Transliteration Guide

Dániel Balogh & Arlo Griffiths

Release Version 4 DRAFT, 2025-##-##

This project has received funding from the European Research Council (ERC)  
under the European Union’s Horizon 2020 research and innovation programme  
(grant agreement No 809994).

Contents

[1. Introduction 1](#_Toc203989356)

[1.1. Version History 1](#_Toc203989357)

[1.2. Coverage 1](#_Toc203989358)

[1.3. Abbreviations 1](#_Toc203989359)

[1.4. Brackets for linguistic notation 2](#_Toc203989360)

[1.5. Terms and definitions 2](#_Toc203989361)

[1.6. Working with Unicode 3](#_Toc203989362)

[1.6.1. Fonts and supported characters 3](#_Toc203989363)

[1.6.2. Entering Unicode characters 4](#_Toc203989364)

[1.6.3. Precomposed characters 4](#_Toc203989365)

[2. Theoretical framework 5](#_Toc203989366)

[2.1. Scripts and writing systems 5](#_Toc203989367)

[2.1.1. Writing system typology 5](#_Toc203989368)

[2.2. Conversion between writing systems 6](#_Toc203989369)

[2.2.1. Transliteration and transcription 6](#_Toc203989370)

[2.3. The elusive grapheme 7](#_Toc203989371)

[2.3.1. A sketch of grapholinguistic concepts 7](#_Toc203989372)

[2.3.2. Refining the concept of the grapheme 8](#_Toc203989373)

[2.3.2.1. The linguistic aspect of the grapheme 9](#_Toc203989374)

[2.3.2.2. The graphic aspect of the grapheme 10](#_Toc203989375)

[2.4. Graphic structures and their elements 11](#_Toc203989376)

[2.4.1. Characters and glyphs 11](#_Toc203989377)

[2.4.2. Polygraphy 12](#_Toc203989378)

[2.4.3. Glyph complexity 12](#_Toc203989379)

[2.4.3.1. Glyph components 13](#_Toc203989380)

[2.4.3.2. Markers 13](#_Toc203989381)

[2.4.3.3. Graphic elements 13](#_Toc203989382)

[2.5. Problematic cases: what is a grapheme, and what is not? 14](#_Toc203989383)

[2.5.1. *Diacritical marks* 14](#_Toc203989384)

[2.5.2. *Virāma* 14](#_Toc203989385)

[2.5.3. *Anusvāra* relatives 15](#_Toc203989386)

[2.5.4. Other signs of vague status 16](#_Toc203989387)

[2.5.5. Non-phonographic signs 16](#_Toc203989388)

[2.5.6. Fuzzy segmentation 16](#_Toc203989389)

[2.5.6.1. Character or component? 17](#_Toc203989390)

[2.5.6.2. Component or element? 17](#_Toc203989391)

[2.6. Revisiting allography 18](#_Toc203989392)

[3. General principles of the DHARMA transliteration scheme 21](#_Toc203989393)

[3.1. Compatibility with other transliteration systems 21](#_Toc203989394)

[3.2. Graphematic entities in transliteration 21](#_Toc203989395)

[3.2.1. Transliterating graphemes 21](#_Toc203989396)

[3.2.2. Representing complex characters and allographs 22](#_Toc203989397)

[3.3. Case sensitivity 22](#_Toc203989398)

[3.3.1. A note on the use of uppercase for independent vowels and consonants 23](#_Toc203989399)

[3.4. The accuracy of transliteration 23](#_Toc203989400)

[3.4.1. Strict transliteration 23](#_Toc203989401)

[3.4.2. Loose transliteration 24](#_Toc203989402)

[3.5. Transliteration and markup 24](#_Toc203989403)

[3.5.1. Disambiguation 25](#_Toc203989404)

[3.5.2. Segmentation 25](#_Toc203989405)

[3.5.3. Sandhi analysis 26](#_Toc203989406)

[3.5.4. Truncation 26](#_Toc203989407)

[3.5.5. Placeholders for split *akṣara*s 26](#_Toc203989408)

[3.5.6. Shorthand 27](#_Toc203989409)

[3.5.6.1. Private shorthand 27](#_Toc203989410)

[3.5.6.2. Public shorthand 28](#_Toc203989411)

[3.5.6.3. Optional shorthand 28](#_Toc203989412)

[3.6. Non-graphematic entities and transliteration 29](#_Toc203989413)

[3.6.1. Marks and imagery peripheral to the text 29](#_Toc203989414)

[3.6.1.1. Scribal marks 29](#_Toc203989415)

[3.6.1.2. Decorative features 29](#_Toc203989416)

[3.6.2. The materiality of the support 29](#_Toc203989417)

[4. Transliterating alphabetic graphemes 30](#_Toc203989418)

[4.1. The basic inventory of Indic graphemes for Old Indo-Aryan 30](#_Toc203989419)

[4.1.1. Digraphs in the transliteration 30](#_Toc203989420)

[4.1.2. Vocalic *r* and *l* 30](#_Toc203989421)

[4.1.3. Transliteration of *e* and *o* 30](#_Toc203989422)

[4.1.4. *Anusvāra* and its relatives 31](#_Toc203989423)

[4.1.4.1. *Anunāsika* or *candrabindu* 31](#_Toc203989424)

[4.1.4.2. Other *anusvāra* variants 31](#_Toc203989425)

[4.1.5. *Visarga* and its relatives 32](#_Toc203989426)

[4.2. Graphemes extending the basic repertoire 32](#_Toc203989427)

[4.2.1. Graphemes of other Indian writing systems 33](#_Toc203989428)

[4.2.2. Graphemes of Southeast Asian writing systems 33](#_Toc203989429)

[4.2.2.1. Graphemic combination of the vowel markers |u| and |i| 34](#_Toc203989430)

[4.3. Vowelless consonants 34](#_Toc203989431)

[4.3.1. Distinguishing final forms from characters with a vowel killer 34](#_Toc203989432)

[4.3.2. Final consonants as simplex characters 35](#_Toc203989433)

[4.3.3. Independent consonants as complex characters involving a vowel killer 35](#_Toc203989434)

[4.3.4. Regular consonant signs for vowelless consonants: the “implicit *puḷḷi*” 35](#_Toc203989435)

[4.4. Independent vowels 36](#_Toc203989436)

[4.4.1. Independent vowels as simplex characters 36](#_Toc203989437)

[4.4.2. Independent vowels involving a vowel support 36](#_Toc203989438)

[4.5. Systemic innovations in the Indic writing system 38](#_Toc203989439)

[4.5.1. Borderline diacritical marks 39](#_Toc203989440)

[4.5.1.1. The |ā| graph as a signifier of length in maritime Southeast Asia 39](#_Toc203989441)

[4.5.1.2. Underdotted *akṣara*s in mainland Southeast Asia 40](#_Toc203989442)

[4.5.2. Repurposed graphic signs 40](#_Toc203989443)

[4.5.3. Variation in glyph composition 41](#_Toc203989444)

[4.5.3.1. Optional shorthand for complex characters 41](#_Toc203989445)

[4.5.3.2. Conjunct consonants in writing systems where they are not the norm 41](#_Toc203989446)

[4.5.3.3. Double *kāl* in Tamil 42](#_Toc203989447)

[4.5.3.4. Independent and dependent *upadhmānīya* and *jihvāmūlīya* 42](#_Toc203989448)

[4.5.3.5. Alternative behaviour of the superscript |r| 42](#_Toc203989449)

[4.5.3.6. Multiple vowel markers within a complex glyph 43](#_Toc203989450)

[4.5.3.7. Independent vowel signs as parts of complex glyphs 44](#_Toc203989451)

[4.6. Glyphs or graphs split by an intervening feature 44](#_Toc203989452)

[4.7. Editorial interpretation 45](#_Toc203989453)

[4.7.1. Silent identification of homographs 45](#_Toc203989454)

[4.7.2. Poorly legible text 46](#_Toc203989455)

[4.7.3. Distinction of long *ē* and *ō* from short *e* and *o* 46](#_Toc203989456)

[4.7.4. Short vowel written where a corresponding long vowel is expected 46](#_Toc203989457)

[4.7.5. Sandhi analysis 46](#_Toc203989458)

[4.7.5.1. Sandhi in Sanskrit and related languages 46](#_Toc203989459)

[4.7.5.2. Sandhi in Tamil and other Dravidian languages 47](#_Toc203989460)

[5. Transliterating non-alphabetic signs 48](#_Toc203989461)

[5.1. Numeral signs 48](#_Toc203989462)

[5.1.1. The digits 0 to 9 48](#_Toc203989463)

[5.1.2. Other numeral signs 48](#_Toc203989464)

[5.1.2.1. Markup for numeral signs transliterated with more than one target character 49](#_Toc203989465)

[5.1.2.2. Signs for numbers greater than 9 49](#_Toc203989466)

[5.1.2.3. Numbers denoted by bars 50](#_Toc203989467)

[5.1.2.4. Fraction signs 50](#_Toc203989468)

[5.2. Signs with special graphematic functions 50](#_Toc203989469)

[5.2.1. *Avagraha* 51](#_Toc203989470)

[*5.2.1.1.* Editorial *avagraha* 51](#_Toc203989471)

[5.2.2. Burmese abbreviation signs 51](#_Toc203989472)

[5.3. Symbols 52](#_Toc203989473)

[5.3.1. Punctuation marks 52](#_Toc203989474)

[5.3.1.1. Transliterating punctuation marks 52](#_Toc203989475)

[5.3.1.2. Supplying punctuation 53](#_Toc203989476)

[5.3.2. Functional symbols 53](#_Toc203989477)

[5.3.2.1. Space filler symbols 54](#_Toc203989478)

[5.3.2.2. Word break indicators 54](#_Toc203989479)

[5.3.3. Generic symbols 54](#_Toc203989480)

[5.4. Blank space 54](#_Toc203989481)

[6. Segmentation in transliterated text 56](#_Toc203989482)

[6.1. Descriptive and interpretive blocks 56](#_Toc203989483)

[6.2. Segmentation versus *akṣara*s and *sandhi* 56](#_Toc203989484)

[6.3. Editorial spacing 57](#_Toc203989485)

[6.3.1. Good practice in editorial spacing 58](#_Toc203989486)

[6.3.1.1. Space and numerals 58](#_Toc203989487)

[6.3.1.2. Space and *avagraha* 58](#_Toc203989488)

[6.3.1.3. Space and punctuation marks 58](#_Toc203989489)

[6.3.1.4. Space and symbols 58](#_Toc203989490)

[6.3.1.5. Space and original space 59](#_Toc203989491)

[6.4. Editorial hyphenation 59](#_Toc203989492)

[6.4.1. Good practice in editorial hyphenation 59](#_Toc203989493)

[6.5. Segmentation guidelines 59](#_Toc203989494)

[6.5.1. Phrases 59](#_Toc203989495)

[6.5.2. Grammaticalised structures 60](#_Toc203989496)

[6.5.3. Multiple function words 60](#_Toc203989497)

[6.5.4. Repetitive structures 60](#_Toc203989498)

[6.5.5. Quasi-compounds 61](#_Toc203989499)

[6.5.6. Verbal formations 61](#_Toc203989500)

[6.5.7. Nominal compounds 62](#_Toc203989501)

[6.5.7.1. Basic compounds 62](#_Toc203989502)

[6.5.7.2. Proper names and styles 63](#_Toc203989503)

[6.5.8. Derivatives of compounds 63](#_Toc203989504)

[6.5.9. Affixes and clitics 63](#_Toc203989505)

[7. References 65](#_Toc203989506)

# Introduction

## 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 | 2 (1.1)[[1]](#footnote-1) | Revision | 2019-12 |
| Balogh, Griffiths | 3 | Revision and expansion | 2020-07-05 |
| Balogh, with Griffiths | 4 | Thorough revision | ####-##-## |

This is the final version of the DHARMA Transliteration Guide. Specific changes to the transliteration scheme since the last release (version 3) are few and minor, but the structure and conceptual background have been thoroughly revised. For these reasons, a list of specific changes since the last release is not included here.

but if shorthand is made fully private, that will need to be noted here

## 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 several 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 as and when needed for future projects.

The contents of this Guide are primarily applicable to digital editions of the text of inscriptions and manuscripts for preservation, dissemination and computer-aided research. Digital editions 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, and that the solutions recommended here gain currency beyond the scope of the DHARMA project. Section 1.6.3 gives some further pointers on what features of the transliteration system can be ignored outside diplomatic editions. We further encourage all project participants to include a reference to the current release version of the Guide (on HAL-SHS) in their publications, in order to make clear that the conventions DHARMA proposes are a published standard, and to disseminate awareness of this standard.

## Abbreviations

In addition to common abbreviations, this Guide uses:

TG for the DHARMA Transliteration Guide (the present document)

EGD for the DHARMA Encoding Guide for Documentary Editions (version 1.0)[[2]](#footnote-2)

## Brackets for linguistic notation

The concepts indicated by these brackets are introduced in §2.3.1.

/a/, /ɑː/ **slashes** indicate phonemic entities

[ɑː], [ɐ] **square brackets** indicate phonetic entities

<a>, <k> **angle brackets** indicate graphematic entities

|र|, |a| **vertical bars** indicate graphetic entities

## Terms and definitions

The following is a list of cursory definitions of script-related terminology, located here for quick reference. Concepts which are vital for working with this guide are underlined in the list. It is strongly recommended that you read the more detailed discussion of grapholinguistic terminology in §1.6 before reading the contents of this section.

* script and writing (§2.1)
  + **writing** is the graphic representation of language
  + a **script** is an inventory of graphic signs which can be used conventionally for writing
  + a **writing system** is a system of rules governing how certain aspects of a particular language can be recorded by means of a particular script
* typology of writing systems (§2.1.1)
  + a **phonographic** writing system is one which predominantly records language by representing (an abstraction of) speech sound
  + an **alphabetic** writing system is a phonographic system which represents every abstract phoneme by a visually independent graphic sign
    - an **aksharic** writing system is a phonographic system where the graphic signs representing abstract phonemes are often visually dependent on primary signs, so that only their combinations are visually independent; moreover, primary consonant signs by default also indicate an ‘inherent’ vowel
* in conversion between writing systems (§2.2.1),
  + **transliteration** is the use of a target writing system to represent how something is written in a source writing system, so that target graphemes correspond to source graphemes
  + **transcription** is the use of a target writing system to represent how something sounds in a source language, so that target graphemes correspond to source phonemes or phones
  + **Romanisation** is transliteration or transcription with Roman as a target writing system
* **graphemes** (§2.3.1, §2.3.2) are a finite set of the minimal functional units of a writing system, conceived of as signs with a graphic feature as their signifier and an abstract linguistic unit as their signified
  + **graphematics** is the study of graphemes
* **graphs** (§2.3.1, §2.4.1) are an infinite set of the concrete graphic instantiations of individual graphemes, such as |A|, |A|, |अ|, |अ|
  + **graphetics** is the study of graphs
* **allographs** (§2.3.1, §2.6) are graphs which instantiate the same grapheme
  + **graphetic allographs** are allographs which do not signify any linguistic information other than that contained in the grapheme, such as Devanagari |अ|, |अ| and |अ|
  + **graphotactic allographs** are allographs of which only one is permitted in a given graphematic context, such as the graphs corresponding to <r> in Devanagari |र्क| and |क्र|
  + **graphematic allographs** are allographs which potentially signify linguistic information in addition to, and at a different level than, that inherent in the grapheme, such as Devanagari |र| versus (|र्क| and |क्र|)
* **homographs** (§2.3.1) are identical or nearly identical graphs which instantiate different graphemes
* **polygraphs** (§2.4.2) are groups of two or more graphemes that together conventionally indicate a particular phoneme
* a **character** (§2.4.1)is the minimal graphetically autonomous unit of a writing system comprised of one or more graphemes
  + for example, <A>, <T>, <t·>, <ka>, <kha> and <rtsnyai> are characters of the Indic writing system
* a **glyph** (§2.4.1) is the concrete manifestation of a character, for example, |अ|, |त्|, |क| and |र्त्स्न्यै| are glyphs of the Devanagari script
* glyphs may be comprised of one or more graphs (§2.4.3)
  + a **simplex glyph** is a glyph comprised of a single graph, such as such as Devanagari |अ| and Bengali |ৎ| as well as Devanagari |त|
  + a **complex glyph** is a glyph comprised of two or more graphs, such as Devanagari |ते| and |क्ल|
  + the terms ‘simplex character’ and ‘complex character’ may be used as shorter expressions for ‘character manifesting as a simplex/complex glyph’, so for instance <A>, <T> and <ta> are simplex characters, while <te> and <kla> are complex characters in the Indic writing system
  + a **conjunct** or **ligature** in an aksharic writing system is a particular kind of complex character, which involves more than one consonant
* terminology for the constituent parts of glyphs, also applicable to characters (§2.4.3)
  + a **component** is a visually discernible graph within a complex glyph, such as the graphs corresponding to <t> and <e> in |ते|
  + a **marker** is a graph which can only manifest as a component, and never as an independent glyph, such as that corresponding to <e> in |ते|
    - we primarily use this term for the Indic writing system’s vowel markers and virāma, but it also includes the dependent graphs for consonants
  + a (graphic) **element** is any salient part of a graph or glyph that is not itself a graph, even though it may be the point of distinction between two graphs, such as the horizontal stroke that distinguishes Brāhmī |𑀓| <ka> from |𑀭| <ra>
    - a **stroke** is a graphic element which may be conceived of as a single stroke of the writing instrument
    - a **diacritical mark** (§2.5.1) is for our purposes nothing more than a kind of element which can be added a graph and thereby change the grapheme associated with that graph
* terminology for graphic signs signifying various kinds of information (§2.5.5)
  + an **alphabetic sign** is one that represents speech sounds in any phonographic writing system
  + a **numeral sign** or cipher is one that denotes a number
  + a **non-alphanumeric sign** is a graphic sign that is neither alphabetic nor numeric

## Working with Unicode

All text documents in the project, transliterated or otherwise, must conform to Unicode. Today, Unicode is the universal default code table for characters, so this requires no special effort. Never use a custom or legacy character encoding, as it will turn into gobbledygook whenever someone attempts to display it in a different font. For disambiguation, we often specify Unicode codepoints in the format U+####, where #### stands for a four-digit hexadecimal code.

### Fonts and supported characters

It is up to you to choose your preferred font for working in a word processor or XML editor, so long as it is Unicode-compliant. Some of the widely current fonts lack support for many of the special characters employed in our transliteration scheme, while others nominally support these characters, but render them awkwardly. When choosing a font, it is advisable to test it for some of the rarer characters relevant to the language you work with (r̥̄, m̐ and ḫ are good candidates).

Some publishers require using a specific font, which may or may not support all the special characters you require. If persuasion does not work, feel free to make any necessary compromises and substitute more widely supported characters for the problematic ones, i.e. to use public transliteration shorthand (§3.5.6.2).

Of the fonts shipped with current versions of Windows and Mac OS, Times New Roman, Tahoma, Arial and Calibri are fair choices, but we preferentially recommend one of the following free fonts:

* Gentium by SIL, <https://software.sil.org/gentium/> (in which the body text of this Guide is typeset)
* Noto Serif and Noto Sans by Google, <https://www.google.com/get/noto/>

### Entering Unicode characters

You probably already have a favourite keyboard layout to access the special characters you need in your work, but some of the characters used in DHARMA transliteration may not be covered. Detailed technical instructions on designing, installing and using keyboard layouts or assigning shortcut keys are beyond the scope of this guide, so we can only offer the following tips.

* for out-of-the-box solutions,
  + on a Mac, try the layouts Easy Unicode or ABC Extended (formerly US Extended)
  + on Windows, there is no suitable keyboard shipped with the system, 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>
* to create your own keyboard for Windows, use the Microsoft Keyboard Layout Creator, <https://www.microsoft.com/en-us/download/details.aspx?id=102134>

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
* insert them from a table of available characters
  + in MS Office, use Insert Symbol
  + on Mac OS (systemwide), use the Character Table
* copy and paste the inaccessible characters from this guide (or from a file you keep at your fingertips, listing each of those characters) each time you need one of them
* 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 enter the code with or without the 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, use private transliteration shorthand (§3.5.6.1)

### Precomposed characters

Unicode caters for Roman letters with diacritical marks in two separate ways. On the one hand, many diacritical characters — such as ü, ā or ṭ — are available in precomposed form, which means that the letter-plus-diacritic is a single Unicode character.[[3]](#footnote-3) On the other hand, diacritical marks — such as ◌̈, ◌̄ or ◌̣ — themselves are available as special combining characters which, when entered after a suitable base character, will be combined with it to render as the base character with the diacritical mark. For example, the character ā may be produced either by entering a (U+0061 Latin Small Letter A) followed by ◌̄ (U+0304 Combining Macron), or by entering ā (U+0101 Latin Small Letter A With Macron). Processing software is generally able to treat the two as equivalent (so, for instance, searching for one will find the other as well), but this cannot be guaranteed in all situations. For this reason we urge you to **always use precomposed characters** whenever they are available, and resort to combining diacritics only when the Unicode codetable does not include a precomposed form. The most common character in our practice for which no precomposed form is available is r̥; others will be noted in the body of the Guide where applicable.

The current version of Oxygen (our preferred XML editor) handles combining diacritics as separate. As a result, when you wish to delete a combined character or to select it in order to place markup around it, you will need two keypresses instead of one, otherwise stray undercircles may be left in unwanted places.

# Theoretical framework

The study of written language, like many areas of linguistics, has been approached by theoreticians in various ways, and many of the relevant technical terms refer to slightly (or radically) different concepts depending on approach. The name of this field of knowledge is itself variable, with *grammatology* and, more recently, *grapholinguistics* being perhaps the most prevalent (Meletis 2020a, 3). Many scholars of writing have worked exclusively or primarily with alphabetic writing, enabling them to ignore features of other writing systems that cannot be readily explained in their terms. We in turn are concerned specifically with the transliteration of Indic writing systems to a Roman one, and are, for pragmatic reasons, at ease to ignore subtle and complicated aspects of written language that are not directly relevant to this. Nonetheless, we must define certain concepts with sufficient accuracy in order to be able to discuss the elements of writing productively and unambiguously. The following discussion, which we recommend you read through at least once, situates the concepts relevant to us in a rough theoretical framework. For future reference, the list of quick-and-dirty definitions §1.4 shall suffice.

## Scripts and writing systems

We start from the premise that **writing** is essentially *glottography*, meaning the graphic representation of language.[[4]](#footnote-4) A **script** is an inventory of graphic signs (a *signary*), which can be used conventionally for the graphic representation of linguistic information. A **writing system** is a system of rules governing how certain aspects of a particular language can be recorded by means of a particular script.[[5]](#footnote-5) It follows from this that scripts are in themselves independent of language, and a language may be associated with several writing systems, or none. The concept of a writing system, on the other hand, presupposes both a particular language and a particular script.[[6]](#footnote-6)

We hasten to add that the word “particular” in this statement may be understood at different points on a scale of concreteness. Indeed, to speak in terms like “the system for writing English in the Roman script,” as current grapholinguistic theorists do, is already a generalisation relative to more specific systems.[[7]](#footnote-7) Accordingly, we also endorse sweeping generalisations such as **Indic script** for all of the Brāhmī-derived scripts and **Roman script** for all of the Latin-derived scripts,[[8]](#footnote-8) and likewise **Indic writing system** and **Roman writing system** understood as the generic set of rules for employing variants of these scripts for any language with which they have been associated historically. In spite of individual differences at lower points of the scale, these rules in the broad sense still share a common systemic framework and prototypical building blocks.

