Transliteration Guide

for members of the

project

Release Version 1.1, 2019-12-18

Dániel Balogh & Arlo Griffiths

Contents

1. Introduction 2
1.1. Version History 2
1.1.1. Summary of changes since the last version 2
1.2. Coverage 2
1.3. Separation of Transliteration and Encoding 2
1.4. Terms and Definitions 3
1.4.1. Abbreviations 3
1.4.2. Script and its elements 4
1.4.3. Script conversion 6
1.4.4. Notation for transliteration and transcription 7
2. General Principles 8
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
2.4.1. A note on the use of uppercase for standalone vowels and consonants 11
2.5. Disambiguation 12
2.6. Editorial Additions for Text Analysis 13
2.6.1. Editorial spaces for word segmentation 13
2.6.2. Editorial hyphenation 14
2.6.3. Representation of *avagraha* 15
2.6.4. Representation of elided overshort final *u* in Tamil 15
3. Alphabetic Characters 16
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 18
3.3.2. Final consonants as complex characters involving a zero vowel marker 18
3.3.3. Independent vowels as special simplex characters 19
3.3.4. Independent vowels as complex characters involving a “vowel support” 19
3.3.5. Multiple vowel markers within an *akṣara* 20
3.3.6. Repurposed vowel markers 20
3.3.7. Short vowel written where a corresponding long vowel is expected 20
3.3.8. Unusually composed complex characters 21
3.3.9. Characters with alternative or optional phonemic values 22
3.3.10. Complex characters split by an intervening feature 23
4. Non-alphabetic Characters 24
4.1. Numerals 24
4.1.1. Numbers denoted by bars 24
4.1.2. Fractions 25
4.2. Symbols 25
4.2.1. Punctuation marks 25
4.2.2. Space filler signs 26
4.2.3. Generic symbols 26
4.3. Space 26
References 27

1. Introduction

1.1. Version History

Author(s)	Version	Changes	Date
Balogh, Griffiths	0.1	First draft	2019-07
Balogh, Griffiths	1	Expansion and revision for first release	2019-09
Balogh	1.1	Revision	2019-12

1.1.1. Summary of changes since the last version

– deprecated ° as a marker for independent vowels and final consonants (§2.4.1)
– added §3.3.10 on *Complex characters split by an intervening feature*
– restricted the use of : to ISO-15919 disambiguation (§2.5) and the Indonesian *tarung/tedong* (§3.3.6)
– introduced the use of = (to replace the former :) for flagging unusual *akṣara* composition (§3.3.8) including Tamil ligatures and varying reading modes of superscript *r*
– extended the use of + in numeral notation to cover fractions (§4.1.2)
– revised the shorthand notation for symbol tokens (§4.2.3)

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.[1] 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 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[2] 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. Terms and Definitions

1.4.1. Abbreviations

In addition to common abbreviations, this Guide uses:

TG the Dharma Transliteration Guide (the present document)

EG the Dharma Encoding Guide (version 1.0)

[1] We follow the TEI Guidelines in using the terms 'markup' and 'encoding' as interchangeable synonyms.

[2] Good practice recommendations/rules for naming files will be distributed in a separate document.

1.4.2. 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)
  - 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 /$t^{ɦ}$/ 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 द्धे are each one character
        - while none of the elements corresponding to the transliterated characters *r*, *d*, *dh* and *e* in the *akṣara* द्धे 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 द्धे 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.)[3]
- 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 = ꣟ = ki *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 क्त 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 ᨠ, म, ᨣ, ४, ᨠ 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.)

[3] The term 'grapheme' is sometimes defined differently, so that polygraphs are considered to be a single grapheme; this definition does not concern us here.

  - 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)
- 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 द्र्धे, 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.

1.4.3. 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)

1.4.4. Notation for transliteration and transcription

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 §1.4 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.

