



# Transliteration Guide for Members of the DHARMA Project

Dániel Balogh, Arlo Griffiths, Axelle Janiak

## ► To cite this version:

Dániel Balogh, Arlo Griffiths, Axelle Janiak. Transliteration Guide for Members of the DHARMA Project. 2019. halshs-02272407v1

**HAL Id: halshs-02272407**

**<https://shs.hal.science/halshs-02272407v1>**

Preprint submitted on 27 Aug 2019 (v1), last revised 2 Jul 2020 (v3)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

# **Transliteration Guide**

**for members of the**



Release Version 1, 2019-08-27

Dániel Balogh, Arlo Griffiths & Axelle Janiak

# Contents

<b>1. Introduction .....</b>	<b>3</b>
1.1. Version Info	3
1.2. Coverage	3
1.3. Separation of Transliteration and Encoding	3
1.4. Symbols	4
1.5. Terms and Definitions	4
1.5.1. Script and its elements	4
1.5.2. Script conversion	7
<b>2. General Principles .....</b>	<b>8</b>
2.1. Character Set and Input Method	8
2.2. Strict and Loose Transliteration	9
2.2.1. Strict transliteration	9
2.2.2. Loose transliteration	10
2.3. Transliteration Scheme	10
2.4. Case Sensitivity	11
2.5. Disambiguation	12
2.6. Editorial Additions for Text Analysis	12
2.6.1. Editorial spaces for word segmentation	13
2.6.2. Editorial hyphenation	14
2.6.3. Representation of <i>avagraha</i>	14
2.6.4. Representation of elided overshoot final <i>u</i> in Tamil	15
<b>3. Alphabetic Characters .....</b>	<b>16</b>
3.1. Some Special Characters	16
3.2. Long and Short e and o	17
3.3. Special Glyph Forms and Compositions	17
3.3.1. Final consonants as special simplex characters	17
3.3.2. Final consonants as complex characters involving a zero vowel marker	18
3.3.3. Independent vowels as special simplex characters	18
3.3.4. Independent vowels as complex characters involving a “vowel support”	19
3.3.5. Multiple and repurposed vowel markers	20
3.3.6. Short vowel written where a corresponding long vowel is expected	20
3.3.7. Superscript <i>r</i> marker versus regular <i>r</i> in conjuncts	20
3.3.8. Reading order of superscript <i>r</i> marker	21
3.3.9. Other unusually composed conjunct glyphs	21
3.3.10. Characters with alternative or optional phonemic values	21
<b>4. Non-alphabetic Characters .....</b>	<b>23</b>
4.1. Numerals	23
4.2. Punctuation	23
4.3. Space Filler Signs	24
4.4. Other Symbols	25
4.5. Space	25
<b>References.....</b>	<b>26</b>

# 1. Introduction

## 1.1. Version Info

- 2019-08-02 Beta version for dissemination to a small group of selected readers for opinions
- 2019-08-22 Release Candidate version for internal discussion
- 2019-08-27 Release Version 1 for use and feedback by all project participants

## 1.2. Coverage

This Guide is essentially intended to cover the scripts relevant to the languages with which the DHARMA project is concerned, i.e., in alphabetical order (omitting the adjective “Old” relevant in most cases): Balinese, Cam, Javanese, Kannada, Khmer, Malay, Prakrit, Sanskrit, Sundanese, Tamil, Telugu. However, the recommendations we give here are certainly intended to be compatible with and extensible to other languages and scripts. We request from colleagues reading and using this text to draw our attention to phenomena in the covered languages/scripts that we have so far failed to address, and to give suggestions on how they might be integrated, as well as to phenomena in languages/scripts so far not covered that may cause issues of compatibility.

The contents of this Guide are primarily applicable to digital editions of epigraphic texts, which must follow these instructions rigorously. We do however hope (and, to some degree, expect) that project members will use the same transliteration method, as far as applicable, in their print publications and other work. Section 2.2 gives some further pointers on what features of the transliteration system can be ignored outside diplomatic editions.

## 1.3. Separation of Transliteration and Encoding

When digitally representing the text of inscriptions (and manuscripts) for preservation and for computer-aided research, we strive to keep recorded content (i.e. what text is written on a certain support) separate, or at least separable, from our annotations *describing* various aspects of that content (for instance how it is written and laid out, how clearly it is readable, or what sort of information it carries). Content is transliterated according to the methods covered in this Guide, while descriptive annotation is added in the form of EpiDoc markup as detailed in the Encoding Guide.<sup>1</sup> The same descriptive annotation also plays a role in determining how our text will be ultimately presented to users on screen and in print, but this is yet another separate concern and will not be addressed here.

Ideally, therefore, no issues that pertain to the description of the physical manifestation of a text should be recorded in the transliterated text itself; and likewise, no issues that pertain to the text content should be omitted from the transliterated text and recorded only in markup. In practice, there are a number of borderline cases that could arguably belong to either of these domains. Given that we are primarily concerned with the faithful documentation of epigraphic texts, some of these issues (such as the use of

---

<sup>1</sup> We follow the TEI Guidelines in using the terms ‘markup’ and ‘encoding’ as interchangeable synonyms.

dedicated signs for independent vowels and final consonants) are addressed at the level of transliteration, while others (such as the possibility of interpreting an ambiguous glyph as either of two or more characters) are dealt with in markup. There is inevitably a certain degree of fuzziness and permeability at the boundary between these domains. Some of the phenomena we cover in transliteration (because we feel that this makes the encoders' job easier) will be universally and automatically converted to markup, and some others may at a later time be likewise converted.

It should be apparent from this that transliteration and markup go hand in hand. We hope that everyone involved in digitising texts will acquire a working familiarity with both Guides, and that even those who will not be creating fully marked-up EpiDoc editions will be willing and able to add snippets of markup to their texts to cover phenomena that cannot be handled through transliteration alone. Cross-references between the Guides should help you find the correct way to deal with each case. If, however, you are absolutely unable to use markup, yet you encounter a phenomenon that the present Guide tells you to handle via markup (for example to record the visual appearance of a symbol or to describe the extent of a blank space in your inscription), then please keep clear notes of the precise location and nature of such problems and include these in the same file as your electronic text or in a separate file kept with your text file and easily recognisable by the filename<sup>2</sup> as belonging with it, so that the person who will later add markup to your electronic text can incorporate the required details in the markup.

## 1.4. Symbols

Partly for use in this guide, and partly as a reminder of the scholarly conventions that we recommend DHARMA team members adopt on the (probably rare) occasions that this will be useful or necessary, we define the use of the following brackets in the following functions:

- <...> graphemic transliteration
- /.../ phonological transcription
- [...] phonetic transcription

See §1.5 below on the concepts of grapheme, transliteration and phonological transcription. We presume team members will rarely have need to offer phonetic transcription, but include the square brackets (which in other contexts may bear other meanings) for completeness. We presume all team members are familiar with the distinction between phonology and phonetics, or if not have the ability to look it up on Wikipedia.

## 1.5. Terms and Definitions

### 1.5.1. Script and its elements

- a **script** may be defined as “a set of conventional graphic signs designed to give visual representation to the elements of a writing system” (Wellisch 1978, 15)
- here, a **graphic sign** is defined as “any conventional mark by which a human being intends to affect the state or behavior of other human beings” (ibid. 10)

---

<sup>2</sup> Good practice recommendations/rules for naming files will be distributed in a separate document.

