Proposal: Joint working group for orthographic sequences

Peter Constable, Unicode Technical Coordination Group
January 16, 2026

The following is a proposed mandate for a new working group, a *joint* working group under CLDR-TC and also UTC. The instigation for this proposed JWG arose from a request by MeitY during UTC #185 for Unicode projects described below.

This document is for consideration by CLDR-TC and UTC, and the following UTC action is proposed:

> [186-Cx] Consensus: UTC supports the formation of a limited duration working group, jointly with CLDR-TC, as described in L2/26-045.

Joint Working Group: Orthographic Sequences

This ad-hoc joint working group is created to explore development of language-specific specifications of orthographically-valid character sequences.

Status

This WG is initially formed on a temporary basis as an ad hoc WG; depending on early findings, it could be reformalized as a standing working group. The scope of work has relevance to both CLDR TC and UTC (see below), hence it is a *joint* working group, with participation from both TCs.

If later formalized as a standing working group, it is anticipated that CLDR TC would be the primary owning TC, with UTC as a partner.

Scope

The WG will explore development of technical documents and machine-readable data that describe valid orthographic character sequences for particular languages, potentially leading to development and publication of such documents. These documents could potentially be published as Technical Reports or as Technical Standards.

It is expected that work would focus primarily on languages of South and Southeast Asia written with alpha-syllabic scripts that have complex grapheme cluster behaviours. Typically, these would be scripts historically derived from Brahmi (sometimes referred to as “Indic” scripts). Work could also encompass scripts from other regions, such as Arabic or Hebrew.

Description of valid character sequences for a language could have some resemblance to line-breaking rules in UAX #14 or to grapheme cluster rules in UAX #29, or could be expressed as regular expressions, but they would be tailored to individual languages. Because they would be language-specific, the nature of such descriptions would resemble what is typically developed within CLDR TC. The work could also lead to findings that might apply to a script, independent of language, in which case, that would lead to recommendations for consideration by UTC. For these reasons, this work has relevance to both CLDR TC and UTC.

Business case / motivations

Scripts that involve complex grapheme cluster behaviours have proven to be challenging to support in applications. Significant complexity can exist for implementation of several different types of processing, such as text display (shaping, fonts), input methods or line breaking.

As a result, different implementations are not always interoperable—that is, Unicode-encoded text that “works” for one implementation does not always “work” for other implementations. This has sometimes hindered adoption of Unicode encoding, and continues to create challenges for end users whose languages are written with these scripts

The reason for a language-specific focus rather than focusing on a script independent of language is that attempting to write a generic specification for a script can be far more complicated, particularly when a script is used for many different languages, or has been used over long periods of time. By focusing first on the specific languages (or, rather, *writing system* or *orthography*), an overly-complicated challenge can be made feasible. It may later be possible to produce “script-based” data that encompasses the languages written with that script, for applications where the language is not known.

> Note: The direct impetus that led to forming this WG was a request from MeitY, presented at UTC #185, for Unicode to publish specifications for valid orthographic sequences for Hindi and other languages of India. Their need was arising from on-going problems with inconsistent application behaviours and content using inconsistent encoded representations for text.