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 ¯ (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. *ṛ* instead of *r̥* 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 §§2.6.4 and 3.2)
  - this applies even when you are certain that a specific *akṣara* was pronounced in a way unlike that dictated by the inherent logic of the script; see §3.3.9 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"[4]) 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 *vvaṁ/vvaṅ* merged into *vvaṅ* (phonologically /wwaŋ/), *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*, *rahiyaṅ* and *ku nu reya* (even when written as *sasṭā*, *ku nu rye* and *rahyiṁ* as in the examples under §3.3.9)
  - 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[5]
  - Wikipedia (https://en.wikipedia.org/wiki/ISO_15919) summarises the essential features
- if you are used to IAST, this means paying attention to using *ṁ*, *r̥*, *r̥̄* and *l̥* rather than *ṃ*, *ṛ*, *ṝ* and *ḷ*

---

[4] 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."

[5] 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 align as much as possible the guidelines on Kannada transliteration drafted by Andrew Ollett and Sarah Pierce Taylor (forthcoming), although at this stage it is unclear whether agreement can be reached on all points

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. कृतमेतत्) but chooses instead to use a final consonant followed by an initial vowel to represent a pause for semantic or metrical segmentation (e.g. कृतम्एतत्)
    - 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.

2.4.1. A note on the use of uppercase for standalone vowels and consonants

- some of us have previously adopted the system of using a ° character before transliterated vowels and after transliterated consonants to denote special initial and final forms
  - the principal investigators have agreed to discontinue using that notation, so henceforth it should not be used in XML files
    - it is also recommended that you adopt the uppercase notation in all your work including printed publications
  - intellectual considerations in favour of adopting the uppercase notation include the following:
    - 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, there is no such equivalence in the case of special character forms, which are more rigorously transliterated using a single Latin character than by a digraph

- 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 *akṣara*s, 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 the sign ° free for the conventional use as a marker of truncation (e.g. when cutting words to be cited in a critical apparatus)

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
  - note that a disambiguation sign is not required if an editorial space or hyphen separates the two characters in question, since the transliteration is already unambiguous in this case without
- in ISO-15919, a disambiguation colon is used between vowels in hiatus to distinguish certain vowel sequences from diphthongs transliterated 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, which ISO-15919 achieves by transliterating them as *pra:uga*, *ca:uttho*, *da:iā*
  - however, our strict transliteration system[6] provides ways of distinguishing independent vowel signs of the original script from vowel markers (see §§3.3.3 and 3.3.4), and thus we can transliterate the above words as *praUga*, *caUttho* and *daIĀ*
  - as a consequence, 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*)
  - accordingly, we have chosen to preserve alternate uses of the colon for some special purposes, namely to indicate the use of the *ā* marker in Indonesian scripts as an indicator of vowel length or consonant doubling (§3.3.6)

[6] We recommend 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*, *caüttho* and *daïā*.

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 EG §XXX
- 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 *putrām̆l lakṣmīḥ* (space goes between the two *l*-s)
      - Old Javanese *tann inaku* (space goes between the *-nn* and the *i-*)
      - Tamil *arit' eṉṟu* (for *aritu + eṉṟu*; 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 *śaraṇaṅ 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
  - in repetitions of Sanskrit inflected pronouns and nouns (*āmreḍita*) expressing a generalised or distributive meaning, e.g. *yasya yasya*, *dine dine*
- **do not**, however, use spaces to 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. *āṉa*, *āy*, *āka*) when used adverbially
  - Old Javanese enclitic pronominal suffixes (*-(ṅ)ku* etc.), possessive constructions built with the linker *-ni* (*-nikaṅ*, etc.); definite article *-ṅ*; conjunction *-n*
  - 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*
- in sub-standard Sanskrit, strings of words without case endings but apparently intended as nominatives should preferably be spaced instead of being treated as compounds (e.g. *dvandva*), unless the latter in fact facilitates interpretation, e.g.
  - *lamvoṣṭha dedamita mahādeva guṇṭhaka ity evam-ādibhyo*
  - *samrāṬ vākātakānāṁ mahārāja-śrī-pravarasenasya*

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
    - in the case of Old Javanese, consider that reduplicated expressions are always compounds, whereas close-knit structures consisting of two different elements only become compounds if any morphological derivation takes place
  - **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)[7]
- 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-makaḷ* (திருமகள் *tiru+makaḷ*)
  - *koṇṭ-āṭu* (கொண்டாடு *koṇṭu+āṭu*)
  - *I-p-peruṅ-kōyil* (இப்பெருங்கோயில் *i+perum+kōyil*)
  - *tiru-mēṉi-y āṭa* (திருமேனியாட *tiru+mēṉi āṭa*)
- some examples of Old Javanese hyphenation:
  - *vulu-vulu*
  - *tahi tikus* > *manahi-tikusa*
  - *bvat haji* > *makabvat-hajya*
- **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 EG §XXX)
  - 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

[7] 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*.

