Linguistic Heritage Research Foundation (LHRF)

Preserving and Advancing Global Linguistic Diversity & Creating Datasets

LHRF Founding Members

Linguistic Heritage Research Foundation

December 3, 2025

Our Vision

- Safeguard global linguistic heritage through cutting-edge research and technology.

- Bridge the gap between official and non-official languages worldwide with modern AI technologies, ensuring no voice is lost.

- Foster inclusive innovation by empowering low-resource, vulnerable, and endangered languages globally.

- Create sustainable research foundation through open-source datasets and paid services.

“Languages are the soul of a nation.”

“Let us not let the languages of our ancestors die out. Let us revive them for tomorrow.”

Mission

- Build comprehensive datasets and tools for computational linguistics in all languages and extend to other languages.

- Collaborate with academia, industry, and communities to drive sustainable preservation.

- Promote open-source resources that enable AI applications in education, healthcare, and culture.

Core Focus

Prioritizing low-resource languages as models for scalable impact across all languages and beyond.

Key Objectives

- Collect and annotate 1,000 hours of conversational speech data per language for robust ASR models in each language.

- Develop datasets tailored for training large language models (LLMs) in each language.

- Create specialized NLP datasets: POS tagging, NER, anaphora resolution, dependency parsing in each language.

- Build OCR datasets for each language scripts, including historical manuscripts in each language.

- Construct multilingual knowledge graphs linking dialects and cultural contexts.

- Design tools for dialectal variation analysis and language revitalization in each language.

- Enable cross-lingual transfer learning for endangered languages.

- Integrate sociolinguistic insights into AI for equitable language tech.

Language Phase 1 (2026-2028)

Priority Languages: English (Indian Accent), Hindi, Marathi, Konkani

- Foundation building with major Indian languages
- Establish data collection and annotation pipelines
- Create baseline AI models and tools
- Pilot applications in education and community engagement

Language Phase 2 (2029-2030)

Additional Languages: Bengali, Telugu, Tamil, Gujarati, Urdu

- Expand to major Dravidian and other Indian languages
- Leverage Phase 1 infrastructure and methodologies
- Cross-lingual transfer learning experiments
- Regional collaboration partnerships

Language Phase 3 (2031-2032)

Additional Languages: Kannada, Odia, Malayalam, Punjabi, Assamese

- Maithili, Santali, Kashmiri, Nepali, Sindhi

- Focus on North and North-East Indian languages

- Special emphasis on tribal and minority languages

- Integration with existing linguistic research networks

Language Phases 4-6 (2033-2038)

Phase 4 (2033-2034): Manipuri, Dogri, Meitei, Bodo, Sanskrit

Phase 5 (2035-2036): Bhili, Gondi, Kurukh, Khandeshi, Tulu

Example

Phase 6 (2037-2038): Khasi, Ho, Garo, Mundari, Tripuri

- Priority on highly endangered and tribal languages
- Community-driven data collection approaches
- Preservation of oral traditions and cultural contexts
- Global collaboration opportunities

Task Phased Approach

Dataset Collection (Parallel across all language phases)

- 1000 hours of conversational speech per language
- Speech-to-Text Translation datasets
- 1000 hours, 4 speakers Text-to-Speech per language
- Text-to-Speech Translation per language
- 1M POS Tagged Sentences per language
- 1M NER Tagged Sentences per language
- 1M Anaphora Tagged Sentences per language
- 1M Sentiment Tagged Sentences per language
- 1M Emotion Tagged Sentences per language
- 1M Bidirectional Multiple Pairs of Translation per language
- 1M Sentences for Transliteration per script

Foundational Model Building

AI Model Development

- ASR (Automatic Speech Recognition) models
- Speech-to-Text Translation models
- Text-to-Speech models
- Text-to-Speech Translation models
- Speech-to-Speech Translation Models
- Language models (SLMs, LLMs)
- Machine Translation models
- POS Tagging and Syntactic Analysis models
- Named Entity Recognition (NER) models
- Sentiment and Emotion Analysis models
- Transliteration models
- Voice AI Agents and conversational systems
- Multimodal models (speech + text integration)