- and a **writing system** is defined as “a system of rules governing the recording of words and sentences of a language by means of conventional graphic signs” (ibid. 13)
- in the usage of this Guide,
  - **Latin script** refers to the family of fully alphabetic scripts used for writing most European and many other languages
    - the term Roman script is sometimes used in an equivalent sense, but we prefer to designate it as Latin here because Unicode and ISO do so, and because Roman is used in typography to designate a specific set of typefaces within the Latin script
  - **Indic script** refers to the family of alpha-syllabic scripts derived from the Brāhmī script and used for writing most historic South and Southeast Asian languages
- the term **character** may be defined in several ways
  - according to Wellisch (1978, 16), “A **character** is an element of a script, representing a phoneme, syllable, word, or prosodic feature of a language by means of graphic signs.”
  - for our purposes we prefer to emphasise, with Ollett and Taylor (2019), that a character is “an element of the writing system that can be used independently according to the logic of that writing system”
  - thus, Latin letters such as *a*, *b*, *c* are each one character, and one character represents no more than one phoneme
    - some phonemes are represented in some writing systems by a combination of several characters, e.g.
      - English *th* (representing either the voiced dental fricative /ð/ as in ‘this,’ or the voiceless dental fricative /θ/ as in ‘thing’)
      - ISO15919-transliterated Indic *th* (representing the aspirated voiceless dental plosive /tʰ/ as in *ratha*)
  - such combinations are technically called **polygraphs** or, when exactly two characters are involved, **digraphs**
- however, in an Indic writing system, one *akṣara* is one character
  - regardless of how many phonemes it represents and how many visually and semantically distinguishable parts it consists of
    - e.g. Devanagari उ, क्, क, कि and ङ् are each one character
    - while none of the elements corresponding to the transliterated characters *r*, *d*, *dh* and *e* in the *akṣara* ङ् are themselves characters (we refer to these as components, see below)
    - to reduce ambiguity, characters such as उ and क may be called **simplex characters**, while characters such as कि and ङ् may be called **complex characters** (and note that characters such as क् could arguably belong to either of these classes)
  - strictly speaking, *anusvāra* and *visarga* are not characters by this definition
    - however, we do not foresee a need to classify them rigorously, and believe that in some circumstances it may be more productive and intuitively correct to think of these signs (especially *visarga*) as characters
- some characters (in any writing system) have a semantic value that does not correspond directly to any phonemes, e.g.
  - numeral signs are definitely characters

- punctuation signs and other symbols used in written text are arguably characters, and we prefer to include them in the scope of the term
- to reduce ambiguity, the terms **alphabetic character** and **non-alphabetic character** may be used to distinguish between these subsets
- a character defined as above is essentially equivalent to a **grapheme**, often defined as “the smallest functional unit of writing on whatever structural level of language the writing system operates” (Coulmas 2006, s.v.)<sup>3</sup>
- in information and computer science, a **Unicode character** is an abstract element of the script, defined as a “member of a set of elements used for the organization, control, or representation of textual data” (ISO/IEC 10646:2017(E), 2)
- this technical definition is not something we need to use regularly, but it is good to be aware that this definition of a character includes:
  - entities with a visual counterpart (graphic characters) that *represent* phonemes or other information (e.g. punctuation)
  - thus, in this sense of character, the *akṣara* कि = कि = की *ki* consists of two characters, the abstract *k* and the abstract *i*
  - as well as functional characters that do not necessarily have a visual counterpart and exercise *organization* and *control* over graphic characters; for instance in Indic scripts
    - conjunct consonants such as Devanagari क्क involve a non-graphic *virāma* character whose function is to tell the computer that the graphic characters are to form a conjunct (ligature)
    - unusually formed conjuncts such as Devanagari द्म include, in addition, a control character called a zero-width non-joiner to tell the computer that this particular *virāma* should not form a conjunct (the expected Devanagari द्म), but manifest as a visible zero vowel marker
- a **glyph** is a concrete graphical representation of any particular character
  - thus the Indic character *ma* may be represented by the glyphs म, म, म, म, म etc.
  - Unicode parlance prefers to use the term **graphic symbol**, defined as the “visual representation of a graphic character or of a composite sequence” (ISO/IEC 10646:2017(E), 5)
  - another roughly synonymous term is **graphic sign**, defined as “any conventional mark by which a human being intends to affect the state or behavior of other human beings” (Wellisch 1978, 10)
  - yet another quasi-synonym is **graph**, defined as “The smallest formal unit of written language on the level of handwriting or print” (Coulmas 2006, s.v.)
  - visually different glyphs representing the same character within a writing system are known as **allographs**
    - e.g. in the Latin script, the glyphs ‘a’ and ‘a’ are allographs (and, for most practical purposes in most languages, a and A are likewise allographs)

---

<sup>3</sup> The term ‘grapheme’ is sometimes defined differently, so that polygraphs are considered to be a single grapheme; this definition does not concern us here.

- to refer to parts of complex Indic characters that are visually distinct and have a semantic value of their own, we use (and encourage the use of) the term **component**; thus,
  - **character components** are elements such as those representing the phonemes *r*, *d*, *dh* and *e* in the Indic character *rddhe*, as well as the zero vowel marker in the Indic character *k* composed with an explicit vowel killer
  - while **glyph components** are particular realisations of character components in any specific script, such as the stroke combinations corresponding to the transliterated characters *r*, *d*, *dh* and *e* in Devanagari र्द्धे, or those representing *ka* and the zero vowel marker in Devanagari क्
  - when no distinction between character and glyph is required, “component” may be used on its own to refer to these entities
  - components which can never occur independently, but which can occur in combination with various other components, may be specifically called **markers** (with Ollett and Taylor 2019)
    - in Indic scripts these include in particular dependent vowel markers and zero vowel markers, but some other signs, such as the *upadhmānīya* and *jihvāmūlīya*, the *repha*, and arguably also the *anusvāra* and *visarga*, may also be included in the scope of this term
  - note that the term “component” is sometimes (e.g. Brookes et al. 2015, 34) also used to refer to distinctive subunits of non-complex characters, i.e. to elements without phonemic correspondence
    - although it is not relevant to this guide, we recommend avoiding the word “component” in this sense and instead encourage the use of **stroke** to refer e.g. in palaeographic descriptions to the visual elements that make up a character and to their graphic manifestations that make up a particular glyph
    - we also encourage the use of biological and architectural analogues to describe particular strokes, e.g. arm, leg, wing, tail, stem, lobe, arch, base, etc.