### Writing system typology

Writing systems typically establish correspondences between graphic signs and linguistic units of a particular kind. We shall refer to this as the *dominant level of representational mapping* (Meletis and Dürscheid 2022, 212, 216). The writing systems we are concerned with are **phonographic**, meaning that they predominantly record speech sounds.[[9]](#footnote-9) Phonographic writing systems may be further classified as alphabetic, aksharic, abjadic or syllabic. An **alphabetic** writing system — such as the Roman one — ideally matches each and every phoneme of the language to exactly one freely combinable graphic sign of the script. An **aksharic**[[10]](#footnote-10) system — such as the Indic one — is distinguished from alphabetic systems by the fact that it involves visual segmentation into chunks we shall refer to as characters (§2.4.1), each of which typically consists of a primary sign and zero or more dependent signs. The latter prominently include signs for postconsonantal vowels and secondary forms of consonant signs for use in consonant clusters. Regular characters based on a primary consonantal sign imply an inherent vowel which has no overt graphic representation.[[11]](#footnote-11)

Our primary concern here is with alphabetic and aksharic writing systems, but cursory references to other systems will be made in the discussion that follows. In a nutshell, the key feature of **abjadic** systems — as the Arabic one — is that not all of the relevant language’s phonemes are represented in writing: typically only the consonants are written. Unlike all of these, **syllabic** (or syllabographic) systems — like the Japanese kana systems — employ signs that represent clusters of speech sounds (syllables),[[12]](#footnote-12) but these signs cannot be analysed into components that individually represent individual phonemes. Writing systems other than phonographic — such as the Chinese one — have been traditionally referred to as *logographic*, ‘word-writing’ or *ideographic*, ‘concept-writing’ (Coulmas 2003, 40–41), while more recently the term *morphographic*, ‘morpheme-writing’ has gained ground.[[13]](#footnote-13)

## Conversion between writing systems

### Transliteration and transcription

For the conversion of one writing system to another,[[14]](#footnote-14) the terms ‘transliteration’ and ‘transcription’ are sometimes used interchangeably. Moreover, ‘transcription’ is widely prevalent in epigraphic and manuscript studies for the process of re-recording (using any writing system) the written text witnessed on an artefact, and in many contexts for that of writing down a text witnessed in a different medium, primarily audio. In this guide, we use (and generally encourage using) these terms in specialised senses.[[15]](#footnote-15) **Transcription** in this technical sense means using a target writing system — often but not necessarily an artificial one intended to record speech sounds unequivocally — to represent how a (written or other) text *sounds*. **Transliteration**, in turn, is the use of a target writing system to represent how a text *is written* in a dissimilar source writing system. The modern Indic name written in Devanagari as बलदेव may be transcribed (in loose Anglicisation) as Baldev or even Buldeo and (in IPA) as [bɐldeːɔ], and it may be transliterated (in the DHARMA standard) as baladeva. The term **Romanisation** can refer to either transcription or transliteration, where the target writing system is (a variety of) the Roman system. Thus, transcription is essentially concerned with the phonemes of a spoken language, without regard to how they might be written in a source script; conversely, transliteration is essentially concerned with the graphemes of a written language, without regard to pronunciation (Wellisch 1978, 18).

## The elusive grapheme

With this, we come to muddier waters. What actually is a grapheme? Wellisch[[16]](#footnote-16) describes transliteration in terms of graphemes but conveniently neglects to define that concept anywhere in his book. The grapholinguistic literature is replete with mutually incompatible, indeed often contradictory, definitions,[[17]](#footnote-17) to the extent that some leading theorists question the usefulness of this concept altogether.[[18]](#footnote-18) In our opinion the concept of the grapheme, if suitably defined, can be productive for grapholinguistics as an abstract object of scientific inquiry, and is certainly relevant for us as the basic unit of transliteration. But before we can proceed to the grapheme itself, we need to situate it in the conceptual framework of grapholinguistics.

### A sketch of grapholinguistic concepts

A language — as an abstract object of scientific study — may be conceptualised as a system comprised of multiple modules,[[19]](#footnote-19) primarily of a grammar and a lexicon. The grammatical module includes phonology (the study of speech sounds as contributing to the distinction of linguistic meaning) and morphology (the study of words and other minimal units which themselves carry meaning), as well as syntax (the functional behaviour of meaning-bearing units) and semantics (meaning and its expression in general). In the study of several such modules, a fruitful distinction may be made between an **etic** and **emic** approach. The former concerns variable concrete realisations, while the latter pertains to invariant abstractions on the basis of shared function.[[20]](#footnote-20)

We assume that our readers are familiar with the basic concepts of phonology, so we begin by summarising these as an entry point. The concrete speech sounds constituting spoken language come in a practically infinite continuum. These etic sounds are referred to as **phones**, and the study of their production (articulation) and description (physical characteristics) is called **phonetics**. According to widespread convention, we use square brackets to represent phones in written discussion, e.g. [ɐ].[[21]](#footnote-21) The phones of any language can be sorted into a relatively small number of abstract entities on the basis of their function, namely their role in distinguishing semantic meaning. These emic sound classes are termed **phonemes**, and the field of their study is called **phonemics**.[[22]](#footnote-22) To represent phonemes in written discussion, we enclose them in slashes, e.g. /a/.[[23]](#footnote-23) The different phones which may manifest a given phoneme are known as **allophones**. Thus, the sound corresponding to the letter |o| in the English word “go” may be realised as the phone [əʊ] by a Brit, as [oʊ] by an American, and as [ɔː] by a Scot or an Indian, and in actual speech each instance of these semi-abstract phones will have slightly different actual physical characteristics. All of these sounds are allophones of the same phoneme, as they distinguish meaning in the same way, for example in making “go” a different word than “gay” or “goo”. **Homophony** is the situation where two linguistic units (e.g. words) are identical with regard to phonemes and/or phones, yet have different meanings. For example, the English words “horde” and “hoard” are homophonous. **Phonotactics** is (the study of) the rules governing how phonemes may combine into sequences in a particular language.

In grapholinguistics, the etic approach — known as **graphetics** — is concerned with **graphs**, which are infinitely variable concrete graphic signs that can be used for the representation of linguistic information. In written discussion, graphs are enclosed in vertical bars, e.g. |a|.[[24]](#footnote-24) In the emic aspect, the field of **graphematics**[[25]](#footnote-25) studies **graphemes**, a limited number of discrete abstract entities systematised on the basis of their role, namely the nature of the linguistic information signified by them. In discussion, we enclose graphemes in angle brackets, e.g. <a>.[[26]](#footnote-26) Various graphic signs which signify the same linguistic information are known as **allographs**. Thus, in the Roman writing system, the graphs |a|, |*a*| or |a| are allographs of the grapheme <a>, since they all are conventionally associated with the same linguistic information. **Homography** is when two graphic entities are visually identical, but have different significations. On the level of words, for instance, English ‘read’ (present tense) and ‘read’ (past tense) are homographs, as they consist of the same graphemes, but have different meanings. More important to us is homography on the level of graphs — as in Roman |l| (lowercase L) and |I| (uppercase i) or |O| (the letter O) and |0| (the digit zero) — which have a very similar or fully identical graphic manifestation, but carry different linguistic information, and are thus different graphemes.[[27]](#footnote-27) **Graphotactics** is (the study of) the rules governing how graphemes may combine into sequences[[28]](#footnote-28) in a particular language.

Continuing the above modular model, a writing system’s two principal constituents are the graphematic and the graphetic module. Many, but not all, writing systems include in addition an optional orthographic module, which is essentially a set of norms restricting how graphemes can correspond to linguistic units.[[29]](#footnote-29) The graphematic module links the writing system to a module of the language system corresponding to the writing system’s dominant level of representational mapping, i.e. in the case of phonographic writing systems, to the phonological module. The graphetic module, conversely, is a component of the writing system that is independent of the language system.

Thus, the Sanskrit phoneme /a/ — which may be phonologically instantiated by several allophones such as [ɐ], [ɔ] or [ə] — is the linguistic information signified by the grapheme <a>. In the writing system that associates modern Devanagari with Sanskrit, this grapheme can be instantiated by several allographs such as |अ| or |अ|. In the DHARMA transliteration scheme, which is itself a writing system for Indic languages, the same grapheme can be instantiated by allographs such as |a|, |*a*| or |a|.

### Refining the concept of the grapheme

The above sketch is sufficient for an ideal alphabetic writing system. However, in actual writing systems, one-on-one correspondence between graphic signs and phonemes is rare, and — especially in non-alphabetic system — the visual segmentation of graphic signs and the functional segmentation of graphemes do not necessarily overlap. In this section, we present a definition of the grapheme that can be productively applied in the context of the Romanisation of Indic script. Our approach has much in common with that advocated by Meletis (e.g. 2019), but we have adapted some of his notions in altered form, and flatly disagree with others.[[30]](#footnote-30) Discussion of where, how and why we agree or disagree with him and other grapholinguists will be kept to a minimum here in order to be able to proceed as quickly as possible to our actual subject matter.

For our purposes, a **grapheme** is a minimal functional unit of a writing system,[[31]](#footnote-31) meaning that a grapheme cannot be subdivided into smaller units which themselves satisfy the definition of a grapheme. The grapheme may be conceptualised as a dyadic sign[[32]](#footnote-32) comprised of a linguistic aspect as its signified and a graphic aspect as its signifier.[[33]](#footnote-33) The linguistic aspect is comprised of abstract information that theoretically pertains to the presence of a linguistic unit on the writing system’s dominant level of representational mapping. The graphic aspect is comprised of any graphic feature that unconditionally signifies a certain item of linguistic information by the conventions of the writing system in question. In the following subsections, we elaborate the terms of this definition and discuss its implications.

#### The linguistic aspect of the grapheme

Many features of writing can have a graphematic function, i.e. signify some sort of linguistic (or arguably linguistic) information. In order to limit the inventory of graphemes to a manageable set, we restrict the scope of the term ‘grapheme’ to sign relations whose linguistic aspect pertains to the presence of linguistic units according to the writing system’s dominant level of representational mapping.[[34]](#footnote-34) Thus, the graphemes of a phonographic writing system are limited to those that signify information pertaining to the presence of speech sounds. In addition to these primary or default graphemes, we of course recognise the existence of other graphematic features, which may be graphically smaller than graphemes (e.g. diacritical marks), coterminous with graphemes (e.g. bold styling), or larger than graphemes (e.g. word underlining, paragraph indentation), and the linguistic information they convey may be at any level from phonological features (e.g. vowel length, locus of articulation, stress accent) to word level (e.g. emphasis) and beyond (e.g. paragraph structure).[[35]](#footnote-35) These other graphematic units can be relegated to a separate stage of analysis,[[36]](#footnote-36) and are generally not relevant to our topic.[[37]](#footnote-37)

We must, however, emphasise that what a grapheme signifies is not the actual linguistic unit — such as a phoneme — but rather some prototypical idea of a linguistic unit. In a writing system whose dominant level of representational mapping is that of phonology, the graphemes contribute information pertaining to the presence of speech sounds, but this information may often be under-specific (referring to a set of phonemes which typically share some phonological features) or over-specific (referring to a particular allophone). Moreover, the reconstruction of actual speech from writing is a complex process requiring input from other graphematic units and from context, as well as reciprocal interaction with higher levels of language processing (such as morphology and lexicon). Therefore, the study of grapheme-phoneme correspondence is a subfield separate from the study of graphemes themselves.[[38]](#footnote-38) Depending on this complex process, the phoneme theoretically represented by any individual grapheme may be realised in speech as a very different phoneme, or not realised at all. When for the sake of brevity we speak of the graphemes of phonographic writing systems as signifying phonemes, we always mean that they signify abstract prototypical phonemes, which often do not correspond with any consistency to the phonemes of the language being represented.[[39]](#footnote-39)

#### The graphic aspect of the grapheme

Our definition does not require the graphic manifestation of a grapheme to be *segmental* (isolable in the linear flow of the script) or *autonomous* (independently combinable).[[40]](#footnote-40) Thus, the components of a complex Indic akṣara are genuine graphemes, since they signify the same kind of linguistic information as the corresponding independent signs (Meletis and Dürscheid 2022, 130). The fact that some of these components cannot occur on their own — i.e. that they are graphetically dependent on the primary forms and usually occupy no separate segmental space — has no bearing on their grapheme status.[[41]](#footnote-41) Graphemes are functional units of writing which do not necessarily coincide with the segmental and autonomous formal units of a writing system. We shall refer to the latter as characters, to be discussed in §2.4.1.

Additionally, we do not limit the graphic aspect of a grapheme to an overt, visually isolable unit of script,[[42]](#footnote-42) but extend it to include any graphic feature that unconditionally signifies a certain item of linguistic information (of the abstract kind discussed above), provided that it does so solely by the conventions of the writing system in question, i.e. without requiring interaction with other modules of the language system, such as the lexicon. Many Indic writing systems use graphic signs — such as Devanagari |क्ष| <kṣa> — that represent a sequence of graphemes in such a way that no isolable graphic components correspond to its constituent graphemes. We advance[[43]](#footnote-43) that the relevant graphemes are individually present in spite of their lack of graphic distinctness, i.e. that we are not dealing with a single syllabic grapheme. Our rationale is that, unlike syllabic writing systems, aksharic systems refer to (abstract) phonemes at their dominant level of representational mapping, and by default use complex signs comprised of isolable components. Signs which holistically correspond to a phoneme sequence are exceptions to the rule, and are moreover palaeographically derived from earlier signs where the components can in fact be isolated.

The above discussion anticipates a further problematic detail: what is the status of the inherent vowel (typically, /a/) of Indic akṣaras? From the way we have phrased the requirement for the linguistic and graphic aspects of a grapheme, it follows that the inherent vowel is also a bona fide grapheme, since its presence is unequivocally signified by a graphic feature: an integral consonant-based akṣara without any additional components that would negate that vowel or replace it with a different one.[[44]](#footnote-44) For the purpose of analysis, we consider it best to recognise the inherent vowel as a grapheme, notwithstanding the fact that the covert graphic manifestation of the inherent vowel is not palaeographically derived from an earlier overt graphic sign, and that the shape of an akṣara consisting of a consonant and its inherent vowel is (as a rule) identical to the shape of that consonant graph when combined with a different vowel.[[45]](#footnote-45)

In summary, Devanagari |क्ष| <kṣa> corresponds in our view to not one, not two, but three graphemes, <k>, <ṣ> and <a>. This conclusion may be somewhat surprising, but it is most expedient for the analysis of transliteration, and we feel that it may also be relevant to graphematics in general.

## Graphic structures and their elements

Having arrived at a concept of the grapheme that appears to be feasible in general and practicable for our purposes, we now need to investigate the status of salient graphic entities which are not graphemes by our definition. We begin by introducing the concept of the character for the graphetically autonomous units of writing, and then work upward and downward from that level. The terms and definitions we advance here may or may not be relevant to graphematics at large, but they are certainly useful for the analysis of aksharic writing. We therefore encourage you not only to read the definitions in order to understand the instructions of this Guide clearly, but also to use these terms — in the senses defined here — in your own philological and palaeographic discussions.

### Characters and glyphs

We define a **character** as a minimal graphetically autonomous unit of a writing system, meaning that the character cannot be broken down into components that can also manifest independently according to the rules of a writing system. Characters are graphematic (i.e. emic) units delineated on the basis of graphetic segmentality.

The term ‘character’ is widely used in a vague non-technical sense, and is rarely defined in a scholarly manner.[[46]](#footnote-46) Importantly, ‘character’ as defined in information technology, specifically in the Unicode standard,[[47]](#footnote-47) is *not* what we define as a character. The grapholinguistic literature rarely uses this term,[[48]](#footnote-48) and tends to deal with the character — as defined by us — under the blanket term ‘segment’. In alphabetic writing systems, characters (as a rule) coincide with graphemes,[[49]](#footnote-49) but this is not necessarily so in other types of writing systems.[[50]](#footnote-50)

For the etic counterparts of characters — their actual graphic manifestations — we shall use the term **glyph**. The concept of glyphs is thus related, but not identical, to that of **graphs**, which we define in a stricter sense as the etic counterparts of (individual) graphemes. Where the distinction between glyphs and graphs is irrelevant, we use the umbrella term **graphic sign**.[[51]](#footnote-51) Thus, in our terms, Devanagari |अ| and Roman |A| (corresponding to the character <A> in their respective writing systems) are simultaneously glyphs and graphs, while the Devanagari glyph |का| (corresponding to the character <kā>) is constituted of two graphs (|क| and |ा|), which individually correspond to the graphemes <k> and <ā>. As is the case with ‘character’, the term ‘glyph’ is rarely used in a technical sense and, when used, is rarely distinguished from ‘graph’.[[52]](#footnote-52)

### Polygraphy

Many writing systems employ established sequences of graphs with a graphematic function, such as English sh, which corresponds to the phoneme /ʃ/ with great consistency. Such combinations are technically known as **polygraphs** (and as **digraphs** when comprised of two members), and have been recognised as graphemes in some approaches.[[53]](#footnote-53) Our definition of the grapheme precludes this: polygraphs are not minimal and can always be broken down into individual graphemes, whose context-dependent joint correspondence to a phoneme is a matter of grapheme to phoneme mapping. Polygraphs are common in alphabetic writing systems, but rare in other phonographic systems.[[54]](#footnote-54) The DHARMA transliteration scheme employs polygraphs for certain source graphemes, the practical consequences of which will be discussed in §3.4.1.

### Glyph complexity

As outlined above, Indic akṣaras can (as characters) incorporate one or more graphemes, and can (as glyphs) be composed of one or more graphs. These two kinds of complexity — graphematic and graphetic — do not necessarily overlap, because the inherent vowel, though a grapheme, never manifests overtly as a graph. When discussing how the graphs of an Indic writing system are composed into glyphs (and accordingly, how the corresponding graphemes cluster into characters), we find it most useful to prioritise graphetic complexity, as follows.

A **simplex glyph** is one that is comprised of a single graph, which may correspond to a single grapheme or may — on the akṣara level— indicate the presence of the inherent vowel in addition to a consonant grapheme. This includes independent (typically initial) vowel signs such as Devanagari |अ| (independent <A>), and independent (typically final) consonant signs such as Bengali |ৎ| (final <T>), as well as basic consonant signs comprising an akṣara with their inherent vowel (e.g. Devanagari |त| <ta>).

Conversely, a **complex glyph** is one that is comprised of two or more graphs, such as Devanagari |ते| <te> and |क्ल| <kla>.[[55]](#footnote-55) This class includes every kind of glyph not classified above as simplex. Glyphs which represent several consonant graphemes without isolable graphic components, such as Devanagari |क्ष| <kṣa> discussed in §2.3.2.2 above, are a special case in this regard. Arguably, they are graphetically simplex, yet as already noted, they are exceptions to a rule and derive palaeographically from modifications of graphically complex glyphs. We therefore classify them with complex glyphs, and note that in any case, the notion of complex glyphs tends to be relevant only in cases where graphetic complexity is unequivocally present. A complex glyph is thus not the same as a **conjunct** or **ligature** in an aksharic writing system,[[56]](#footnote-56) since a conjunct includes two or more consonant components, while a complex glyph may also consist of a consonant and a vowel marker.

We have provided some actual examples of simplex and complex glyphs above, but generally when discussing the Indic writing system, and especially when the script involved cannot be easily represented in a digital document, we find it more practical to represent them in transliteration, i.e. in their graphematic aspect as characters. That is to say, we endorse the use of simplified phrases such as “the Devanagari glyph <ta>” or even “the complex character <rdha>”, where the former would be meticulously put as “the Devanagari glyph manifesting the character <ta>”, and the latter as “the character <rdha>, which is represented by a complex glyph”.

#### Glyph components

We use the term **component** for a visually discernible graph that is part of a complex glyph.[[57]](#footnote-57) As above, we endorse describing components in graphematic terms even though they are graphetic phenomena. A palaeographic description or a discussion of reading difficulties may thus mention, for instance, “the <dh> component of the glyph <rddhe>” (where a meticulous phrasing would be something like “the component which instantiates the grapheme <dh> in the glyph manifesting <rddhe>”).

#### Markers

We employ the term **marker** for a special set of graphs that can only appear in a graphetically dependent position, i.e. which can only manifest as components and never as glyphs. ‘Marker’ is not an established term in grapholinguistics,[[58]](#footnote-58) but is defined here in agreement with Ollett and Taylor (forthcoming). In actual usage we find ‘marker’ most useful for referring to graphs which modify the default vowel of an akṣara (i.e. to the dependent vowel signs and the virāma, for which see §2.5.2), and prefer to speak of consonantal markers as dependent consonants.[[59]](#footnote-59)

#### Graphic elements

The scope of the term ‘component’, as defined above, is restricted to graphs, i.e. to manifestations of graphemes. A single graph may itself be graphically quite complex, for instance Devanagari |औ| <au>. For any visually discernible constituent parts of individual graphs, i.e. for salient shapes which are not themselves graphs, we recommend the term **element** or, for extra clarity, graphic element. Elements can be distinctive features of graphs — as the horizontal strokes that distinguish Brāhmī |𑀓| <ka> from |𑀭| <ra> or Roman |E| from |F| — but they do not in themselves carry linguistic information. Much as glyphs may be comprised of one graph or several graphs, graphs too may be comprised of a single element (as the graph for <e> in Devanagari |के|) or made up of several elements, which may be connected (as in the graph for <k> in Devanagari |के|) or unconnected (as in the graph for <o> in Tamil |கொ|).[[60]](#footnote-60)

We further suggest the term **stroke** for elements which are uninterrupted lines and may be conceived of as a single stroke of the writing instrument. To refer to particular elements of a graph, chiefly in palaeographic description or a discussion of reading difficulties, we encourage the use of intuitive biological and architectural analogues such as arm, leg, wing, tail, stem, lobe, arch, base, etc.

## Problematic cases: what is a grapheme, and what is not?

### Diacritical marks

**Diacritical marks** (or, in short, diacritics) are, in our terms, a special class of element (##) which can combine with specific (individually meaningful) graphs to alter their connotation. Most diacritical marks can be added to a range of graphs and often (though not always) alter the connotation of those graphs in a somewhat consistent manner (for instance, in the DHARMA transliteration, the macron makes vowels long, and the underdot makes consonants retroflex). Diacritical marks may thus be perceived as independently combinable bearers of linguistic information, giving rise to the question whether they should be considered graphemes in their own right. While we acknowledge that diacritical marks possess a graphematic function, and that this deserves attention, we emphasise that they are not graphemes by our definition, since the information they represent does not pertain to the *presence* of a particular linguistic unit on the writing system’s dominant level of representational mapping, but rather to some *feature* of a unit whose presence is signified by a proper grapheme. We emphasise that the markers of the Indic writing system (as defined in §2.4.3.2 above) are *not* diacritical marks by our definition, even though secondary literature often refers to them as such.

