L2 Document Number:	L2/26-260 (Status Report)
Date:	October 20, 2026
To:	Unicode Technical Committee (UTC #189, Nancy, France)
From:	Cibu Johny (Chair, Orthographic Sequences Working Group)
Subject:	**Joint CLDR/UTC Orthographic Sequences Working Group (SEQ) Activity & Status Report**

Status Report: Joint CLDR/UTC Orthographic Sequences Working Group (SEQ)

1. Executive Summary & Formation Mandate (L2/26-045)

The Orthographic Sequences Working Group (SEQ) was established following the mandate set forth in proposal **L2/26-045** and formal UTC consensus decision **[186-Cx]**. The creation of this Joint Working Group responded directly to a request presented by the Ministry of Electronics and Information Technology (MeitY), Government of India, during UTC #185 for standardized, machine-readable specifications of valid orthographic character sequences across Indian languages.

Administratively, SEQ is chartered as a Joint Working Group with the **CLDR Technical Committee (CLDR-TC) as the primary owning TC** and the **Unicode Technical Committee (UTC) as a partner**. This structural placement reflects the dual nature of the work: language-specific orthographic specifications fit naturally within CLDR locale structures, while script-wide invariants feed directly into UTC core standards and the UCD DoNotEmit.txt.

2. Group Operations & Leadership

The working group operates on a bi-weekly cadence (Mondays, 8:30–9:30 AM PT). Leadership confirmation has been routed through Unicode executive leadership (CEO Toral and CTO Kathy) and CLDR-TC:

- **Chair:** Cibu Johny (Google)
- **Vice-Chairs:** Lawrence Wolf-Sonkin (Google), and Karan Miśra (Apple)
- **Active Stakeholder Engagement:** Broad cross-industry and institutional representation including Google, Apple, Microsoft, Typotheque, Adobe, Code Hive, and the Government of India.
- **Indian Government Liaison:** Full legal and Contributor License Agreement (CLA) clearance achieved for key C-DAC / MeitY representatives (Neha, Vainateya, Tarun, and Vijay).

3. Technical Architecture & Three-Tier Functional Framework

The working group has been discussing a declarative LDML XML schema structure with rules restricted to a portable regular expression subset ensuring 100% loss-free compilation into both ICU Transliterators and Pynini Finite State Transducers (FSTs).

To resolve operational ambiguities across software stacks, SEQ has articulated three explicit functional tiers for orthographic processing:

- **1. Well-Formedness Validation [Primary Priority]:** Context-dependent structural assertions to detect and block ill-formed, malicious, or unrenderable character sequences at the point of input or application validation.
- **2. Display Normalization:** Context-aware, display-intended rewriting that fixes malformed or legacy sequences (e.g., reordering out-of-order matras or converting legacy sequence fragments) while strictly preserving visual appearance.
- **3. Reading Normalization (Searching & Indexing):** Language-dependent orthographic unifications for phonetically identical spelling variants (e.g., Hindi vs. Nepali nasalization, Amharic homophones) to support search engines, corpus indexing, and AI/ML model training without altering source display data.

4. Key Technical Consensus & Layer Boundaries

During recent bi-weekly technical sessions, SEQ seems to be aligning on critical architectural boundaries:

- **Script-Wide (und) vs. Language Locales:** Universal script-level validation grammars (e.g., und_Deva.xml) act as a superset to accommodate untagged text and creative orthographies across operating systems, while language-specific tailorings (e.g., hi_Deva.xml vs. ne_Deva.xml) reside in locale files.
- **Font Rendering Engine Boundaries:** Font shaping engines (HarfBuzz, CoreText, DirectWrite) should NOT enforce strict language-specific orthography via dotted circles. Inserting dotted circles for non-standard language spellings causes major text corruption across historical web archives (such as pre-2010 Marathi news archives). Dotted circles are reserved strictly for script-level spoofing or unrenderable structural failures or to display incomplete parts of an akshara.
- **Layer Responsibilities:** Language-specific orthographic typing constraints must be enforced at the Keyboard / Input Method layer, application spellcheckers, or editor validation APIs rather than font rendering engines.
- **Unicode NFC Stability:** Recognizing that Unicode Normalization Form C (NFC) cannot be modified due to IETF/IDNA stability policies, the Normalization forms defined by this working group will be an updateable equivalence layer that operates outside of NFC without compromising core Unicode normalization stability.

5. External Standards Coordination & Ongoing Work

SEQ is actively reviewing and synthesizing external standards to ensure interoperability:

- **MeitY / C-DAC 'AKSHAR' Standards:** Conducted detailed technical reviews of C-DAC's draft national Hindi composition rules (Doc No: LITD 20 26504 / BIS) to map ABNF syllabic grammars into CLDR WellFormedness assertions.
- **ICANN Label Generation Rules (RFC 7940):** Adapting ICANN's script-level LGR consensus grammars and joiner preservation contexts (ZWJ/ZWNJ) for general-purpose text.
- **CLDR Keyboard Reordering (UTS #35 Part 7):** Integrating canonical element reordering passes into Visual Normalization to simplify combinatorial transformation rules.