- 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 EG §XXX 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 EG §XXX)
    - 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 *avagraha*s in a typed text will be assumed to be supplied by the editor, and markup signifying this (for which see EG §XXX) will be added automatically
  - the supplying of *avagraha*s 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
    - this way, the *avagraha* in question will not be automatically marked up as supplied, and the ! will be removed after marking up all other *avagraha*s 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 overshort final *u* in Tamil

- in the transliteration of Tamil text, use an apostrophe followed by a space to represent the elided overshort *u* at the end of an independent word, e.g.
  - *arit’ eṉṟu* (அரிதென்று for *aritu + eṉṟu*)
- but do not use an apostrophe for the elided overshort *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)
- these apostrophes are understood to be integral parts of our transliteration system, and not as something supplied by the editor for the sake of normalisation

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
- **vocalic *r* and *l***
  - these are not available in Unicode as pre-composed characters, so to create them, you may need to enter an *r* or *l* as applicable, followed by ◌̥ (U+0325 Combining Ring Below) and, if needed, by ◌̄ (U+0304 Combining Macron) in this order
  - alternatively, since none of the languages we work with require the use of *ṛ* to represent a consonant, you may optionally use *ṛ* and *ṝ*, which will later be automatically converted to *r̥* and *r̥̄* in your files (but note that this does not apply to *ḷ*)
- ***anunāsika/candrabindu***
  - *m̐* (this character is not available as a precomposed glyph, so it must be composed of a regular *m* and a ◌̐ 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 *tāl̐ lakṣmīm* and write *tām̐l lakṣmīm* instead)
- ***upadhmānīya*** (if distinguished in the script from *visarga*)
  - *ḫ* (U+1E2B Latin Small Letter H with Breve Below)
- ***jihvāmūlīya*** (if distinguished in the script from *visarga*)
  - *ẖ* (U+1E96 Latin Small Letter H with Line Below)
- **Tamil *āytam***, ஃ
  - *ḵ* (U+1E35 Latin Small Letter K with Line Below)
- **retroflex lateral**, Tamil ழ Kannada/Telugu ಳ
  - *ḷ* (U+1E37 Latin Small Letter L with Dot Below)
- **alveolar trill/stop**, Tamil ற Kannada/Telugu ಱ
  - *ṟ* (U+1E5F Latin Small Letter R with Line Below)
- **retroflex approximant / frictionless continuant**, Tamil ழ Kannaḍa/Telugu ೞ
  - *ḻ* (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 ◌̃ 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
  - *ṁ°* (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); uppercase Ə (U+018F Latin Capital Letter Schwa)
    - you may type *ĕ* instead of *ə* if it is easier for you; since *ĕ* is not used for any other purpose in our transliteration, it can be automatically converted to *ə*
  - long, *ə:* (with length-mark represented by a colon as per §3.3.6) in strict transliteration
    - *ə̄* in loose transliteration (not available as a precomposed character: add U+0304 Combining Macron to the plain character)
- ! **Khmer (and Mon-Burmese) glottal stop**
  - *q* (the Latin letter q)
  - see also §3.3.4 about the representation of independent vowels involving this character component
- ! **special signs for Mon and Pyu:**
  - barred/dotted variant of *b*
    - *ḃ* (U+1E05 Latin Small Letter B with Dot Below)
  - *akṣara*s with underdot
    - *ṃ* (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 கெ *ke*, கொ *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 கே *kē*, கோ *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 employ some mandatory distinctions set out in the following subsections

