

IEEE SA: PRE-STANDARDIZATION STUDY FOR INDIAN LANGUAGES REPORT

Authored by

IEEE SA Pre-Standardization Workstream Team for Study for Indian
Languages (2020–22)

TRADEMARKS AND DISCLAIMERS

IEEE believes the information in this publication is accurate as of its publication date; such information is subject to change without notice. IEEE is not responsible for any inadvertent errors.

The ideas and proposals in this specification are the respective author's views and do not represent the views of the affiliated organization.

The Institute of Electrical and Electronics Engineers, Inc. 3 Park Avenue, New York, NY 10016-5997, USA

Copyright © 2023 by The Institute of Electrical and Electronics Engineers, Inc.

All rights reserved. 30 June 2023. Printed in the United States of America.

PDF: STDVA26259 978-1-5044-9844-9

IEEE is a registered trademark in the U.S. Patent & Trademark Office, owned by The Institute of Electrical and Electronics Engineers, Incorporated. All other trademarks are the property of the respective trademark owners.

IEEE prohibits discrimination, harassment, and bullying. For more information, visit <http://www.ieee.org/web/aboutus/whatis/policies/p9-26.html>.

No part of this publication may be reproduced in any form, in an electronic retrieval system, or otherwise, without the prior written permission of the publisher.

Find IEEE standards and standards-related product listings at: <http://standards.ieee.org>.

NOTICE AND DISCLAIMER OF LIABILITY CONCERNING THE USE OF IEEE SA INDUSTRY CONNECTIONS DOCUMENTS

This IEEE Standards Association (“IEEE SA”) Industry Connections publication (“Work”) is not a consensus standard document. Specifically, this document is NOT AN IEEE STANDARD. Information contained in this Work has been created by, or obtained from, sources believed to be reliable, and reviewed by members of the IEEE SA Industry Connections activity that produced this Work. IEEE and the IEEE SA Industry Connections activity members expressly disclaim all warranties (express, implied, and statutory) related to this Work, including, but not limited to, the warranties of: merchantability; fitness for a particular purpose; non-infringement; quality, accuracy, effectiveness, currency, or completeness of the Work or content within the Work. In addition, IEEE and the IEEE SA Industry Connections activity members disclaim any and all conditions relating to: results; and workmanlike effort. This IEEE SA Industry Connections document is supplied “AS IS” and “WITH ALL FAULTS.”

Although the IEEE SA Industry Connections activity members who have created this Work believe that the information and guidance given in this Work serve as an enhancement to users, all persons must rely upon their own skill and judgment when making use of it. IN NO EVENT SHALL IEEE OR IEEE SA INDUSTRY CONNECTIONS ACTIVITY MEMBERS BE LIABLE FOR ANY ERRORS OR OMISSIONS OR DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS WORK, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

Further, information contained in this Work may be protected by intellectual property rights held by third parties or organizations, and the use of this information may require the user to negotiate with any such rights holders in order to legally acquire the rights to do so, and such rights holders may refuse to grant such rights. Attention is also called to the possibility that implementation of any or all of this Work may require use of subject matter covered by patent rights. By publication of this Work, no position is taken by the IEEE with respect to the existence or validity of any patent rights in connection therewith. The IEEE is not responsible for identifying patent rights for which a license may be required, or for conducting inquiries into the legal validity or scope of patents claims. Users are expressly advised that determination of the validity of any patent rights, and the risk of infringement of such rights, is entirely their own responsibility. No commitment to grant licenses under patent rights on a reasonable or non-discriminatory basis has been sought or received from any rights holder. The policies and procedures under which this document was created can be viewed at <http://standards.ieee.org/about/sasb/iccom/>.

This Work is published with the understanding that IEEE and the ICCC members are supplying information through this Work, not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought. IEEE is not responsible for the statements and opinions advanced in this Work.