Diacritical marks thus have no special relevance to transliteration: like any distinctive element, they are only relevant to us inasmuch as they distinguish one graph from another.[[61]](#footnote-61) If a source grapheme is distinguished by a diacritical mark from another source grapheme, then the distinction itself must be preserved in transliteration, but it need not involve a diacritical mark, or indeed, any feature corresponding to the diacritical mark.[[62]](#footnote-62)

### Virāma

A moot point we have been carefully avoiding is the status of the virāma, the “vowel killer” sign of aksharic writing systems. The virāma negates the inherent vowel of an akṣara, contributing nothing positive regarding a phoneme, so it is difficult to recognise it as a grapheme. Simultaneously, it is also difficult to recognise as a diacritical mark, as it does not serve for the distinction of one grapheme from another: an akṣara without a virāma is comprised of at least two graphemes, and if a virāma is added to the same akṣara, one or more of those graphemes remain[[63]](#footnote-63) and one ceases to be present, but none are changed to a different grapheme.

All things considered, we are willing to grant grapheme status to the virāma, and have phrased our definition of the grapheme accordingly: we require a grapheme to represent “information pertaining to the presence” of linguistic units, where this information may be that a unit is *not* present even though its presence has been indicated by another grapheme (that manifest in the integral akṣara). The rationale is admittedly tortuous, but it is essential for a transliteration scheme used in philology and palaeography to make a distinction between a vowelless consonant expressed using a virāma and one expressed in a different way (see also §3.2.2).[[64]](#footnote-64) In further support for recognising the virāma as a grapheme we emphasise that functionally it is fully analogous to vowel markers, which are definitely graphemes in their own right. A basic akṣara signifies the presence of an inherent vowel, a vowel marker signifies the presence of another vowel in place of that inherent vowel, and the virāma signifies a zero-vowel in place of that inherent vowel.[[65]](#footnote-65)

It follows from the above that from a grapholinguistic perspective, a basic consonant graph combined with a vowel killer is an independent and integral character (akṣara), even if it is employed in a non-final position as an alternative to part of a ligature (as e.g. in Devanagari |द्‍म| <dma> or Tamil |க்க| <kka>), and thus does not correspond to a phonological syllable.[[66]](#footnote-66)

### Anusvāra relatives

The typical usage of the anusvāra is to indicate the presence of a nasal consonant, whose locus of articulation is determined by the subsequent context. Used in this manner, the anusvāra is beyond doubt a grapheme by our definition. However, in certain languages in certain contexts, it can also be a diacritical mark representing the nasal quality of a vowel. Some standardised modern writing systems employ different graphic signs (the bindu |ं| and the candrabindu |ँ|) for these purposes, but such a distinction is by no means universal.

In the diplomatic transliteration of primary texts, we wish to make no assumptions about whether such signs were employed by their writers to represent a phoneme or to qualify a phoneme represented by another graph, and prefer to err on the side of caution. An additional rationale for doing so is that anusvāra variants may be perceived as always representing an abstract nasal phoneme, and the fact that it is sometimes realised as a nasalisation of the preceding vowel phoneme may be viewed as an issue of complex grapheme to phoneme correspondence. Therefore, our general approach is to treat all anusvāra variants as graphemes on their own right, and thus to represent them with a corresponding target grapheme.[[67]](#footnote-67)

### Other signs of vague status

When in a particular variant of the Indic writing system the status of a graphic sign vis-à-vis graphemes is poorly understood or difficult to establish in spite of being well understood, we — as in the case of anusvāra variants — we prefer to err on the side of caution and recognise such signs as “honorary” graphemes. This approach allows us to represent the problematic sign with a single target grapheme, instead of having to introduce a series of target graphemes for each combination of the problematic sign with a basic grapheme. An example of a poorly understood graphic sign is the underdot in Mon, Pyu and Burmese (#§), while the Gurmukhi addak may exemplify signs which are well understood but difficult to classify.[[68]](#footnote-68) The DHARMA transliteration scheme does not include a specific provision for the addak, and is open to extension with additional target graphemes (often already conventionalised in the relevant subfield) for use in any linguistic or regional context that we have not yet catered for.

### Non-phonographic signs

Texts written in a phonographic writing system often include graphic signs which signify non-phonographic information. The most common examples of such signs are punctuation marks, which primarily convey information about the syntactic, semantic and/or prosodic segmentation of the text. There are also signs for abstract concepts: ciphers signify numbers, and certain signs signify miscellaneous concepts, such as the section sign, the percent sign or currency signs. Some may pertain to language in an abstract way, like the avagraha of the Indic writing system or the apostrophe of the Roman one, which are often used to indicate elision. Others may be quite beyond the domain of glottography, such as “auspiciousness” presumably conveyed by many opening and closing symbols used in inscriptions or “irony” conveyed by a winking smiley in a text message.

Whether such signs are to be considered graphemes — and if yes, exactly how the grapheme ought to be defined to accommodate them — is a vexed question (cf. Meletis and Dürscheid 2022, 132–33). We feel that the best way to reconcile such signs with the framework presented above is to understand them as graphemes belonging to one or more secondary writing systems, which are separate from (though generally correlated with) the primary writing system used for a particular language-script pair, and which are not in themselves capable of representing any language in its entirety. But whatever their exact grapholinguistic status, all such signs need to be reflected somehow in transliteration.

It would, however, not be practicable to give full justice to the practically infinite variety of such signs by transliterating each with a different target grapheme. We therefore classify written signs along the following lines.[[69]](#footnote-69) An **alphabetic sign** or alphabetic grapheme is one that represents speech sounds in any phonographic writing system. Thus, ‘alphabetic’ in this phrase does not imply an alphabetic writing system; in fact, we prefer referring to the alphabetic signs of alphabetic systems as letters. All other graphic signs are non-alphabetic signs. Among these, a **numeral sign** or **cipher** is one that denotes a number, transliterated as per #§. Alphabetic and numeral signs together are referred to as **alphanumeric signs**, and all others are non-alphanumeric. Among the latter, we recognise some special signs including the avagraha (§5.2), and refer collectively to all other signs as **symbols** (§5.3, q.v. for further classification). Our transliteration scheme comes close to giving full justice to alphabetic signs, but becomes increasingly insensitive as we progress away from phonography. Some of the variety of non-phonographic signs may, however, be captured in the encoding of our digital editions.

### Fuzzy segmentation

When a writing system in the broad sense is considered with an extensive diachronic and/or synchronic scope, the class of complex characters inevitably exhibits a smooth transition to character sequences on one side and to simplex characters on the other. Put in terms of graphic entities, character components cannot always be clearly distinguished from independent characters on one side, and from graphic elements on the other. Such fuzzy boundaries are more prominent in alphabetic writing systems,[[70]](#footnote-70) but they can also occur in the Indic system, and require either special cases applicable to the writing systems of particular languages, or an arbitrary blanket decision. As a general rule of thumb for the latter, category assignment may be based on the “dominant level of graphic segmentation”, much as we have defined graphemes above on the basis of the dominant level of representational mapping. Thus, for borderline cases in the Indic writing system, we tend to assume by default that we are dealing with a complex glyph consisting of multiple graphs, rather than with a simplex glyph involving merely distinctive elements.

#### Character or component?

The status of the visarga and the anusvāra (along with its variants) remains open: they are traditionally considered to be dependent on akṣaras and can never occur in an initial position; nonetheless, they are always graphetically distinct, and can in many specific writing systems occupy a separate segmental space of their own, be physically separated from their “parent” akṣara by a physical feature, or even be graphetically associated with the following character. For the former reason, they are thus arguably components in our terms (forming a single complex character with the preceding akṣara), while for the latter reason it may be argued that they are independent simplex characters.[[71]](#footnote-71) As far as transliteration is concerned, the question is irrelevant: in either case they are graphemes in their own right, and are thus represented in transliteration by an equivalent grapheme, which in turn is always an independent segment in transliteration.

#### Component or element?

The crucial difference between components (i.e. graphs) and elements (i.e. graph constituents) is that the graphs making up a complex glyph are individually “meaningful”,[[72]](#footnote-72) whereas the elements making up a multi-element graph are not. Most sign inventories, however, utilise some recurring graphic elements that can occur either alone or in various combinations to constitute different graphs. Importantly, even if such an element can comprise a graph (i.e. manifest a grapheme) on its own (such as Devanagari |आ| <Ā>, |े| <e> or |ा| <ā>), it is still merely an element when it combines with other elements to constitute a different graph (such as Devanagari |ओ| <O> or |ो| <o>).[[73]](#footnote-73) In other words, Devanagari |ओ| is graphematically simplex in spite of its graphetic complexity. It is not a combination of the graphemes <Ā> and <e>, merely of elements which, incidentally, manifest those graphemes when they are not combined with other elements.[[74]](#footnote-74) Analogous dual patterning is present in other aspects of language as well: some morphemes (i.e. meaningful[[75]](#footnote-75) units) consist of a single phoneme (e.g. English “I” or “o!”), but the same phonemes (in themselves meaningless) can be constituents of other morphemes (e.g. “hi” or “go”).

That said, some phenomena in some of the specific writing systems we are concerned with resist a clear classification as either (graph) elements or (glyph) components. Above (§2.5.2) we have accorded grapheme status to the virāma, but this grapheme appears to have evolved out of strokes whose status is ambiguous. We discuss good practice pertaining to these proto-virāmas in §4.3.1. In some Southeast Asian writing systems, we see the evolution of the original vowel marker for the grapheme <ā> into a diacritical mark signifying length in association with any vowel,[[76]](#footnote-76) and even with consonants. The transliteration of this usage is treated in §###. Also in specific Southeast Asian systems, the original graph for the independent vowel <A> has evolved into a sign representing a “zero consonant” or a “minimal consonant” (a glottal stop or a /h/-like glottal fricative).[[77]](#footnote-77) In this role, it can function as a *vowel support*, taking on vowel markers in combinations which replace the original independent vowel graphs, as described in §###. Moreover, in some writing systems it can even form conjuncts with consonant graphs, indicating that it was perceived by the users of those systems as a fully fledged consonantal grapheme, addressed in §###.

## Revisiting allography

By our definition, different graphs that signify the same linguistic information as far as the dominant level of representational mapping (i.e. in our case, the level of abstract phonemes) is concerned cannot be considered different graphemes. Therefore, the graphs expressing the grapheme <r> in Devanagari |र| (<ra>), |र्क| (<rka>) and |क्र| (<kra>), those expressing <u> in Devanagari |उ| (<u>), |कु| (<ku>) and |रु| (<ru>), and those expressing <t> in Bengali |ত| (<ta>) and |ৎ| (final <t>) must be deemed allographs. We feel, however, that the alternation of such forms is different, and more fundamental, than the “simple” allography exemplified by the alternation of Devanagari |अ|, |अ| and |अ| (for <a>) or |झ|, |झ| and |झ| (for <jha>). In other writing systems, allography of the more fundamental kind may be exemplified by Roman |a| and |A|, Greek |σ| and |ς| (allographs of the grapheme <s>, the former being used in medial positions, the latter in final positions), or the isolated, initial, medial and final forms of Arabic script. The distinction might be grasped in terms of supplementary linguistic information carried by the graphs in question. We have restricted the linguistic information of a grapheme (in the strict sense) to linguistic units at the dominant level of representational mapping, but we have recognised that other kinds of linguistic information may also be expressed by various elements of writing, which thus also have a graphematic role.

Accordingly, we advance that allography may be viewed as being of three kinds: graphetic, graphotactic and graphematic.[[78]](#footnote-78) **Graphetic allographs** are alternative graphs or glyphs[[79]](#footnote-79) which do not signify any linguistic information other than that contained in the grapheme, as in Devanagari |अ|, |अ| and |अ| or |झ|, |झ| and |झ|.[[80]](#footnote-80) **Graphotactic allographs** are alternative graphs of which the rules of the writing system permit only one in a given graphotactic context, as in the <r> of |र्क| <rka> and |क्र| <kra> or the <u> of |कु| <ku> and |रु| <ru>.[[81]](#footnote-81) Such allographs carry implicit information about the grapheme’s context, i.e. linguistic information other than that present in the grapheme proper. This information may provide clues for reconstructing the context of a legible graph in a damaged inscription, but the writer does not normally have the option of substituting one graphotactic allograph for another. Finally, **graphematic allographs** are alternative graphs which potentially signify linguistic information in addition to, and at a different level than, that inherent in the grapheme, as in Devanagari |र| versus (|र्क| and |क्र|) or |उ| versus (|कु| and |रु|), or in Roman |a| and |A|.[[82]](#footnote-82) This choice is determined (or at least influenced) by linguistic context, but not by the graphematic context. The writer can deliberately employ one allograph or the other in order to express in writing an aspect of language that is not conveyed in the graphemes themselves. Consider the Sanskrit words kr̥tam etat written in Devanagari once as |कृतम्एतत्| and once as |कृतमेतत्|, or in early Brāhmī as |𑀓𑀾𑀢𑀫𑀋𑀢𑀢| (involving the final allograph of <m> and the initial allograph of <e>) and as |𑀓𑀾𑀢𑀫𑁂𑀢𑀢| (involving the akṣara <me>).[[83]](#footnote-83) Both are legitimate alternatives in either writing system, but the former in both cases carries additional linguistic information: it implies a pause (and hence, a syntactical or semantic boundary analogous to that expressed by punctuation) between the two words. Likewise, the use of a Roman capital letter instead of a lowercase one can, among other things, indicate the beginning of a sentence or the beginning of a proper name.

As usual, our definitions leave us with fuzzy borderline cases. To be sure, practically any graphic feature can be a *potential* carrier of additional linguistic (or sort-of-linguistic) information, and thus have a graphematic role. Switching to a different graph inventory — such as italics for emphasis in a Roman text, Grantha for a Sanskrit word in a Tamil text, or a more ornate script for the royal signature on an Indic inscription — often does have a graphematic function. With Meletis (2020b, 256–57) we consider suprasegmental variation, i.e. the use of a different inventory for chunks larger than a graphetic segment (as in all of the above examples), to be outside the domain of allography.[[84]](#footnote-84) Epigraphic examples of allographic variation on or below the level of segments include the use of alternate glyphs (e.g. a northern-class Middle Brāhmī <ma> in one spot while other instances of that grapheme in the same inscription are southern-class glyphs; or the use of a cursive form of <lo> while other instances of that grapheme involve a separable consonant base and vowel marker) and the use of alternate glyph composition (e.g. the conjunct <rya> written with the body allograph of <r> and a subscript <y> instead of the standard superscript <r> and body <y>). The semantic implications of such variation, if any, are vague and unlikely to be expressible in terms of language. On the whole, therefore, the only phenomenon we recognise as graphematic allography in the Indic writing system is the alternation of in-akṣara allographs with independent allographs.

# General principles of the DHARMA transliteration scheme

## Compatibility with other transliteration systems

Aside from increasingly obsolete ASCII-based transliteration schemes optimised for computer processing, the only internationally prevalent schemes for the Romanisation of Indic scripts are ISO-15919 and IAST (the International Alphabet of Sanskrit Transliteration).[[85]](#footnote-85) The latter is the most widely used among Sanskritists, but provides no distinction for some graphemes associated with other Indian languages. For this reason, the DHARMA transliteration scheme is based on the ISO-15919 scheme, extending it in order to handle graphemes of specific writing systems not covered in the standard, and modifying it in some technical details in order to allow for the representation of certain graphematic allographs (§3.2.2) and editorial markup (§3.5). Among writing systems not or not wholly covered by ISO-15919, our transliteration scheme is by and large compatible with current recommendations for the transliteration of Kannada (Ollett and Taylor forthcoming) and Burmese (Lammerts and Griffiths 2016).

If you are primarily used to working with IAST, keep in mind a key difference between that scheme and ISO-15919 (as well as the DMARMA scheme): namely that the former employs underdotted ṃ for the anusvāra and underdotted ṛ, ṝ and ḷ for vocalic consonants, while in the latter, anusvāra is transliterated ṁ with a dot above, and the vocals r̥, r̥̄ and l̥ with an undercircle.

## Graphematic entities in transliteration

### Transliterating graphemes

As stated in §2.2.1, the basic unit of transliteration is the grapheme, and the aim of transliteration is to faithfully reflect the graphemes of the source. That said, transliteration systems in actual practice include a number of transcription-like features, at least for the sake of pronounceability and ease of acquisition.[[86]](#footnote-86) Moreover, transliteration cannot practicably aim to represent each and every graphic detail of the source. 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 particular, a transliteration system for philological and epigraphic purposes should strive toward “a one-to-one relationship between each grapheme in a target script and a corresponding […] grapheme in the source script” (Wellisch 1978, 322) in order to be fully reversible.[[87]](#footnote-87)

Establishing a one-to-one relationship between source and target graphemes means not only that no graphemes must be added or dropped in the process of transliteration, but also that the specific identity of graphemes must be preserved. A particular target grapheme must correspond to a particular source grapheme rather to any of a set of related source graphemes, and vice versa.

### Representing complex characters and allographs

The Indic writing system by default employs complex characters for the representation of certain phonological sequences, while the Roman writing system of our transliteration lacks complex characters. Which adjacent graphemes combine into a complex character, and how that complex character is formed, is largely determined by the inbuilt rules of the Indic writing system. These rules, however, are somewhat malleable, and therefore our transliteration scheme provides some measures for reflecting exceptions to the general rule in order to allow for a more accurate diplomatic reconstruction of the originally inscribed source.

Most importantly, the independent graphematic allographs (as defined in §2.6) of vowel and consonant graphemes universally have the potential to represent important linguistic information (typically the presence or absence of a pause) in addition to that inherent in the grapheme. Since the alternation of Indic independent graphs and in-akṣara graphs is superficially parallel to the alternation of uppercase and lowercase Roman letters, the latter can be conveniently used to represent the former in transliteration (§3.3.1).

The Indic writing system at large has strictly deterministic rules for the grapheme sequences that can combine into a complex character, and for the selection of graphotactic allographs (§2.6) depending on position within a complex character. However, some specific varieties of the Indic system deviate from these rules more or less systematically, and idiosyncratic deviations also occur in specific written source texts. To be able to give these variations some degree of justice in transliteration, we use text-based (§4.5.3) or XML-based (EGD §###) editorial markup to indicate which target graphemes belong to a single complex source character, when this is not straightforward on the basis of the general rules of the writing system. Only XML markup (EGD §###) can reflect situations where the graphetic allograph employed is not the one expected in a given position (as in Figure 3.2.A/1, where <rya> is written with a body <r> and a subscript <y> instead of a superscript <r> and a body ><>), and where the spatial configuration of a ligature is unusual (as in Figure 3.2.A/2, where a prescript <g> has been joined to the left of a body <gh> instead of joining a subscript <gh> to a body <g>).

Attempting to reflect graphetic allography (§2.6) in transliteration would complicate the transliteration scheme by an order of magnitude, without much tangible gain in return. Therefore, we have deliberately chosen to ignore graphetic allographs in our transliteration. This includes some conspicuous variation, such as the use of two alternative glyphs within the same inscription for the same simplex character, or different forms of a vowel marker which may be attached to a consonant in different ways (as in Figure 3.2.A/3, 4). When deemed interesting or potentially relevant to scholarly study, such allographic variation can be noted in your palaeographic description, but will not be directly represented in the transliteration or the markup.

|  |  |  |  |
| --- | --- | --- | --- |
| Figure .. Allographs ignored in transliteration | | | |
| 1 | 2 | 3 | 4 |
|  |  |  |  |
| rya | rggha | ko | mo |

## Case sensitivity

The ISO-15919 scheme is case-insensitive (International Organization for Standardization 2001, 16, §8.1.1). Strict DHARMA transliteration, in contrast, is case sensitive in order to give distinction to graphematic allographs in transliteration. Uppercase letters in our scheme stand for the independent forms of Indic graphemes, while lowercase letters represent only the in-akṣara forms. The advantage of this is not only that a potentially graphematic feature of the original written text is preserved in transliteration, but also that in the transliteration of poorly legible and lacunose text it makes clear whether what can be made out is a dependent or independent graph, thereby facilitating conjectural restoration. Incidentally, this distinction also reduces the range of situations in which disambiguation (§3.5.1) becomes necessary for the digraphs of the transliteration system. The inevitable drawback of this is that in strict transliteration, uppercase cannot and must never be used for the purposes to which capitalisation is traditionally put in the Roman writing system, such as marking the initials of proper names, sentences and metrical units.

### A note on the use of uppercase for independent vowels and consonants

Some of DHARMA’s predecessor projects have used a the character ° (degree sign) before transliterated vowels and/or after transliterated consonants to denote their special independent forms. DHARMA has decided on a projectwide level to discontinue using that notation, and to endorse and propagate the use of uppercase Roman letters instead. Intellectual considerations in favour of doing so include the following:

* unlike final consonants written with a vowel killer graph (transliterated with a middle dot · as per §0), the special independent forms do not involve an additional grapheme, so it is better to use a single character for their transliteration
* the default (unmarked) allographs of Indic graphemes are the in-akṣara forms, while the independent forms are the exception from the norm (marked), and this situation is superficially analogous to that of lowercase and uppercase letters in modern Roman writing
* uppercase letters are easy to enter on any keyboard, so their inclusion in the transliteration scheme helps productivity
* text-based search algorithms can easily be switched to case-insensitive mode and thus retrieve, for instance, taM Eva when searching for tam eva, whereas in order to retrieve tam° °eva for the same search string, special programming would be required
* using uppercase letters for the special forms allows us to keep the ° sign free for conventional use as a marker of truncation (§3.5.2)

## The accuracy of transliteration

### Strict transliteration

The DHARMA transliteration system is a **strict transliteration** scheme inasmuch as it aims to represent with diplomatic accuracy as much detail of the original written text as possible, and to do so as uniformly as possible across the diverse specific writing systems in our project’s ambit. Its objective, as stated above, is to represent every source grapheme in such a way that it is separable from every other source grapheme and distinguishable not only from any other kind of source grapheme, but also, when applicable, from graphemic allographs of the same grapheme. In addition, the scheme strives for consistency in always using the same transliteration for a given source grapheme, regardless of how that grapheme may be pronounced in any given language and graphemic context.

As noted above (§3), transliteration schemes in practice inevitably fall short of ideal “pure” transliteration, and the DHARMA system is no exception to this. In order not to break with universally known conventions including the ISO-15919 standard, it employs digraphs (§2.4.2) for aspirated consonants and diphthongs. In this respect, it fails to establish a one-to-one relationship between the graphemes of the source and target scripts: the transliterated word ratha has five (target) graphemes, but there are only four (source) graphemes in the same word written in an Indic script, where, <th> is a single grapheme which, in Sanskrit, corresponds to the phoneme [tʱ]. It is therefore very important to keep in mind that in any discussion of Indic graphemes, *source* graphemes are meant, but some of these may be represented by digraphs (i.e. two graphemes) in transliteration. Furthermore, it may be necessary to disambiguate (§3.5.1) target graphemes which can appear in digraphs but are present on rare occasions in their individual roles.

In addition to this slight inconvenience, we deliberately diverge from the maxim “the truth, the whole truth, and nothing but the truth” in permitting certain editorial additions and interpretative alterations to our transliterated text (§##), with the understanding that these are *always* editorial.

