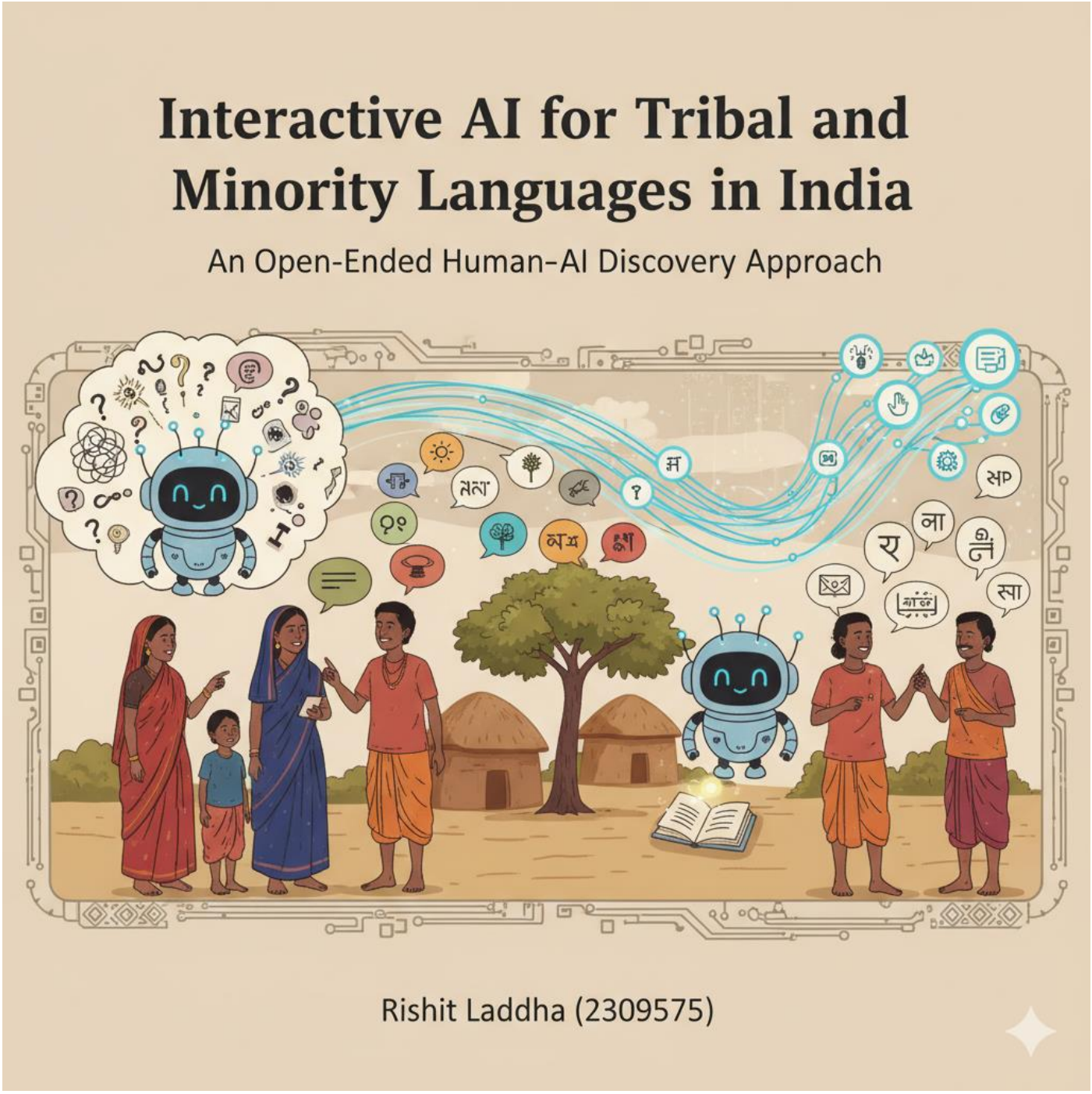
Interactive AI for Tribal and Minority Languages in India:

An Open-Ended Human–AI Discovery Approach

Team: Rishit Laddha (2309575)



ABSTRACT

India has hundreds of tribal and minority languages that are spoken by small communities and often passed down only through speech. These languages usually do not have large written datasets, standard spelling, or digital resources. Because of this, modern AI systems such as translators, voice assistants, and educational tools completely ignore them.

Most existing AI language systems depend on large, fixed datasets collected before training. This approach does not work for tribal and minority languages, where data is scarce or does not exist at all. As a result, millions of speakers remain digitally excluded.

This project proposes a new direction: **interactive and open-ended language discovery**. Instead of learning only from static datasets, the AI system will learn by interacting with humans. It will detect when it is unsure, ask simple questions, receive feedback, and slowly build knowledge of the language over time.