- certain further distinctions set out in the following subsections may be optionally made using markup or a shorthand notation that will be auto-converted to markup
- other potentially interesting allographs – for instance the use of two alternative glyphs within the same inscription for the same simplex character, or different ways in which a vowel marker is attached to a consonant – will need to be described in your metadata, and will not be directly represented in the transliteration or the markup[8]

3.3.1. Final consonants as special simplex characters

- special character forms representing consonants without a vowel (called *halanta* consonants in Sanskrit) are typically a miniature and/or subscript rendering of a simplex consonant *akṣara*
- such special final forms shall be mandatorily transliterated as a corresponding uppercase Latin consonant, e.g. *T*
- the criterion by which to distinguish special final forms from complex characters involving a zero vowel marker (§3.3.2) is the use of a glyph distinct (in size, shape or both) from 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. ᮃᮏᮃᮊᮢᮦᮑᮢᮦᮑᮢᮦᮑ

  - we may at a later point decide to automatically convert such transliterations into markup signifying that a *puḷḷi* 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 *puḷḷi*
  - note that where an actual ligature occurs in Tamil script, this must be treated as unusual *akṣara* composition, for which see §3.3.8
- 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 tag 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 *akṣara*s (initial vowels and vowels in hiatus), these shall be mandatorily 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)

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 mandatorily 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[9]
  - hence, we prefer to transliterate all "vowel supports" with the dedicated character *q*

combination	glyph	phoneme	translit.
A with *taling*	ᬍᬾᬅ	/e/	*qe*
A with *suku*	ᬅᬼ	/u/	*qu*
A with *ulu*	ᬅᬺ	/i/	*qi*
A with *taling tedong*	ᬍᬾᬅᬾ	/o/	*qo*

- note in particular that this applies equally to Sanskrit text written in this manner:
  - not only to isolated words and names, but even to (monolingual) Sanskrit compositions written in the given script (thus, emphatically, to Sanskrit inscriptions from Cambodia)

[9] This systemic change is complete in mainland Southeast Asia, where the "vowel support" (e.g. Khmer អ) 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 ᬅ can be combined with a *pasangan* consonant, e.g. ᬕᬺᬲᬶ *qhi* and ᬕᬼᬲᬼ *qhu*, but cannot itself become a *pasangan*, while in Balinese writing neither possibility exists.

3.3.5. Multiple vowel markers within an *akṣara*

- multiple vowel markers may be used deliberately
  - to represent a particular phoneme or modification (§3.3.6)
  - to mark segments as deleted (this belongs in the domain of markup, not that of transliteration; see EG §XXX)
  - if you encounter multiple vowel markers that appear to be used deliberately for a purpose other than the above, please contact us to discuss how best to represent them
- other appearances of multiple vowel marks are likely to be cases where the scribe erroneously engraved more than one explicit vowel mark, neither of which appears to be deliberately cancelled
  - if one of these vowels is expected in the context and the other is not, it is acceptable and preferable to encode this as a premodern correction as per EG §XXX
  - in other cases, which we expect to be very rare, transliterate all vowels in an order you deem suitable
    - the fact that the transliterated vowels are lowercase indicates in our system that none of them are independent vowel *akṣara*s (cf. §2.5)
    - nonetheless, the unusual fact that multiple vowel markers are present in a single *akṣara* may optionally be made explicit using an = (equals) sign between the transliterated vowels belonging to a single *akṣara*

3.3.6. Repurposed vowel markers

- for the ***ā* marker** (Javanese *tarung*, Balinese *tedong*) used as a marker of vowel length or consonant doubling in Indonesian texts:
  - when used in conjunction with another vowel marker to transform the latter into a long vowel, mandatorily type a colon (:) after the short vowel to transliterate the length marker
  - when representing a doubling of the consonant component of the *akṣara* to which it is attached, mandatorily transliterate this by typing a colon (:) after the transliterated consonant
    - e.g. Old Sundanese ᮃᮔᮤᮃᮕᮤᮊᮥ᮪ (pronounce *gənəp pipitu* "fully seven") is to be transliterated as *gnəp:ipitu*
- for the **vowel markers *u/ū* and *i*** used together to represent a particular phoneme in Khmer, Burmese and Mon (as in the Khmer character shown in the image):

