Project BoLI – Terms and Conditions

Bolstering Linguistic Resources in the Indigenous Languages of India

Attribution

This document is authored and published by:

UnReaL-TecE LLP
and
Council for Diversity and Innovation (CoDIN)

In connection with the BoLI initiative.

License

You are free to:

- Share — copy and redistribute the material in any medium or format
- Adapt — remix, transform, and build upon the material for any purpose, even commercially

Under the following terms:

- Attribution — You must give appropriate credit to *UnReaL-TecE LLP* and *CoDIN*, provide a link to the license, and indicate if changes were made. License link: https://creativecommons.org/licenses/by/4.0/

Machine-readable attribution:

Document Version

Version: 1.0 (2025)
Part of the BoLI Guidelines

Project BoLI Terms and Conditions

I. These terms and conditions govern the partnership between the researchers, speakers and experts of given languages with specialisation and training in Linguistics (referred to as “linguists”) and the Project BoLI by UnReaL-TecE LLP (referred to as “BoLI”).

II. What is BoLI?

a. BoLI is a network of linguists, language experts (community members), and NLP experts working towards building linguistic resources and technologies in the Indian languages, languoids and varieties.

b. BoLI is a collection of resources in different Indian languages, languoids and varieties consisting of

i. Audio Recordings

ii. Transcription of audio recordings (generally in IPA)

iii. Inter-linear glossing

iv. Free Translation in English / other language

III. Objectives of BoLI

a. Ensure Fairness in Data Collection

i. Ensure that the speakers and researchers get fair compensation and retain IPR through transparent pricing slabs, data royalty and Creative Commons licensing.

ii. Ensure visibility and availability of the dataset through different means.

b. Build the following Linguistic Resources for numerous Indian languages

i. Speech recordings with transcriptions

ii. Translated into English

iii. Grammatically rich representation at the morph level

iv. Ensure representation of different linguistic structures in the datasets for fair training and evaluation.

IV. Joining BoLI

a. Access to BoLI is currently only via invitation. The invitation for proposals is generally sent to the departments or the supervisors of the prospective members/linguists or directly to the linguists.

b. The proposal for joining BoLI involves two steps -

    i. Step 1: Complete the proposal form, as shared by the BoLI team. The BoLI team will review this form. If accepted, an account on the LiFE app or any other app used for BoLI will be created for submitting the first sample dataset. Prospective members will be trained on using the app to submit the dataset and will be required to submit it within the given timeframe.

    ii. Step 2: The submitted sample dataset will be reviewed by the BoLI team for quality and other requirements. The dataset may be accepted without revisions, sent for further revisions, edits, or rejected. If accepted, the dataset will become available as a sample dataset for all members of the BoLI network and commercial and non-commercial sales and access. If sent for revisions, the linguists will need to update the dataset as per the comments from the BoLI team and resubmit. The resubmission will again undergo the review and may be accepted, rejected or asked for further revisions. If the dataset gets rejected, the linguist will not be inducted into the BoLI network. Still, they can apply again with a new or improved version of the dataset submitted earlier.

c. The following terms and conditions will come into effect only after a linguist has been approved to join BoLI.

d. The datasets rejected during the review process will not be governed by these Terms and Conditions. The BoLI team will destroy such rejected datasets within 6 months of declaring the result.

V. By partnering with BoLI, the linguists agree to the following.

a. Make available a sample set of at least 200 basic sentences (translations) and 2 narrations of the language that the linguist is working on (including your native language), recorded by at least 2 speakers (which may include the linguist), within 15 days of signing up for the project, failing which their proposal for joining the BoLI network may be rejected.

b. The recordings will be collected from speakers with different demographic profiles.

c. The linguists will follow all the ethical guidelines as laid down by BoLI for collecting the recording from the community members, including, but not limited to, ensuring the linguistic rights of the community members, adhering to their cultural sensitivities and strictly adhering to the consent-related norms set by BoLI.

d. The recordings will be transcribed into IPA and the official/native/most widely used script, interlinearly glossed as per the Leipzig Glossing Rules and translated into English / other appropriate language(s) using the LiFE App or any such other app as suggested by the UnReaL-TecE team.

e. Whenever a request for additional data comes in for commercial usage or research, linguists will ensure the collection and preparation of such datasets within the given timeline at the mutually agreed-upon rate.

f. The linguists and their associates will directly contact the community members for data collection, and it will not be further subcontracted to any third-party organisations or institutions.

g. The linguists and their associates will directly compensate the community members as per the mutual agreement between them, the community members, and BoLI.

h. The community members will also have a share of the data royalty at the mutually agreed-upon rate.

VI. BoLI promises the following to the linguists.

a. BoLI reserves the right to accept or reject any proposal by any linguist for joining the network without specifying any reason.

b. UnReaL-TecE will provide all necessary training and resources to the linguists to conduct fieldwork with the community for data collection, which includes but is not limited to training in linguistic fieldwork, providing necessary prompts and questionnaires, training in phonetic, phonological and morphosyntactic analysis of linguistic data and training to promote ethical conduct in linguistic fieldwork and research, including guidelines on informed consent, confidentiality and respectful engagement with speech communities.

c. UnReaL-TecE will provide consultancy on appropriate pricing and help you negotiate data royalty provisions whenever the data is resold for commercial usage.

d. The dataset will be hosted on the UnReaL-TecE server for controlled access, and links will be made available on public channels such as GitHub, HuggingFace Hub and others for visibility.

e. All BoLI partners and collaborators will have unlimited, continuous access to all the datasets (including those in other languages) for research.

f. All BoLI partners and collaborators will have unlimited, continuous access to the LiFE App for use in their own research, including those related to their coursework and PhD research, subject to the UnReaL-TecE fair usage policy as decided at different times.

g. All BoLI partners and collaborators will get priority engagement and employment in any project related to the languages they have contributed to, subject to fulfilling the other eligibility criteria for the positions they apply for.