The goal of this research is to design, test, and evaluate such an interactive AI system for Indian tribal and minority languages. The project builds on recent ideas in open-ended discovery and human-in-the-loop learning, and adapts them specifically to the Indian context.

LITERATURE REVIEW

A. OPEN-ENDED AND INTERACTIVE DISCOVERY FOR LOW-RESOURCE LANGUAGES

1. Dossou & Aïdasso (2025) – *Towards Open-Ended Discovery for Low-Resource NLP*

This paper introduces the idea of open-ended discovery for languages that do not have much data. Instead of training AI on a fixed dataset, the authors suggest that AI should learn like humans. The model should notice when it is unsure, ask questions to humans, and slowly build its understanding of the language through interaction.

The key idea is that for many languages, especially low-resource ones, it is impossible to first collect a big dataset and then train a model. So the learning process itself must involve humans in the loop.

Results / contributions

- The paper clearly explains why current NLP methods fail for low-resource languages
- It proposes a new learning style where the model keeps learning over time
- It provides a conceptual framework for uncertainty, questioning, and discovery

Limitations

- The work is mostly conceptual, not a full working system
- It does not focus on a specific country or language group
- There is no clear evaluation method defined yet

Relevance to our research
This paper is the main foundation of the proposed research. Our work directly builds on this idea and tries to apply it to Indian tribal and minority languages, which the paper itself does not explore.

2. Masakhane – Community-Driven NLP for Low-Resource Languages

Masakhane is a large research community focused on building NLP systems for African low-resource languages. Instead of relying on large companies or pre-existing datasets, the community brings together native speakers, students, and researchers to jointly create datasets, models, and benchmarks. This work clearly

shows that human collaboration can overcome the lack of data and enable meaningful NLP systems even for under-represented languages. Masakhane has successfully produced machine translation and other NLP models while promoting ethical and inclusive AI practices. However, the approach still depends on manual data creation, and the learning process is not interactive during model training. The AI does not decide when or what to ask humans. This work is important because it proves that human involvement works; the proposed research builds on this idea by enabling the **AI itself to decide when and how to ask humans for help**.

B. HUMAN-IN-THE-LOOP AND INTERACTIVE MACHINE TRANSLATION

3. Knowles & Koehn (2016) –> Neural Interactive Translation Prediction

This paper studies interactive machine translation, where a human corrects the system while it is generating a translation. The system then immediately updates its predictions using this feedback, improving the output step by step rather than producing a fixed translation. The authors show that human feedback can reduce errors, speed up corrections, and improve translation quality with fewer edits. However, the work focuses on well-resourced languages and assumes that the language structure is already known. Human feedback is applied only after translation begins, and the method does not support language discovery or learning missing grammar. This work shows **the value of human feedback**, but the proposed research applies this idea to learning the language itself, not just fixing translations.

4. Peris & Casacuberta (2018) –> Active Learning for Interactive NMT

This work combines interactive translation with active learning. Instead of asking humans to correct everything, the model selects only the most useful sentences for human feedback, reducing effort and cost. The results show that better translation quality can be achieved with fewer human corrections. However, this method assumes that some initial dataset already exists and remains focused on translation rather than discovering new language structure. It is also not suitable for languages without a written standard. This work supports the idea that **AI should ask only useful questions,** which is central to the proposed research.

C. ACTIVE LEARNING FOR MACHINE TRANSLATION

5. Haffari et al. –> Active Learning for MT

This line of research studies how machine translation models can select the most informative examples for training instead of relying on random data. By measuring uncertainty, models can reduce annotation cost and improve efficiency. While effective, these methods assume the availability of labeled data and do not involve real-time human interaction. They are also not designed for oral or undocumented languages. The proposed research extends this idea by using **dialogue-based active learning, where questions are asked directly to humans during interaction.**

D. UNCERTAINTY ESTIMATION AND CLARIFICATION QUESTIONS

6. Clarification Question Generation

Research in clarification question generation explores how models can detect uncertainty or ambiguity in input and ask questions to resolve it. These methods show that asking questions improves understanding and reduces incorrect assumptions. However, they are mostly applied to text understanding tasks and are not used for language learning or documentation. This capability is critical for the proposed system, where the AI must recognize when it does not understand a word, meaning, or structure and ask humans for clarification.

7. UncertaiNLP –> Uncertainty-Aware NLP

UncertaiNLP workshops focus on how uncertainty should be measured and handled in NLP models. This work highlights that uncertainty estimation is essential and introduces evaluation tools and methods. However, the research is largely theoretical or benchmark-focused and has not been applied to low-resource language learning. These ideas are useful for deciding **when the AI should ask humans** in the proposed system.