### 1.5.2. Script conversion

- for the conversion of one script to another, the words ‘transliteration’ and ‘transcription’ are often used interchangeably in non-specialist parlance, but they have more restricted, and distinct, meanings in the usage we encourage
- **transcription** is “when the **phonemes** of a source language written in a dissimilar script (or not written at all) are represented more or less faithfully by the characters (letters and other graphic signs) of a dominant script” (Wellisch 1978, 18, emphasis added)
- **transliteration** is “when the **graphemes** of a source script are converted into graphemes of a target script without any regard to pronunciation and also, at least in the strictest sense, without either adding or deleting any graphemes that are not present in the source script” (ibid.)
- by the same author’s definition, **Romanisation** is “used as a neutral term to denote both methods of script conversion ... into the Roman script” (ibid., 19)



## 2. General Principles

### 2.1. Character Set and Input Method

- always use the Unicode code table (<https://www.unicode.org/standard/standard.html>),
  - never a custom/legacy encoding (i.e. one that turns into gobbledygook if you change the font to a Unicode font for the same script)
- wherever available, type using Unicode precomposed characters
  - e.g. for *ā* use the Unicode character U+0101 Latin Small Letter A With Macron, not a combination of *a* (U+0061 Latin Small Letter A) and *ˉ* (Unicode 0304 Combining Macron)
- the notation U+#### means a Unicode character identified by the four-digit hexadecimal code ####
- the font you use in your texts is irrelevant so long as it is Unicode-compliant
  - freely available fonts supporting all or nearly all of the special characters we require include:
    - Gentium, <https://software.sil.org/gentium/> and several other fonts by SIL
    - Google’s Noto Serif (and Sans Serif) fonts, <https://www.google.com/get/noto/>
    - several of the fonts shipped with Windows 10, e.g. Times New Roman, Tahoma, Calibri
    - several of the fonts shipped with Mac OS, e.g. Times New Roman, Arial, Calibri
- you probably already have a favourite keyboard layout to access the special characters you need in your work
- if not, and you are a Mac user, you may want to try the out-of-the-box layouts Easy Unicode or ABC Extended (formerly US Extended)
  - there is, unfortunately, no out-of-the-box general solution for a Windows platform, but you may be able to use and/or adapt John Smith’s keyboard layout and Word macros, available at <http://bombay.indology.info/software/fonts/induni/index.html>
- if you can access most of the characters you need via your keyboard, but there are a few that you need occasionally and cannot access, one of the following solutions may help:
  - assign a shortcut key or sequence to the inaccessible characters in your editing software
  - copy and paste the inaccessible characters from this guide each time you need one of them (or save a separate document with those characters, keep it at your fingertips, and copy-paste from that)
  - insert them from a table of available characters
    - in MS Office, use Insert Symbol
    - on Mac OS (systemwide), use the Character Table
  - use Unicode codes to enter special characters
    - in MS Office you can type the code, then press ALT + x to convert the code into the corresponding character
    - you can omit prefix U+, but using it will make certain the software recognises where the code begins, so the last characters you typed before the code will not interfere with what you want to produce

- on Mac OS (systemwide), you need to enable Unicode Hex Input in Language Preferences
- once you have done this, whenever you switch to this keyboard layout, you can press and hold Option while you type the character code (without the prefix U+) then release Option
- if all else fails, then consistently type one and the same particular alternative character throughout your corpus (e.g. *r* instead of *ṛ* or *š* instead of *ś*, etc.)
- do not use that particular sign for any other purpose than representing the character you cannot type
- make clear note of what you are doing, so your custom character can then be auto-converted to the correct one
- please note that detailed technical instructions on installing and using keyboard layouts or assigning shortcut keys are beyond the scope of this guide

## 2.2. Strict and Loose Transliteration

- as Wellisch (1978, 314) points out, “there is no single ‘scientific’ system whose principles can be applied uniformly to all scripts and for all purposes ... Rather, there is a plurality of more or less justified but mutually incompatible requirements ... so that a choice must be made among those requirements that are *optimally* needed to make the system work for a particular purpose or task.” (emphasis original)
- in addition to the notion that no single Romanisation system can be applied in a practicable manner to all known scripts and languages, this implies that for actual Romanisation systems to work, they need to find an optimal point on the continuum between ideal transliteration and ideal transcription

### 2.2.1. Strict transliteration

- as our aim in epigraphic editions is to faithfully reflect the graphemes (characters) of the original script, the Romanisation system prescribed in this guide is very close to the transliteration end of the spectrum, and therefore we refer to it as “strict transliteration”
- the same aim, and thus the same Romanisation system, applies to diplomatic editions of single manuscripts, and for readings of specific manuscripts cited in the apparatus of a critical edition
- when strict transliteration is called for, fully prioritise transliteration over transcription except in specific cases where this guide explicitly calls for the use of Romanisation more akin to transcription (such as §3.2 and 3.3.8)
- this applies even when you are certain that a specific *akṣara* composition was pronounced in a way unlike that dictated by the inherent logic of the script; see §3.3.10 for some specific examples

### 2.2.2. Loose transliteration

- however, in other contexts, a method of Romanisation closer to the transcription end of the spectrum (which we term “loose transliteration”<sup>4</sup>) is acceptable and recommended, primarily in the following situations
  - in the text of a critical edition of multiple manuscripts, especially where there is a mismatch between script and language (e.g. over- or underspecificity of the script for the phonemic system)
  - when citing isolated words, names or passages from an inscription in a modern-language discussion
- the Romanisation scheme you use in such contexts is to be guided by your preference and the conventions of your field, and may differ from strict transliteration for instance in
  - avoiding specific representation of certain features of the writing system such as initial vowels, final consonants or the particular way a ligature is composed
  - normalisation by reducing graphic diversity in a writing system that has more characters than the phonology of the language needs, i.e. merging alternative notations of a single phoneme into one sign (that must also be a member of the larger subset of signs used in ISO-15919), e.g.
    - substitution of the class nasal for *anusvāra* or vice versa
    - Old Javanese *vvaṃ/vvaṇ* merged into *vvaṇ* (phonologically /wwaṇ/), *luraḥ/lurah* merged into *lurah* (phonologically /lurah/)
  - disambiguation where a language uses one feature of a writing system to represent more than one phonological feature, e.g.
    - Old Sundanese *sastra* and *ku nu reya* (even when written as *saṣṭā* and *ku nu rye* as in the examples under §3.3.10)
  - normalisation of orthography, e.g.
    - elimination of consonants doubled in conjunction with *r* in Sanskrit
    - distinction of *e/ē* and *o/ō* even if not present in the original writing

### 2.3. Transliteration Scheme

- in general, use the **ISO-15919** transliteration system for all languages written in an Indic script
  - the standard, published as a pamphlet, is accessible in the form of a pdf file in the PDF Library on Sharedocs<sup>5</sup>
  - Wikipedia ([https://en.wikipedia.org/wiki/ISO\\_15919](https://en.wikipedia.org/wiki/ISO_15919)) summarises the essential features
- if you are used to IAST, this means paying attention to using *ṁ*, *ṛ*, *ṙ* and *ḷ* rather than *m̐*, *r̐*, *ṛ̐* and *ḷ̐*

---

<sup>4</sup> Loose transliteration is a generic term that allows for the possibility that certain non-phonological features are retained in Romanisation while others are transcribed phonologically. In many practical applications, our “loose transliteration” can be justifiably called (and has been called) simply “transcription.”

