International Linguistic Heritage Research Foundation (LHRF)

Preserving and Advancing Global Linguistic Diversity & Creating Datasets

ILHRF Founding Members

International 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.

The Language Crisis: Low-Resource Languages

Current Reality

40% of the world's 7,000+ languages are endangered [3]

2,500 languages have fewer than 1,000 speakers [1]

Languages are disappearing at rate of 1 every 2 weeks

90% of languages lack basic digital resources

Indigenous knowledge systems at risk of permanent loss

UNESCO Classification: Language Vitality

Safe Languages

Spoken by all generations with intergenerational transmission

Used in education, administration, media, and daily life

No immediate endangerment threat

Vulnerable Languages

Most children speak the language, but it may be restricted to certain domains

Used orally by all generations but increasingly by adults only

UNESCO: Languages with developing endangerment [2]

UNESCO Classification: Endangered Languages

Definitely Endangered

Children no longer learn the language as mother tongue

Spoken mainly by parental generation and up

UNESCO: Languages with fewer than 100 speakers in some cases [2]

Severely Endangered

Language spoken by grandparents and older generations

While parents may understand it, they do not speak it to children

UNESCO: Languages spoken by fewer than 100 people [3]

UNESCO Classification: Extinct Languages

Critically Endangered & Extinct

Critically Endangered: Spoken only by few members of oldest generation

Extinct: No speakers remaining, but may survive in recorded form

UNESCO: Languages with 0 speakers or only ceremonial use [3]

UNESCO Assessment Factors [2]

Intergenerational transmission

Absolute number of speakers

Proportion of speakers within the total population

Trends in existing language domains

Response to new domains and media

Materials for language education and literacy

Why Immediate Action is Critical

Cultural & Knowledge Loss

- Traditional ecological knowledge disappears with languages
- Medicinal plant knowledge lost forever
- Cultural narratives and oral histories vanish
- Indigenous governance systems undermined
- Community identity and social cohesion threatened

AI & Digital Exclusion

- Low-resource languages excluded from AI advancements
- Digital divide widens between majority and minority languages
- Educational opportunities limited for speakers of endangered languages
- Healthcare access reduced due to language barriers in AI systems

The Foundation's Urgent Mission

Why LHRF Must Start Immediately

Languages are disappearing faster than we can document them

AI technology evolves rapidly - low-resource languages risk permanent exclusion

Community elders are aging - oral knowledge must be preserved now

Global collaboration needed before critical mass is lost

Foundation must be established to coordinate worldwide efforts

Our Commitment

Establish sustainable infrastructure for language preservation

Build global community of researchers and speakers

Create open-source tools accessible to all communities

Ensure no language is left behind in the AI revolution

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

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

Web Platforms for Data Collection

Crowdsourcing Platform (Launch: December 2026)

Open to everyone worldwide for voluntary contributions

Supports spontaneous speech, scripted speech, token classification

Enables translation, transliteration, and affect tagging

Languages and scripts added based on community needs

Language Experts Platform (Invite-Only)

Restricted access - admins approve participants only

Higher quality annotations and reviews than crowdsourcing

Focus on annotation and review of uploaded datasets

Datasets collected offline by data collection teams

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)

Revenue Streams for Foundation Sustainability

Multi-Channel Revenue Strategy

Commercial Services: API access, custom training, consulting

Data Licensing: Premium datasets, research partnerships

Government Grants: Cultural preservation, research funding

Corporate Partnerships: Tech companies, language services

Educational Products: Training programs, certifications

Membership Programs: Institutional access, premium support

Commercial Technology Services

API & Technology Access

Language model APIs with usage-based pricing

Speech-to-text/speech APIs for multiple languages

Machine translation APIs with quality guarantees

Custom model hosting and inference services

Real-time language processing services

Enterprise Solutions

Custom model training for specific domains

Multi-language chatbot development services

Voice assistant integration for businesses

Linguistic consulting for AI product development

Data annotation services for enterprises

Data Licensing & Research Partnerships

Dataset Monetization

- Premium access to curated language datasets
- Research partnerships with data licensing agreements
- Commercial dataset licensing for AI companies
- Specialized corpora for industry applications
- Historical and cultural text collections

Academic & Research Collaborations

- Sponsored research projects with industry partners
- Joint publications and IP sharing agreements
- Technology transfer partnerships
- Collaborative grant applications
- Student internship and research programs

Government & Institutional Funding

Grant-Based Revenue

UNESCO cultural preservation grants

Government research and development funding

Indigenous language revitalization programs

Educational technology innovation grants

International development agency funding

Policy & Advisory Services

Language policy consulting for governments

Impact assessments for language technology projects

Cultural preservation strategy development

Educational curriculum design services

Regulatory compliance consulting

Educational Products & Training

Training Programs

Low-resource language technology certification courses

AI for linguistic research workshops

Cultural preservation technology training

Data annotation and quality assurance training

Computational linguistics bootcamps

Educational Technology

Language learning platform licensing

Educational content development services

Assessment and evaluation tools

Teacher training materials and programs

Digital language preservation toolkits

Membership Programs & Community Support

Institutional Memberships

Tiered membership levels for universities and research institutions

Premium access to datasets and research tools

Priority support and technical assistance

Collaborative research opportunities

Early access to new language models and tools

Community Funding

Crowdfunding campaigns for specific language projects

Individual and community donations

Sponsorship programs for language preservation

Legacy giving and endowment programs

Corporate social responsibility partnerships

Publishing & Content Monetization

Academic Publishing

Research publications and conference proceedings

Open-access journals on linguistic technologies

Books and monographs on language preservation

Technical reports and white papers

Case studies and best practice guides

Digital Content

Online courses and educational materials

Language learning mobile applications

Cultural content licensing and distribution

Podcast series on linguistic diversity

Webinar series and virtual events

Technology Licensing & IP Management

Software Licensing

Open-source software with commercial support options

Proprietary tools for enterprise use

Algorithm licensing for language processing

Platform licensing for educational institutions

API licensing with service level agreements

Intellectual Property Strategy

Patent filings for novel language technologies

Trademark protection for foundation brands

Copyright management for educational materials

Collaborative IP development with partners

Technology transfer agreements

Global Partnerships & Sponsorships

Strategic Alliances

Technology partnerships with major AI companies

Academic collaborations with universities worldwide

NGO partnerships for cultural preservation

Government agency collaborations

International organization memberships

Sponsorship Opportunities

Language-specific sponsorship programs

Event sponsorship and naming rights

Technology showcase sponsorships

Research project sponsorships

Community outreach program sponsorships

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

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

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 ilhrf.org (coming soon).

Thank You!

SIL International.
Ethnologue: Languages of the world.
Ethnologue, 2024.

UNESCO.
Unesco language vitality and endangerment framework.
Technical report, UNESCO, Paris, France, 2003.

UNESCO.
Atlas of the world's languages in danger.
https://www.unesco.org/en/languages-atlas, 2024.
Accessed: December 2024.