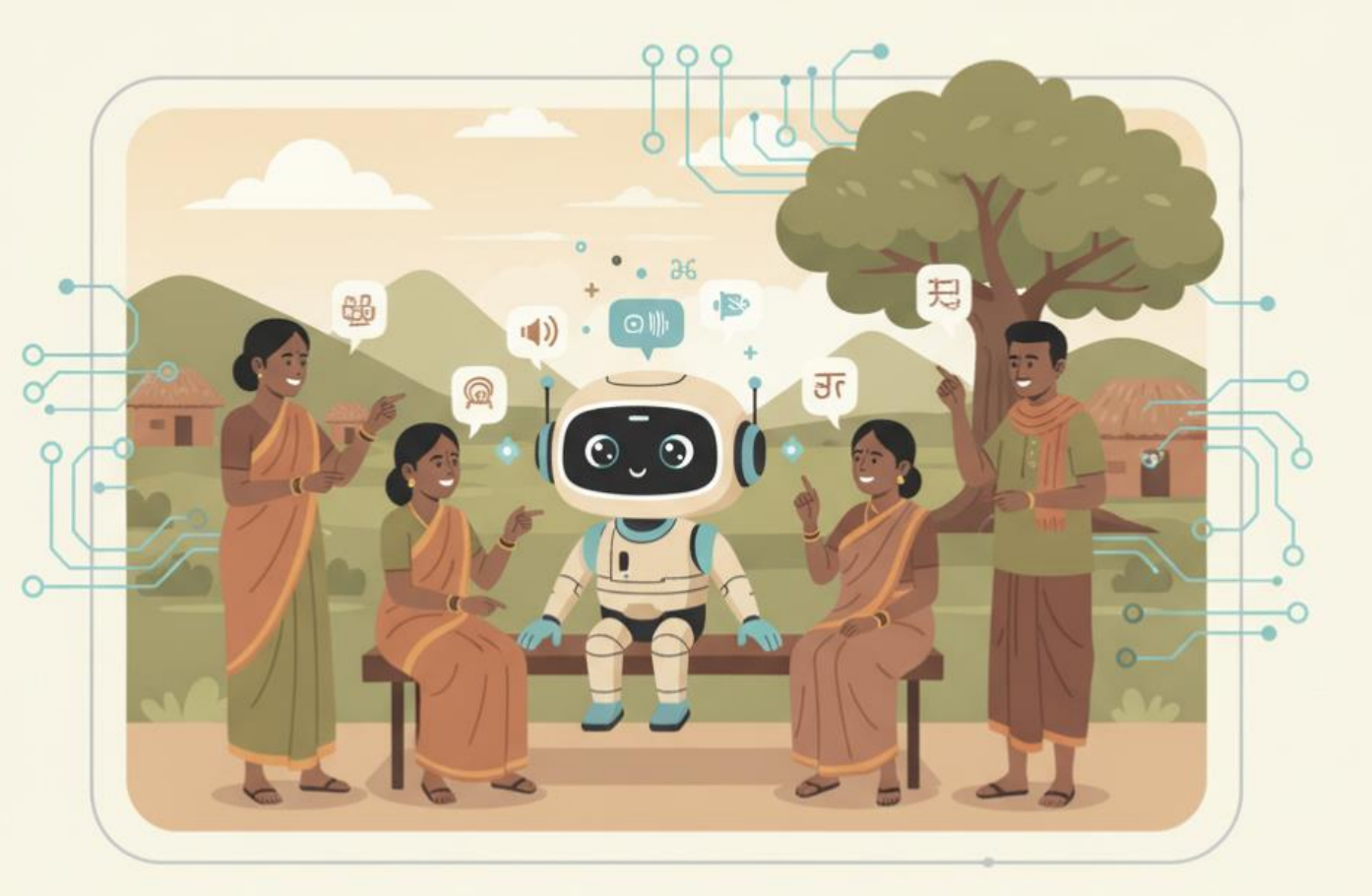
E. LANGUAGE DOCUMENTATION AND TOOLING FOR ENDANGERED AND TRIBAL LANGUAGES

Systems such as **ELPIS** and **PERSEPHONE** focus on building practical AI tools for endangered and very low-resource languages. They provide pipelines for speech recognition and basic NLP, allowing linguists and non-experts to work with small datasets, especially for oral languages. These tools demonstrate that AI systems can function with extremely limited data. However, they still rely on manually labeled data, do not support interactive or continuous learning, and do not allow the AI to

ask questions or adapt during use. This work shows that low-resource language technology is possible, but it does not address **language discovery through interaction**, which is the core focus of the proposed research.

F. INDIAN CONTEXT – RAMESH ET AL. (2022), SAMANANTAR

SAMANANTAR is a large parallel dataset covering 11 major Indian languages and has enabled strong machine translation and broader progress in Indic NLP. However, it does not include tribal or minority languages and assumes the existence of large written corpora. This highlights a major gap: while resources exist for major languages, **India's tribal and minority languages remain unsupported**, making them ideal candidates for an open-ended, interactive learning approach.