<sup>5</sup> <https://sharedocs.huma-num.fr/wl/?id=3y8R1K48Budcn6HjZdWcQV88xooR66kv>

- if you are used to the scheme of the Madras *Tamil Lexicon*, rest assured that it is identical to ISO-15919 on all fundamental points
- for Kannada, we will on the whole align with the guidelines on Kannada transliteration drafted by Andrew Ollett and Sarah Pierce Taylor (2019), but differ in some details such as the representation of initial vowels and final consonants

## 2.4. Case Sensitivity

- in general principle (as per ISO-15919 Rule 8.1.1), our transliteration is case insensitive
- however, we propose to supplement ISO-15919 and – in strict transliteration – use certain uppercase letters to distinguish final consonant characters (see §3.3.1) and independent vowel characters (see §3.3.3) of the original script
  - this distinction may in some cases be redundant, but it can be particularly useful
    - where the original inscription could have used a regular *akṣara* (e.g. कृतमेत्तत्) but chooses instead to use a final consonant followed by an initial vowel to represent a pause for semantic or metrical segmentation (e.g. कृतम्एत्तत्)
    - where part of the original is not legible, and a lacuna is preceded by a consonant or followed by a vowel, this notation makes it clear to the reader whether
      - the preceding consonant is a final form or a partial *akṣara* (with an illegible vowel component)
      - the following vowel is an independent form or a partial *akṣara* (with an illegible consonant component)
    - it also eliminates the need for a special disambiguation character (for which see §2.5) to distinguish vowel hiatus involving an *a* followed by an *i* or a *u* from the diphthongs *ai* and *au*
  - therefore, in strict transliteration use uppercase only for these special features, and use **only lowercase** letters everywhere else, including
    - the initials of proper names, and
    - the beginnings of paragraphs, sentences, metrical units, etc.
- some of us have already adopted the system of using a ° character before transliterated vowels and after transliterated consonants to denote special initial and final forms
  - at least for the time being, we retain that notation as an option; however, the system using uppercase Latin letters is to be preferred in future, including printed publications
- additional considerations in favour of adopting this particular notation include the following:
  - while an additional character in transliteration corresponds well to a zero vowel marker (“vowel killer”) attached to regular consonants, there is no such equivalence in the case of special character forms, which are more rigorously transliterated using a single Latin character
  - if we postulate that the ideal type of an *akṣara* is a combination of consonant(s) + vowel, then our rules mean using lowercase for normal *akṣaras*, while uppercase is used for vowels which are special by lacking a consonant, and for consonants which are special by lacking a vowel (and an explicit *virāma*)

- uppercase letters are pre-existing special forms of Latin letters, which are easy to type on all keyboards and can be readily co-opted for our purposes as case is not used for any other purpose in ISO-15919
- search algorithms will find text written with special forms without requiring special provisions (e.g. a search for *tad eva* will also find *taD Eva*, but fail to find *tad° °eva*), whereas if only a specific orthography is desired, a case sensitive search will find only the desired string
- using uppercase letters for special forms allows us to keep free the sign ° for the conventional use as a marker of truncation (e.g. when cutting words to be cited in a critical apparatus)
- whereas our use of the middle dot · to transliterate explicit “vowel killers” (see §3.3.2) allows us to add markup to such markers as separate from the consonants to which they are attached, transliterations to the model of °C and V° are digraphs which cannot be broken up by markup in spite of their appearance

## 2.5. Disambiguation

- since our transliteration standard includes digraphs (e.g. *kh*, *au*), it occasionally happens that such digraphs must be distinguished from juxtapositions of the characters transliterated by the individual components of a digraph (e.g. *k* followed by *h*; *a* followed by *u*)
- in accordance with ISO-15919 (Rule 8.1.15), we use the colon (:) as a disambiguation sign where our transliteration would be ambiguous without such a sign
- however, unlike the standard, our transliteration system provides a way of distinguishing independent vowel signs of the original script from vowel markers (see §§3.3.3 and —), and therefore we only need a disambiguation sign
  - to distinguish consonant + *h* combinations from aspirated consonants
    - e.g. *p:h* for *p* conjoined to *h* to distinguish it from the aspirate *ph*
- and, again unlike the standard, we extend the use of the disambiguation sign to facilitate the recording of some particular features of the writing systems we are concerned with, namely:
  - the presence of multiple vowel markers within one *akṣara* (for which the standard does not provide; see §3.3.5)
  - conjuncts composed in a way other than the default for a given language and writing system (see §3.3.6 and §3.3.9)

## 2.6. Editorial Additions for Text Analysis

- as a general rule, do not add anything to your transliteration that is not already present in the original text
- the way to handle editorial additions and alterations goes through markup; see the relevant chapters of the Encoding Guide

- however, this general rule comes with the following exceptions, which serve as a low-level editorial markup to facilitate the analysis and segmentation of a text for human readers, and which will (or may at a later stage) be converted to machine-readable XML markup

### 2.6.1. Editorial spaces for word segmentation

- **words** should be **separated** from one another with a space wherever Romanised transliteration allows, notwithstanding that the original inscription or a published edition, whether in Indic or Latin script, does not do so
- emphatically, **do add spaces**
  - where the end of one word and the beginning of the next word constitute a single *akṣara* in the original
    - even if such an *akṣara* involves a sandhi modification, e.g.
      - Sanskrit *tad dhi* (for *tat + hi* – space goes between *d* and *dh*)
      - Sanskrit *gacchaty eva* (space goes after the *y*)
      - Sanskrit *putrāñl lakṣmīḥ* (space goes between the two *l*-s)
      - Old Javanese *tann inaku* (space goes between the *-nn* and the *i*-)
      - Tamil *arit' enru* (for *aritu + enru*; see also §2.6.4 for elision of overshort *u* in Tamil)
    - including non-standard sandhi and orthographic practice, e.g.
      - nasals used where standard orthography would employ an *anusvāra*, e.g. Sanskrit *uktañ ca* or *śaraṇaṇ gataḥ*
      - Sanskrit *dīnārair ddaśabhiḥ*
      - Old Javanese *darpaṇa ryy avakta*
  - before an *avagraha*, unless it occurs within a compound (so *'bhūt* instead of *so'bhūt*, but e.g. *saro'nte*)
  - in close-knit structures such as *atha vā*, *kiṁ ca* and *kiṁ tu* (even if spelt *kiñ ca* and *kin tu*), *tad yathā*; including grammaticalised structures such as
    - Sanskrit periphrastic perfects, e.g. *varayāṁ cakāra* (especially since other words may intrude inside such a construction, e.g. *saṁraṁjayāṁ ca prakṛtīr babhūva*)
    - Sanskrit formations with *-sāt* prefixed to a verb such as *brāhmaṇasād gatāḥ*
    - Sanskrit prepositions such as *ā samudrāt*, *anu gaṅgām*
    - note that some editors prefer to hyphenate certain collocations; please avoid this
- **do not**, however, separate
  - successive words where the final vowel of the first and the initial vowel of the second are fused in vowel sandhi, e.g.
    - *tasyāyam* stays as is, though *so yam* is separated
    - *gacchatīva* stays as is, though *gacchaty eva* is separated
  - Tamil enclitic particles (e.g. *ē*, *ō*) and forms of the verb *āku-tal* (e.g. *āṇa*, *āy*, *āka*) when used adverbially
  - for Sanskrit close-knit structures borrowed into other languages, follow the spelling with or without space (generally the latter) of the relevant dictionaries, if there are any
    - e.g. Old Javanese *kimuta*, Old Cam *kintu*