  - mandatorily transliterate the vocalisation as *ui* or *ūi*
  - however, the deliberate use of *u* and *i* markers in conjunction to signify deletion belongs in the domain of markup (see EG §XXX), not that of transliteration

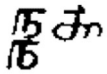
3.3.7. Short vowel written where a corresponding long vowel is expected

- where a short vowel is written in place of an otherwise identical long vowel, optionally 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 EG §)

3.3.8. Unusually composed complex characters

– in order to highlight certain formations that deviate from the standard glyph composition for any particular language and writing system, our transliteration scheme permits the use of the dedicated character = (equals sign), in the specific cases set out below
  – this notation is optional, but if do you employ it anywhere within an edition, please attempt to use it consistently throughout that edition wherever applicable
  – the = sign 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
  – should you need to add an editorial space or hyphen between such characters, put the space *after* the = sign
  – this notation will be auto-converted to markup (EG §XXX)
– **where a Tamil text written in Tamil script employs a ligature** such as *nna* and *kka* in the image



  – use an = sign between the transliterated consonants to distinguish the ligature from the script's default method of writing conjunct consonants as two glyphs with an explicit or implicit vowel killer (for which see also 3.3.2), e.g.
    – *n=na* as distinct from *n·na* and *nna*
    – *k=ka* as distinct from *k·ka* and *kka*
    – *t=ta* as distinct from *t·ta* and *tta*
  – it is **strongly recommended** that whenever feasible, you should make Tamil ligatures explicit in this way
  – however, never add an = sign where ligatures are a writing system's default method of representing conjunct consonants (including Tamil written in Grantha)
– **where an Indonesian text employs the superscript *r* marker** (*repha*, *layar*, *surang*) **in two modes**,
  – namely
    – the "Indian" mode, i.e. to be read before the consonant it is attached to, as in ᨿᨾᩕ *sarva*; and
    – the "Indonesian" mode, i.e. to be read after the rest of the *akṣara* it is attached to, as in ᨿᨾᩮᩕ *samar*
  – then this fact must be noted in your commentary to the text, including a specification of which mode is the default (dominant) one for that text
  – in addition, you may optionally use an = sign between the transliterated *r* and the other characters transliterating the same *akṣara* in instances of the non-default mode, i.e.
– 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, i.e.
    – if the "Indonesian" mode is dominant, transliterate ᨿᨾᩕ as *sar=va*
    – if the "Indian" mode is dominant, transliterate ᨿᨾᩮᩕ as *sama=r*

- for the sake of explicitness you may also choose to use the above notation redundantly for all occurrences of a superscript *r*, not only for ones that deviate from the default
- **where a superscript *r* marker is combined with an independent vowel sign**, as in the Javanese example on the right

  - use an = sign between the transliterated characters to make it explicit that the transliteration is not a mistake (for an original text involving a final consonant, a vowel killer marker or a dependent vowel)
  - thus, the text in the image is *Umiṅsor= I* (note the editorial space after the = sign)

3.3.9. Characters with alternative or optional phonemic values

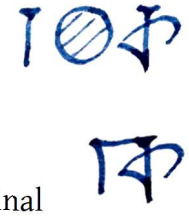
- some writing systems may use certain glyphs to represent more than one phoneme or sequence of phonemes, or may use a non-alphabetic character in an alphabetic function[10]
  - **in strict transliteration**, always prioritise the primary value of such glyphs
  - **in loose transliteration**, however, it is preferable to transcribe the phonemic value intended in the context
  - **in an EpiDoc edition**, you may add markup to normalise the transliterated primary value to a transcription of the intended value (see EG §XXX 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: *sasṭā*; 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]")
    - strict: *rahyiṁ*; loose: *rahiyaṅ* (with *anusvāra* normalised to *ṅ*)
  - 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 EG §XXX), but use *ro* in loose transliteration
    - e.g. – strict: *di jə2niṁ vavaṁṅun:an·* ; loose: *di jəroniṅ vavaṅunan* ("in the interior of the building")[11]

---

[10] 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.