GAPS IDENTIFIED IN EXISTING RESEARCH

The literature survey shows strong progress in low-resource NLP, human-in-the-loop learning, active learning, and uncertainty estimation. Prior work has clearly demonstrated that human feedback improves model quality, that uncertainty can be measured, and that language technologies can be built even with limited data.

However, these ideas largely exist in isolation. Most systems still depend on pre-collected datasets, fixed training pipelines, or post-hoc human correction. Very few approaches treat language learning itself as a **continuous discovery process**, especially for languages that lack written form, standard grammar, or existing corpora.

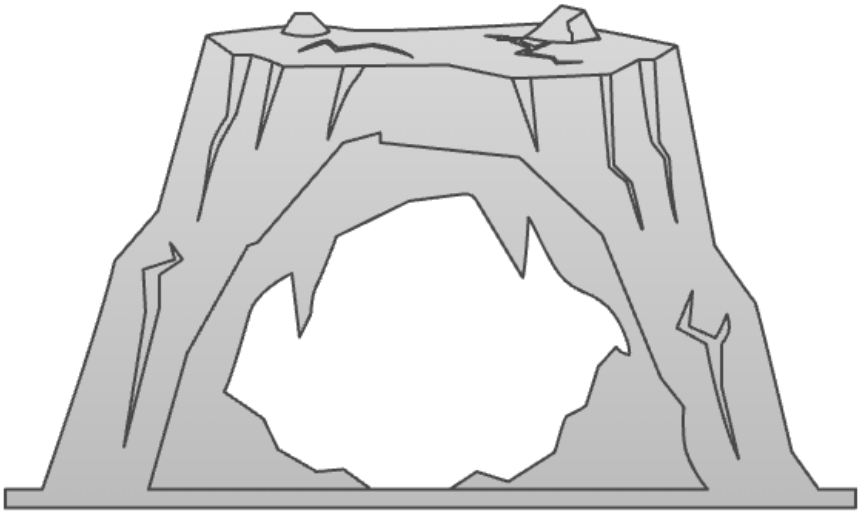
This gap is especially visible in the Indian context, where tribal and minority languages remain almost entirely outside current NLP research.

The key gaps identified across existing work are:

- No end-to-end system that **learns a language through interaction**, rather than training once on a static dataset
- Open-ended discovery is proposed conceptually, but **not implemented or evaluated** in real low-resource settings
- Human-in-the-loop methods focus on **correcting outputs**, not on discovering unknown words, meanings, or grammar
- Active learning assumes **some labelled data already exists**, which is not true for many tribal languages
- Uncertainty estimation is studied, but **not connected to question-asking strategies** for language learning
- Clarification question research is applied to understanding tasks, **not to building a language model itself**
- Existing endangered language tools rely on **manual labelling** and do not support adaptive learning
- Indian NLP resources focus on **major written languages**, leaving tribal and minority languages unaddressed
- There is **no clear evaluation framework** for measuring discovery-based, interactive language learning
- Reinforcement learning and interaction-driven improvement are **rarely applied to low-resource language discovery**

Taken together, these gaps show that while many components needed for interactive language learning already exist, they have not been integrated into a single system aimed at **discovering and documenting languages through**

human interaction. In particular, there is no practical framework tailored to Indian tribal and minority languages, where data scarcity, oral usage, and variation are the norm. This creates a clear opportunity for a new approach that combines **minimal initial training**, **uncertainty-aware questioning**, and **reinforcement learning from human feedback** to enable open-ended language discovery. The proposed research directly addresses these gaps by designing and evaluating such a system in the Indian context



Static Datasets
No learning through interaction

Conceptual Discovery
Not implemented or evaluated

Output Correction
Focus on correcting, not discovering

Labelled Data Assumption
Active learning needs existing labels

Unconnected Uncertainty
Not linked to question strategies

Task Application
Clarification questions not for language model

Manual Labelling
Endangered language tools lack adaptive learning

Major Language Focus
Indian NLP ignores tribal languages

Unclear Evaluation
No framework for interactive learning

Rare Application
Reinforcement learning rarely applied

PROBLEM DEFINITION

The central problem addressed by this research is:

How can an AI system learn, document, and improve understanding of Indian tribal and minority languages through continuous human interaction, without relying on large static datasets?

India is home to hundreds of tribal and minority languages that are spoken by small communities, often passed down orally, and rarely supported by written records or digital resources. These languages play a crucial role in cultural identity, education, healthcare communication, and local governance. However, most of them remain completely absent from modern AI systems.

The core problem is that **current NLP systems cannot learn or support languages that lack large, pre-collected datasets**. Modern language models depend heavily on large volumes of labelled text for training. For tribal and minority languages in India, such datasets either do not exist or are extremely limited due to factors such as non-standard spelling, oral transmission, dialect variation, and limited literacy.

