

Keyboards for the Endangered Idu Mishmi Language

Akhilesh Kakolu Ramarao

Department of English Language and Linguistics
Faculty of Arts and Humanities, Heinrich Heine University Düsseldorf
akhilesh.kakolu.ramarao@uni-duesseldorf.de

Abstract

We present mobile and desktop keyboards for Idu Mishmi, an endangered Trans-Himalayan language spoken by approximately 11,000 people in Arunachal Pradesh, India. A Latin-based orthography, the Idu Azobra, was developed in 2018, but no digital input tools existed to use it. The orthography requires characters absent from standard keyboards, including schwa (ə), retracted vowels (ɘ, ɔ, ɤ), nasalized vowels, and accented forms, several of which involve multi-codepoint Unicode sequences that default keyboards do not support. Developed with the Idu Mishmi community, the keyboards comprise: (1) an Android mobile keyboard, published on the Google Play Store, and (2) a Windows desktop keyboard distributed as a single portable executable. Both tools support the complete character inventory, and operate fully offline with zero network permissions. The Android keyboard has been adopted by community leaders and teachers who currently know and actively use the Idu Azobra orthography. The Windows keyboard is currently undergoing testing with community leaders. We describe the design, implementation, and deployment as a replicable model for other endangered language communities.

Keywords: Endangered Languages, Language Technology, Keyboard, Idu Mishmi, Language Revitalization

1. Introduction

Idu Mishmi, also known as Kera'a (Peck, 2020) (ISO 639-3: `cl̥k`), is an endangered Trans-Himalayan language spoken in five districts of Arunachal Pradesh, India, mainly in the Dibang Valley and Lower Dibang Valley regions. The number of speakers is estimated at 11,000 (Lewis et al., 2015), covering two main dialects, Midu and Mithu (Reinöhl, 2022). The language is endangered for social, economic, and political reasons: Hindi and English serve as prestige varieties that speakers shift to (Kaland et al., 2023), and Idu Mishmi is mostly no longer the primary language of the home. For the community, this shift threatens not only the language but the cultural identity embedded in it, a concern shared by many language communities across northeast India.

Community-led efforts to reverse this shift have produced significant results. The Idu Language Development Committee (ILDC) formally developed the Latin-based *Idu Azobra*, the official and standard script for Idu Mishmi, in 2018. That same year, the first Idu Mishmi dictionary was completed (Blench et al., 2018), comprising both dialects, alongside a grammar (Blench, 2018) and the school textbook *Asahi*. In 2021, Idu Mishmi entered classrooms as an optional third language in Idu-inhabited districts.

However, while the orthography exists on paper, no digital input tools existed to use it. The Idu Azobra requires characters such as schwa (ə), retracted vowels (ɘ, ɔ, ɤ), and their accented combinations, none of which are available on most default keyboards. Without a way to type their language, teachers cannot prepare instructional ma-

terials, community organizations cannot write official communications, and young speakers cannot use Idu Mishmi in the digital spaces where they increasingly communicate.

We address this gap with two tools, developed in collaboration with the Idu Mishmi Cultural and Literary Society (IMCLS), the community's apex body for language and cultural affairs: (1) an **Android mobile keyboard**, published on the Google Play Store and actively used in teacher training programs, and (2) a **Windows desktop keyboard** currently undergoing community testing. The character inventory, layout decisions, and testing were carried out jointly with community leaders and teachers, ensuring that the tools reflect the orthography as sanctioned by the community. Both keyboards support the full Idu Mishmi character inventory, operate entirely offline, and are open-source. We describe the design, and deployment, and present this as a model that may be adapted by other endangered language communities.

2. Background

2.1. The Idu Mishmi Writing System

The Idu Azobra uses Latin letters extended with special characters to represent sounds that English does not have (Blench, 2015). These include the schwa (ə), retracted vowels (ɘ, ɔ, ɤ) where the tongue root is pulled back, five nasalized vowels (ã, ê, ĩ, õ, ü), and three accent marks for tonal contrasts: grave for low tone, acute for high tone, and macron for mid tone. Table 1 lists the full set of characters that need dedicated keyboard support.

