. Aikuma: A mobile app for collaborative language documentation. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–5, Baltimore, Maryland, USA. Association for Computational Linguistics.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Lynnika Butler and Heather Volkinburg. 2007. Review of fieldworks language explorer (flex). *Language Documentation and Conservation*, 1.

Anna Luisa Daigneault and Gregory D. S. Anderson. 2023. Living dictionaries: A platform for indigenous and under-resourced languages. *Dictionaries: Journal of the Dictionary Society of North America*, 44(02):57–74.

Elodie Gauthier, David Blachon, Laurent Besacier, Guy-Noel Kouarata, Martine Adda-Decker, Annie Rialland, Gilles Adda, and Grégoire Bachman. 2016. Ligaikuma: A mobile app to collect parallel speech for under-resourced language studies. pages 381–382.

Valérie Guérin and Sébastien Lacrampe. 2007. Lexique pro. *Language Documentation and Conservation*, 1(2):293 – 300.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico.

Ritesh Kumar, Meiraba Takhellambam, Bornini Lahiri, Amalesh Gope, Shyam Ratan, Neerav Mathur, and Siddharth Singh. 2023. Collecting speech data for endangered and under-resourced indian languages. In *Proceedings of the 2nd Annual Meeting of the ELRA/ISCA Special Interest Group on Under-resourced Languages (SIGUL 2023)*, pages 31–38.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, and Metze Florian. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE.

Sarah Ruth Moeller. 2014. Saymore, a tool for language documentation productivity. 08:66–74.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

NVIDIA. 2025. Nvidia nemo: Open-source toolkit for conversational ai. `https://developer.nvidia.com/nvidia-nemo`. Accessed: 2025-05-11.

Ross Perlin. 2012. Wesay, a tool for collaborating on dictionaries with non-linguists. *Language Documentation & Conservation*, 6:181 – 186.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356.

Stuart Robinson, Greg Aumann, and Steven Bird. 2007. Managing fieldwork data with toolbox and the natural language toolkit. *Language Documentation and Conservation*, 1.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Silero Team. 2024. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. `https://github.com/snakers4/silero-vad`.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Nic Vries, Marelie Davel, Jaco Badenhorst, Willem Basson, Etienne Barnard, and Alta de Waal. 2014. A smartphone-based asr data collection tool for under-resourced languages. *Speech Communication*, 56:119–131.