Existing approaches attempt to address low-resource settings by collecting more data, using transfer learning, or involving humans to manually label examples. While helpful, these methods still assume that a language is already documented and that sufficient data can be collected before learning begins. This assumption does not hold for many Indian tribal languages, where documentation itself is incomplete or missing.

As a result, there is a **fundamental mismatch between how current AI systems learn and how these languages exist in reality**. Humans learn languages gradually through interaction by asking questions, receiving clarification, and refining understanding over time. In contrast, AI systems are trained in a static manner and lack the ability to actively explore, ask questions, or adapt through real-time human interaction.

This problem has three key challenges:

1. **Language Discovery:**
   The AI must learn linguistic structure, vocabulary, and meaning without assuming prior documentation or complete grammatical knowledge.
2. **Uncertainty Awareness:**
   The system must recognize when it does not understand a word, sentence, or meaning and avoid making incorrect assumptions.
3. **Human–AI Interaction:**
   The AI must be able to ask meaningful questions to humans and use their responses to improve its internal language model over time.

Current NLP systems do not jointly address these challenges. There is no existing framework that allows AI to *discover* a language through interaction, decide *what* to ask humans based on uncertainty, and *learn continuously* from those interactions in a low-resource Indian context.

This research aims to address this gap by developing an **open-ended human–AI discovery approach**, where the AI system learns tribal and minority languages through small initial data, active questioning, and reinforcement from human feedback, making language learning possible even in the absence of large datasets.

OUR PROPOSED ARCHITECTURE

Overview

This research proposes an **interactive learning system** for tribal and minority languages in India, where an AI model learns language knowledge through **direct interaction with humans** rather than relying on large pre-collected datasets. The system is designed for languages that lack written resources, standardized grammar, or large digital corpora, and where linguistic knowledge primarily exists within native-speaking communities.

Instead of treating learning as a one-time training process, the proposed system follows an **open-ended learning paradigm**. The AI continuously interacts with speakers, identifies gaps in its understanding, asks clarification questions when needed, and updates its internal knowledge based on human feedback. This process allows the system to gradually acquire vocabulary, grammar, and meaning over time, closely mirroring how humans learn new languages.

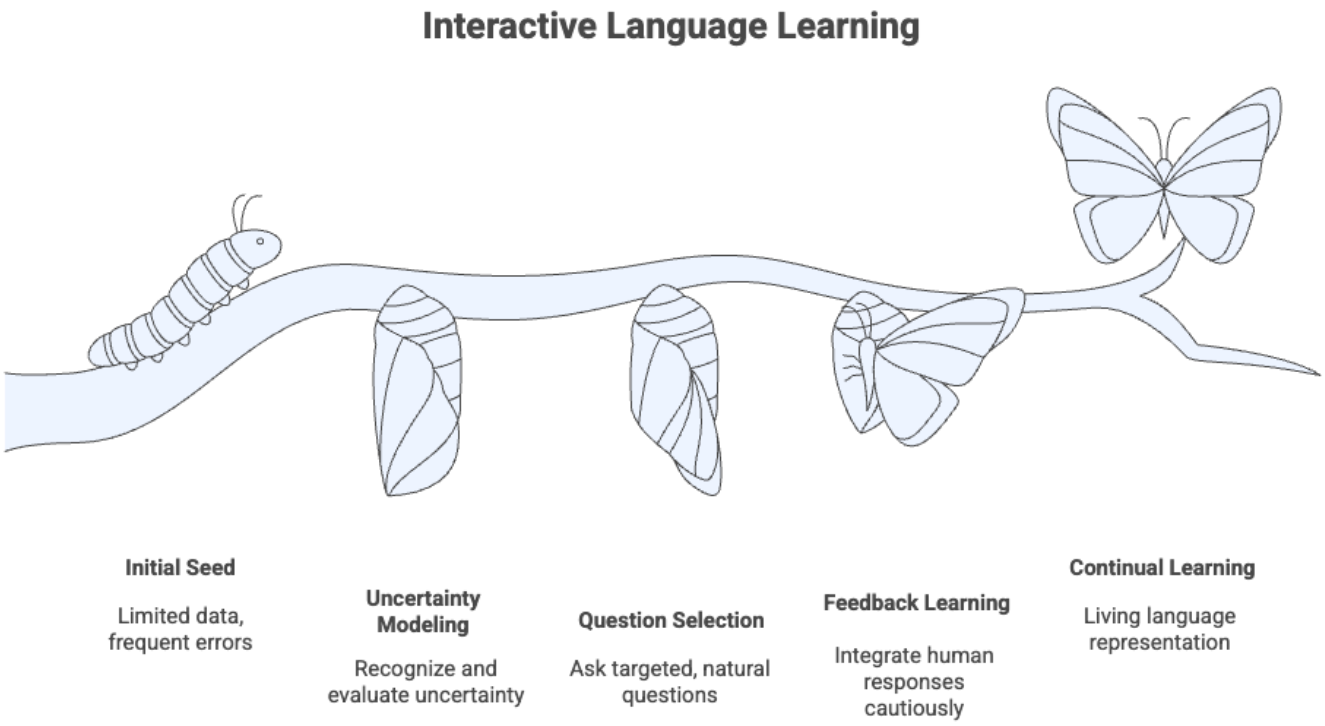
At a high level, the system operates as a **closed interactive loop**: it observes language input, attempts an interpretation, evaluates its own uncertainty, requests clarification if necessary, incorporates human feedback, and refines its knowledge. This loop repeats throughout the lifetime of the system, enabling continuous learning and language discovery.