In Unicode, common accented vowels like à

Category	Characters	Unicode sequence
Schwa	ə, ɘ	U+0259, U+018F
Retracted ə	ɘ̠, ɘ̡	base + U+0331
Retracted o	ɔ̠, ɔ̡	base + U+0331
Retracted u	ɯ̠, ɯ̡	base + U+0331
Grave	à, è, ì, ò, ù	precomposed (e.g. U+00E0)
Acute	á, é, í, ó, ú	precomposed (e.g. U+00E1)
Nasalized	ã, ã̃, ï, õ, ù̃	precomposed (e.g. U+00E3)
Macron	ā, ē, ī, ō, ū	precomposed (e.g. U+0101)
Accented ə	è̃, é̃, ì̃, ò̃	U+0259 + accent
Accented ɘ	ɘ̃, ɘ̃̃, ɘ̃̃̃	U+0259 + U+0331 + accent
Accented ɔ	ɔ̃, ɔ̃̃, ɔ̃̃̃	o + U+0331 + accent
Accented ɯ	ɯ̃, ɯ̃̃, ɯ̃̃̃	u + U+0331 + accent

Table 1: Idu Mishmi characters requiring keyboard support. U+0331 is Combining Macron Below (retraction); accents are U+0300 (grave), U+0301 (acute), U+0303 (tilde/nasalization), U+0304 (macron).

(U+00E0) or ã (U+0101) each have a single precomposed codepoint. However, accented schwas must be built by combining separate codepoints: for example, è̃ is U+0259 (Latin Small Letter Schwa) followed by U+0300 (Combining Grave Accent). Retracted vowels use U+0331 (Combining Macron Below): for example, ɘ̃ is U+0259 + U+0331. Accented retracted schwas require three codepoints in canonical order: è̃̃̃ is U+0259 + U+0331 (CCC,220) + U+0300 (CCC,230), where lower combining class values must come first. Getting this order wrong produces text that looks identical on screen but is treated as different by software, causing failures in search and copy-paste. Both keyboards enforce this ordering for all output.

2.2. Related Work

Keyboards are recognized as foundational infrastructure for digital language use, and community control over such technology is essential for meaningful revitalization (Meighan, 2021). Several projects have developed keyboards for endangered languages, including Livonian (Hämäläinen and Alnajjar, 2021) and Plains Cree syllabics (Santos and Harrigan, 2020). A recurring design challenge is what Paterson (2015) calls the “diacritic density problem”: tone languages use accent marks far more often than European languages, and standard keyboards are not designed to handle this. Idu Mishmi shares this challenge, as a single vowel can carry a tone mark, nasalization, and retraction at the same time.

On the desktop side, existing tools for building custom keyboards each have drawbacks for our setting. Microsoft Keyboard Layout Creator (MSKLC) can only output one Unicode code per key, so it cannot produce characters like ɘ̃ that require two codes. Windows Text Services Frame-

work (TSF) integrates with the system language bar but requires administrator privileges to install, which is not feasible on shared school and government computers. Keyman¹ is the most capable alternative: it supports multi-codepoint output, dead key composition, cross-platform deployment, and is widely used for endangered language keyboards. However, Keyman requires users to install a runtime engine, and on shared school and government computers where users lack installation privileges, this creates a barrier. Our hook-based approach offers a single executable that runs without requiring administrator access. A longer-term path is to submit an Idu Mishmi keyboard layout to the Unicode CLDR² (Common Locale Data Repository). Layouts accepted into CLDR can eventually be shipped natively by operating systems, removing the need for any third-party software. However, CLDR inclusion requires a stable, community-approved layout and a formal locale definition, both of which are preconditions we are working toward.