### Loose transliteration

In many contexts, written text in a source language comes under scrutiny as abstract text rather than as a particular physical instance. Such situations call for **loose transliteration**, where some constraints of the strict system are relinquished, resulting in a shift toward the transcription end of the spectrum.

As a baseline, in any context where the graphic details of a particular written instance are irrelevant, loose transliteration dispenses with the distinction between the independent and in-akṣara allographs of a grapheme. Relaxing this constraint allows us to use uppercase in transliteration for traditional purposes, such as the capitalisation of the initials of loosely transliterated proper names. Thus, whenever a proper name in a source language is mentioned in discussion (in an international language), we recommend using capital initials. Also, when featuring names or terms of a source language in such discussion, we recommend silent normalisation of the orthography and, where applicable, a shift toward transcription.

Over and above this, depending on your own judgement of the context, relevance and the specific subfield, loose transliteration may involve leniency in further details, such as those below.

* 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 our transliteration scheme), 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/)
  + Old Javanese R̥ interpreted as expressing the syllable rǝ, as in sǝR̥ḥ (in strict transliteration) corresponding to sǝrǝh in loose transliteration
* interpretive transcription where a writing system uses a script sign in more than graphematic function, e.g.
  + Old Sundanese sastra, rahiyaṅ and ku nu reya (when written as sasṭā, ku nu rye and rahyiṁ as in §4.5.2)
* normalisation of orthography, e.g.
  + simplification of consonants doubled in conjunction with r in Sanskrit
  + simplification of consonants doubled at morpheme boundaries in Old Javanese (e.g., lavann ika simplified to lavan ika, muvaḥhakan to muvahakan, at thana simplified to at hana)
  + distinction of e/ē and o/ō even if not present in the original writing

## Transliteration and markup

When creating a digital edition, we strive to keep the transliteration *representing* original content (i.e. what is written on a certain support) separate, or at least separable, from our annotations *describing* and *interpreting* 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 annotation is added primarily in the form of TEI-compliant markup introduced in the Encoding Guides.[[88]](#footnote-88) The terms ‘markup’ and ‘encoding’ are fully synonymous in the context of digital editions, but elsewhere, ‘markup’ can also mean a non-electronic annotation system, such as the various brackets employed in printed epigraphic editions and the use of footnotes.

Ideally, this separation of concerns should be complete: no issues pertaining to the description of the physical manifestation of a text should be recorded in the transliterated text itself; and likewise, no issues pertaining to the 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, with a certain degree of inevitable fuzziness and permeability. Given that we are primarily concerned with the faithful documentation of primary texts, the most fundamental of these details — the distinction of graphematic allographs — is addressed at the level of target graphemes (§3.2.2). Less tangible details are normally dealt with in markup, and our transliteration scheme includes some elements of “low-level markup” (i.e. markup comprised of characters at the level of the transliterated text rather than on the separate tier of XML encoding). Such markup may pertain to the transliteration itself (disambiguation, §3.5.1), to the editorial interpretation of the transliterated text (segmentation, §3.5.2; sandhi analysis, §3.5.3; and the truncation of words extracted from the source, §3.5.4), or to the description of the physical arrangement of the source (split akṣaras, §3.5.5). In addition, we foresee that our transliteration scheme will also be used in situations where TEI markup is not applicable, and therefore suggest some text-based markup alternatives (shorthand, §3.5.6) for certain phenomena.

### Disambiguation

Since our transliteration scheme involves some digraphs (§3.4.1), it is conceivable that in some cases the target graphemes constituting these digraphs might appear together in their own individual functions, rather than as a digraph. The ISO-15919 scheme (International Organization for Standardization 2001, 17, §8.1.15) employs the colon : as a disambiguation sign in such situations, as well as in sundry contexts where the graphetic composition of an original akṣara is unusual. The latter is in our perception a very different kind of situation, for which we prefer a different notation (§4.5.3). As for the former, the need to distinguish vowels in hiatus from diphthongs is eliminated by our use of uppercase for independent vowels in strict transliteration. The disambiguation colon is therefore only needed in the exceedingly rare cases where an <h> follows a stop consonant grapheme, in order to indicate that it is this combination, and not the aspirated consonant grapheme, that is present in the original.[[89]](#footnote-89)

* in strict transliteration, use the colon to disambiguate any occurrences of h preceded by a non-aspirated consonant which has an aspirated counterpart
  + e.g. p:h as opposed to ph
  + the disambiguation colon is not necessary and should not be used for the disambiguation of vowel clusters in strict transliteration, where vowels in hiatus are represented by uppercase target vowels
    - e.g. praUga, caUttho and daIĀ rather than pra:uga, ca:uttho and da:iā as in ISO-15919
* in loose transliteration, the colon may also be employed for the disambiguation of the diphthongs <ai> and <au> from the corresponding vowel clusters
  + but we recommend instead that you follow the established convention of using a diaeresis (pair of dots) above the second vowel in settings where you do not use uppercase for independent vowel graphemes
    - e.g. praüga, caüttho and daïā
  + do not use the diaeresis (or any other disambiguation) when one of the vowels has a macron, since in this case there is no ambiguity
    - e.g. āicca (strict āIcca), not āïcca
* in both strict and loose transliteration, the disambiguation colon and the diaeresis are both unnecessary and must not be used wherever editorial segmentation (§3.5.2) by a space or hyphen intervenes between the target graphemes that might otherwise be read as a digraph, e.g.
  + strict mr̥ta Iva, loose mr̥ta iva; never mr̥ta: iva or mr̥ta ïva
  + strict mana-Icchā, loose mana-icchā; never mana:-icchā or mana-ïcchā

### Segmentation

The editorial segmentation of transliterated texts is strongly recommended in order to help the reader and to show how you analyse the text into words. Since the texts we work with typically use *scripto continua* — i.e. continuous writing without interword spaces — any and all spaces in our transliterated texts will be considered editorial. In other words, the spaces in our texts constitute interpretive markup. The same applies to hyphens — which have no equivalent in the source writing systems[[90]](#footnote-90) — and to new lines. The technical details and semantics of editorial segmentation will be discussed in §6, the essence of which is summarised here.

* in transliterated text which is not marked up in XML, new lines (§6.1) may be created
  + to correspond to original inscribed lines
  + to reflect the semantic or prosodic structure of the text
  + in XML documents, both of these functions are served by computer markup
* in all texts, with or without XML markup,
  + editorial spaces (§6.3) should always be inserted between independent words wherever possible
  + editorial hyphens (§6.4) should be inserted between words in compound where expedient
  + editorial hyphens coinciding with a line break must be placed after the break (§6.4.1)
* as discussed further in §6.2, editorial spacing and hyphenation
  + can and should be used between transliterated characters whose corresponding source graphemes belong to the same akṣara
  + cannot be used and must be avoided at word boundaries which are obscured by sandhi involving the complete fusion of one word’s final vowel with the next word’s initial vowel

### Sandhi analysis

Since transliteration is concerned with the graphemes found in the source text, sandhi should not normally be resolved. Phonographic graphemes must never be added, removed or altered in the process of transliteration. However, to accommodate the conventions of specific subfields, we permit adding a hyphen or an apostrophe to indicate the presence of particular kinds of sandhi, as detailed in §4.7.5.

### Truncation

In accordance with widespread convention, we endorse the use of ° (U+00B0 Degree Sign) as a marker of truncation, primarily when citing individual words verbatim from a source text, in discussion or a critical apparatus (EGD §###).

* the degree sign must not be used in the actual editions of source texts
  + see §3.3.1 for a deprecated use of ° in transliteration by some of DHARMA’s predecessor projects
* the degree sign for truncation should be used when a string of characters must be cut at a point where segmentation with a space or hyphen (§3.5.2) is not applicable, i.e.
  + when only part of a word is lifted from a text, for example at the beginning or end of a line
    - e.g. pari° (cited as a lemma from the word parihāra broken across epigraphic lines)
  + when a word lifted from a text is fused in vowel sandhi to an adjacent word
    - e.g. °oddyotita (for the word uddyotita fused in vowel sandhi to a preceding a)
* when words in source languages are cited in a modern-language discussion, it is fully acceptable to cite them without a truncation sign in stem (“dictionary”) form whenever the focus is on the word as a lexical item rather than on the verbatim received reading

### Placeholders for split akṣaras

An akṣara is normally an atomic unit of the Indic writing system. However, akṣaras interacting with a material feature of the support (such as a line break, or a defect or other feature on the writing surface) are sometimes graphetically separated into their constituent parts. When separation occurs in a way that cannot be reflected through the separation of the corresponding transliterated graphemes, we introduce dedicated characters we call placeholders. The placeholder characters ⌈ (left ceiling, U+2308) and ⌉ (right ceiling, U+2309) stand for detached parts of akṣaras which cannot be transliterated on their own. Their use is discussed and illustrated in §4.6.

### Shorthand

In the context of this Guide, ‘shorthand’ refers to the provisional use of something simple in place of something more complex. Specifically, transliteration shorthand refers to an accessible substitute for a special character that you cannot easily produce on your keyboard. Markup shorthand refers to simple text-based markup in place of more complicated TEI encoding. Shorthand can be helpful for productivity because it takes less time to enter and because markup shorthand can be scanned more easily by the human reader than computer markup. Indeed, the practices introduced in §3.5.1 to §3.5.3 are also a kind of shorthand inasmuch as the phenomena they represent could also be represented in XML markup. We call them ‘essential shorthand’, because — for the time being at least — they constitute an essential part of our transliteration scheme. We have not devised XML encoding solutions for these (although this could be done in the future) because requiring them to be encoded would complicate our XML editions inordinately.

Conversely, the shorthand practices introduced in this section always have a “proper” alternative. This latter should be used whenever feasible, while the shorthand alternative may be used when warranted by circumstances. Depending on these circumstances, we distinguish between ‘private shorthand’ (§3.5.6.1), which may be used in the production stage but must always be replaced with the proper alternative; ‘public shorthand’ (§3.5.6.2), which may be used in the production stage and in contexts not involving XML markup, but must always be replaced with the proper alternative in XML editions; and ‘optional shorthand’ (§3.5.6.3) which may be used in any context including XML editions in order to reduce code clutter and ease your work. No shorthand is mandatory: you are always welcome to stick to the proper solutions.

Since a proper alternative always exists for shorthand, the recommended shorthand solutions are intended to be unambiguous or minimally ambiguous in order to facilitate their conversion to the proper alternative by batch search-and-replace operations involving no or minimum human attention. In order for this to be possible, shorthand must be used rigorously in the proposed format. If you deviate from the recommendations or create your own private shorthand, make every effort to ensure that it can be likewise replaced efficiently. The specific shorthand solutions recommended by us are distributed in the main part of this Guide, but it is important that you familiarise yourself with the concept of shorthand and its private, public and optional varieties introduced below.

#### Private shorthand

Private shorthand may be used in the production stage to speed up the preparation of electronic texts, including collaborative text preparation, but it must always be eventually replaced with the proper transliteration or XML encoding. If any special character required for transliterating your texts is not easily accessible in your editing environment, feel free to substitute a character or sequence that is simple to enter on your keyboard. The substitute character or sequence should never occur in your documents in any other function. If it does, then it should only occur in contexts which clearly differentiate it from occurrences in another function. It is **your own responsibility** to keep careful track of what shorthand solutions you have used, and to replace them with the proper diacritical character or encoding when your work is completed, or to seek technical help for doing so. This Guide contains recommendations for the following kinds of private shorthand, shown with a red label in the relevant sections:

* private shorthand for transliteration
  + ĕ instead of ə (§4.2.2) DO WE REALLY NEED THIS? IF YES, SHOULD IT BE PUBLIC SHORTHAND?
  + \* instead of · (§4.3.3)
* private shorthand for markup
  + ’ or ' or +’ for editorial avagraha (§5.2.1.1; see also §5.2.1)
  + . (full stop) for supplied punctuation (§5.3.1.2)
  + §abc for space fillers with abc replaced by a symbol token (§5.3.2.1) I WOULD PREFER TO DEPRECATE THIS
    - to be distinguished from the public shorthand § (§5.3.2.1)
  + $abc for symbols with abc replaced by a symbol token (§5.3.3) I WOULD PREFER TO DEPRECATE THIS

#### Public shorthand

Public shorthand may be used in the production stage to speed up work, and is also applicable in contexts not involving XML markup. Public transliteration shorthand may be necessary in the course of publication in an environment where a font supporting all the required special characters is not available (§1.6.1). Public markup shorthand is intended mainly for print or on-screen publications not involving XML markup (and to some degree imitates the standard display rendering of our XML editions). All public shorthand must always be replaced with the proper alternative in XML editions. It is **your own responsibility** to keep careful track of what shorthand solutions you have used, and to replace them with the proper diacritical character or encoding when your work is completed, or to seek technical help for doing so. This Guide contains recommendations for the following kinds of public shorthand, shown with a blue label in the relevant sections:

* public shorthand for transliteration
  + underdot instead of undercircle diacritic (§4.1.2)
  + f and x instead of ḫ (upadhmānīya) and ẖ (jihvāmūlīya) (§4.1.5)
* public shorthand for markup
  + + to mark up numeral signs transliterated with more than one target character (§5.1.2.1)
  + \*abc for logographic symbols [NO FINAL PLACE, NOT YET WRITTEN UP, NOT SURE IF INTRODUCING]
  + | (vertical bar) and other graphetically similar Unicode characters for punctuation marks (§5.3.1.1) IF ALL OF THESE ARE PUBLIC SHORTHAND, THAT MEANS WE’LL HAVE TO REPLACE ALL SUCH SIGNS INCLUDING PLAIN | WITH MARKUP IN OUR XML EDITIONS – MAKE SOME OR ALL OF THESE OPTIONAL INSTEAD?
  + § for space filler symbols (§5.3.2.1)
    - to be distinguished from the public shorthand § IF RETAINING THAT
  + ¬ (U+00AC Not Sign) for word breaker signs (§5.3.2.2) NEWLY INTRODUCED, MAY OR MAY NOT WANT IT
  + graphetically similar Unicode dingbats for generic symbols (§5.3.3)
  + \_ (underscore) for original space (§5.4)
  + uppercase C and V for illegible consonant and vowel (§4.7.2.1)
  + - (hyphen) for words cut across inscribed lines (§6.1)

#### Optional shorthand

Optional shorthand may be used in any context including XML editions in order to reduce code clutter and ease your work. Although proper alternatives for these shorthand solutions do exist and are preferred, the slight inconsistency resulting from their use is acceptable if they make your work easier. You do not have to replace optional shorthand with the proper implementation. In the future, we may make a projectwide decision to make these replacements automatically. This Guide contains recommendations for the following kinds of optional shorthand, shown with a green label in the relevant sections:

* optional shorthand for transliteration
  + ' (U+0027 Apostrophe) instead of right single quote (§4.7.5.2 in the analysis of Dravidian sandhi; §5.2.1 as transliteration for the avagraha)
    - this will be processed in the same way as the right single quotation mark and may also be displayed as such
    - but for the sake of rigorous homogeneity in our editions, it is preferable to replace this sign with the right single quotation mark when finalising a digital edition
* optional shorthand for markup
  + = (the equals sign) to indicate that a pair of transliterated characters belong to the same akṣara in the source (§4.5.3.1)
    - there is therefore no straightforward way for automatic conversion between this shorthand and XML markup
    - as a baseline, we recommend using the = sign in the specific situations addressed in §4.5.3 and the XML encoding elsewhere
  + ē and ō in Dravidian languages where long and short versions of these phonemes are not distinguished in writing (§4.7.3)
    - since graphic distinction of this kind is rare in the texts we work with, we choose to add explicit XML markup only where the distinction is actually present in the source
  + ă, ĭ or ŭ in Southeast Asian languages, where a short vowel is written instead of an expected long vowel (§4.7.4) [OLD TG PROMISED AUTO-CONVERSION FOR THESE]

## Non-graphematic entities and transliteration

### Marks and imagery peripheral to the text

Non-alphanumeric symbols integrated into the normal flow of text, i.e. occupying the same kind of segmental space as alphanumeric signs, are to be treated in the same way as graphemes proper (§2.5.5) and represented accordingly in transliteration (§5.3). However, signs peripheral to the inscribed text, including scribal marks and decorative imagery, belong to a separate echelon and may be represented in XML markup (§3.5) or described outside the transliterated text.

#### Scribal marks

Marks of an editorial nature added to a source text by a premodern scribe shall not be represented in transliteration. When scribal deletion or insertion involves such marks, these are to be encoded in XML as per EGD §###.

#### Decorative features

Large symbols, abstract designs and figural elements which are not integral to the linear flow the inscribed text are to be regarded as decorative features. They are neither to be represented in transliteration, nor encoded in XML editions, only to be described for human readers in the metadata of the text.

### The materiality of the support

On the whole, the characteristics of a source text’s layout, including the ways in which the materiality of the support may affect the physical arrangement of the writing, fall outside the domain of transliteration. Such characteristics can be described in general in the metadata, and some of them can be specifically encoded in XML markup. This Guide contains some instructions pertaining to layout (§6.1) and specifically addresses the special case where complex glyphs are split across a feature in such a way that the linear sequence of graphic components does not correspond to the logical sequence of graphemes (§4.6).

# Transliterating alphabetic graphemes

## The basic inventory of Indic graphemes for Old Indo-Aryan

We assume that every project member is familiar with the standard transliteration equivalents of the basic graphemes of the Indic writing system, but for the sake of completeness we list them in Figure 4.1.A below. Additional considerations applicable to this basic repertoire are discussed in the following subsections.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Figure .. The basic inventory of Indic graphemes | | | | | | | | | | | |
| a | ā | i | ī | u | ū | r̥, r̥̄ | l̥, l̥̄ | e (ē) | ai | o (ō) | au |
| k | kh | g | gh | ṅ |  | c | ch | j | jh | ñ |  |
| ṭ | ṭh | ḍ | ḍh | ṇ |  | t | th | d | dh | n |  |
| p | ph | b | bh | m |  | y | r | l | v |  |  |
| ś | ṣ | s | h |  |  | ṁ | ḥ |  |  |  |  |

### Digraphs in the transliteration

Our transliteration scheme employs digraphs (§2.4.2) for diphthongs and for aspirated consonants. Keep in mind that, as pointed out in §3.4.1, these digraphs always correspond to a single source grapheme, and are therefore never separable (for instance by spaces or hyphenation). In the rare situations where the Roman letters which normally constitute a digraph appear side by side in their own role, add a colon between the letters to disambiguate (§3.5.1).

### Vocalic r and l

If you are used to working with IAST rather than ISO-15919, note that vocalic r and l are transliterated in the DHARMA scheme with undercircles rather than underdots. The corresponding diacritical letters are not available in Unicode as pre-composed characters and will need to be composed by entering 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, copy and paste the composed characters from Figure 4.1.A above, or use shorthand (§3.5.6).

* if necessary, you may use underdots instead of undercircles as public shorthand for vocalic r and l, i.e.
  + ṛ instead of r̥
  + ṝ instead of r̥̄
  + ḷ instead of l̥
  + ḹ instead of l̥̄
* note that depending on the language you work with, ḷ and ṛ may be required for the transliteration of consonant sounds (§4.2.1)
  + if you use shorthand, it is your responsibility to avoid ambiguity

### Transliteration of e and o

According to Option 9.1 of the ISO15919 standard, for languages that do not make a distinction between the phonemes /e/ vs. /ē/ and /o/ vs. /ō/, we use the letters e and o to transliterate the corresponding phonemes, without adding a macron, regardless of whether they are normally pronounced long in the source language. Importantly, we apply the same rule to any **writing systems** that do not make this distinction, and this includes many historic Dravidian writing systems where a length-based opposition exists in the language, but is not reflected in writing. Since the specific signs for the long variants of these vowels (such as those in Tamil |கே| kē and |கோ| kō) are usually palaeographically derived from the original signs for the generic — and, in Sanskrit, long — vowel (such as those in Tamil |கெ| ke and |கொ| ko), the default transliteration in our scheme is always the unmarked Roman character e or o. See also §4.7.3 about the editorial distinction of the long phonemes /ē/ and /ō/ when written with the basic graphemes <e> and <o>.

### Anusvāra and its relatives

The basic anusvāra grapheme of Indic script, typically represented by a dot or a circle, is always transliterated as ṁ, with a dot above as in ISO-15959 rather than with a dot below as in IAST. In the linear sequence of transliterated characters, position the anusvāra where it logically falls in language, regardless of whether it is graphically above an akṣara, in a separate position inline, or even above the next akṣara. The graph representing the anusvāra may vary somewhat within a source, but unless this variation systematically corresponds to different linguistic information (i.e. if the various graphs are not allographs but separate graphemes), the transliteration should remain identical. Graphemes related to the anusvāra are discussed in the following subsections.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Figure .. Anusvāra relatives | | | | |
| 1 | 2 | 3 | 4 | 5 |
| anusvāra | anunāsika/ candrabindu | ulu ricem | anusvāra-candra | Bengali variant anusvāra |
|  |  |  |  |  |
| saṁ | Om̐ | vr̥ttaṁ\* | tad-vam̃śe | m\* |

#### Anunāsika or candrabindu

If a candrabindu (crescent-and-dot) graph (Figure 4.1.B/2) is distinguished from the basic anusvāra in a way you consider, based on your knowledge of the source text and its writing system, to be graphematically relevant (i.e. to represent vowel nasalisation or a nasal sound other than the standard Indic nasal consonants), then distinguish it from the anusvāra in your transliteration:

* **anunāsika or candrabindu** → m̐
  + not available as a precomposed glyph: enter a regular m and ◌̐ (U+0310 Combining Candrabindu)
* do not use the candrabindu diacritic in combination with any other Roman letter, thus transliterate tām̐l lakṣmīm rather than tāl̐ lakṣmīm

#### Other anusvāra variants

pending <https://github.com/erc-dharma/project-documentation/issues/387>

* **Cam anusvāra-candra** → m̃
  + not available as a precomposed glyph: enter a regular m and ◌̃ (U+0303 Combining Tilde)
  + not covered by ISO-15919
* if your text employs a special glyph for anusvāra in addition to the regular form, then occurrences of the special glyph shall be transliterated as ṁ\*, including in particular
  + the Javanese/Balinese special anusvāra with a small stroke beside it (to indicate pronunciation as /m/), called ulu ricem in Balinese, as shown on the right in nāśaṁ\*
  + the alternative anusvāra character used in some mediaeval Bengali documents, shown in the image on the right
  + note that if you use an asterisk for this purpose, then you are advised not to use asterisks as shorthand for a zero vowel marker (§4.2.2.1)
* extension to others not covered in our corpora, e.g.
  + Gurmukhi ṭippī
  + Telugu half-anusvāra (also called candrabindu, but it is not an anunāsika, does not have a candra+bindu shape, and can co-occur with proper candrabindu in some texts, <https://unicode.org/L2/L2010/10392r2-chandrabindus.pdf> )

### Visarga and its relatives

Some varieties of the Indic writing system make a graphic distinction corresponding (with more or less regularity) to allophones of the visarga, using the sign called upadhmānīya before palatals and another sign called jihvāmūlīya before velars (Figure 4.1.C). When a distinction is made in the text you are transliterating, then preserve this distinction in transliteration regardless of whether the visarga alternatives are used systematically and correctly or not.