Key Design Principles

The proposed architecture is guided by the following principles:

- **Learning with minimal data**: The system must function even when only a very small amount of initial data is available.
- **Uncertainty awareness**: The AI must recognize when it does not fully understand an input.
- **Human-centered learning**: Humans are treated as teachers and collaborators, not passive data labelers.
- **Learning from interaction**: The system improves through feedback rather than static supervision.
- **Open-ended growth**: Language learning and documentation continue over time without a fixed endpoint.

SYSTEM COMPONENTS



The proposed approach consists of five tightly connected components, each responsible for a specific part of the learning process.

1. Initial Seed Learning

The system begins with a **small initial learning phase** using a limited amount of data collected from native speakers or linguists. This data may include short example sentences, basic vocabulary lists, or rough translations. The purpose of this phase is not to achieve high performance, but to give the AI a minimal ability to process the language and produce tentative outputs.

At this stage, the AI is expected to make frequent mistakes. These errors are intentional and valuable, as they allow the system to identify areas of uncertainty and drive future interaction. The seed model simply provides a starting point from which interactive learning can begin.

2. Modeling Interactional Uncertainty

A central aspect of the proposed approach is the AI's ability to **recognize when it is unsure**. Rather than always producing an answer, the system continuously evaluates its confidence in its own predictions.

Uncertainty may arise from several factors, such as unfamiliar words, multiple possible meanings, inconsistent grammatical patterns, or unstable outputs across

similar inputs. By monitoring these signals, the system builds an internal understanding of where its knowledge is incomplete or unreliable.
This uncertainty awareness is critical because it determines whether the AI should proceed independently or seek human guidance. Without this capability, the system would risk reinforcing incorrect assumptions and learning errors over time.

3. Interactive Question Selection

When uncertainty exceeds an acceptable level, the system transitions from passive prediction to **active questioning**. Instead of asking generic or frequent questions, the AI carefully selects questions that are most likely to improve its understanding while minimizing effort for the human speaker.

The system generates **simple, natural questions**, such as asking for the meaning of a word, confirming whether a sentence is correct, or requesting an alternative phrasing. Questions are designed to be easily answered by non-experts and to fit naturally into conversation.

Importantly, the system avoids asking questions when humans themselves appear uncertain or confused. This ensures that interaction remains cooperative rather than burdensome. By prioritizing questions that offer high learning value, the system maintains efficiency and respect for human time.

4. Learning from Human Feedback

Once a human responds to a question, the system treats the feedback as a learning signal. Feedback may be clear and direct, such as a confirmed meaning, or it may be partial or ambiguous, reflecting natural variation in human responses.

The AI does not treat all feedback as equally reliable. Instead, it considers both the confidence of the human response and its own prior understanding before updating its knowledge. Clear and confident feedback has a stronger influence, while uncertain responses are integrated more cautiously.

This allows the system to learn incrementally while avoiding overfitting to noisy or ambiguous input. Learning remains flexible and adaptive, reflecting the realities of human communication.

5. Continual Learning and Knowledge Consolidation

Language learning does not happen in a single interaction. To support long-term learning, the system stores past interactions in a memory structure that includes the original input, the human response, and a measure of confidence associated with that interaction.

Over time, the system revisits stored interactions to reinforce reliable knowledge and re-examine uncertain cases. Highly reliable examples strengthen the model's understanding, while low-confidence examples remain available for future clarification when additional context becomes available.

Through this process, the AI gradually builds a **living representation of the language**, including vocabulary, sentence structures, and variations across speakers or regions. This representation evolves continuously as new interactions occur.

End-to-End Learning Loop

The complete system operates through the following repeating cycle:

1. Observe language input from a speaker
2. Attempt to interpret or respond
3. Evaluate uncertainty in the interpretation
4. Ask a clarification question if needed
5. Receive human feedback
6. Update internal knowledge
7. Store the interaction for future learning
8. Repeat with new inputs

This loop never fully terminates, allowing the system to adapt as language use evolves.

Why This Approach Is Suitable for Tribal and Minority Languages

The proposed architecture is well suited for tribal and minority languages because it does not assume the existence of large datasets, written standards, or fixed grammar rules. It works naturally with spoken language, leverages community knowledge, and adapts to variation across speakers.