3. System Design

3.1. Android Mobile Keyboard

The Android keyboard is built on HeliBoard,³ an open-source, privacy-focused keyboard app (GPL-3.0). Rather than adding Idu Mishmi support to HeliBoard directly, we released a separate app for three reasons: (1) a community-branded app is easier to find on the Play Store and builds trust with speakers; (2) Idu Mishmi-specific settings, such as which characters appear under each key, are hard-coded rather than left as user options; and (3) the community retains full control over updates without depending on the upstream project’s release schedule. Building on an established keyboard gave us text prediction, gesture typing, clipboard management, and theming out of the box, so we could focus entirely on Idu Mishmi customizations.

Layout and input method. The keyboard uses a standard QWERTY layout, since Idu Mishmi speakers are bilingual in English. To type Idu Mishmi characters, the user **long-presses** a base key, which opens a popup strip of related characters (Figure 1). For example, long-pressing ə reveals all schwa variants (ə, ɘ) and their accented forms alongside standard accented ə characters. Similarly, o and u include retracted ɔ̠ and ɯ̠ with their accented forms, while a and i provide accented and nasalized variants. This grouping was

¹<https://github.com/keymanapp/keyman>

²<https://cldr.unicode.org/>

³<https://github.com/Helium314/HeliBoard>

decided in consultation with community members, with the most frequently used Idu-specific characters placed at the beginning of each strip for quick access. No memorization is needed; users simply browse the popup visually, making the keyboard accessible to first-time users. We registered the app with a locale code for Idu Mishmi, so that Android recognizes it as an Idu Mishmi input method. The current release uses English dictionaries for basic word completion; an Idu Mishmi dictionary is planned as future work (see Section 5).

Privacy and compatibility. The app requests no internet permissions; all text processing is on-device. It targets SDK 35 with minimum SDK 21, covering approximately 98% of active Android devices.

3.2. Windows Desktop Keyboard

The Windows keyboard is a standalone program written in Go, distributed as a single portable executable that requires no additional software. It works by intercepting keystrokes at the operating system level using a low-level keyboard hook (`WH_KEYBOARD_LL`). When the user presses a mapped key combination, the program suppresses the original keystroke and sends the corresponding Idu Mishmi character to whichever application is active. Normal typing, including standard shortcuts (Ctrl+C, Alt+Tab, etc.), is unaffected.

Input method. The keyboard provides Idu Mishmi characters through the **AltGr** key (Right Alt) in two ways. First, *direct mappings* produce special characters in a single step: AltGr+E → ē, AltGr+R → ě, AltGr+O → օ, AltGr+U → ŭ (with Shift for uppercase). Second, *dead keys* allow two-step composition for accented characters: the user first presses AltGr plus an accent key (backtick for grave, apostrophe for acute, tilde for nasalization, hyphen for macron), then presses a vowel key to produce the accented form. For example, AltGr+~ followed by AltGr+E produces ê. Dead keys compose with both plain and retracted schwas. The implementation ensures that multi-codepoint characters are always output in the correct Unicode order.

Compatibility with other input methods. The keyboard coexists with other input methods (e.g., Hindi InScript, Devanagari phonetic) without conflict, as it only activates on AltGr combinations and passes all other keystrokes through unchanged. This is important in practice as many Idu Mishmi speakers also type in Hindi and English, and the keyboard must not interfere with those workflows.

Limitations. The hook-based approach was chosen for portability but has known limitations. Some enterprise security policies may block low-level hooks. When deleting a composed character like è (which consists of three Unicode codepoints), most modern applications remove the entire character in one backspace press, but some older applications delete one codepoint at a time, requiring multiple presses. The keyboard guarantees correct ordering for new input but cannot fix incorrectly ordered sequences in pasted text, as it has no access to the text field itself.