### 2.6.2. Editorial hyphenation

- editorial hyphens may be optionally added for the following purposes
  - **segmentation of compounds** in Sanskrit and other compound-heavy languages
    - such segmentation need not be exhaustive; feel free to hyphenate only long or difficult compounds and leave others intact
  - **sandhi analysis** when hyphens are conventionally used for this purpose in your field
    - specifically, epenthesis in Tamil may be indicated by joining the added letter to the preceding word with a hyphen (see examples below)<sup>6</sup>
- as with editorial spacing, feel free to add hyphens between transliterated characters that belong to a single *akṣara* of the original, but do not use a hyphen at points where the final and initial vowels of two successive compound members are fused in sandhi
- some examples of Tamil hyphenation:
  - *tiru-makal* (திருமகல் *tiru+makal*)
  - *koṇṭ-āṭu* (கொண்டாடு *koṇṭu+āṭu*)
  - *I-p-peruṇ-kōyil* (இப்பெருங்கோயில் *i+perum+kōyil*)
  - *tiru-mēṇi-y āṭa* (திருமேனியாட *tiru+mēṇi āṭa*)
- **do not use hyphens** for any other purpose, e.g. to show that a word has been broken into two parts by the end of an inscribed line
  - this should be noted in markup (see §XXX of the Encoding Guide)
  - if you are not adding any markup, please use the character ¬ (U+00AC Not Sign; do not use a hyphen), which will be auto-converted into the proper markup
- if you use hyphens for editorial compound analysis, and
  - **a physical line break** coincides with such a hyphen, then
    - first encode the physical line break as one inside a word (as per the Encoding Guide § or with the shorthand ¬)
    - then put the editorial hyphen at the beginning of the new line
  - **a verse line break** coincides with such a hyphen, then
    - first encode verse line break as one inside a word (as per the Encoding Guide §)
    - then put the editorial hyphen at the beginning of the new line

### 2.6.3. Representation of *avagraha*

- since our inscriptions very rarely use an *avagraha* sign, any and all *avagrahas* in a typed text will be assumed to be supplied by the editor, and markup signifying this (for which see §XXX in the Encoding Guide) will be added automatically
- the supplying of *avagrahas* is optional, but recommended especially in cases where the text would be meaningful (and often contradictory in meaning) both with and without an *avagraha* (e.g. the inscribed sequence *sohataḥ* may stand for *so hataḥ* or *so 'hataḥ*)
- for supplying *avagraha*, use ' (right single quote) or, alternatively, ' (plain apostrophe)
- in the exceptional cases where there is an original *avagraha* in your texts, use '! (right single quote with an exclamation mark) to transliterate it

---

<sup>6</sup> Non-standard Sanskrit sandhi involving an epenthetical *m* may be indicated in the same way, should the need arise, e.g. *mleccha-rājye-m apūjitaḥ*.

- this way, the *avagraha* in question will not be automatically marked up as supplied, and the ! will be removed after marking up all other *avagrahas* as supplied
- note that an apostrophe representing an *avagraha* must never be followed by a space (though it may or may not be preceded by one, see §2.6.1), in order to distinguish it from the apostrophe used to represent elision in Tamil (q.v. §2.6.4)

#### 2.6.4. Representation of elided overshoot final *u* in Tamil

- in the transliteration of Tamil text, use an apostrophe followed by a space to represent the elided overshoot *u* at the end of an independent word, e.g.
  - *arit' enru* (அரிதென்று for *aritu + enru*)
- but do not use an apostrophe for the elided overshoot *u* inside a lexicalised compound, e.g.
  - *koṇṭ-āṭu* (for *koṇṭāṭu*)
- note that an apostrophe used for this purpose must always be followed by a space (and not be preceded by one), in order to distinguish it from the apostrophe used to represent *avagraha* (q.v. §2.6.3)



### 3. Alphabetic Characters

#### 3.1. Some Special Characters

- most of the characters below are covered by ISO-15919, but are specifically mentioned here because their transliteration may not be self-evident to all of us
- ! transliterations not covered by ISO-15919 will be marked in this section by an initial exclamation mark
- ***anunāsika/candrabindu***
  - *m̃* (this character is not available as a precomposed glyph, so it must be composed of a regular *m* and a <sup>˘</sup> sign: U+0310 Combining Candrabindu)
  - use **only** if distinguished in the script from *anusvāra*
    - but, conversely, **always** make the distinction in transliteration if the distinction is made in the original
  - *candrabindu* signs enlarged and embellished for ornamentation do not receive a different treatment in transliteration
  - only add the Candrabindu sign to *m* (i.e. avoid using *tāl lakṣmīm* and write *tāml lakṣmīm* instead)
- ***upadhmāñīya*** (if distinguished in the script from *visarga*)
  - *ḥ* (U+1E2B Latin Small Letter H with Breve Below)
- ***jīhvāmūliya*** (if distinguished in the script from *visarga*)
  - *ḥ* (U+1E96 Latin Small Letter H with Line Below)
- **Tamil āytam**, ூ
  - *ḱ* (U+1E35 Latin Small Letter K with Line Below)
- **retroflex lateral**, Tamil ண Kannada/Telugu ಳ
  - *ḷ* (U+1E37 Latin Small Letter L with Dot Below)
- **alveolar trill/stop**, Tamil ற Kannada/Telugu ೞ
  - *ṛ* (U+1E5F Latin Small Letter R with Line Below)
- **retroflex approximant / frictionless continuant**, Tamil ೞ Kannada/Telugu ೞ
  - *ḷ* (U+1E3B Latin Small Letter L with Line Below)
- ! **Cam anusvāra-candra**
  - *m̃* (this character is not available as a precomposed glyph, so it must be composed of a regular *m* and a <sup>˘</sup> sign, U+0303 “Combining Tilde”)
- ! **Javanese/Balinese special anusvāra** with a small stroke beside it (to indicate pronunciation as /m/), called *ulu ricem* in Balinese
  - *m°* (the regular transliteration of *anusvāra* followed by a degree sign, U+00B0)
- ! **Javanese/Balinese pepet** (expressing the vowel schwa)
  - short, ə (U+0259 Latin Small Letter Schwa)
  - long, ɐ: (with length-mark represented by a colon as per §3.3.5) in strict transliteration
    - ā in loose transliteration (not available as a precomposed character; use U+0259 Latin Small Letter Schwa, followed by U+0304 Combining Macron)
- ! **Khmer (and Mon-Burmese) glottal stop**
  - *q* (the Latin letter q)
  - see also §3.3.3 about the representation of independent vowels

- **! special signs for Mon and Pyu:**
  - barred/dotted variant of *b*
    - ʙ (U+1E05 Latin Small Letter B with Dot Below)
  - *akṣaras* with underdot
    - ṁ (U+1E43 Latin Small Letter M with Dot Below)