By shifting the focus from data collection to **interactive discovery**, the system enables language learning in contexts where traditional NLP methods fail.

Architectural Contribution

The key contribution of this approach lies in combining:

- uncertainty-aware decision making
- targeted human interaction
- continual learning over time

into a single unified system designed specifically for low-resource, under-documented languages in India. Rather than building static models, the proposed architecture builds a **learning system** that grows alongside the language and its speakers.

To illustrate the concept of open-ended, interactive language learning, we refer to an example interaction originally presented in prior work on open-ended discovery for low-resource languages. The figure demonstrates how an AI system, initially unfamiliar with a language, gradually acquires linguistic knowledge through uncertainty-aware questioning and human feedback.

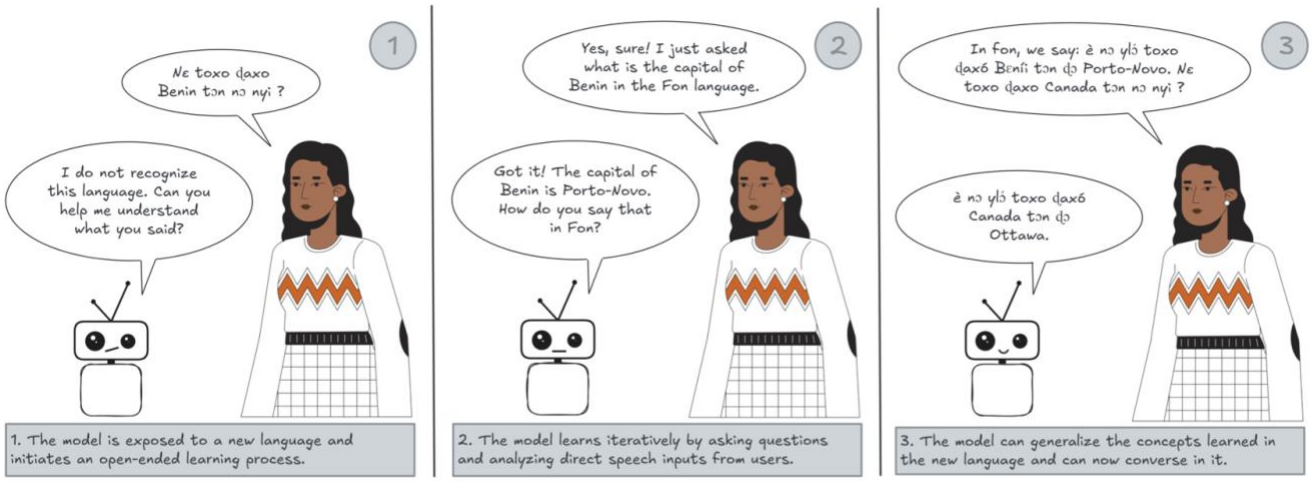


Figure 1: Illustration of open-ended language learning through human–AI interaction. Adapted from Dossou and Aïdasso (2025).

While the original example uses the Fon language, our proposed architecture generalizes this interaction paradigm to Indian tribal languages such as Bhil.

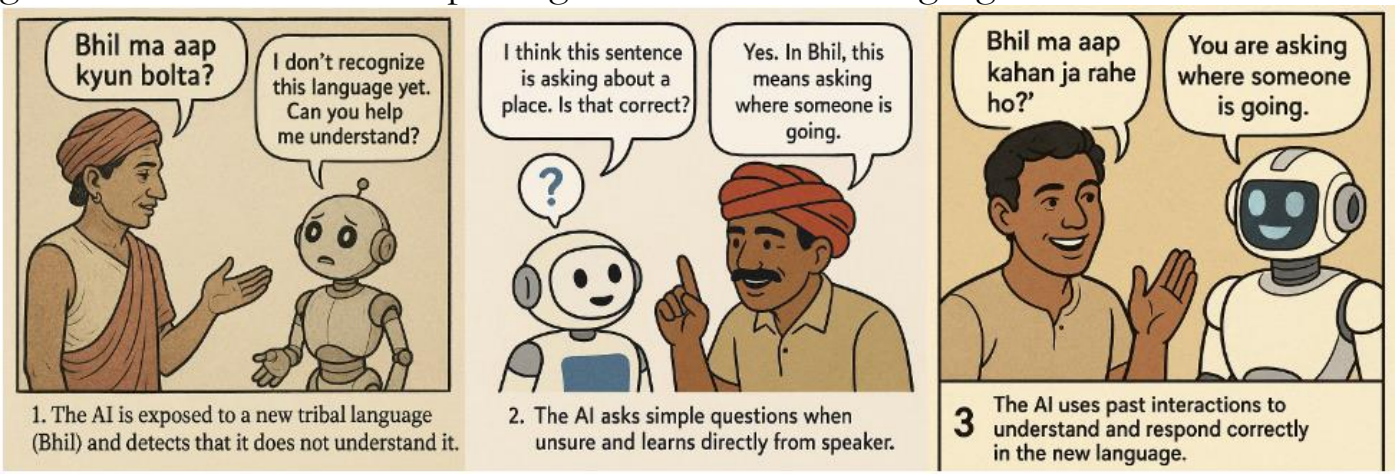


Figure 2: Example of open-ended human–AI interaction, where the system learns the Bhil language by asking questions and improving through feedback.
Source: *Sora by OpenAI*