Paid Services Offerings

Revenue Generation for Sustainability

- Commercial API access to language models and datasets
- Custom language model training services
- Linguistic consulting for AI companies
- Data annotation services for enterprises
- Speech technology integration for businesses
- Educational technology licensing
- Cultural preservation technology solutions

Global Implementation Strategy

Phase 1: Foundation Building (2026–2028)

- Establish data collection infrastructure for Language Phase 1 languages

- Build annotation pipelines and quality assurance processes

- Develop foundational AI models and open-source tools

- Create global collaboration network and partnerships

Key Milestones

- 4 language datasets completed and publicly released

- Functional AI models for all supported tasks

- Revenue generation through paid services begins

- International partnerships established

Global Scaling Strategy

Phase 2: Expansion (2029 onwards)

- Scale operations to Language Phases 2-6 simultaneously
- Leverage parallel funding and international collaborations
- Adapt methodologies for diverse linguistic contexts worldwide
- Build comprehensive global language technology platform

Example

Global Impact Vision

- Unified platform serving 100+ languages worldwide
- Sustainable funding through commercial services
- Global network of linguistic researchers and communities
- Preservation of endangered languages and cultures

Academic Division - Tasks & Roles

Tasks

- Optical Character Recognition (OCR)
- Machine Translation (MT)
- Speech-to-Text (STT)
- Token Classification (POS, NER tagging)
- Affect (Sentiment, Emotion tagging)

Roles for each task

- Data Collection Team
- Annotation Team
- Review Team

Academic Division - TTS & Experts

Text-to-Speech (TTS)

1 Male & 1 Female Voice Artist (20–40 years)

1 Male & 1 Female Voice Artist (41–60 years)

Total: 1000 hours per language, 250 hours per speaker

Example

Additional Academic Roles

Language Experts (5 per language)

Education Experts (1 per language)

Non-Academic Division

Support Functions

Support Staff

Management (HR, Product Management)

Finance & Administration

Legal & Compliance Assistance

Development Division

Technical Development Teams

Software Developers

NLP Engineers

DevOps Engineers

MLOps Engineers

LLMOps Engineers

Data Engineers

Research Division

Research Specializations

Computer Science Research

Linguistic Researchers

Sociolinguistic Researchers

Psycholinguistic Researchers

Computational Linguistic Researchers

Roadmap & Milestones

Timeline	Milestones
Q1-Q2 2026	Legal registration as Global Foundation; Initial team onboarding and infrastructure setup.
Q3-Q4 2026	Launch data collection for Phase 1 languages (English, Hindi, Marathi, Konkani).
2027	Complete foundational models and release open-source datasets for Phase 1 languages.
2028	Launch commercial services; Begin Phase 2 language data collection.
2029-2038	Parallel expansion to Phases 2-6 languages with international collaborations.
2038+	Global linguistic technology platform fully operational with 100+ languages.

Next Steps

Secure foundational partnerships for ethical data sourcing and technological infrastructure.

Partnerships & Sustainability

Academic Collaborations: Universities for expertise and student involvement.

Industry Ties: Tech firms for tools and compute resources.

Community Engagement: Local stakeholders for authentic data and advocacy.

CSR Synergies: Aligning with corporate social responsibility goals in education and heritage preservation to fuel annotation and scaling efforts.

Together, building a legacy of linguistic innovation.

Join Us in Preserving India's Voices

Become a partner: Contribute expertise, resources, or funding.

Volunteer or collaborate: From data annotation to research pilots.

Follow our journey: Updates via lhrf.org (coming soon)[1].

Thank You!

Author.
Dummy entry, 2024.
This is a dummy entry to satisfy BibTeX processing.