### 3.2. Long and Short *e* and *o*

- when transliterating a language that does not make a distinction between long and short *e* and *o*, use these Latin characters without a macron
  - this corresponds to Option 9.1 of the ISO15919 standard, applicable to languages that do not make a distinction between the phonemes *e/ē* and *o/ō*
- however, for Dravidian **languages that distinguish long and short *e* and *o***, you have the option to record that distinction even if it is not present in the script you are working with
  - in this case, transcribe long vowels as *ē/ō* even in strict transliteration
  - subsequently, markup will be automatically added to these, signifying that *e* or *o* was in fact inscribed, but the spelling has been normalised by the editor
    - that is to say, the palaeographically primary generic vowel marker, e.g. that in के *ke*, को *ko*, may represent either a short or a long vowel; when it represents a long vowel, this will be shown as an editorial normalisation, e.g. to के *kē*, को *kō*
  - should your inscription (or manuscript) explicitly distinguish between short and long *e/o*, please contact us to devise a solution for handling this

### 3.3. Special Glyph Forms and Compositions

- ideally, transliteration would not be concerned with what allograph is used in a particular instance to represent a particular grapheme
  - however, we find that it may be important for our research interests to preserve in the transliterated text some alternative ways of representing the same character or character combination
- for this reason, in strict transliteration we shall always make the distinctions set out below
- other potentially interesting allographs, for instance the use of two alternative glyphs within the same inscription for the same simplex character, will need to be described outside your edition of the text (e.g. in the header element `<handDesc>`, for which see the Encoding Guide, §XXX), and will not be directly represented in the transliteration or the markup
  - this is a conscious decision of the authors of this Guide, who consider that we need to impose a limit on the granularity of our representation of potentially interesting phenomena

#### 3.3.1. Final consonants as special simplex characters

- special character forms representing consonants without a vowel (called *halanta* consonants in Sanskrit) shall be transliterated as follows

- type a corresponding uppercase Latin consonant, e.g. *T*
- **or, optionally**, type a ° (degree sign, U+00B0) after the transliterated (lowercase) consonant, e.g. *t°*
- note: the former transliteration is recommended for future use and for printed publications, but at least for the time being we retain the latter as an option; we may also devise a way to distinguish such characters in markup
- these forms are typically a miniature and/or subscript rendering of a simplex consonant *akṣara*
- the criterion by which to distinguish this case from complex characters involving a zero vowel marker (§3.3.2) is the use of a glyph unlike the regular simplex character employed for that consonant with an inherent *a*
- if this criterion is met, then the character in question should be transliterated with an uppercase consonant even if the special form includes a component that may be perceived as a zero vowel marker, e.g.
  - a horizontal dash above a miniature consonant sign in an Indian inscription, which may be viewed as a proto-*virāma*, but which we treat as part of the special consonant form, not as an explicit vowel killer
  - a special vowel killer attached to a special form of *ka* in Old Sundanese, e.g. 𑌕𑌃𑌔𑌃𑌔𑌃 *AnaK rahyīm* (compare the regular vowel killer in 𑌕𑌃𑌔𑌃𑌔𑌃𑌔𑌃 *gadiṁ manik°*)

### 3.3.2. Final consonants as complex characters involving a zero vowel marker

- complex characters involving a regular simplex form and an explicit zero vowel marker (“vowel killer”: *virāma*, *pulli*, *patén/pangkon*, etc.) shall be transliterated as follows
  - type the character · (U+00B7 Middle Dot) after the Latin consonant, e.g. *t·*
    - if you have difficulty typing this sign, optionally use an asterisk \* in its place; this will be replaced later on with the middle dot
- use the same method to represent a **Tamil *pulli*** that is explicitly present in your original (e.g. *t·ta* to transliterate த் த)
- where *pulli* is not present in an inscription but is to be understood implicitly, simply type the transliterated consonant cluster without any additional characters (e.g. *tta* to transliterate த் த understood as த் த)
- we may at a later point decide to automatically convert such transliterations into markup signifying that a *pulli* has been supplied by the editor, but for the time being our default assumption is that any consonant cluster in transliterated Tamil involves an implicit *pulli*
- representing zero vowel markers by a separate character in the transliteration has the added advantage of being able to apply markup to this sign, e.g. to label it as unclear, restored or supplied

### 3.3.3. Independent vowels as special simplex characters

- if the original script employs a distinct character for vowel-only *akṣaras* (initial vowels and vowels in hiatus), these shall be transliterated as follows
  - type a corresponding uppercase Latin consonant, e.g. *A*
    - thus, इति becomes *Iti*, whereas कृतमिति becomes *kṛtam iti*

- for the initial forms of the diphthongs *ai* and *au*, capitalise only the first character of the digraph in your transliteration, i.e. use *Ai* and *Au* (whereas *AI* and *AU* would transliterate अइ and अउ, should these combinations occur)
- **or, optionally**, type a ° (degree sign, U+00B0) before the transliterated (lowercase) vowel, e.g. °a
  - thus, इति becomes °iti, whereas कृतमिति becomes *kṛtam iti*
- note: the former transliteration is recommended for future use and for printed publications, but at least for the time being we retain the latter as an option; we may also devise a way to distinguish such characters in markup
- note on the **disambiguation of vowels in hiatus** from diphthongs:
  - normally, certain vowel sequences need to be distinguished in transliteration from diphthongs represented by the same Latin vowels, e.g. Sanskrit प्रउग and Prakrit चउत्थो and दइआ must be kept distinct in transliteration from प्रौग, चौत्थो and दैआ
  - ISO-15919 uses a colon (see also §2.5) between such vowels in hiatus (thus: *pra:uga*, *ca:uttho*, *da:iā*), but our consistent use of uppercase for independent vowels dispenses with that need in strict transliteration (where प्रउग, चउत्थो and दइआ are represented as *praUga*, *caUttho* and *daIĀ*)
  - however, it is recommended that **in loose transliteration** you follow the established convention of using a diaeresis (pair of dots) above the second vowel, thus प्रउग, चउत्थो and दइआ become *praüga*, *caütttho* and *daïä*

### 3.3.4. Independent vowels as complex characters involving a “vowel support”

- if the original script employs a “vowel support” character with a vowel marker attached to it, then transliterate this with the letter *q* followed by the applicable (lowercase) Latin vowel
- see the table on the right for examples in Balinese
- the character used as a “vowel support” may otherwise represent a glottal stop, be only a zero consonant sign, or represent the independent vowel A
- we find that the *function* of this character component as a vowel support is distinct from and more relevant to research than its *derivation* from a vowel sign and that in its function as a “vowel support,” these characters can behave as regular consonants<sup>7</sup>
- hence, we prefer to transliterate all “vowel supports” with the dedicated character *q*

combination	glyph	phoneme	translit.
A with <i>taling</i>	ꦠꦺꦁ	/e/	<i>qe</i>
A with <i>suku</i>	ꦱꦸ	/u/	<i>qu</i>
A with <i>ulu</i>	ꦸ	/i/	<i>qi</i>
A with <i>taling tedong</i>	ꦠꦺꦁꦠꦺꦢꦺꦁ	/o/	<i>qo</i>

<sup>7</sup> This systemic change is complete in mainland Southeast Asia, where the “vowel support” (e.g. Khmer ឃ) functions fully as a consonant sign, but has been carried through to varying degrees in the case of Java-Bali-Lombok. According to Ida Bagus Komang Sudarma (personal communication, 16 Aug. 2019), in Sasak writing ꦠꦺꦁ can be combined with a *pasangan* consonant, e.g. ꦠꦺꦁꦏꦲ and ꦠꦺꦁꦏꦸ, but cannot itself become a *pasangan*, while in Balinese writing neither possibility exists.