* **upadhmānīya** → ḫ
  + (U+1E2B Latin Small Letter H with Breve Below)
  + the public shorthand f may, when necessary, be used instead
* **jihvāmūlīya** → ẖ
  + (U+1E96 Latin Small Letter H with Line Below)
  + the public shorthand x may, when necessary, be used instead

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Figure .. Visarga relatives | | | | |
| 1 | 2 | 3 | 4 | 5 |
| visarga | combining upadhmānīya | combining jihvāmūlīya | inline upadhmānīya | inline jihvāmūlīya |
|  |  |  | ᳶ |  |
| yaḥ | ḫpu | traẖka | ḫ | yoẖka |

While the basic visarga is always written inline as a practically independent graph (§2.5.6.1), the jihvāmūlīya or upādhmānīya can, depending on the specific writing system, either appear independently (Figure 4.1.C/4,5) or behave like a consonant graph and form a conjunct with the following consonant (Figure 4.1.C/2, 3). Since all members of the visarga family are recognised as graphemes in our scheme and are thus transliterated with a separate character, it is neither possible nor essential for the transliteration to indicate which of these behaviours is present in a given source text. However, whenever deemed expedient, the markup described in §4.5.3.4 may be used to represent deviations from the expected behaviour.

## Graphemes extending the basic repertoire

This section is concerned with alphabetic graphemes of Indic writing systems used primarily for languages other than Middle Indo-Aryan and Classical Sanskrit. Systemic changes (i.e. innovations which do not consist in the introduction of novel graphs for specific graphemes) will be discussed further on, in §4.5.

In addition to the transliteration signs listed below, which are relevant to the DHARMA project, the inventory of transliteration equivalences may be expanded freely according to the needs of future projects. While adding new signs to the inventory, the following rules of thumb should be considered.

* **compatibility**: unless there are strong reasons to the contrary, use the signs prescribed by ISO-15919 when applicable
* **uniqueness**: if at all possible, the Roman character or digraphs selected for any particular grapheme should not be identical to any transliteration character already in use for another grapheme in the DHARMA scheme
  + deviate from this rule only when it is reasonable to assume that the two different source graphemes transliterated by an identical target grapheme or digraph will never co-occur in any source text

### Graphemes of other Indian writing systems

The transliteration equivalences listed in this section are conformant with ISO-15919.

* **Dravidian (and other) short e and o**, as distinct from long ē and ō, are discussed in §4.1.3 above
* **special variants of the anusvāra** are discussed in §4.1.4.2 above
* **Tamil āytam**, |ஃ| → ḵ
  + (U+1E35 Latin Small Letter K with Line Below)
* **Dravidian, Vedic and New Indo-Aryan retroflex lateral**, Tamil |ள|, Kannada/Telugu |ಳ|, Vedic/NIA |ळ| → ḷ
  + (U+1E37 Latin Small Letter L with Dot Below)
  + **Vedic aspirated retroflex lateral** |ळ्ह| → ḷh
* **Dravidian retroflex approximant / frictionless continuant**, Tamil |ழ|, Kannada/Telugu |ೞ| → ḻ
  + (U+1E3B Latin Small Letter L with Line Below)
* **Dravidian alveolar trill/stop**, Tamil |ற|, Kannada/Telugu |ಱ| → ṟ
  + (U+1E5F Latin Small Letter R with Line Below)
* **New Indo-Aryan retroflex flap**, Hindi |ड़| → ṛ
  + (U+1E5B Latin Small Letter R with Dot Below)
  + **New Indo-Aryan aspirated retroflex flap**, Hindi |ढ़| → ṛh

### Graphemes of Southeast Asian writing systems

The transliteration equivalences listed in this section are not covered by ISO-15919 (which is targeted at Indian writing systems).

* **special variants of the anusvāra** are discussed in §4.1.4.2 above
* **short mid-central vowel, “schwa”**, Javanese/Balinese pepet |ᭂ| → ə
  + U+0259 Latin Small Letter Schwa
  + if this character is difficult to produce, you may use the private shorthand ĕ
* **long mid-central vowel**, Javanese/Balinese pepet tedung |ᭃ| → ə:
  + U+0259 followed by a colon[[91]](#footnote-91)
  + in loose transliteration, ə̄ may be used instead
    - not available as a precomposed glyph: enter U+0259 and ◌̄ (U+0304 Combining Macron)
* **Khmer (and Mon-Burmese) glottal stop** → q
  + the Roman letter q
  + see also §### about the representation of independent vowels involving this character component and §### about its use as a consonantal grapheme
* **barred/dotted variant of b in Mon and Pyu** → ḅ
  + U+1E05 Latin Small Letter B with Dot Below

#### Graphemic combination of the vowel markers |u| and |i|

|  |
| --- |
| Figure .. Combined ui |
|  |
| gui |

In Khmer, Burmese and Mon, the vowel markers which independently signify <u> or <ū> and <i> may be combined so that together they represent a particular phoneme. Since this is an established feature of the writing system, in principle the combination of these graphic elements is a single graph (cf. §2.5.6.2[[92]](#footnote-92)). Nonetheless, in accordance with the conventions of the field and in order to highlight the palaeographic precedent of this graph, we transliterate them individually. This transliteration approach may also be perceived as the use of a target digraph for a diphthong.

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):

* the vowel marker |u| or |ū| used in combination with the marker |i| to represent a phoneme shall be transliterated as ui and ūi
  + for example, the Khmer glyph in Figure 4.2.A is transliterated gui
* note that in certain writing traditions, a combination of the markers |u| and |i| may signify scribal deletion
  + scribal marks (§5.3.3) are not alphabetic graphemes and must not be transliterated as such, only represented in XML markup
* see also §4.5.3.6 about other circumstances in which an akṣara may have more than one vowel marker

## Vowelless consonants

Many historic variants of the Indic writing system employ special final (halanta) forms of the regular consonant graphemes to represent vowelless consonants (Figure 4.3.A). We consider these to be graphematic allographs (§2.6) of the consonant graphemes and deem it important to distinguish them from the default in-akṣara allographs in philologically accurate transliteration (§3.2.2), as described in §4.3.2. Conversely, many Indic writing systems (especially more recent ones) employ an explicit zero vowel marker (virāma, puḷḷi, patén/pangkon, etc.; also referred to as a vowel killer) in combination with the regular consonant graphs in order to represent vowelless consonants. In our view, the vowel killer is, for all practical purposes, a grapheme (§2.5.2), which must thus be transliterated with a corresponding target grapheme as per §4.3.3. Representing zero vowel markers by a separate character in the transliteration has the added advantage of allowing the application of markup to that sign.[[93]](#footnote-93) Considerations for distinguishing a final (simplex) form from (a complex) one involving a vowel killer are given in §4.3.1. In some source texts, a vowelless consonant may also be represented by the regular consonant graphs, as discussed in §4.3.4.

### Distinguishing final forms from characters with a vowel killer

Final forms may differ from the regular akṣara-forming graphs in any combination of size reduction, subscript positioning, stroke simplification and the absence of a headmark. Final consonant forms often also involve an additional “diacritical” stroke — such as a horizontal dash, a vertical squiggle above the primary graph, or a curl below it — which may be interpreted as a proto-virāma (see especially Figure 4.3.A/4). These transitional forms constitute a fuzzy boundary between simplex and complex characters.

The **primary criterion** for assigning a particular written instance to one of these classes is whether the body of the consonant is graphically identical to the form seen in regular akṣaras. If it is, and a stroke that may be interpreted as a zero-vowel marker is present, then transliterate it as a complex character with a virāma (§4.3.3). If, on the other hand, the body is graphically altered in one or more of the ways mentioned above, then transliterate it as a simplex final character (§4.3.2), even if a stroke that might be interpreted as a zero-vowel marker is present. Being fairly objective, this is a practicable criterion, but it does not necessarily do justice to all occurrences. In ambiguous cases, base your decision on your knowledge of the practice of the source text in question and of the writing system of the time and region. A relevant consideration is the productivity of these nascent vowel killer signs. If they appear only in final positions, with a small range of consonants, and especially if their form varies depending on the consonant, then it is better not to recognise them as virāma graphemes. Conversely, if they can appear word-internally (where ligatures would be a feasible alternative), are combined with a variety of consonants, and they have the same form regardless of the consonant, then it is more prudent to interpret them as actual virāmas.

### Final consonants as simplex characters

* **simplex final consonant forms must always be transliterated** as the corresponding uppercase Roman consonant, e.g. N, M, etc.
  + should you encounter simplex final forms of consonants for which our transliteration uses digraphs, use uppercase only for the first letter of the digraph, e.g. Th
  + guidelines for deciding what constitutes a special final form are given in §4.3.1 above

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Figure .. Final consonant graphs | | | | |
| 1 | 2 | 3 | 4 | 5 |
|  |  |  |  |  |
| yaN | dattaM | ṇāM|| | yaT | dr̥K |

### Independent consonants as complex characters involving a vowel killer

* complex characters involving a regular consonant form and an explicit zero vowel marker must always be transliterated as follows
  + transliterate the consonant component normally (with the lowercase equivalent)
  + transliterate the vowel killer as ·
    - U+00B7 Middle Dot
  + e.g. Devanagari |त्| → t·
  + if you need to transliterate vowel killers frequently but have difficulty entering the middle dot, you may use an asterisk \* as private shorthand
    - this shorthand must not co-occur with asterisks used for special forms of the anusvāra, §4.1.4.2
* the **Tamil puḷḷi**, when explicitly present in the source (cf. §4.3.4), is not to be treated differently from other vowel killers
  + e.g. |த்த| → t·ta

### Regular consonant signs for vowelless consonants: the “implicit puḷḷi”

Tamil inscriptions often use the regular consonant forms (which would imply the presence of the inherent vowel) for writing the non-final members of consonant clusters, as in Figure 4.3.B/1. For complete accuracy in the spirit of a universal transliteration of all Indic writing systems, these instances should be transliterated so that the inherent vowel is present in the transliteration. However, because this practice is widely prevalent, while conjunct akṣaras (ligatures) are rare in Tamil, we choose to make an exception for primary sources in this particular language,[[94]](#footnote-94) and prioritise transcription over accurate transliteration in the case of consonant clusters. That is to say, whenever a consonant cluster is present in a DHARMA-transliterated Tamil text, it is to be understood by default that the non-final consonants of that cluster were written in the source with the regular (akṣara-constituting) consonant glyphs. As a consequence of this decision, any cases where a consonant cluster is written differently in the source must be made explicit in the transliteration.

* where puḷḷi is not present in the source but is to be understood implicitly, as in Figure 4.3.B/1, simply transliterate the consonant cluster without any additional characters
  + e.g. |தத| (understood as த்த /tta/) → tta
* when puḷḷi is present in the source, as in Figure 4.3.B/2, always transliterate it like any other virāma (§4.3.3)
* when an actual ligature occurs in Tamil script, as in Figure 4.3.B/3, this must be treated as unusual akṣara composition, according to the instructions in §4.5.3.2

|  |  |  |
| --- | --- | --- |
| Figure .. Consonant clusters in Tamil | | |
| 1 | 2 | 3 |
| நந கக | ந்ந க்க | **C:\GoogleDrive\DHARMA project\Encoding Guide\Encoding Guide Images\Tamil ligatures from Manu\nnakka_SII_6_p_166.png** |
| nna, kka | n·na, k·ka | n=na, k=ka |

## Independent vowels

### Independent vowels as simplex characters

The aksharic Indic writing system normally represents vowels with markers (§2.4.3.2) attached to consonant graphs, or as inherently implied by the consonant graphs themselves. This method is incapable of representing vowels in an initial position (i.e. at the beginning of a graphic sequence) and vowels in hiatus (i.e. those following another vowel in a sequence). Therefore, aksharic writing systems include independent vowel graphs for use in such situations. As is the case with special vowelless consonant forms (§4.2.2.1), these independent vowels are in our view graphematic allographs (§2.6) of the vowel graphemes, and we deem it important to distinguish them from the default in-akṣara allographs in philologically accurate transliteration (§3.2.2), in the manner described in §4.4.1.

* if the original script employs a distinct character for vowel-only akṣaras (initial vowels and vowels in hiatus), these shall be mandatorily transliterated as follows
  + enter the corresponding uppercase Roman consonant, e.g. A
    - thus, इति → Iti, whereas कृतमिति → kr̥tam iti
    - for the independent 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)

### Independent vowels involving a vowel support

As noted in §2.5.6.2, the graphs for certain independent vowels are in many Indic writing systems derived from other independent vowel signs and a vowel marker; however, some Southeast Asian systems have a full set of independent vowel signs derived from a single basic graph (the independent allograph of the neutral vowel) and appropriate vowel markers. In this role, the basic graph can be referred to as a “vowel support”, meaning a graph whose only function is to provide a “zero consonant” to which vowel markers can be anchored. Some of these systems go further and additionally employ this basic graph to represent an actual consonantal phoneme (a glottal stop, /ʔ/) rather than a zero consonant. In these latter, the graph in question has the capacity to form ligatures with other consonant graphs. This systemic change is complete in Khmer, where |អ| functions fully as a consonantal graph, but has been carried through to varying degrees in the writing traditions of Java, Bali and Lombok.[[95]](#footnote-95)

When this graph functions only as a vowel support, it is questionable whether it can be properly called a graph (i.e. the representation of a grapheme), or merely a graphic element. However, in a consonantal role it is definitely a graph, representing a grapheme. For this reason, we prefer to recognise it as a graph regardless of the role it plays, and always transliterate it explicitly. The following instructions for the strict and loose transliteration (§3.3) of vowels composed with a vowel support are summarised and illustrated for the Khmer and Balinese writing systems in Figure 4.4.B below.

|  |
| --- |
| Figure .. Vowel support |
|  |
| qət r̥ṅyəkən tikiṁ |

* in the **strict transliteration of all writing systems in which a vowel support exists**, regardless of whether it corresponds to a consonantal phoneme or is used only as a vowel support
  + transliterate the vowel support with the Roman letter q, as in Figure 4.4.A
  + when the same graph is used on its own to represent independent <A>, transliterate it as A (and not as qa)
  + likewise, independent vowel graphs which do not involve a vowel support must always be transliterated with the corresponding uppercase vowel, even if they clearly signify a consonant phoneme in addition to the vowel
    - for example, Khmer អេត → qeta but ឯត្ត → Etta
* the same transliteration is to be used when the graph representing the glottal stop combines into a ligature with other consonantal graphs
  + in a syllable-initial position, e.g. |អ្នក| → qnaka
  + in post-consonantal position, e.g.
    - |ផ្អក| → phqaka
    - |ល្អិត| → lqita (as opposed to |ល្ឥត| → lIta)
    - |ប្អូន| → pqūna (as opposed to |ផ្ឧន| → phUna)
    - |ក្អេក| → kqeka (as opposed to |ក្ឯក| → kEka and |ក្អែក| → kqaeka[[96]](#footnote-96))
* however, in the **loose transliteration** of characters involving a vowel support, **transcription** of the language has priority over transliteration of the script, therefore
  + in languages **where the vowel support is silent** (i.e. it stands for a zero consonant), as in Balinese in Figure 4.4.B, ignore it in loose transliteration and simply transcribe the relevant vowels
    - e.g. Balinese |ᬅᬾ| <qe> → e
  + in languages **where the vowel support signifies a consonant** (normally the glottal stop), as in Khmer in Figure 4.4.B, always transcribe it explicitly
    - in loose transliteration, either q or ’ (the right single quote) may be used to represent the glottal stop
    - note that this sound should be transcribed even when the vowel support is on its own (representing the independent vowel <A>)
      * e.g. Khmer |អ| <A> → qa or ’a
    - except for Sanskrit names and loanwords occurring in an Old Khmer text, whose loose transliteration should transcribe the Sanskrit and not involve the glottal stop (e.g. Amoghapura rather than qAmoghapura or ’Amoghapura)
      * even though Khmer speakers would have pronounced a glottal stop in these names and words

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Figure .. Independent vowels composed with a vowel support | | | | | | | | | |
| language and writing system | IPA  transcription | simplex character | | | complex character | | | | |
| glyph | transliteration | | glyph | components | transliteration | | |
| strict | loose | |
| strict | loose | lowercase | uppercase |
| Khmer | /ʔa/ | អ | | | | |A| | A | qa / ’a | qA / ’A |
| Balinese | /a/ | ᬅ | | | | a | A |
| Khmer | /ʔi/ | ឥ | I |  | អិ | |A| + |i| | qi | qi / ’i | qI / ’I |
| Balinese | /i/ | ᬇ |  | ᬅᬶ | i | I |
| Khmer | /ʔiː/ | ឦ | Ī |  | អី | |A| + |ī| | qī | qī / ’ī | qĪ / ’Ī |
| Khmer | /ʔu/ | ឧ | U |  | អុ | |A| + |u| | qu | qu / ’u | qU / ’U |
| Balinese | /u/ | ᬉ |  | ᬅᬸ | u | U |
| Khmer | /ʔuː/ | ឪ | Ū |  | អូ | |A| + |ū| | qū | qū / ’ū | qŪ / ’Ū |
| Khmer | /ʔe/ | ឯ | E |  | អេ | |A| + |e| | qe | qe / ’e | qE  / ’E |
| Balinese | /e/ | ᬏ |  | ᬅᬾ | e | E |
| Khmer | /ʔaj/ | ឰ | Ai |  | អៃ | |A| + |ai| | qai | qai / ’ai | qAi / ’Ai |
| Khmer | /ʔo/ | ឱ | O |  | អោ | |A| + |o|[[97]](#footnote-97) | qo | qo / ’o | qO / ’o |
| Balinese | /o/ | ᬑ |  | ᬅᭀ | o | O |
| Khmer | /ʔao/ | ឳ | Au |  | អៅ | |A| + |au| | qau | qau / ’au | qAu / ’Au |
| Balinese | /ə/ |  |  |  | ᬅᭂ | |A| + |ə| | qə | ə | Ə |
| Balinese | /əː/ |  |  |  | ᬅᭃ | |A| +  |ə| + |ā|[[98]](#footnote-98) | qə: | ə̄ | Ə̄ |

## Systemic innovations in the Indic writing system

The adaptation of the Indic writing system to languages whose phonology and phonotactics differed considerably from Old and Middle Indo-Aryan languages entailed a number of innovations in specific varieties of the writing system. We have attempted to group these according to the manner in which we approach their transliteration, which in turn is based on our understanding of their graphematic nature, but many of these innovations are difficult to classify in graphematic terms, so the structuring of this section is not entirely rigorous. Additional graphemes (among which we include the combination of the graphs for |u| and |i| to represent a particular phoneme, §4.2.2.1) have been covered above in §0. Herein, §4.5.1 treats graphic signs which are, or may be, diacritical marks in the source writing system, but are treated in transliteration as graphemes. In §4.5.2 we cover cases where pre-existing graphic signs are repurposed for the writing of sequences that would be written in a more complex way by the standard conventions of the Indic writing system. Finally, §4.5.3 discusses what we can do in transliteration to reflect unusual ways of combining graphs into complex characters.

### Borderline diacritical marks

As explained in §2.5.1, diacritical marks in the strict sense are not graphemes and therefore should not be individually represented in transliteration. What matters is that the graphemes distinguished by diacritics in the source must also be distinguished — with diacritics or otherwise — in the transliteration. Thus, §0 above provides transliteration equivalences for a few diacritically modified source graphemes in our scope, while others may be added to the transliteration scheme as and when the need arises.[[99]](#footnote-99) However, as observed in §2.5.6.2, ambiguous situations do arise in the course of the adaptation and evolution of writing systems. This section contains provisions for graphic elements of some specific writing systems which are arguably diacritical marks, but which we nonetheless prefer to transliterate individually, as if they were graphemes.

#### The |ā| graph as a signifier of length in maritime Southeast Asia

The graph originally serving as the marker for the dependent vowel <ā> (Javanese tarung |ꦴ|, Balinese tedong |ᬵ|) is also used in some Indonesian texts as a signifier of vowel length or consonant doubling (Figure 4.5.A). In this function, it is a sort of diacritical mark, but treating it as such in transliteration would require an inordinate number of additional target graphemes, while transliterating it strictly according to its original graphematic value (like the adaptations discussed in §###) would obscure its function. We therefore grant special treatment to this phenomenon.

Taking advantage of the facts that the colon is widely used in transcription as an indicator of length and that the disambiguation colon of ISO-15919 is hardly ever needed in our transliteration scheme (§3.5.1), we dedicate the colon to the transliteration of this “honorary grapheme”, and use it as described in the following subsections. This is a slight inconsistency in the transliteration system, which we deem acceptable when weighed against the gain of providing easy keyboard access and intuitive interpretability for the length marker. If in any text or corpus an ambiguity should arise (i.e. both the disambiguation colon and the length marker colon need to be present), use the IPA triangular colon ː (U+02D0) for the length marker.

|  |  |  |
| --- | --- | --- |
| Figure .. The |ā| graph as a signifier of length | | |
| 1 | 2 | 3 |
|  | Macintosh HD:Users:username:Documents:PHILOLOGY PROJECT:Edisi Bhīmaswarga:BS Cahier d'Archipel:Figure4.jpg |  |
| qə:bni pilaṁ | gnәp:ipitu | turut:vaḥna |

* when the graph originally denoting <ā> is used in conjunction with another vowel marker to transform the latter into a long vowel, enter a colon (:) after the short vowel to transliterate the length marker
  + e.g. Figure 4.5.A/1 → qə:bni pilaṁ (see also §4.4.2 about the transliteration of the vowel support)
* when the graph originally denoting <ā> is attached to an akṣara to indicate the lengthening (doubling, gemination) of the consonant component of that akṣara, enter a colon (:) after the transliterated consonant to which the doubling pertains
  + e.g. Figure 4.5.A/2 → gnәp:ipitu (pronounce /gәnәp pipitu/, Old Sundanese “fully seven”)
  + the colon shall be next to the transliterated consonant even if it is not adjacent in the original
    - e.g. Figure 4.5.A/3 → turut:vaḥna (pronounce /turut tvaḥna/, “you should follow his behaviour”)

#### Underdotted akṣaras in mainland Southeast Asia

to be written up once <https://github.com/erc-dharma/project-documentation/issues/387> is decided

* **akṣaras with underdot in Mon, Pyu and Burmese** → ṃ
  + U+1E43 Latin Small Letter M with Dot Below
  + to be added where? after the akṣara? after the consonant? e.g.?
  + because the function of this underdot is poorly understood, we prefer to transliterate it as if it were a separate grapheme, even though it may be merely a diacritical mark

### Repurposed graphic signs

The present section is concerned with innovations in specific writing systems which repurpose a simple (and not necessarily alphabetic) graph or glyph of the generic Indic writing system to optionally represent phonological information that would be expressed in a more complex way according to the usual conventions of the Indic system. A similar phenomenon in modern international writing is the use of the numeral sign |2| to represent the English word “to” or the corresponding sequence of phonemes, as in the phrase “go 2 bed” or the word “2gether”. The graphic sign can thus have two significations: the conventional one and the innovative one. Situations where an identical graphic representation corresponds to several different graphemic significations (the way |2| can mean either the number 2 or the phonemes /tʊ/) may be regarded as homography (§2.3.1). Proper homographs are essentially different graphemes which are, incidentally, manifested by identical graphs, and should accordingly be transliterated on the basis of their graphematic value (cf. §4.7.1). However, in this case homography is combined with heterography, meaning that a single graphematic value can be expressed in two ways: the conventional (complex) way and the innovative (simpler) way which is homographic to a different signification. Transliterating only the graphematic value would entail loss of information (equating “together” to “2gether”) about the alternative that was employed in any given instance in a source text.