TABLE OF CONTENTS

Abstract.....	5
1. Introduction	6
2. Scope	7
3. Key areas of investigations.....	8
3.1 Introduction	8
3.2 Accessibility	8
3.3 Evaluation metrics.....	9
3.4 Speech processing	9
3.5 Text processing.....	10
3.6 Script processing	11
4. Standards survey	11
4.1 Introduction	11
4.2 Survey of accessibility standards.....	12
4.3 Survey of evaluation standards.....	16
4.4 Survey of speech processing standards.....	18
4.5 Survey of text processing standards.....	19
5. Standards gaps	20
6. Resources and key gaps.....	23
7. References	24
Appendix A	25

IEEE SA: PRE-STANDARDIZATION STUDY FOR INDIAN LANGUAGES REPORT

ABSTRACT

The IEEE Standards Association (IEEE SA) pre-standards workstream for under-resourced Indian languages was initiated to identify use cases that could lead to proposals for standards to govern language resources for Indian languages. Language resources include speech, language data, and descriptions that are made available in a machine-readable form and used for developing, evaluating, and improving algorithms in natural language and speech processing. Participants in the workstream were carefully gathered based on their body of work in furthering deployment of technology for Indian languages. The team grew to include their network of language experts, as well as representatives from other studies on Indian languages. The workstream commenced in August 2020 and was completed in September 2022. Participants in this pre-standards workstream who are the authors of this paper are listed in Appendix A. The pre-standards workstream led to a detailed assessment of the leading use cases driving the deployment of technology, a list of existing standards that can be used to develop technology for these use cases, and standards gap assessment that recognizes the limits to solutions deployment with Indian languages. The workstream recommendation contains a detailed assessment of use cases and standards in four focus areas: accessibility, evaluation metrics, speech, and text. The same assessment for scripts could not be completed. The output of the workstream in each of the focus areas is a detailed report each of which is being released independently. The recommendation from the workstream should be deliberated in a workshop, leading to the definition of concrete steps to address the identified gaps.

1. INTRODUCTION

India has 1369 classified mother tongues. Of these classified languages, 121 are spoken by 10,000 or more speakers. Twenty-two of these languages comprise nearly 97% of the Indian population and are termed *scheduled languages*. The remaining 99 non-scheduled languages account for the remaining 3% of the people (Office of the Registrar General, India [2]). The scheduled languages of the country are Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Odia, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu, and Urdu.

Language-enabled products are gaining wide traction in every sector of the economy. It is acknowledged that the next billion internet users in India will come because of the localization of content in Indian languages. Automatic translation, speech and text recognition, context-aware automation, and extraction, etc., are the key use cases driving the deployment. The biggest challenge, and hence an area of active innovation for Indian languages, is the reality that most of these languages are under-resourced. This limits the deployment of technology support to overcome the language diversity challenge. This challenge opens a big opportunity for standardization and fits very nicely with why IEEE drives standardization activities. This document is intended to serve as a study of the standardization landscape for Indian languages and highlight the key standardization gaps that need to be addressed on priority. Standardization helps ensure that language products and materials are tailor-made for their purpose, promotes interoperability of products and services in a multilingual country like India, promotes common understanding of a product irrespective of the language to which it caters, and assures consumer safety and privacy through security requirements. Standardization will spur innovation, fuel new growth of technology consumers, expand customer choice, reduce cost structures, and support interoperability between multiple languages, among other benefits.

The study covers relevant use cases that drive technology deployment, identify gaps in standardization that need to be addressed for wider adoption of technology, and provide proposals justifying the need for standards to govern language resources for Indian languages. Language resources include speech, text, language data, language descriptions (available in a machine-readable form and used for developing solutions), and metrics used for the evaluation of solutions and algorithms in natural language and speech processing. Standards can be used for language studies, localization of software, electronic publishing, and other purposes by researchers, subject area specialists, etc. Examples of language resources include spoken and written corpora, computational lexica, and terminology databases.

2. SCOPE

The scope of the pre-standardization study is as follows:

1. This series of documents is a pre-standardization study focusing on identifying the gaps in the standards. The series does not offer any recommendation on prioritization of the gaps nor defines any standards to address the identified gaps.
2. The study covers only the listed 22 scheduled languages. The series offers no comments on the validity of its content for the non-scheduled languages. However, readers are free to take the contents and apply them to the non-scheduled languages.
3. The working group was divided into specific focus groups that together dominate the technology landscape. The working groups are accessibility, evaluation metrics, speech, text, and script. Each focus group output is available as a separate report.
4. Each focus group's individual scope of investigation supersedes the scope listed in this section.
5. Identification of required standards is based on the use cases driving the deployment. This analysis should lead to a detailed description of the gaps that exist. The study can adopt methodologies and incorporate results from similar undertakings from around the world. A good example is the work undertaken by the European Union through [ELRA](#).
6. The series of documents can also propose workshops, conferences, detailed discussions, and deliberations to trigger research initiatives, contests, etc., to promote investigations to detail the gaps.
7. AI is driving most of the technological innovations in language-enabled products. Although this series is not limited to AI-driven technology, there will be a distinct bias toward gaps as determined by the development of language products with strong AI components. Readers are requested to highlight gaps in the series related to the non-AI-based language-enabled products.