### 3.3.5. Multiple and repurposed vowel markers

- **in general**, if you encounter **two or more vowel markers** attached to a single consonant, transliterate each vowel in the order you deem logical, and type a colon between them to mark them as belonging to the same *akṣara*
- this method is primarily for cases where the scribe erroneously engraved more than one explicit vowel mark, neither of which appears to be deliberately cancelled
  - for the deliberate use of multiple vowel markers to record certain features of a non-Indian language, see the special cases below
  - the deliberate use of multiple vowel markers (e.g. a combination of *u* and *i*) to signify deletion belongs in the domain of markup (see §XXX in the Encoding Guide), not that of transliteration
- note: the colon is redundant, since the fact that the transliterated vowels are lowercase indicates in our system (cf. §2.5) that none of them are independent vowel *akṣaras*
- nonetheless, the colon is useful to highlight the presence of an unusually composed character (see also §3.3.6 and §3.3.9 below)
- **special cases** for specific languages:
  - when an **extra ā marker** (Javanese *tarung*, Balinese *tedong*) appears deliberately in combination with another vowel marker, for instance as length marker in Javanese:
    - type a colon (:) after the short vowel to transliterate the length marker
    - unlike the general case above, do not type an ā after the colon in this case
  - when *tarung* represents a doubling of the consonant component of the *akṣara* to which it is attached, transliterate this by typing a colon after the transliterated consonant
    - e.g. Old Sundanese ꦒꦤꦥꦶꦥꦶꦠꦸ (pronounce *gənəp pipitu* “fully seven”) is to be transliterated as *gnəp:ipitu*
    - in this case too, do not type an ā after the colon
  - when the **vowel markers for u/ū and i** appear deliberately together, for instance to represent a particular phoneme in Khmer (as well as in Burmese and Mon):
    - transliterate the vocalisation as *ui* or *ūi*
    - unlike the general case above, do not use a colon in this case

### 3.3.6. Short vowel written where a corresponding long vowel is expected

- where a short vowel is written in place of an otherwise identical long vowel, you have the option to add a breve to the transliterated short vowel in order to highlight the fact that the short vowel is not an editorial mistake
  - i.e. use *ă*, *ĭ* or *ŭ* when *a*, *i* or *u* is used for expected *ā*, *ī* or *ū*
- this option is especially recommended for Sanskrit loanwords in Javanese and Balinese text, following Damais (1955, 15)
- this notation will be converted to markup involving the tag `<orig>` (for which see Encoding Guide §)

### 3.3.7. Superscript *r* marker versus regular *r* in conjuncts

- when in transliteration an *r* is followed by another consonant (e.g. *rya*), we assume by default that this transliterates an *akṣara* composed of a superscript marker (*repha*, *layar*, *surang*) and a regular consonant glyph

- note: this means that we do not use *ṙ* (or any other dedicated Latin character) to transliterate the *r* marker as distinct from *r*
- however, in some cases conjuncts with an initial *r* will be composed of a regular *r* plus a subscript second consonant
  - in strict transliteration, use a colon (:) after the *r* in such cases, as disambiguation from the default way of composition, e.g.
    - *ry* by default transliterates *y* with a *repha* (רֿ)
    - *r:y* as a special case transliterates *r* with a subscript *y*

### 3.3.8. Reading order of superscript *r* marker

- the representation of the “Indonesian” (versus “Indian”) positioning of the *r* marker is handled via markup
  - thus, even in strict transliteration, transcribe the text in the order the script components were meant to be pronounced, e.g.
    - $\text{ᮊᮥᮒ}$  *samar*; but  $\text{ᮊᮥᮒ}$  *sarva*
  - and see §XXX of the Encoding Guide for the applicable markup

### 3.3.9. Other unusually composed conjunct glyphs




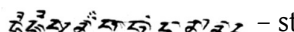
- the above case involving preconsonantal *r* may be generalised to highlight other conjunct formations that deviate from the standard glyph composition for any particular language and writing system, e.g.
  - *kṣa* formed with a superscript *k* attached to a regular *ṣ* (instead of the default regular *k* with a subscript *ṣ*)
  - conjuncts formed with the consonant components arranged horizontally when vertical composition is the norm for the language and writing system
- in such cases, *optionally* use the disambiguation marker : between any two consonants that are conjoined in an unusual way
  - this colon will be ignored by search and processing software, but serve as a marker that something strange is going on in the text here, and may be used as a starting point for future analysis or harvesting of such cases

### 3.3.10. Characters with alternative or optional phonemic values

- some writing systems may use certain glyphs to represent more than one phoneme or sequence of phonemes, or may use a non-alphabetic character in an alphabetic function<sup>8</sup>
  - **in strict transliteration**, always prioritise the primary value of such glyphs
  - **in loose transliteration**, however, it is preferable to transcribe the phonemic value intended in the context

---

<sup>8</sup> It would be possible to view these as different characters that happen to manifest as identical-looking glyphs: this phenomenon is known as synoglyphy and happens e.g. in the case of *I* (uppercase *i*) and *l* (lowercase *L*) in Latin script printed in a sans-serif font. In this case we would transliterate them according to their function in context and disregard their appearance. However, we find that this approach would cover up palaeographically interesting and important features of the original. For a modern Western parallel, the text “going 2 bed” is not entirely equivalent to “going to bed”, and “2” and “to” are not synoglyphic.

- **in an EpiDoc edition**, you may add markup to normalise the transliterated primary value to a transcription of the intended value (see § of the Encoding Guide on editorial normalisation)
- some specific examples:
  - when the glyph  *ṭā* is used in Old Sundanese to represent the phonemes /tra/, transliterate it as *ṭā*, but in loose transliteration transcribe it as *tra*
    - e.g.  – strict: *saṣṭā*; loose: *sastra*
  - when a vowel marker added to a ligature with subscript *y* is intended to be pronounced before the *y* (and after the primary consonant) in Old Sundanese, write the transliterated vowel after the *y* in strict transliteration, but transcribe in the intended order when using loose transliteration
    - e.g.  – strict: *ku nu rye*; loose: *ku nu reya* (“by many [people]”)
  - when the numeral 2 is used in Old Sundanese to represent the phonemes /ro/, transliterate it strictly as 2 (without adding numeral markup as per § of the Encoding Guide), but use *ro* in loose transliteration
    - e.g.  – strict: *di ja2nim vavamñun:an* ; loose: *di jaroniñ vavañunan* (“in the interior of the building”)<sup>9</sup>

---

<sup>9</sup> Note that in the loose transliteration of this example, *m* and *mñ* are both be represented by *ñ* (being the sign selected to represent phoneme /ŋ/). Simultaneously, *n*: (theoretically denoting /nn/) is simplified to *n*, and *mñ* (theoretically /ññ/) is simplified to *ñ*, because consonant gemination is not considered to be a phonemic feature of the language, but rather an orthographic particularity.