VII. IPR, Licensing and Reuse

a. The sample and full datasets will be accessible only through the LiFE App or any other portal developed by UnReaL-TecE under a Creative Commons CC BY-NC-SA 4.0 License.

b. Under the above-mentioned license, the sample dataset will be freely available for research by non-commercial entities and individuals.

c. The full dataset will be made available only to members of the BoLI network, including linguists and community members, for non-commercial research under the above-mentioned license.

d. The sample and the full dataset may be used to develop different kinds of language technologies for these languages. These technologies will remain freely accessible to the members of the BoLI network but might be made available at a charge to non-members by UnReaL-TecE and BoLI.

e. The request for non-commercial usage of the full dataset by researchers, linguists and other stakeholders will be individually handled jointly by the community members and linguists (the primary stakeholders) associated with the specific dataset and languages. The primary stakeholders will decide upon and communicate the terms and conditions, including license and cost, if any, of such datasets. BoLI may provide non-binding consultations to the primary stakeholders in such cases on request. Suppose a cost is associated with the dataset for non-commercial usage. In that case, it will be treated on par with the dataset for commercial usage (even if the cost is lower than that for commercial usage), and all conditions related to the request for commercial usage will override these conditions for non-commercial usage.

f. The request for commercial usage of the sample and full dataset will be managed through a centralised portal/platform (including but not limited to mailing groups, instant messaging groups, etc.) accessible to all partners, and regular updates will be made available to them on these platforms via emails and other means. In case of a potential purchase agreement, the community members, linguists, and BoLI will work together to develop a purchase agreement in which they uphold the agreed-upon terms. Such purchase agreements will be developed separately for each dataset and for each purchaser every time there is a requirement from the purchaser.

g. All copyright and intellectual property rights, including the right to sell, redistribute, and use the data shared with BoLI, will rest with the respective speakers, speech communities, and linguists. They are free to use, sell, and distribute the datasets through other channels, including their own internal sales

mechanisms. BoLI acts only as a non-exclusive distribution partner for these datasets, which makes the data and its creators visible to a wider audience through its distribution channels.

VIII. Termination, Exit and Rejoining

a. Linguists can exit the BoLI network at any point by sending a written intimation to BoLI. They are not bound to give any explanation or reason for their exit, although BoLI prefers to receive feedback in such instances to improve its services.

b. Linguists can rejoin the BoLI network by providing at least one sample dataset to the project - this could be an existing dataset by the linguist or a new dataset.

c. BoLI may terminate the membership of any linguist from the BoLI network for violation of its terms and conditions or any other ethical or legal violation or malpractice, either related to BoLI or otherwise, as decided by UnReaL-TecE. BoLI is not bound to explain the reason for the termination of the membership. However such a decision of termination will be taken after at least two warnings issued to the concerned linguists

d. The termination from the BoLI network is perpetual such that linguists once terminated from the BoLI network will not be able to rejoin the network under any condition.

e. Upon exit or termination, linguists will cease to have access to all benefits of the BoLI network as outlined in Clause V of this document.

f. The sample and full datasets contributed by linguists who are no longer members of the BoLI network will continue to be hosted and made available to the other members of the BoLI network for non-commercial research purposes. However, such datasets will no longer be available for commercial sale or use by entities or individuals outside the network.

g. The linguists (both the current and the past members of the BoLI network) can remove one or more of their sample and/or full dataset to all members of the BoLI network at any point by sending a written intimation to BoLI. They are not bound to give

any explanation or reason for their action, although BoLI prefers to receive feedback in such instances to improve its services.

h. If a current member of the BoLI network removes all their sample and full datasets, they will automatically seize to be a member of the BoLI network. In such instances, they will be considered at par with the members who have exited the network.

i. The removed datasets will seize to be hosted by BoLI and it will no longer be made available to anyone outside of the BoLI network or any member joining the BoLI network after the removal of the dataset. However, the dataset will continue to be made available to the earlier members of the network and anyone to whom the dataset was made available for commercial or non-commercial purposes prior to its removal. Access to the dataset, once granted, remains perpetual within the license under which the access was granted.

j. No dataset can be made available to the entities outside of the BoLI network if its access is not available for the members of the BoLI network.

k. Access to the dataset to members of the BoLI network is independent of the membership of the BoLI network. Thus exit from the network and removal of access to the dataset must be intimated individually.

IX. Audits and Compliance Monitoring

a. BoLI will conduct regular audits and assessments of data privacy and ethical data collection practices by all partners and associates of the project to ensure compliance with internal policies and external regulations.

b. Wherever necessary, corrective measures and action, including but not limited to that mentioned in VIII.c and VIII.d of this document, will be taken to address any identified deficiencies or non-compliance issues.

c. Such corrective measures may also be taken at the request or complaint of any stakeholder of the project.

X. Amendments and Conflict

a. These terms and conditions may be amended at any point by a majority of the existing members of the BoLI including members from UnReaL-TecE, linguists, community members and other stakeholders that are part of the project. The co-founders of UnReaL-TecE will have a veto power in such instances.

b. In case of any conflict, resolutions may be sought through arbitration among the members of the BoLI network.