[11] Note that in the loose transliteration of this example, *ṁ* and *ṁṅ* are both represented by *ṅ* (being the sign selected to represent phoneme /ŋ/). Simultaneously, *n:* (theoretically denoting /nn/) is simplified to *n*, and *ṁṅ* (theoretically /ṅṅ/) is simplified to *ṅ*, because consonant gemination is not considered to be a phonemic feature of the language, but rather an orthographic particularity.

3.3.10. Complex characters split by an intervening feature



- certain character components are treated as separable in some scripts, such as the prescript and postscript vowel markers in Tamil ஒகர *ko* or the *pṛṣṭhamātrā e* in varieties of Nagari (as in the images to the right)
- while the separation of a postscript *ā* marker from its consonant could be represented accurately in transliteration, separations involving prescript markers are impossible to duplicate due to the non-linear nature of the original script
- we therefore introduce two *placeholder characters* into our transliteration scheme:
  - ⌈ (left ceiling, U+2308) to represent a prescript component split off from the following original character
  - ⌉ (right ceiling, U+2309) to represent a postscript component split off from the preceding original character
  - if you have difficulty entering these characters, you can instead use [[and]] respectively, which will be automatically converted to the above special characters
- in transliteration, put all of the transliterated characters belonging to the split original character on that side of the interruption where the consonant body is located, and add the applicable placeholder character on the other side of the interruption, thus:
  - கா<>◌ா as *kā*<>⌉
  - ஒ◌<>க as ⌈<>*ke*
  - ஒக<>◌ா as *ko*<>⌉ (likewise for split *au*)
  - ஒ◌<>கா as ⌈<>*ko* (likewise for split *au*)
- in the above examples, ignore the dotted circle representing the body associated with dependent vowel signs
- in the above examples, <> represents the interruption, which must be encoded appropriately or, if you are only creating an e-text for later markup, clearly indicated in the transliteration:
  - line break: EG §XXX; failing that, start a new line in the e-text
  - space imposed by a physical feature of the support: EG §XXX; failing that, use an _ character (TG §4.3)
- if you encounter a character with a split-off part other than a prescript or postscript vowel marker, please contact us to discuss its most suitable representation
- see also EG §XXX about encoding lacunae and reading difficulties in combination with split characters, including in particular situations where an original glyph (component) may be either the Tamil postscript vowel marker *kāl* (◌ா) or the character *ra* (ர)
- the use of these placeholder characters is **optional, but strongly recommended** in all cases where you have access to the original or a surrogate
  - if you only have access to a printed edition or choose not to employ placeholder characters, you should still put all your transliterated characters pertaining to a single *akṣara* on one side of the interruption, i.e. avoid transliterations such as *k*<>*ā*, *k*<>*e*, *k*<>*o*

4. Non-alphabetic Characters

4.1. Numerals

- numbers written in **decimal place-value notation** in the original shall be transliterated straightforwardly, e.g. 876
- numbers recorded in an **additive system** must be transliterated so that each separate character of the original is represented separately in the transliteration, and never simply "translated" into the Arabic place-value notation of the corresponding number[12]
  - type a + sign after each transliterated number sequence of two or more Arabic numerals that represents a single numeral character in the original
    - 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)
  - for example:
    - 10+ – "10" written as a character for 10, e.g. Brahmi ∝
    - 10+2 – "12" written as a character for 10 followed by one for 2, e.g. Brahmi ∝=
    - 80+10+ – "90" written as a character for 80 followed by one for 10 , e.g. Brahmi ⊕∝
    - 300+50+2 – "352" written as a character for 300, one for 50, and one for 2
  - this notation will be automatically converted to markup indicating that these Arabic digits transliterate a single original numeral character (see EG §XXX)
  - note that the above notation differs slightly 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
    - therefore, sequences
    - thus, in our system a final + sign is required in cases such as the following
      - 10+ (rather than 10) – "10" written as a character for 10
      - 300+50+ (rather than 300+50) – "350" written a character for 300 and one for 50

4.1.1. Numbers denoted by bars

- to transliterate numerals represented in Cambodian inscriptions by bars (*daṇḍa*) instead of numeral characters (as in the image, showing the number 3):
  - type as many I characters as there are bars in the original (NB: uppercase i characters, not vertical bars |)
  - 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

[12] The actual value of the number must, however, be represented in the markup added to your text, for which see EG §XXX.