4. Deployment and Community Impact

The Android keyboard was launched in December 2025 and has approximately 100 installations. Its users are mainly community leaders and teachers, the people who currently know and actively use the Idu Azobra orthography. This reflects where orthography adoption stands today: since 2021, Idu Mishmi has been taught as an optional third language in Idu-inhabited districts, and the first textbook, *Asahi (Idu Azobra Level I)*, was developed by the Idu Tho Ahie Committee (ITAC) under the IMCLS and submitted to the State Council of Educational Research and Training (SCERT). Community leaders now use the keyboard to train school teachers in digital literacy for this program, and the user base is expected to grow as more people learn the orthography. The Windows keyboard is being tested with community members, targeting schools and government offices where Windows desktops are standard.

Both tools work fully offline. This is necessary because internet connectivity in the Dibang Valley is limited, and it is also a deliberate choice to keep language data under community control rather than sending it to external servers (Meighan, 2021). For Idu Mishmi speakers living in other Indian cities, the Play Store listing makes the keyboard easy to find, and the offline design means it works the same way everywhere.

Sustainability and governance. Long-term maintenance is managed by the IMCLS, which has decision-making authority over the orthography and its digital tools. The source code is publicly available under GPL-3.0, so the community is not dependent on any single developer. Teacher training sessions, run as part of the Idu Azobra school program, include instruction on installing and using the keyboard, making it part of existing educational infrastructure rather than a separate tool. Adopting the keyboards in schools and government offices requires no special IT setup: the Windows program is a single portable

6. Bibliographical References

- Roger Blench. 2015. Draft proposal for a writing system for Idu Mishmi, a language of Arunachal Pradesh. Cambridge: Kay Williamson Educational Foundation.
- Roger Blench. 2018. A grammar of Idu, a language of Arunachal Pradesh. Cambridge: Kay Williamson Educational Foundation.
- Roger Blench, Mite Lingi, and Hindu Meme. 2018. A dictionary of Idu, a language of Arunachal Pradesh: Comprising both Midu and Mithu dialects. Cambridge: Kay Williamson Educational Foundation.
- Mika Hämäläinen and Khalid Alnajjar. 2021. Developing keyboards for the endangered Livonian language. In *Proceedings of the Fifth Workshop on Widening Natural Language Processing*, pages 18–20, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Constantijn Kaland, Naomi Peck, T. Mark Ellison, and Uta Reinöhl. 2023. An initial exploration of the interaction of tone and intonation in Kera'a. In *Proceedings of the 1st International Conference on Tone and Intonation (TAI)*, pages 132–136.
- M. Paul Lewis, Gary F. Simons, and Charles D. Fennig. 2015. *Ethnologue: Languages of Asia*. SIL International, Dallas.
- Paul J Meighan. 2021. Decolonizing the digital landscape: The role of technology in Indigenous language revitalization. *AlterNative: An International Journal of Indigenous Peoples*, 17(3):397–405.
- Hugh J. III Paterson. 2015. Keyboard layouts: Lessons from the Me'phaa and Sochiapam Chinantec designs. In Mari C. Jones, editor, *Endangered Languages and New Technologies*, pages 49–66. Cambridge University Press.
- Naomi Peck. 2020. Kera'a (Arunachal Pradesh, India) – language snapshot. *Language Documentation and Description*, 19:26–34.
- Uta Reinöhl. 2022. Locating Kera'a (Idu Mishmi) in its linguistic neighbourhood: Evidence from dialectology. In Mark W. Post, Stephen Morey, and Toni Huber, editors, *Ethnolinguistic Prehistory of the Eastern Himalaya*. Brill, Leiden.
- Eddie Antonio Santos and Atticus Harrigan. 2020. Design and evaluation of a smartphone keyboard for Plains Cree syllabics. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced Languages and Collaboration and Computing for Under-Resourced Languages (SLTU-CCURL 2020)*, pages 88–96, Marseille, France. European Language Resources Association.