In view of this complication, in the strict transliteration of such graphemes we choose to adhere strictly to the primary graphematic value that is universal to Indic writing systems, disregarding their innovative phonetic value in accordance with the principle that transliteration is concerned primarily with graphemes, and not with their pronunciation in any particular language (§2.2.1). However, in loose transliteration (§3.4.2) we prioritise the phonetic value applicable in any given context, i.e. actually transcribe such graphemes rather than transliterating them. In XML editions, moreover, the transliterated primary signification may be normalised through computer markup (EGD §###) to the transcribed secondary signification.

The actual cases of repurposing that we are aware of (Figure 4.5.B) are all Old Sundanese, but analogous phenomena in other specific writing systems should be handled in the same way.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Figure .. Repurposed graphic signs | | | | |
|  | 1 | 2 | 3 | 4 |
|  | Macintosh HD:Users:username:Pictures:Naskah Nusantara:Perpusnas:Bima Swarga:Bhīma Swarga 623 P 16:sastra.png | Macintosh HD:Users:username:Pictures:Naskah Nusantara:Perpusnas:69 L 626:di jeroning wawangunan.png | Macintosh HD:Users:username:Pictures:Naskah Nusantara:Perpusnas:L 621 P15:ku nu rye.png |  |
| strict | sasṭā | di jә2niṁ vavaṁṅun:an· | ku nu rye | rahyiṁ |
| loose | sastra | di jәroniṅ vavaṅunan | ku nu reya | rahiyaṅ |

* when the glyph Macintosh HD:Users:username:Documents:PHILOLOGY PROJECT:Edisi Bhīmaswarga:BS Cahier d'Archipel:tra_b.jpg ṭā is used in Old Sundanese to represent the phonemes /tra/
  + transliterate it as ṭā
  + in loose transliteration, transcribe it as tra
  + e.g. Figure 4.5.B/1 → sasṭā; loose transliteration: sastra
* when the numeral sign |2| is used in Old Sundanese to represent the phonemes /ro/ (presumably because the word for the number 2 is ro or roro in Javanese, whose writing tradition heavily influenced Sundanese)
  + transliterate it as 2
    - since the sign does not represent a number in this case, the XML markup for numbers (EGD §###) must not be used
  + in loose transliteration transcribe it as ro
  + e.g. Figure 4.5.B/2 → di jә2niṁ vavaṁṅun:an·; loose transliteration: di jәroniṅ vavaṅunan (“in the interior of the building”)[[100]](#footnote-100)
* when subscript <y> is used in Old Sundanese to make an akṣara bisyllabic (so that the body consonant and the attached vowel marker are to be pronounced first, followed by /ya/)
  + transliterate this as dictated by the default logic of the writing system at large, i.e. transliterate only the vowel marker and place it after the <y>
  + in loose transliteration, follow the sequence of the pronunciation
  + e.g. Figure 4.5.B/3 → ku nu rye; loose transliteration: ku nu reya (“by many [people]”)
  + e.g. Figure 4.5.B/4 → rahyiṁ; loose transliteration: rahiyaṅ[[101]](#footnote-101)

### Variation in glyph composition

As noted in §3.2.2, the universal rules of the Indic writing system determine which graphemes combine into a single complex character and which do not, but there are some systematic as well as idiosyncratic deviations from these rules. This section introduces text-based editorial markup involving the = (equals) sign, to be used when graphemes combine into a complex source in a way other than what would be expected on the basis of the general rules of the Indic system.

#### Optional shorthand for complex characters

We use the = (equals) sign as optional shorthand in the role of a character joiner, as simple text-based markup for use between two target graphemes that are manifested as parts of a single character in the source, even though they would normally manifest as (parts of) separate characters according to the general rules of the Indic writing system. This markup is a simple alternative to the XML-based markup introduced in EGD §###. The latter can replace the former in any function and has a broader scope than the shorthand. Nonetheless, because the shorthand is simpler to produce and to read, we approve of employing it in XML editions, especially for cases of systematic variation that the informed reader will be able to interpret. Prominent cases of such variation are discussed with specific recommendations in the following subsections.

* when the character joiner sign is placed between two graphemes which are separated by an editorial space (§###) or hyphen, always put the = sign first, and the space or hyphen second

#### Conjunct consonants in writing systems where they are not the norm

Some specific writing systems, notably for the Tamil language, allow the writing of consonant clusters with regular consonant signs, which are then to be understood as if they were accompanied by a vowel killer (puḷḷi). As per §4.3.4, we treat this as the default mode of writing consonant clusters in these writing systems. Occasionally, however, ligatures composed according to the standard manner of the Indic writing systems may also be used in texts employing this system. The possibilities for writing a consonant cluster in Tamil are illustrated in Figure 4.3.B. In the transliteration of consonant clusters written with a ligature, it is recommended that you use the optional character joiner sign = to indicate that the graphemes in question belong to a single akṣara.

* where a Tamil text written in Tamil script employs a ligature, as in Figure 4.3.B/3
  + use the = sign between the corresponding transliterated consonants to distinguish the ligature from the script’s default method of writing conjunct consonants as two glyphs with an explicit or implicit zero vowel marker, e.g.
    - n=na (as distinct from |ந்ந| n·na and |நந| nna)
    - *k=ka* (as distinct from |க்க| k·ka and |கக| kka )
* Tamil ligatures should be made explicit in this manner whenever feasible
* however, the = sign must never be used in ligatures of a writing system where conjoining is the default method of representing consonant clusters, and this includes Tamil written in Grantha

#### Double kāl in Tamil

|  |  |
| --- | --- |
| Figure .. Double kāl | |
|  | |
| rā, r=ā | ā=r |

* when in Tamil script a fusion of two kāl signs appears, as in Figure 4.5.C, which may be interpreted either as <rā> or as <ār>, proceed as follows
  + make an interpretive decision as to what grapheme sequence was intended (§4.7.1)
  + transliterate that sequence as follows
    - for <ār>, add the character joiner sign to indicate that these graphemes belong to a single glyph in original
    - for <rā>, the joiner sign is not necessary, since these graphemes would naturally constitute one akṣara, but you may choose to add the = sign nonetheless to make it explicit that a fused double kāl is present
  + if ambiguity is present, then record it (in an apparatus note or through XML encoding)

#### Independent and dependent upadhmānīya and jihvāmūlīya

As noted in §4.1.5 (with Figure 4.1.C), upādhmānīya and jihvāmūlīya can either appear independently or behave like a consonant graph and form a conjunct with the following consonant. If these two behaviours alternate within a single text, or if the behaviour of these signs in a particular text is not what would be expected on the basis of related texts (in the same language, from the same region and time), then note the default behaviour in your palaeographic description and record the deviations from it. The character joiner sign = may in this case be optionally used in the edition to indicate the idiosyncratic association of the visarga variants.

* if the default behaviour is visarga-like, then idiosyncratic combining upadhmānīya and jihvāmūlīya may be transliterated ḫ= and ẖ=, indicating that it combines with the following consonant
  + e.g. Figure 4.1.C/2: ḫ= pu; Figure 4.1.C/3: traẖ= ka
* if the default behaviour is consonant-like, then idiosyncratic inline upadhmānīya and jihvāmūlīya may be transliterated =ḫ and =ẖ, indicating that it, like a visarga, is associated more closely with the preceding akṣara
  + e.g. Figure 4.1.C/5: yo=ẖ ka

#### Alternative behaviour of the superscript |r|

In most varieties of the Indic writing system, the grapheme <r> has, in addition to its basic form, two graphotactic allographs (§2.6) for use in complex characters. The superscript |r| (e.g. in Devanagari |र्क| <rka>) indicates that the phoneme /r/ is to be pronounced before the body consonant of a complex akṣara, while the subscript |r| (e.g. in Devanagari |क्र| <kra>) means that it is to be pronounced after the body consonant. Indonesian writing systems, however, often employ the superscript graph (called layar or surang) in an alternate mode in order to cater for closed syllables, where the /r/ phoneme represented by it is to be pronounced at the end of the sequence corresponding to the akṣara with the superscript r, as in |ᬲᬫᬃ| <*samar*> and in Figure 4.5.D. We shall refer to this as the Indonesian mode, as distinct from the Indian mode where the subscript |r| is to be interpreted in the traditional way, as in |ᬲᬯᬃ| <*sarva*>.

Because this is a systematic feature of certain writing systems, we prefer to transliterate both modes in the sequence in which they were intended to be pronounced. If a given text uses only one of the two modes, then it is sufficient to note this in the text’s palaeographic description. If both modes are present within a single text, then you must declare one of the modes (the dominant one) to be the default for the text and note the exceptions which use the other mode. In addition, you can use the optional character joiner sign = the graphic attachment of the <r> grapheme to the preceding or the following grapheme. This is recommended for the non-dominant mode in a text that uses both modes, and may (for the sake of explicitness) be used with every instance of a superscript |r| in any text where alternates modes might be applicable.

|  |
| --- |
| Figure .. Indonesian superscript |r| |
|  |
| Ina=rpaṇakan· |

* if the Indonesian mode is dominant in a text,
  + preferably add = to instances of the Indian mode
    - e.g. ᬲᬯᬃ → sar=va
  + for maximum precision, optionally also add = to instances of the Indonesian mode
    - e.g. ᬲᬫᬃ → sama=r
* if the Indian mode is dominant in a text,
  + preferably add = to instances of the Indonesian mode
    - e.g. ᬲᬫᬃ → sama=r
    - e.g. Figure 4.5.D → Ina=rpaṇakan·
  + for maximum precision, optionally also add = to instances of the Indian mode
    - e.g. ᬲᬯᬃ → sar=va

#### Multiple vowel markers within a complex glyph

The only systematic usage of more than one vowel marker graph in a complex glyph in our scope is the simultaneous presence of |u| and |i| in some writing systems of Mainland Southeast Asia. Since we prefer to see this as an established graphic sign for the distinct grapheme <ui>, it has been treated above under §4.2.2.1. When the function of the same vowel markers is to indicate deleted characters, then we are not dealing with a grapheme proper, but with a scribal mark, for which see §3.6.1.1.

Any other instance of more than one vowel marker in an akṣara is likely to be a result of non-standard orthography or simple scribal error (e.g. intended scribal correction from one marker to the other, without deletion of the former). XML markup is available for normalising non-standard usage (EGD §###) and correcting scribal error (EGD §###) or encoding scribal correction (EGD §###). Since our transliteration system employs lowercase Roman vowels for source vowels materialising as markers (and uppercase for those materialising as independent vowel signs), the transliteration of the received text in itself indicates unambiguously when more than one vowel marker is present in a source akṣara. Therefore, simply transliterate both vowel markers, in an arbitrary order that seems most plausible. Nonetheless, when deemed desirable, such occurrences in the source can be made explicit by adding the optional character joiner = between the two transliterated vowels. For example:

|  |  |
| --- | --- |
| Figure .. Multiple vowel markers | |
| 1 | 2 |
|  |  |
| du=ā | mr̥=i |

* Figure 4.5.E/1 → du=ā
  + this instance is probably a scribal mistake for an intended ddhā, and its editorial correction is to be encoded as per EGD §###
* Figure 4.5.E/1 → mr̥=i
  + the simultaneous use of the vowel markers |r̥| and |i| (as well as of subscript |r| and the |i| marker) to represent the sonant /r̥/ is not uncommon, so we prefer to perceive it as non-standard orthography (rather than a scribal mistake), and encode its normalisation as per EGD §###

#### Independent vowel signs as parts of complex glyphs

In some Southeast Asian varieties of the Indic writing system, some of the graphs which traditionally represent independent vowels can combine into complex characters with consonantal graphs. Situations where complex graphs representing independent vowels (i.e. combinations involving a vowel support) form ligatures with consonants have already been mentioned under §4.4.2 and are not relevant here, since the “vowel support” graph in that case represents a regular consonantal grapheme. This section is concerned with simplex vowel graphs which are normally standalone glyphs, but which form ligatures in certain specific writing systems, as in Figure 4.5.F.

|  |  |
| --- | --- |
| Figure .. Independent vowel graphs in ligatures | |
| 1 | 2 |
|  |  |
| maR̥k= R̥mpva | Umiṅsor= I |

Our transliteration scheme represents such graphs with uppercase Roman characters (§4.4.1). The same transliteration is to be used for the identical graphs when they enter into ligatures. However, since this would imply that the graphs transliterated in this manner are also independent glyphs (§2.4.1), it is preferable in this case to indicate with the character joiner sign that they form a complex character with an adjacent grapheme.

* where an independent vowel graph is combined with regular a consonant graph or a superscript |r| graph to form a complex character
  + transliterate the vowel in uppercase as usual
  + preferably add the = sign between the consonant and the vowel sign to indicate that the two belong to the same akṣara
  + e.g. Figure 4.5.F/1 → maR̥k= R̥mpva
  + e.g. Figure 4.5.F/2 → Umiṅsor= I

## Glyphs or graphs split by an intervening feature

finalise as per <https://github.com/erc-dharma/project-documentation/issues/284>

add mention of eventual encoding alternative as per <https://github.com/erc-dharma/project-documentation/issues/336>

and also inadvertent splits as in <https://github.com/erc-dharma/project-documentation/issues/237>

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Figure .. Split glyphs and graphs | | | | |
| 1 | 2 | 3 | 4 | 5 |
|  |  |  |  |  |
| malaṁka | dr̥vya | sa | ke |  |
|  |  |  |  |  |
| mala<>ṁka | dr̥v<>ya | sa<>⌉ | ⌈<>ke | A⌈<>horātri |

* certain glyph components are treated as separable in some scripts, such as the prescript and postscript vowel markers in Tamil கொ *ko* or the pr̥ṣṭhamātrā 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:
  + க<>ா as kā<>⌉
  + ெ<>க as ⌈<>ke
  + கெ<>ா as ko<>⌉ (likewise for split au)
  + ெ<>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) as follows:
  + line break: EGD §3.2.1 (if you are not using XML tags, start a new line in the e-text)
  + space imposed by a physical feature of the support: EGD §4.3.5 to §4.3.8 (if you are not using XML tags, use an \_ character as per TG §5.3.3)
    - e.g. *A⌈\_horātri* for the second line in the above copper-plate image
* 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 EGD §4.1.4 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

## Editorial interpretation

Transliteration is ideally a mechanical representation of the source graphemes, to be kept clearly apart from any editorial interpretation. However, the identification of those source graphemes is in fact already an interpretive act, and in the specific cases discussed below, we endorse adding a limited amount of editorial interpretation to the transliterated text.

### Silent identification of homographs

When determining the graphematic function of a graphic sign, it is sometimes impossible to collapse the solution space to a single point. The best practice in such cases is to establish the most likely reading and transliterate it, then address alternatives and uncertainties in XML markup (EGD §###) and/or a commentary (apparatus note), as applicable.

* note that alternative readings of a graphic sign may be of disparate nature, such as
  + the Tamil postscript vowel marker kāl (ா) or the character ra (ர)
  + a visarga or a punctuation mark
  + a consonant glyph with innate a or a final consonant
* when interpreting homographs and near-homographs, we recommend that you prioritise the alternative which is (in your, sometimes subjective, judgement) the most appropriate to the context
  + this includes extending the benefit of doubt to the scribe or engraver in the case of homographs and near-homographs such as
    - dependent short or long i
    - subscript ṇ or n
  + in such cases, prefer to transliterate the grapheme expected in the context unless the distinction is clearly made by the same scribe elsewhere in the original
    - in the latter case, transliterate the inferior grapheme and encode editorial correction (EGD §###)

### Poorly legible text

When transliterating text that is ambiguous or not confidently legible in the original due to damage to the support or unusual execution of graphic signs, the most likely reading must be selected as primary and transliterated. All uncertainty in reading is to be encoded in XML as per EGD §###. For partially legible akṣaras, feel free to use the shorthand introduced below.

#### Wildcards for unidentified consonants and vowels

* as public shorthand, you may use uppercase C for an illegible consonant component and uppercase V for an illegible vowel component of an akṣara
  + since final <c> and <v> are extremely unlikely to occur in the texts, this will not interfere with our use of uppercase Roman letters for final consonants
  + nonetheless, in documents encoded in XML, this shorthand must be converted to the appropriate encoding (EGD §###)

### Distinction of long ē and ō from short e and o

As noted in §4.1.3, writing systems of the period we work with very rarely make a graphic distinction between the short and long versions of /e/ and /o/, even when this is a phonemic distinction in the language being written. When the palaeographically primary graph is employed for writing a phoneme known to be long, the rigorous approach would be to transliterate the primary (short) grapheme and optionally add an editorial normalisation in XML (EGD §###). However, since distinct graphs for the long counterparts of these vowels do not, as a rule, occur in the texts in our purview, we have made a projectwide decision to adopt shorthand involving transliteration with a macron. The target graphemes ē and ō will therefore always be understood in our practice as editorial, equivalent to the XML markup for the normalisation of an originally written short vowel. As a consequence, any graphically distinct long vowels present in the source will have to be marked up explicitly as such in XML. This makes the editing of texts that do make this distinction more cumbersome, but greatly simplifies the editorial markup of the majority of the texts we edit.

* pending decision on github shorthand issue
* optional shorthand for cases where the generic (short) graph for |e| or |o| stands for the long phoneme /ē/ or /ō/ in a language where the length of these vowels is a phonemic distinction that has no graphic reflection in the writing system
  + use ē or ō instead of e or o to transliterate (actually, transcribe) these instances
* **if the distinction is absent in the source**, i.e. the graphs representing the phonemes /ē/ or /ō/ are graphically indistinguishable from those representing the phonemes /e/ or /o/, then
  + the default transliteration for all instances of the generic graph is e or o
  + for any instances of the generic graph which in your opinion represent the long phoneme, you may choose to add editorial markup, which may take the form of
    - either the above optional shorthand
    - or XML markup for editorial normalisation of a short vowel to a long one (EGD §###)
* **if the distinction is present in the source**, i.e. the graphemes <ē> or <ō> are graphically distinguishable from the graphemes <e> or <o>, then
  + the observation that the distinction is present must be explicitly made in the palaeographic description of the text
  + the default transliteration for all instances of the (theoretically) short grapheme is e or o
  + for any instances of the (theoretically) short grapheme which in your opinion represent the long phoneme, you may choose to add editorial markup, which may take the form of
    - either the above optional shorthand
    - or XML markup for editorial normalisation of a short vowel to a long one (EGD §###)
  + all instances of the long grapheme (whatever phoneme it may represent) must be transliterated with ē or ō and **mandatorily marked up** in XML as original (EGD §###)

### 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 may choose to add a breve to the transliterated short vowel as shorthand indicating that the short vowel in question is the original spelling (and not a typographic mistake in your edition). This is especially recommended for Sanskrit loanwords in Indonesian vernacular documents, following Damais (1955, 15), but need not be applied in such documents if they never make a graphic distinction between short vowels and their long counterparts. This shorthand is equivalent to XML markup for originally inscribed non-standard usage (EGD §###).

* optional shorthand for short vowels written where a corresponding long vowel is expected
  + use ă, ĭ or ŭ when a, i or u is used instead of expected ā, ī or ū

### Sandhi analysis

Since we are concerned with the transliteration of what is written in an original source, sandhi should, as a rule, never be resolved in our texts. However, when the addition of simple text-based markup can serve to highlight the presence of certain kinds of sandhi, we approve of using hyphens to indicate epenthesis (§4.7.5.1) and apostrophes to indicate elision (§4.7.5.2) when such markup is in accordance with the conventions of the subfield.

#### Epenthetic consonants

When an epenthetic consonant is inserted into certain sequences, you may use a hyphen joining that consonant to the preceding grapheme. The use of this analytical hyphen is optional. The epenthetic consonant is regarded as belonging to the former word, and should be separated from the next word by a space (§6.3) or, if in compound, by a hyphen (§6.4) or, optionally, nothing.

* examples in Tamil:
  + I-p-peruṅ-kōyil (இப்பெருங்கோயில் i+perum+kōyil)
  + tiru-mēṉi-y āṭa (திருமேனியாட tiru+mēṉi + āṭa)
* example in Kannada: samavr̥tti-y appavōl (ಸಮವೃತ್ತಿಯಪ್ಪವೋಲ್ samavr̥tti + appavōl)
* non-standard Sanskrit sandhi involving an epenthetic m, s or r may be indicated in the same way, e.g.
  + paṁca-s-triṁśottaratame
  + mleccha-rājye-m apūjitaḥ
* note the above hyphenation must not be used in the following cases
  + when epenthesis takes place on the phonological level, but its graphic reflection involves the alteration of graphemes rather than the insertion of an additional grapheme (as in Sanskrit tad dhi from tat + hi), then we have no choice but to treat the altered grapheme as part of the latter word, and an analytical hyphen must not be inserted
  + when a normally single consonant is doubled in non-standard Sanskrit orthography when in conjunction with an r (or, occasionally, in other formations), this is not regarded as epenthesis; analytical hyphenation is not applicable, and the doubled consonant is understood to belong with the word to which the single consonant belongs
    - e.g. tair ggatvā

#### Elision of final vowels

The elision of a final vowel, in particular of the “overshort” final u of Tamil and short final vowels of other Dravidian languages (lōpasandhi), may be indicated by adding an apostrophe in place of the elided vowel. The use of this analytical apostrophe is optional. The elided vowel is regarded as belonging to the former word, and so the apostrophe should be separated from the next word by a space (§6.3) or, if in compound, by a hyphen (§6.4). The apostrophe should not be used if elision takes place inside a compound that you do not segment with hyphens.

The mandatory use of a space after this apostrophe also serves as a mechanistic distinction from the apostrophe used for the avagraha (§5.2.1). Such a distinction is important because the apostrophe representing elision is a feature of linguistic analysis and constitute essential markup shorthand (§3.5.6), whereas the avagraha in Sanskrit is an actual grapheme which could conceivably be present in the original, and absent, it may be supplied by the editor for the sake of normalisation

* example in Tamil: arit’ eṉṟu (அரிதென்று aritu + eṉṟu)
* examples in Kannada:
  + enag’ īge (ಎನಗೀಗೆ enage + īge)
  + aṁt’ āgi (ಅಂತಾಗಿ aṁtu + āgi)
* do not use the apostrophe in lexicalised compounds, as well as in any compounds you choose not to segment with hyphens
  + e.g. in Tamil, koṇṭ-āṭu (கொண்டாடு koṇṭu + āṭu)
* the apostrophe should preferably be ’ (U+2019 Right Single Quotation Mark)
  + if necessary, the optional shorthand ' (U+0027 Apostrophe), which is accessible on most keyboards, may be used as an alternative

# Transliterating non-alphabetic signs

## Numeral signs

As outlined in §2.5.5, numeral signs can be considered graphemes for all practical purposes, and as such, we strive to transliterate them one-on-one. The Arabic numerals[[102]](#footnote-102) 0 to 9 provide straightforward equivalents for the corresponding ciphers in source writing systems. This takes care of decimal digits, i.e. the numbers 0 to 9 represented in any notation, as well as of all numbers represented in a decimal place-value notation,[[103]](#footnote-103) to be transliterated as per §5.1.1. However, no such equivalent offers itself for most numeral signs employed in sign-value notation (§5.1.2.2) with the exception of some fractions (§5.1.2.4), nor for numbers denoted by vertical bars (§5.1.2.3). To transliterate such signs, we must resort to deploying several target characters (usually, several Arabic numerals). In this case, the transliterated text requires markup (§5.1.2.1) expressing the fact that the target characters in question represent a single numeral sign of the original. The relevant markup is ideally in XML, but we provide shorthand options which may be used instead. Be mindful of the use of editorial spaces (§6.3.1.1) around numerals, regardless of whether their representation involves encoding, shorthand, or neither.