  - this notation will be automatically converted to markup indicating that these characters are not alphabetic and constitute a single meaningful character

4.1.2. Fractions

- for any fractional numbers represented in the original by a single character, use one of the following Unicode characters:
  - ½ (U+00BD Vulgar Fraction One Half)
  - ⅓ (U+2153 Vulgar Fraction One Third)
  - ⅔ (U+2154 Vulgar Fraction Two Thirds)
  - ¼ (U+00BC Vulgar Fraction One Quarter)
  - ¾ (U+00BE Vulgar Fraction Three Quarters)
- should you encounter fraction signs not covered by the above, transliterate them as a common fraction using a slash and add a + sign after the denominator
  - e.g. 1/8+ to transliterate a numeral sign meaning "one eighth"
  - this notation will be automatically converted to markup as per EG §XXX
- the Khmer fraction sign in the shape of a cross (with a single or a double bar, see the images) shall always be transliterated as ½ (U+00BD Vulgar Fraction One Half)

4.2. Symbols

4.2.1. Punctuation marks

- always transliterate all **original punctuation**, but **never add punctuation marks** not already present in the original
  - editorial punctuation may be supplied in markup, see EG §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 dedicate the following **basic set of characters** to transliterating punctuation
  - | (U+007C Vertical Line): for signs comprised of a single (more or less) plain and vertical line (of whatever length)
  - / (regular slash): for signs comprised of a single (more or less) vertical line (of whatever height) with a hook, crossbar or ornamental addition
  - , (regular comma): for dots and strokes shorter than a full *daṇḍa* and normally floating at or above median height, including half-sized *daṇḍa*s and the raised comma-like sign that is the basic punctuation sign on Java and Bali (modern Balinese ᬄ)
- the above characters will be retained as such and will not be converted to XML markup
- all of the above characters may be iterated as needed to transliterate multiple marks
- these characters (or the last iteration of a kind, when they are grouped together) should be followed by a space in transliteration
  - typing a space before (or between) them is unnecessary but acceptable
- punctuation signs corresponding to the above generic categories but exhibiting special features should be described in human-readable terms in your metadata

  - at a later stage, we may harvest such descriptions and use them as a starting point for a more nuanced classification of punctuation marks
- punctuation signs taking **other shapes** (including simple circles and dashes as well as complex patterns) shall be represented in transliteration as generic symbols, as per §4.2.3 below

4.2.2. 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
  - this character will be automatically tagged in XML (EG §XXX)
  - any symbols that do not function as space fillers must be encoded as generic symbols, even if they appear visually identical to space fillers
- the appearance of space fillers shall be described in your metadata

4.2.3. Generic symbols

- symbols other than space fillers and basic punctuation marks as classified under §4.2 shall be represented as symbol tokens
- a symbol token is a simple description of the symbol's visual appearance or its traditional name, prefixed with a $ (dollar sign)
  - the token must contain no spaces (if necessary, use an _ underscore instead), but it may contain any combination of letters and numbers
  - letters with diacritic signs should preferably be avoided
  - these tokens will be automatically converted to markup as per EG §XXX
  - at a later stage, we intend to harvest such tokens and use them as a starting point for a controlled vocabulary for symbol description and for display
  - some examples: $trefoil; $svastika; $siddham; $clockwise_spiral; $floret; $double_circle
    - see §Appendix D of the EG for some further recommendations
  - note that auspicious (*maṅgala*) symbols should never be transliterated as the words *siddham* or *oṁ*
- so long as you are transliterating a text without adding XML markup, **Unicode characters** approximating the original glyph (e.g. ◊卐✤) may be used instead, if this is more convenient in your practice
  - when marking up your files in XML, you will need to replace such characters either with tokens as above, or with markup as per EG §XXX
- more detailed descriptions of your symbols may be included in your metadata

4.3. Space

- where an inscription employs an interword space (large enough to be called a space but smaller than the width of two average characters) between, 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 EG §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.

Ollett, Andrew & Sarah Pierce Taylor. 2019. *Representing Kannada Text*. Draft document, 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.