## 4. Non-alphabetic Characters

### 4.1. Numerals

- numbers written in **decimal place-value notation** in the original shall be transliterated straightforwardly, e.g. 876
- for numbers recorded in an **additive system**, type a + sign after each transliterated number sequence of two or more “Arabic” numerals that represents a single numeral character in the original
  - this notation will be automatically converted to markup indicating that these Arabic digits transliterate a single original numeral character (see \$XXX in the Encoding Guide)
    - numerals transliterated with a single Arabic digit must not be followed by a + sign, since they are understood by default to represent a single original numeral character
    - arguably, most Indic numerals in the 100s range could be viewed as combinations of several characters rather than as a single character, but we foresee no useful purpose that such a complex distinction could serve and therefore treat all these Indic numerals as single characters (with distinguishable components)
- note that this notation differs from that of older printed publications such as *Epigraphia Indica*, which used a + sign only to indicate actual addition, whereas we use it to mark the end of every sequence of two or more Arabic numerals that transliterate a single numeral character in the original
  - thus, our transliteration is identical to the conventional notation in cases such as
    - 10+2 – representing a character for 10 followed by a character for 2; and
    - 300+2 – representing a character for 300 followed by a character for 2
  - but differs from it in cases such as
    - 10+ (rather than just 10) – representing a character for 10; and
    - 300+50+ (rather than just 300+50) – representing a character for 300 followed by a character for 50
- for numerals represented in Cambodian inscriptions by bars (*daṇḍa*) instead of numeral characters, use the following transliteration method:
  - type as many I (NB: uppercase i, not vertical bar |) characters as there are bars in the original
  - and type a + sign after the last I
    - note that unlike regular numerals, the + sign must be used in this case even after a single I representing the numeral 1
    - this notation will be automatically converted to markup indicating that these characters are not alphabetic and constitute a single meaningful character

### 4.2. Punctuation

- transliterate all **original punctuation** but **do not add punctuation marks** not already present in the original (editorial punctuation may be supplied in markup, see Encoding Guide \$XXX)



- we consider the diversity of punctuation signs used in inscriptions to deserve preservation and investigation, but acknowledge the challenges of reproducing them using characters commonly accessible on computers, and therefore suggest the following basic set of signs for transliterating punctuation
- | (U+007C Vertical Line): for signs comprised of a single (more or less) plain and vertical line (of whatever length)
- / (regular slash character): for signs comprised of a single (more or less) vertical line (of whatever height) with a hook, crossbar or ornamental addition
- , (regular comma): for the single, raised dot-like sign that is the basic punctuation sign on Java and Bali (modern Balinese \ but the downward tail is generally not pronounced in inscriptions)
- – (U+2013 En Dash): for signs comprised of a single horizontal line (of whatever length, including very short lines like a dot), with or without a curve or an ornamental addition
  - please take care not to type a hyphen instead of this character
- = (equal sign): for signs comprised of a double horizontal line (of whatever length, including very short lines like a dot), with or without a curve or an ornamental addition
- all of these characters shall be followed by a space in transliteration; typing a space before them is unnecessary but acceptable
- all of these characters may be iterated as needed to transliterate for double (or multiple) marks; do not put spaces between the iterations
- for punctuation signs taking other shapes, use the following notation:
  - type a | followed immediately (without an intervening space) by any Unicode character that you deem to be a passable approximation of the shape of the original sign
  - this notation will be converted to markup, which will retain only the second character you type, but tag it as a punctuation sign
    - e.g. your transliteration /◊ will become a ◊ character marked up as a punctuation sign
  - use this notation only if you are certain that the symbol you are dealing with is a punctuation sign; see §4.4 below for other symbols
  - the shape of such punctuation signs shall be described in human-readable terms in the <handDesc> element of the TEI header, see §XXX of the Encoding Guide
    - at a later stage, we may harvest such descriptions and use them as a starting point for a controlled vocabulary for symbol description

### 4.3. Space Filler Signs

- for symbols whose function is clearly and unambiguously to fill up space in a line to the margin, use the character § (U+00A7 Section Sign) in transliteration
- the shape of space fillers shall be described in human-readable terms in the <handDesc> element of the TEI header, see §XXX of the Encoding Guide
  - at a later stage, we may harvest such descriptions and use them as a starting point for a controlled vocabulary for symbol description

## 4.4. Other Symbols

- handle miscellaneous symbols (those not clearly identifiable as punctuation signs or space fillers) in the following manner:
  - type a \$ character followed immediately by any Unicode character that you deem to be a passable approximation of the shape of the original sign
  - this notation will be converted to markup which will retain only the second character you type, but tag it as a symbol
    - e.g. your transliteration \$ꣳ will become a ꣳ character marked up as a symbol
- the shape and presumable function of such symbols shall be described in human-readable terms in the `<handDesc>` element of the TEI header, see \$XXX of the Encoding Guide
  - at a later stage, we may harvest such descriptions and use them as a starting point for a controlled vocabulary for symbol description
- note that auspicious (*maṅgala*) symbols should never be transliterated as the words *siddham* or *oṃ*

## 4.5. Space

- where an inscription employs a space between words, transliterate that space explicitly with a \_ (underscore) character
  - you may also add a regular space before and/or after the underscore, but this is not required
- any other spaces — such as space left blank for filling later, or because of a defect or feature of the material — will need to be handled in the markup, see Encoding Guide \$XXX

## References

- Brookes, Stewart, Peter A. Stokes, Matilda Watson, and Débora Marques de Matos. 2015. 'The DigiPal Project for European Scripts and Decorations'. In *Writing Europe, 500-1450*, edited by Aidan Conti, Orietta Da Rold, and Philip Shaw, NED-New edition, 25–58. Texts and Contexts. Boydell and Brewer.
- Coulmas, Florian. 2006. *The Blackwell Encyclopedia of Writing Systems*. 4th ed. Oxford: Blackwell.
- Damais, Louis-Charles. 1955. 'II. Etudes d'épigraphie indonésienne, IV : Discussion de la date des inscriptions'. *Bulletin de l'École française d'Extrême-Orient* 47 (1): 7–290.  
<https://doi.org/10.3406/befeo.1955.5406>.
- ISO15919:2001 = International Standard ISO 15919. Information and Documentation — Transliteration of Devanagari and Related Indic Scripts into Latin Characters. Geneva: International Organization for Standardization.  
<https://www.iso.org/standard/28333.html>.
- ISO/IEC 10646:2017(E) = International Standard ISO/IEC 10646. Information Technology — Universal Coded Character Set (UCS). 5th ed. Geneva: International Organization for Standardization.  
[https://standards.iso.org/ittf/PubliclyAvailableStandards/c069119\\_ISO\\_IEC\\_10646\\_2017.zip](https://standards.iso.org/ittf/PubliclyAvailableStandards/c069119_ISO_IEC_10646_2017.zip).
- Ollett, Andrew & Sarah Pierce Taylor. 2019. *Representing Kannada Text*. Draft document, [https://docs.google.com/document/d/18YNbAJIuxOicnyGTeUzNPq\\_GjX3Fg7Ka5liw92ZZBXk/](https://docs.google.com/document/d/18YNbAJIuxOicnyGTeUzNPq_GjX3Fg7Ka5liw92ZZBXk/) (accessed 23 July 2019)
- Wellisch, Hans H. 1978. *The Conversion of Scripts—Its Nature, History, and Utilization*. New York: Wiley.