A critical disclaimer is that this study is only about the standard gaps and will not make any effort to define the required standards. The authors do not make any commitments that the work will equally apply to other non-official languages of India.

3. KEY AREAS OF INVESTIGATIONS

3.1. INTRODUCTION

To enable a detailed study of the standardization gaps in Indian languages, it was decided to split the study into smaller workgroups, namely: accessibility, evaluation metrics, speech processing, and text and script processing. Experts in relevant domains were assigned to each of the workgroups. The workgroups were driven by a respective group lead with full operational freedom to decide how the group went about its activities. This meant that each workgroup had its own meeting schedule and assignment of experts to various sub-aspects of the workgroup investigation. All the workgroups met in a common monthly meeting that was moderated by the chair of the overall study. The report of each workgroup is published as an independent document.

The workgroups were free to choose their focus areas within the broad umbrella of the area they were investigating for gaps. Each workgroup and its focus areas are further described in the following sections.

3.2. ACCESSIBILITY

Enabling universal access to the web also includes access to those with disabilities. About 2.2% of India's population lives with some form of disability. These include disabilities of vision, hearing, speech, movement, and intellect, as well as mental illness, multiple disabilities, and others. India is a signatory to the United Nations Convention on the Rights of Peoples with Disabilities and has an obligation to provide equal opportunities and facilities to everyone, irrespective of any disabilities.

Accessibility enables people with disabilities to perceive, understand, navigate, interact with, and contribute to the web and other information technology—based activities. Accessibility is all pervasive and needs to be present as a built-in feature rather than as an add-on. Apart from standards development, its implementation and compliances are important and need to be respected. Accessibility deserves a place in the full ecosystem of standards (and guidelines and best practices) creation, awareness creation, capacity building and training, test suites and test tools, compliance matrixes, certification labs, and research and development on new technologies and products.

The Accessibility workgroup was chartered with highlighting the state of affairs on standards that govern the availability of accessibility-related features. The workgroup baselined the available standards from bodies like the World Wide Web Consortium (W3C), Quality Assurance International (QAI), and Web Content Accessibility

Guidelines (WCAG) and highlighted the gaps that exist. The gaps were then mapped to the standardization effort by the Bureau of Indian Standards (BIS), thereby highlighting the consolidated set of gaps that exist for Indian languages. The gaps include both missing standards and the customization required to existing languages due to the unique characteristics of Indian languages. The study looked at dominant use cases that require a person with disability (PwD) to interact with technology, and the study highlighted the gaps.

The study is entitled *IEEE SA: Indian Language Resources—Accessibility Subcommittee Report*.

3.3. EVALUATION METRICS

Evaluation is a crucial aspect of any software system, be it an end user-facing system, an intermediate tool, or a model used to build other end-to-end systems. In the context of language technologies, evaluation includes three distinct parts: the evaluation protocol or process, the evaluation test bench (or test data), and evaluation metrics. All three components require standardization to ensure the evaluation is independent, intentional, transparent, reproducible, impartial, high quality, timely, and useful. This requires carefully crafted standards of metrics, datasets, and protocols of evaluation.

The Evaluation workgroup was chartered with highlighting the state of affairs on standards that govern evaluation metrics required for Indian languages. The workgroup baselined the available metrics, standardization approaches, protocols, and evaluation test benches for Indian languages. As a result, the gaps that exist were pointed out. The metrics were grouped as those applicable to end users and developers and those metrics that apply to both. The workgroup focused primarily on text processing and multimodal language processing. Areas that were not looked at were speech processing, basic text processing, and script/input method evaluation—these were left to respective workgroups. Further, the quality of available datasets and the reporting of numbers from evaluation drives and studies were not covered in the investigation. This study appears as the document entitled *IEEE SA: Indian Language Resources—Evaluation Subcommittee Report*.

3.4. SPEECH PROCESSING

Speech processing deals with computer systems that recognize spoken words. Speech solutions are a combination of all speech technology components to form a complete solution. The technology components that go into a speech solution include voice activity detection, language detection, speech enhancement, speaker diarization, speech recognition, speech synthesis, speaker identification and verification, and speaker profiling.