In XML editions, but nowhere else, all numbers written in numeral signs must, and all numbers written in words may be encoded so that the value of the number as a whole is recorded in the encoding, as described in EGD §###. This encoding is not applicable to (originally) numeral signs which are used in a function other than to denote numbers, such as the cipher <2> for the phonemes /ro/ (§4.5.2) and the cipher <1> used as an auspicious symbol (§###).

* @still need somewhere to treat alphanumeric signs repurposed as symbols, e.g. tha, cha and 1 as closers or auspicious marks
  + put this in 5.2, repurposed graphic signs?
    - but then it’s increasingly ill-placed before the numeral signs
    - so move numeral signs to a top-level section, and create another top-level section for special graphematic functions after that, which would then include all repurposed graphemes, 2-ro as well as these?
  + if they are in the same script as the rest of the text, then they are to be transliterated at face value regardless of assumed function
    - if they are in a different script (including ornamental modifications), then they are to be treated as logograms like oṁ

### The digits 0 to 9

|  |  |
| --- | --- |
| Figure .. Numeral signs 0-9 | |
| 1 | 2 |
|  |  |
| 1218 | 8 |

Any numeral sign which denotes an integer between 0 and 9 is to be transliterated plainly with the equivalent Arabic digit. The numeral signs themselves require no markup, but the value of the number is to be marked up in XML editions. These rules apply not only to multi-digit numbers represented in place-value notation (Figure 5.1.A/1), but equally to any digit that stands on its own (Figure 5.1.A/2), as well as to such digits additively combined with other numeral signs (Figure 5.1.B/2,3,5).

### Other numeral signs

Signs denoting numbers other than integers between 0 and 9 — unless they can be conventionally represented by a single target character as some fraction signs (§5.1.2.4) can — require markup (§5.1.2.1) in order to disambiguate them from sequences of graphemes. The various kinds of such numeral signs in our scope are discussed in the following subsections along with the relevant instructions.

#### Markup for numeral signs transliterated with more than one target character

Whenever a single numeral sign in the source is represented by a group of two or more characters in transliteration, the latter must be marked up to indicate that the target characters are to be understood together. Markup optimally takes the form of XML encoding, as per EGD §###. In contexts where XML is not involved (including the preparatory stage of XML editions, if desired), we recommend using the + (plus) sign for this purpose.

* as public shorthand, use the + (plus) sign after any group of target characters which together represent a single numeral grapheme of the source
  + although the use of the + sign is inspired by the conventional transliteration of sign-value ciphers in additive notation, in our scheme it does not signify addition, and its use is not identical to that conventional transliteration

#### Signs for numbers greater than 9

Integers 10 and above, when represented by a single source character, shall be transliterated with the corresponding multi-digit Arabic number followed by a + sign. Multiples of 100 and of higher powers of 10 are in many Indic writing systems represented by glyphs that are (or may arguably viewed as being) composed of two numeric graphs (for the respective power of 10 and a multiplier). For our purposes, this is irrelevant, and these glyphs shall be transliterated according to their final value.[[104]](#footnote-104)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Figure .. Numbers greater than 9 | | | | |
| 1 | 2 | 3 | 4 | 5 |
|  |  |  |  |  |
| 10+ | 80+ 2 | 200+ 80+ 2 | 100+ 20+ | 1000+ 4 100+ 4 10+ |

* in Figure 5.1.B/1, |10|, i.e. the number ten written with a single glyph, such as Brahmi |𑁛| → 10+
  + note that the + must be present even though nothing is added to this 10
* in Figure 5.1.B/2, |80| |2|, i.e. the number eighty-two written with a glyph denoting eighty and one denoting two, such as Brahmi |𑁢𑁓| → 80+ 2
* in Figure 5.1.B/3, |200| |80| |2|, i.e. the number two hundred and eighty-two written with glyphs respectively denoting two hundred, eighty, two → 200+ 80+ 2
* in Figure 5.1.B/4, |100| |20|, i.e. the number one hundred and twenty written with a glyph denoting one hundred and one denoting twenty → 100+ 20+
  + note that the + must be present after 10
* |80| |10|, i.e. the number ninety written (unusually) with a glyph denoting eighty and one denoting ten, such as Brahmi |𑁢𑁛| → 80+ 10+
  + note that the + must be present after 10
* in Figure 5.1.B/5, |1000| |4| |100| |4| |10|, i.e. the number one thousand four hundred and forty written in Tamil notation with glyphs respectively denoting one thousand, four [times], one hundred, four [times], ten → 1000+ 4 100+ 4 10+
  + note that the + must be present after 100, 100 and 10, but not after the instances of 4
* for the editorial spacing of such numbers, see §5.1.2.1

#### Numbers denoted by bars

|  |
| --- |
| Figure .. Number bars |
|  |
| III+ |

Cambodian inscriptions may denote numbers by groups of vertical bars (daṇḍa). This notation shall be transliterated by an identical number of I (uppercase i) characters to distinguish it from the use of regular Indic numeral signs. The transliteration must be followed by a + character in order to disambiguate these target characters from the transliteration of the independent vowel <I>. This applies even when there is only a single bar for the number 1.

* in Figure 5.1.C, three vertical bars denoting the number three → III+
* for the editorial spacing of such numbers, see §5.1.2.1

#### Fraction signs

|  |  |  |
| --- | --- | --- |
| Figure .. Numbers greater than 9 | | |
| 1 | 2 | 3 |
|  |  |  |
| ½ | ½ | 1/40+ |

The Unicode codetable provides for a number of “vulgar fraction signs”, some of which enjoy wide font support, while others are absent from most fonts. For a limited set of fractions, we therefore prefer to use the vulgar fraction signs, because they provide a single target character for the transliteration of a single source character, just as Arabic numerals do. For any other fractions which are represented by a single sign in your source text, use Arabic numbers to record the nominator and the denominator, separated by a slash and marked up as a single source character with the + sign.

* use the vulgar fraction signs for the following fractions:
  + one half → ½ (U+00BD Vulgar Fraction One Half), as in Figure 5.1.D/1
    - double-barred variants of the cross-shaped Khmer fraction sign, as in Figure 5.1.D/2, shall always be transliterated likewise, as ½
  + one third → ⅓ (U+2153 Vulgar Fraction One Third)
  + two thirds → ⅔ (U+2154 Vulgar Fraction Two Thirds)
  + one fourth → ¼ (U+00BC Vulgar Fraction One Quarter)
  + three fourths → ¾ (U+00BE Vulgar Fraction Three Quarters)
* for any fraction signs other than the above, transliterate as a common fraction using a slash, and add a + sign after the denominator, e.g.
  + one eighth → 1/8+
  + one fortieth → 1/40+ as in the Tamil fraction sign of Figure 5.1.D/1
* for the editorial spacing of such numbers, see §5.1.2.1

## Signs with special graphematic functions

refer in intro to the Pyu underdot treated among borderline diacritics

@abbreviation markers (like Devanagari ॰ also called lāghava sign) could be treated either here or with the space fillers and word breakers depending on whether classifying their form matters to us

* + I’m thinking their form matters no more than the form of an avagraha
  + they should be transliterated as ° and marked up in XML as <am> to disambiguate from the truncation sign
  + or rather, do encode them as symbols, add <am>, and mention under Functional symbols below
* @oṁ symbols could be treated like the Burmese abbreviation markers and in encoding, be transliterated as oṁ but tagged with g type om or suchlike
  + this doesn’t work for siddham, which doesn’t contains the graphemes for <siddham> while <oṁ> does, only not in the same script as the rest of a text
  + actually, the Burmese abbreviations should also be encoded as the generally accepted expansion of the abbreviation, tagged with g
  + and the generic type for that could be logogram (to include oṁ)
  + this could also be expanded to Tamil signs for words

### Avagraha

Any avagraha actually present in the source must always be transliterated. Avagrahas not present in the source may be supplied by the editor, marked up in XML as editorial (§5.2.1.1). Be mindful of spaces around avagrahas (both original and editorial), as instructed in §6.3.1.2.

* use the transliteration ’ (U+2019 Right Single Quotation Mark) to represent original avagraha
  + e.g. Devanagari |ऽ| → ’
* if necessary, the optional shorthand ' (U+0027 Apostrophe), which is accessible on most keyboards, may be used as an alternative

#### Editorial avagraha

Where an avagraha is expected according to the modern conventions of this sign’s usage, but none is present in the source, it may be supplied by the editor in order to facilitate the interpretation of the edited text. Editorial avagrahas require XML markup as per EGD §###. However, since the texts in our scope hardly ever contain an original avagraha, you may want to use shorthand instead.

Guidelines for supplying avagraha:

* supply avagraha only to indicate the elision of initial <a> after <e> or <o> in sandhi
  + never supply avagraha to indicate the merging of initial <a> or <ā> with a preceding <a> or <ā>, as is occasionally done in Devanagari texts
* supplying avagraha is recommended in general, and especially in cases where the received text would be meaningful (and often contradictory in meaning) both with and without an avagraha
  + for example, the inscribed sequence sohataḥ may stand for so hataḥ (“he was killed”) or so ’hataḥ (i.e. saḥ ahataḥ, “he was not killed”), so if you interpret the text as the latter, then supply an avagraha to make this clear
  + if you feel that such ambiguity is a deliberate poetic device (bitextuality, śleṣa), then the decision whether or not to supply an avagraha should be based on what you consider to be the prima facie meaning of the text
* the recommendation of supplying avagraha applies likewise to words in compound, whether hyphenated (§6.4) or not
  + e.g. yaśo’mr̥tam or yaśo-’mr̥tam and saro’nte or saro-’nte are preferable to yaśomr̥tam or yaśo-mr̥tam and saronte or saro-nte
* as private shorthand, you may use only ’ or ' (§5.2.1) for editorial avagrahas, and eventually supplement them with XML markup
  + if original avagrahas may also be present, then we recommend the shorthand +’ for editorial avagrahas
  + should your text, or some of the texts you work with, also contain apostrophes in other functions, be especially careful when replacing editorial avagrahas with markup
    - if the apostrophe for Dravidian elision sandhi (§4.7.5.2) may also be present, you can distinguish it mechanically from an avagraha by the fact that the elision apostrophe is always followed by a space, while the avagrha never is

### Burmese abbreviation signs

* @better switch back to \*n etc. as per the referenced guide, this would work better as generic shorthand for logograms such as \*oṁ
* Burmese abbreviation signs shall be transliterated by an alphabetic letter followed by an asterisk, according to the conventions of the field (Lammerts and Griffiths 2016, 3), e.g.
  + ၌ → n\*
  + ၍ → r\*
  + ၎ → r\*
  + ၏ → e\*
* if such abbreviations occur in your corpus, especially within the same text, then you must be careful in using the asterisk as shorthand for any other function
* @add reference to spacing (§6.3.1.4) once the status of logographic signs has been sorted out

## Symbols

In this Guide, ‘symbol’ specifically means a graphic sign which is non-alphanumeric and does not fulfil a special graphematic function as per §5.2. When transliterating a source text, we are only concerned with signs which are integral to the text. This criterion is occasionally ambiguous, but in general, signs which occupy the same kind of segmental space as alphanumeric signs, as well as signs which are set apart from the flow of the text but have a conventional linguistic signification, may be regarded as integral parts of the text. Although their precise graphematic function may not be definable, these signs are considered on a par with proper graphemes and are to be represented explicitly in transliteration. Conversely, signs peripheral to the inscribed text of an inscription or manuscript are to be ignored (§3.6.1).

Symbols integral to the text exhibit great graphic diversity in the source texts and can serve a variety of graphematic functions that cannot always be identified with precision. In our editions, we attempt to capture some of both kinds of variation by means of machine-readable XML encoding (EGD §###), and further encourage all editors to describe the visual appearance of symbols found in their texts in human-readable terms in the palaeographic description accompanying the edition. Nonetheless, we also wish to represent at least some of the phenomenal and functional variation already at the level of transliteration. The following subsections distinguish, and give instructions for the transliteration of, the functional categories of punctuation marks in a strict sense (§5.3.1), functional symbols (§5.3.2) and generic symbols (§5.3.3).

### Punctuation marks

In the terms of this Guide, ‘punctuation mark’ is used in a sense restricted to graphic signs employed in the original for syntactic or metrical segmentation into relatively small units, similar in function to a modern comma, full stop, question mark, exclamation mark, colon or semicolon.

* the scope of punctuation marks **generally includes** signs
  + whose shape is simple and abstract (non-figural), such as the vertical bars, dots, circles and dashes used widely for punctuation in the Indic writing system
    - or which are ornamental elaborations of such shapes
  + whose linguistic function is primarily to segment the text into relatively small units such as sentences, clauses, list items or metrical units
  + which occur repeatedly in the body of a single text
* the scope of punctuation marks **generally excludes** signs
  + whose shape is figural or complexly ornamental, and is not derived from one of the simple basic shapes used for punctuation
  + whose linguistic function may be to mark the beginning or end of an entire inscription or to segment a text into a small number of large units
    - but which, in addition to or instead of this linguistic function, usually have a non-linguistic (semasiographic, cf. note 4 in section §2.1) meaning such as auspiciousness, or serve as decorative elements
  + which occur only once per text or once per major section of text
  + signs of this nature are to be treated as generic symbols (§5.3.3)

#### Transliterating punctuation marks

In all editions, original punctuation must always be preserved in transliteration, but any punctuation supplied by the editor must be clearly distinguished from original punctuation (§5.3.1.2). For both original and supplied punctuation marks, be mindful of the use of editorial spacing (§6.3.1.3).

* editorial punctuation may, however, be supplied using XML markup, see EGD §6.3.6
* as outlined in the introduction to this section, punctuation marks shall be represented
  + 1. at the level of transliteration, by the dedicated character . (full stop, period), which by our convention shall be understood as an abstract punctuation mark without any assertion as to its physical appearance
  + 2. at the level of XML encoding, with a relatively simple classification of their shapes
  + 3. at the level of human-readable metadata, in additional human-readable detail
* as an intermediate step between levels 1 and 2 above, we recommend the use of shorthand markup for the basic forms of common punctuation characters, as follows:
  + all of the following shorthand characters should be followed by a space in transliteration, but not preceded by one
  + | (U+007C Vertical Line): for signs comprised of a single plain vertical bar (corresponding to the symbol token “danda”)
    - when transliterating two or more iterations of single vertical lines, make sure you add a space between them to differentiate them from double daṇḍas
  + || (U+007C Vertical Line, twice): for signs comprised of a double plain vertical bar (corresponding to the symbol token “ddanda”)
  + / (regular slash): for signs comprised of a single vertical bar with a hook, crossbar or ornamental addition (corresponding to the symbol token “dandaOrnate”)
  + // (two regular slashes): for signs comprised of a double vertical bar with a hook, crossbar or ornamental addition (corresponding to the symbol token “ddandaOrnate”)
  + , (comma): for short, predominantly vertical and often curved strokes normally floating at or above median height, including half-sized daṇḍas and the raised comma-like sign that is the basic punctuation mark on Java and Bali (modern Balinese ᭞) (corresponding to the symbol token “comma”)
  + ~ (U+223C Tilde Operator): for signs comprised of a single horizontal dash, plain or with ornamentation (corresponding to the symbol token “dash”)
  + @ (“at” sign) for any punctuation mark that does not easily fall under any of the above categories (such as more or less complex dots and circles)
    - the transformation of this character into XML markup will definitely not be automated and will have to be handled by you manually if you use this shorthand

#### Supplying punctuation

In your XML editions, never supply punctuation at the ends of stanzas and hemistichs, nor at the end of large semantic units which you encode as paragraphs (EGD §###). Any original punctuation at such points must be preserved in transliteration as it is. Supplying punctuation at the ends of smaller semantic units, such as sentences, is permitted and recommended whenever you feel that this is helpful to the reader of your edition. Punctuation supplied by the editor must always be marked up in XML as such (EGD §###).

* as private shorthand, you may use a . (full stop, period) for supplied punctuation while preparing your edition, then convert it to the proper encoding

### Functional symbols

To the category of ‘functional symbols’, we assign non-alphanumeric signs which have a confidently identifiable graphematic function[[105]](#footnote-105) other than punctuation, in particular space fillers (§5.3.2.1) and word breakers (§5.3.2.2). Our treatment of such symbols differs from that of the avagraha (§5.2.1) in that we wish to represent both their function and their graphic appearance in our editions. When transliterating or encoding such symbols, be mindful of the use of editorial spaces around them (§6.3.1.4), bearing in mind that space fillers may, and word breakers by default do, occur inside words.

#### Space filler symbols

@write when symbol encoding finalised

* in the terms of this Guide, “**space filler**” is used in a sense restricted to symbols whose function is clearly and unambiguously to fill up space in a line to the binding-hole or margin
  + symbols that do not meet this functional criterion shall be encoded as generic symbols even if they are identical in visual appearance to symbols used as space fillers elsewhere in the document or the corpus
  + @keep or discard §abc as shorthand
    - ADD PRIVATE SHORTHAND LABEL IF NOT DISCARDING
  + instead or in addition, introduce § as public shorthand

#### Word break indicators

@write when symbol encoding finalised

* see discussion in Taxonomy
* difficult to distinguish from space fillers
  + if the same sign occurs in unbroken words, space filler
  + if narrow sign, probably word breaker
* @introduce public shorthand ¬ (U+00AC Not Sign) for this purpose instead of former editorial hyphen at end of line

### Generic symbols

@write when symbol encoding finalised

We use the term ‘generic symbol’ for any symbol that has not been confidently identified as fulfilling the function of a punctuation mark (§5.3.1) or other functional symbol (§5.3.2). When transliterating or encoding symbols, be mindful of the use of editorial spaces around them (§6.3.1.4).

* note that auspicious (maṅgala) symbols should never be transliterated as the words siddham or om̐
* as outlined in the introduction to this section, generic symbols shall not be represented at the level of transliteration and should ideally always be handled in XML markup (supplemented by human-readable description)
  + the essential feature of the relevant XML markup is the use of symbol tokens (EGD §4.2.3)
* to simplify your work, especially when you are creating an e-text that will not (yet) be marked up in XML, you may choose either of the following shorthand methods for representing generic symbols
  + as tokens, using $abc
    - where “abc” (any sequence of letters, followed by a space) will be converted into a symbol token in the XML tag representing the symbol
    - ADD PRIVATE SHORTHAND LABEL IF NOT DISCARDING
  + as public shorthand, feel free to use dingbats of your choice, i.e. any Unicode character approximating the original glyph (e.g. ◊ 卐 ✤ ⁜)

## Blank space

As noted in §3.5.2, all spaces in our transliterated texts will be considered editorial. Conversely, spaces present in the source and deemed significant must always be represented in XML markup as per EGD §###. To speed up work and to provide a means of representing original spaces in text that will not be encoded in XML, we suggest using the \_ (underscore) sign as shorthand. Be mindful of the use of editorial spaces (§6.3.1.5) around both encoded and shorthand representations of original space.

* whether a space in the source is significant is up to your discretion, but here are some rules of thumb
  + spaces for layout are never significant, including
    - blank space at the left of a right-aligned line or at the right of a left-aligned line
    - spacing between most or all characters of a line justified to the two margins
    - spacing between metrical units when these result in a column-like arrangement for an entire text or section
    - layout may be represented by various encoding methods (EGD §###)
  + small spaces (less than a typical character width) around numerals, punctuation marks and other symbols are generally not significant
  + spaces (including small spaces) used in lieu of punctuation (i.e. for semantic segmentation) are generally significant
* as public shorthand, use the \_ (underscore) sign to represent original spaces in transliterated text
  + if you plan later to batch convert the shorthand to XML encoding, then it is recommended that you use this only for basic interword spaces, which are large enough to be called a space but smaller than the width of two average characters
  + any other spaces — such as space left blank for filling later, or because of a defect or feature of the material — can only be handled in XML
    - if you frequently encounter such spaces in your work, feel free to devise and employ private shorthand for them

# Segmentation in transliterated text

In addition to transliterating a source text, you will usually want to segment it in one or more ways. Division into larger blocks may be descriptive — on the basis of the physical layout of the original — or interpretive — on the basis of the semantic and/or prosodic structure of the text. Both of these concerns are to be addressed primarily through XML markup, but they are discussed briefly in §6.1.

However, for a lower level of semantic analysis, we prefer simple editorial spacing (§6.3) of independent words and hyphenation (§6.4) of words in compound. The same kind of segmentation could be achieved, in some cases more rigorously, through computer markup, but at present we see no advantage in doing so. Conversely, due to the ubiquity of such segmentation, using XML for this purpose would greatly increase code clutter and impede working with our files. Using basic markup consisting in spaces and hyphens keeps the files more human-readable, but it may in the future serve to facilitate the electronic lemmatisation of texts. Furthermore, it not only helps the reader understand the texts you edit, but, in ambiguous cases, expressly communicates your parsing of the text. Finally, it is in many cases advantageous for on-screen display and typesetting.

The insertion of editorial segmentation is facilitated by Romanised transliteration, but hampered in some cases by vowel sandhi. We discuss the consequences of both in §6.2. We also provide some general considerations for cases when, and when not, you should use editorial spaces and hyphens (§6.5).

## Descriptive and interpretive blocks

Editions encoded in XML must accurately represent both the extrinsic structure of the source (EGD §###) — i.e. how the source text is broken up into inscribed lines and, if applicable, into pages or other distinguishable inscribed zones — and its intrinsic structure (EGD §###) — i.e. how it is constituted of semantic or prosodic units (paragraphs, stanzas and verse lines). When preparing a draft for an XML edition or transliterating a text that will not be encoded, you may create new lines for either or both of these purposes and, if applicable and desired, add line numbering and/or headers for specific groups of lines. When converting to XML encoding, all such segmentation and labelling must be removed.

Whether you are using new lines or XML tags for epigraphic lines and zones (i.e. descriptive blocks) and for units of intrinsic structure (i.e. interpretive blocks), always be mindful of the proper use of spaces (§6.3.1) and hyphens (§6.4.1) at such points.

* when a line break, whether representing extrinsic or intrinsic structure, falls inside a word
  + this must normally be encoded in XML
    - see EGD §### about inscribed lines cutting words, and §### about verse lines cutting words
  + when XML encoding is not involved, a hyphen may be used as public shorthand at the end of the former line
* when an editorial hyphen (for compound segmentation) coincides with a line break, put it at the beginning of the latter line (§6.4.1)

## Segmentation versus akṣaras and sandhi

While the boundaries of morphemes are frequently hidden inside an akṣara of Indic script and thus not separable, Romanisation allows us to separate any two adjacent graphemes for the purpose of analysis. Thus, wherever interpretive segmentation of any kind is applicable, akṣara boundaries can and must be disregarded. However, some of our source languages feature vowel fusion sandhi, whereby the final vowel of one word and the initial vowel of the next word merge into a single vowel phoneme, represented in writing by a single grapheme. In such cases we transliterate the resulting grapheme faithfully and forego interpretive segmentation. We furnish some specific cases and examples below.