In our system, uncertainty detection, question selection, and reinforcement learning jointly drive this interaction loop, enabling continuous learning without relying on large pre-existing datasets.

SUMMARY

The proposed system follows a learning strategy in which a small amount of initial training is used only as a starting point, after which the AI learns primarily through reinforcement learning from human interaction, allowing it to improve continuously by asking questions and incorporating feedback.

This project addresses a critical gap in current AI and NLP research: the absence of effective methods for learning and supporting India's tribal and minority languages. These languages are often oral, lack written documentation, and are spoken by small communities, making traditional dataset-driven AI approaches unsuitable.

The literature review shows that important progress has been made in several related areas, including low-resource NLP, human-in-the-loop learning, active learning, uncertainty estimation, and endangered language tooling. Research has demonstrated that human feedback improves model quality, that uncertainty can be measured and used for decision-making, and that language technologies can be built even with limited data. However, these ideas largely exist in isolation.

Most existing systems still rely on static datasets, fixed training pipelines, or post-hoc human correction. Open-ended discovery has been proposed conceptually but not implemented or evaluated in real low-resource settings. Human-in-the-loop methods focus on correcting outputs rather than discovering unknown language structure. Active learning assumes the existence of labeled data, and uncertainty estimation is rarely connected to question-asking strategies for language learning. In the Indian context, NLP resources focus almost entirely on major written languages, leaving tribal and minority languages unaddressed.

To address these gaps, this research proposes an interactive, open-ended human–AI discovery approach. Instead of learning from large static datasets, the AI system learns through continuous interaction with native speakers. It starts with minimal initial data, detects uncertainty in its understanding, asks targeted clarification questions, and improves over time using human feedback. Learning is treated as an ongoing process rather than a one-time training phase.

The proposed architecture combines uncertainty-aware decision-making, targeted human interaction, and continual learning into a single unified system tailored for under-documented languages. By shifting the focus from data collection to interactive discovery, this approach makes it possible to learn, document, and support tribal and minority languages even in the absence of large datasets.

Overall, this work contributes a new direction for AI-based language learning that is human-centered, adaptive, and suitable for the realities of low-resource languages in

India. It has the potential to support language preservation, improve digital inclusion, and enable the development of AI tools for communities that are currently excluded from modern language technologies.

REFERENCES –

1. **Dossou, B. F. P., & Aïdasso, H.** (2025). *Towards Open-Ended Discovery for Low-Resource NLP*. arXiv preprint arXiv:2510.01220. https://arxiv.org/abs/2510.01220
2. **Masakhane Community.** (2020–). *Masakhane: Machine Translation for African Languages*. https://www.masakhane.io
3. **Knowles, R., & Koehn, P.** (2016). *Neural Interactive Translation Prediction*. In Proceedings of the AMTA Conference. https://aclanthology.org/2016.amta-papers.12/
4. **Peris, Á., & Casacuberta, F.** (2018). *Active Learning for Interactive Neural Machine Translation*. Computer Speech & Language, 52, 201–220. https://doi.org/10.1016/j.csl.2018.06.003
5. **Haffari, G., Roy, M., & Sarkar, A.** (2009). *Active Learning for Statistical Machine Translation*. In Proceedings of NAACL-HLT. https://aclanthology.org/N09-1046/
6. **Rao, S., et al.** (2019). *Asking Clarification Questions in Conversational Systems*. In Proceedings of ACL. https://aclanthology.org/P19-1362/
7. **UncertaiNLP Workshop Series.** (2019–2023). *Uncertainty-Aware Natural Language Processing*. https://uncertainty-in-nlp.github.io/
8. **Anastasopoulos, A., et al.** (2020). *ELPIS: Endangered Language Pipeline and Inference System*. In Proceedings of LREC. https://aclanthology.org/2020.lrec-1.469/
9. **Adams, O., et al.** (2018). *Persephone: A Speech Recognition Toolkit for Low-Resource Languages*. In Proceedings of Interspeech. https://doi.org/10.21437/Interspeech.2018-1390
10. **Ramesh, G., et al.** (2022). *Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages*. Transactions of the ACL (TACL). https://aclanthology.org/2022.tacl-1.52/