* where interpretive segmentation is called for, it can and must be used even if a morpheme boundary is within an akṣara in the source text, including
  + independent words, e.g. Sanskrit तदेव → tad eva
  + words in compound, e.g. Sanskrit तत्पुरुष → tat-puruṣa
* this continues to apply if graphemes on either side of the boundary are altered due to sandhi or orthographic convention, so long as the graphemes can be allocated to one side of the boundary or another, including
  + the reduction of vowels to semivowels, e.g. in Sanskrit
    - गच्छत्येव → gacchaty eva (from gacchati + eva)
    - मन्वादि → manv-ādi (from manu+ādi)
    - महर्षि → maha-rṣi (from mahā+r̥ṣi)
  + the use of the class nasal where standard orthography would employ an anusvāra, e.g. in Sanskrit
    - उक्तञ्च → Uktañ ca
    - शरणङ्गतः → śaraṇaṅ gataḥ
  + the assimilation of initial /h/ to a preceding stop, altering the /h/ to a corresponding aspirate, e.g.
    - Sanskrit तद्धि → tad dhi (for tat + hi) and तद्धित tad-dhita (for tat+hita)
    - Old Javanese buAt-thajyanya (derived from bvat haji, §6.5.1)
    - note that in these cases, the phonetic sequences may be more appropriately segmented into /tadd hi/ and /buatt-hajyanya/, but the written form involves the graphemes <dh> and <th>, which, though transliterated by digraphs, are indivisible
  + the generation of graphemes not originally present in either of the morphemes, to be segmented as in the examples below
    - Sanskrit putrām̐l lakṣmīḥ (from putrān + lakṣmīḥ)
    - Sanskrit dīnārair ddaśabhiḥ (orthographic variation of dīnārair daśabhiḥ)
    - Old Javanese darpaṇa ryy avakta (orthographic variation of darpaṇa ry avakta)
    - Old Javanese tann inaku (from tan + inaku)
    - Tamil இப்பெருங்கோயில் → I-p-peruṅ-kōyil (from I + perum + kōyil)
    - Tamil திருமேனியாட → tiru-mēṉi-y āṭa (from tiru + mēṉi + āṭa)
      * see §4.7.5.1 for the use of the hyphen in Tamil sandhi analysis
  + the complete elision of one of the phonemes (and the corresponding grapheme)
    - Tamil arit’ eṉṟu (for aritu + eṉṟu)
      * see §4.7.5.2 about the elision of final u in Tamil
    - the elision of initial a after a preceding o or e in Sanskrit does not involve morpheme boundaries inside an akṣara, so it can be segmented without any difficulty, e.g. tato pi (for tataḥ + api)
* conversely, the fusion of a final and an initial vowel into a single vowel must be distinguished from elision, and can never be segmented, including the following cases in Sanskrit
  + kr̥tārtha (from kr̥ta+artha) and sāpi (from sā + api)
    - note that editorial avagraha (§5.2.1.1) must never be added to indicate the merging of initial a or ā into a preceding a or ā
      * should one or two original avagrahas be present in such a case, they must of course be transliterated (§5.2.1)
  + gacchatīva (from gacchati + iva)
  + seyam (from sā + iyam)
  + gataujas (from gata+ojas)

## Editorial spacing

Editorial spaces should normally be inserted between independent words, while words in compound should be segmented with hyphens. In addition to obviously independent words — including inflected nouns, finite verbs, verbal derivatives, invariant particles, etc. — fixed expressions should in general be regarded as phrases rather than compounds, and their members should accordingly be separated by spaces. Suggestions for segmentation in specific situations are given in §6.5, especially §6.5.1 to §6.5.4. In spacing your texts, make sure you are familiar with §6.2 about the interaction of segmentation with akṣaras and sandhi.

The good practice guidelines in §6.3.1 improve consistency across the project and, in some cases, prevent conflict, redundancy or ambiguity when a text is encoded in XML and rendered for display. The additional subsections contain practical suggestions and rules for using spaces in various contexts and situations. These apply equally to text which is encoded in XML and to text which is not.

### Good practice in editorial spacing

Avoid redundant spaces; in particular:

* never start a line with a space
  + when creating a draft for an XML edition, where you simply enter line numbers at the beginning of each line of the source, make sure you remove the intervening space when converting the numbering to XML markup
* never use more than one space at any point

#### Space and numerals

Around numeral signs, use spaces as follows.

* numeral digits 0-9 (§5.1.1) must be separated by an editorial space from any adjacent text, symbols or numeral signs
  + except that additional digits belonging to the same multi-digit number written in place-value notation must not be separated by spaces, as in Figure 5.1.A
* all other numeral signs (§5.1.2) must be separated by an editorial space from any adjacent text, symbols or numeral signs
  + this includes groups of numeral signs that represent a number together, as in Figure 5.1.B/2-5
  + no spaces must be added
    - between the target characters that stand together for a single numeral sign in the source
    - between the target characters and a + sign used as shorthand (§5.1.2.1)

#### Space and avagraha

The avagraha, editorial or original (§5.2.1), is to be spaced as follows.

* avagraha must never be separated by a space from the following word, to which the elided vowel belongs
* avagraha must always be separated by a space from a preceding independent word, e.g. so ’bhūt and not so’bhūt
* if the word preceding the avagraha is compounded to the next word, then a space must not be inserted, but a hyphen may be used for compound segmentation, e.g. saro’nte or saro-’nte

#### Space and punctuation marks

Around punctuation marks (as defined in §5.3.1), spaces should be deployed as follows.

* never add a space before a punctuation mark
  + even though editions of Indic texts (both in transliteration and in Indic scripts) often do so
* always add a space after punctuation marks if they are followed by text
  + except when a punctuation mark is (for whatever reason) inside a word
* should several punctuation marks appear in a group, do add spaces between them
  + this is particularly important if you use shorthand to transliterate certain punctuation marks

#### Space and symbols

Around all other symbols including logographic symbols (§###), functional symbols (§5.3.2) and generic symbols (§5.3.3), use spacing as follows.

* + symbols must generally be separated by a space from any other characters adjacent on either side
    - including alphabetic graphemes, numeral signs and other symbols
    - except when a symbol is (for whatever reason) inside a word, in which case there should be no spaces around it

#### Space and original space

Around a significant space in the source, whether it is encoded in XML or represented by shorthand (§5.4), use editorial spaces as follows.

* original spaces must generally be separated by an editorial space from any characters adjacent on either side
  + including alphabetic graphemes, numeral signs and other symbols
* except when an original space is (for whatever reason) inside a word, in which case there should be no editorial spaces around it

## Editorial hyphenation

Compound words should generally be segmented with hyphens, particularly in Sanskrit and other compound-heavy languages, while complex morphemes other than compounds — such as new words derived by affixation — should not be segmented analytically. Suggestions for segmentation in specific situations are given in §6.5, especially §6.5.7 on nominal compounds. The hyphenation of compounds is optional (though encouraged) and need not be exhaustive. Hyphenation is primarily intended to help the reader reconstruct how you analyse compounds, and should be used at your discretion. In hyphenating your texts, make sure you are familiar with §6.2 about the interaction of segmentation with akṣaras and sandhi.

The good practice guidelines in §6.4.1 improve consistency across the project and, in some cases, prevent conflict, redundancy or ambiguity when a text is encoded in XML and rendered for display. They apply equally to text which is encoded in XML and to text which is not.

### Good practice in editorial hyphenation

* use hyphens only for the purposes endorsed by this guide, i.e.
  + normally, only for the segmentation of compounds
  + for a specific kind of sandhi analysis (§4.7.5.1)
  + in specific circumstances, as public shorthand for words cut across inscribed lines (§6.1)
* editorial hyphens will normally have alphabetic graphemes on both sides
  + in rare cases, an editorial hyphen may have to be placed next to a different feature, which is normally represented by XML encoding, but may in some circumstances be represented by shorthand
    - such features include symbols, original spaces, the ends of epigraphic lines, and the ends of verse lines
    - in all of these cases, the editorial hyphen must be placed after the feature

## Segmentation guidelines

It is in some cases difficult to decide whether certain morphemes of a source language are to be considered dependent (bound) or independent (free), and whether a dependent morpheme is a compound member or an affix. The prevailing conventions vary considerably, so — depending on your background — you may be tempted to segment differently than what we suggest. We cannot provide exhaustive rules, much less a rationale for such rules, but in order to increase homogeneity of practice across the project, we propose some rules of thumb and considerations for editorial analysis and recommend that you override these only with good reason. Most of our discussion primarily concerns Sanskrit, while some additional (and partly different) rules are given for Tamil and Old Javanese. When working with other languages, make an informed choice after considering any parallelisms with the guidance given here.

### Phrases

On the whole, words constituting a phrase must always be separated with spaces. However, phrases may become parts of compounds, or may be subject to morphological derivation with affixes. In both of these cases, they are to be treated as compounds. Thus,

* in Sanskrit:
  + iti kartavyam → iti-kartavyatā
* in Old Javanese:
  + tahi tikus → manahi-tikusa
  + bvat haji → buAt-thajyanya
    - for the positioning of the hyphen in this second example, see also §6.1

### Grammaticalised structures

Combinations of verbs and substantives with other words should be understood as phrases wherever reasonably possible. This includes grammaticalised structures (performing a function like that of conjugation or declension). For example, in Sanskrit,

* periphrastic perfects, e.g. varayāṁ cakāra
  + especially since other words may intrude inside such a construction, e.g. saṁraṁjayāṁ ca prakr̥tīr babhūva
* past tense formed with imperfect and sma, e.g. samādiśati sma
* verbal prefixes used as prepositions with substantives, e.g.
  + ā samudrāt
  + anu gaṅgām

### Multiple function words

Pairs or groups of function words (mainly conjunctions) are to be considered separate, even when frequently used together in a meaning that is not evident from the individual meanings of these words. For example, in Sanskrit,

* atha vā
* kiṁ ca and kiṁ tu (even if spelt kiñ ca and kin tu)
* api ca and api vā
* anyac ca
* tad yathā
* na hi
* etc.

However, for such **Sanskrit structures borrowed into other languages**, forego segmentation if the structure is listed as a single word in the relevant dictionaries, where applicable. For example,

* Old Javanese kimuta
* Old Cam kintu

### Repetitive structures

When words (chiefly pronouns or substantives) are iterated with the same inflectional ending in order to express a generalised or distributive meaning, each iteration is to be treated as an independent word. Repetitions of this kind should therefore not be hyphenated, even though they are sometimes classified as a special type of compound (āmreḍita).

* words iterated with the same inflectional ending shall be spaced, for example
  + yasya yasya
  + dine dine
* when the first iteration does not have an inflectional ending, the formation is of course to be treated as a compound and accordingly hyphenated (if sandhi allows), for example
  + ekaikam (from eka+eka, not segmentable)
  + pūrva-pūrvāḥ
* in more complicated cases, proceed according to your discretion
  + hyphenation may be preferred in the following cases
    - if the word forms used may be either inflected or uninflected, as in ahar-ahar
    - if the iterations have different inflectional endings, as in ekam-ekāḥ
  + hyphenation or no segmentation may be preferred when such a doubled word is followed by other compound members or a suffix, e.g. dinaṁ-dinaṁ-vivardhamāna
  + no segmentation is to be preferred when an iterative formation is well established in a particular meaning, as in paramparā
* however, **repetitions in Old Javanese** are to be considered compounds (§6.5.8)

### Quasi-compounds

In non-standard Sanskrit, strings of nouns may appear without (or with occasional) inflectional endings. Such strings constitute neither a proper compound nor a syntactically correct phrase without emendation. They are much like compounds, but in order to avoid an interpretation that their author probably did not mean, we generally prefer to separate them with spaces.

* for example,
  + lamvoṣṭha dedamita mahādeva guṇṭhaka ity evam-ādibhyo
    - compare lamvoṣṭha-dedamita-mahādeva-guṇṭhakety-ādibhyo, lamvoṣṭho dedamito mahādevo guṇṭhaka ity evam-ādibhyo, etc.
  + samrāṬ vākāṭakānāṁ mahārāja śrī-pravarasenasya
    - compare samrājo vākāṭakānāṁ mahārājasya śrī-pravarasenasya, samrāḍ-vākāṭaka-mahārāja-śrī-pravarasenasya, etc.
* do feel free to hyphenate such structures if you feel that this is helpful for the reader
* such structures may be normalised or emended in XML markup (EGD §###), but this is only recommended if minor editorial intervention can render them into standard language, or if their interpretation would be ambiguous without intervention

### Verbal formations

Sanskrit structures whose posterior member is a verb (typically from the roots kr̥, as, bhū and gam) or its derivative may often be perceived either as verbs with special prefixes, as compounds involving verbs, or as adverbial phrases. The segmentation of such structures is left to your discretion, based on whether you deem the structure to be separable. We offer the following specific considerations.

* structures whose anterior member is an adverbial morpheme that only occurs when prefixed to verbs — such as tiras, tiras, puras and āvis — may be hyphenated or treated like regular verbal prefixes and left unsegmented, e.g.
  + āvir-bhavati or āvirbhavati
  + tiro-bhūta or tirobhūta
* structures whose anterior member is a noun without any affix — such as alam, namas and śrad — should normally be hyphenated
  + e.g. namas-kr̥tya, śrad-dadhāmi
  + but may at your discretion be treated as phrases (and thus spaced) when the nominal part could occur without the verbal part
    - e.g. astaṁ gacchati, astaṁ gataḥ (compare astaṁ ca savitā yātaḥ)
* formations expressing becoming or making into something, formed with a nominal anterior member derived with the ending -ī or -ū (the taddhita suffix cvi), should be preferably hyphenated, e.g.
  + svī-karoti, svī-kr̥tya
  + vaśī-bhavati, vaśī-bhūta
  + tanū-karoti
* structures with adverbial anterior members formed from nouns with suffixes (taddhita) other than those listed above are best treated as phrases and thus segmented with spaces, e.g.
  + brāhmaṇasād gatāḥ
  + khaṇḍaśaḥ karoti

### Nominal compounds

Nominal compounds should, as a rule, be segmented with hyphens unless they are proper names (§6.5.7.2). Compound hyphenation is optional and intended mainly to help the reader understand the text and follow your interpretation. General considerations for doing so are addressed below, while §6.5.7.1 discusses particular cases where hyphenation is not always desirable. Note that words derived from nominal compounds (§6.5.8) should be left unsegmented.

* regular nominal compounds are to be segmented with hyphens where sandhi permits (cf. §6.2)
  + Sanskrit mahā-devī, nānā-śāstrābhyāsopabr̥ṁhita-niśita-vimala-buddhiḥ
  + Tamil tiru-makaḷ
* the scope of Sanskrit nominal compounds includes those where the anterior member retains a declensional ending (aluk samāsa)
  + e.g. ante-vāsin, bhayaṁ-kara
  + however, in classical Sanskrit, most such compounds tend to be basic compounds (§6.5.7.1) and names (§6.5.7.2), so it is often preferable not to segment them
    - e.g. dhanaṁjaya, puraṁdara
* in **long compounds**, it is generally preferable to leave smaller, close-knit subunits — especially basic compounds (§6.5.7.1) — without analysis, e.g.
  + aneka-tulābhāra-dāna-dhārā-prakṣālita-kalmaṣaḥ (not an-eka or tulā-bhāra)
  + paramabhaṭṭāraka-mahārājādhirāja-rājaparameśvaraḥ (not parama-bhaṭṭāraka-mahā-rājādhirāja-rāja-parameśvaraḥ)
  + snāna-pavitrīkr̥ta-śirasāṁ (not -pavitrī-kr̥ta-)
* in **complex compounds** involving subordinate relationships between some of the members, hyphenate in the way you deem best suited for interpreting the compound, including
  + hyphenating where you would otherwise prefer not to hyphenate, e.g.
    - a-cāṭa-bhaṭa-prāveśya (not acāṭa-)
    - aśva-gaja-śāstra-jña (not -śāstrajña)
    - asi-carman-vat (not -carmanvat, cf. §6.5.8)
  + not hyphenating where you would otherwise prefer to hyphenate
    - e.g. brahmadeyī-kr̥tya (not brahma-deyī-)
* in a poetic text where a **bitextual reading** (śleṣa) depends on two alternative segmentations of the same sequence of graphemes, choose one of the following options at your discretion
  + forego all or part of the segmentation so as not to impose either segmentation on the text
  + prioritise the meaning you translate as primary, segment according to that, and optionally mention the alternative segmentation in a note to your translation

#### Basic compounds

By ‘basic’ compounds we refer to short compound nouns whose meaning is either straightforward or widely known. Such compounds do not necessarily have to be analysed into their members, and are in fact better left unsegmented when they are part of a longer compound. They include in particular the following categories.

* compounds with a conventional meaning that cannot be derived straightforwardly from the meaning of the members, e.g.
  + mahā-rāja (‘great king’) or mahārāja (a particular kind of ruler)
  + nr̥-pati (‘man-lord’) or nr̥pati (a king)
  + dina-maṇi (‘day-jewel’) or dinamaṇi (the sun)
  + turaṁ-gama (‘quickly-goer’) or turaṁgama (a horse)
* compounds whose anterior member cannot occur on its own and is thus rather a prefix, i.e. a-/an-, sa-, su- and ku-
  + e.g. an-artha, sa-putra, ku-rājan
* compounds whose posterior member is a verbal root (with or without an explicit kr̥t suffix such as -t)
  + e.g. śāstra-jña, śatru-jit
  + especially if such a compound has a non-evident conventional meaning, e.g.
    - dvi-ja (‘twice-born’) or dvija (a bird; a member of the upper classes)
    - madhu-kara (‘honey-maker’) or madhukara (a bee)
* compound cardinal numerals, e.g.
  + dvā-daśa
  + pañca-viṁśati
  + see also §6.5.8 about ordinals
* nominal derivatives of verbal formations for which hyphenation is recommended in §6.5.6, e.g.
  + namas-kr̥ta
  + āvir-bhūta
  + svī-kr̥ta

#### Proper names and styles

Proper names should not normally be segmented into compound members.

* e.g. jayasiṁha, puṇḍravardhana, not jaya-siṁha, puṇḍra-vardhana
* compound proper names may be segmented on a case by case basis when this is deemed helpful for interpretation, for example when the literal meaning of a name or part of a name is foregrounded
* do use hyphens to separate honorifics and titles from names
  + e.g. śrī-jayasiṁha-vallabha, viṣṇuvardhana-mahārāja
  + long sequences of such items may be treated as quasi-compounds (§6.5.5)
  + when such words seem to be used as part of a name rather than styles attached to it, segment on the basis of what you interpret to be the name, e.g.
    - bhīmarājākhyaḥ (a person named Bhīmarāja rather than a king named Bhīma)
    - śrīdharaśarman (a Brahmin named Śrīdharaśarman rather than ‘the honourable’ Dharaśarman)

### Derivatives of compounds

Secondary derivatives of Sanskrit compound nouns are not themselves compounds and should therefore not be hyphenated.

* for example, cāturvarṇya is not a compound
  + it is derived from the compound catur-varṇa, but is itself not composed of the members \*cātur and \*varṇya and should not be segmented
* when the derivation does not involve a vr̥ddhi alteration of a vowel, it is still better to forego hyphenation in order to help with the interpretation
  + e.g. asicarmanvat, ‘having a sword and a shield’, consists of the suffix -vat added to the compound asi-carman as a whole, rather than being a compound of asi and carmanvat
    - at your discretion, if you feel this is helpful for the analysis of a complex compound, instead make an exception and separate the suffix, as well as all members, with a hyphen (cf. §6.5.7)
* compound ordinal numbers are also derivatives of compounds and should not be hyphenated, e.g.
  + caturviṁśa
  + caturviṁśatitama

### Affixes and clitics

Morphemes which are not considered independent in a given language should normally not be separated by any segmentation from the word or stem with which they are associated. In particular:

* the following Tamil formations **must not be separated** from the words to which they are attached
  + enclitic particles (e.g. ē, ō)
  + forms of the verb āku-tal (e.g. āṉa, āy, āka) when used adverbially
* the following Old Javanese formations **must not be separated** from the words to which they are attached
  + enclitic pronominal suffixes (-(ṅ)ku etc.)
  + possessive constructions built with the linker -ni (-nikaṅ, etc.)
  + the definite article -*ṅ*
  + the conjunction -*n*

# References

@hard old

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. forthcoming. *Representing Kannada Text*. [consulted in a draft stage]

Wellisch, Hans H. 1978. The Conversion of Scripts—Its Nature, History, and Utilization. New York: Wiley.

@soft new

Coulmas, Florian. 2003. *Writing Systems: An Introduction to Their Linguistic Analysis*. Cambridge University Press.

———. 2006. *The Blackwell Encyclopedia of Writing Systems*. 4. Oxford: Blackwell.

Daniels, Peter T. 2018. *An Exploration of Writing*. Sheffield: Equinox.

Fedorova, Liudmila. 2013. ‘The Development of Graphic Representation in Abugida Writing: The Akshara’s Grammar’. *Lingua Posnaniensis* 55 (2): 49–66. [DOI: 10.2478/linpo-2013-0013].

Gnanadesikan, Amalia E. 2017. ‘Towards a Typology of Phonemic Scripts’. *Writing Systems Research* 9 (1): 14–35. [DOI: 10.1080/17586801.2017.1308239].

International Organization for Standardization. 2001. *International Standard ISO 15919: Information and Documentation, Transliteration of Devanagari and Related Indic Scripts into Latin Characters = Information et Documentation, Translittération Du Devanagari et Des Écritures Indiennes Liées En Caractères Latins*. Geneva: International Organization for Standardization.

ISO/IEC 10646. 2020. *International Standard ISO/IEC 10646. Information Technology — Universal Coded Character Set (UCS)*. 6. Geneva: ISO. [https://standards.iso.org/ittf/PubliclyAvailableStandards/c076835\_ISO\_IEC\_10646\_2020(E).zip].

Iyengar, Arvind. 2024. ‘The Akshara as a Graphematic Unit’. In *Proceedings: Grapholinguistics in the 21st Century 2022*, 419–36. Fluxus Editions. [https://hdl.handle.net/1959.11/63505].

Lammerts, D. Christian & Arlo Griffiths. 2016. ‘Standardized and Simplified Systems for the Transliteration of Old Burmese, Burmese, Pali, and Sanskrit Written in the Burmese Script (Version 2.2)’. [https://sites.rutgers.edu/dclammerts/transliteration/].

Meletis, Dimitrios. 2019. ‘The Grapheme as a Universal Basic Unit of Writing’. *Writing Systems Research* 11 (1): 26–49. [DOI: 10.1080/17586801.2019.1697412].

———. 2020a. *The Nature of Writing: A Theory of Grapholinguistics*. Grapholinguistics and Its Applications 3. Brest: Fluxus Editions. [DOI: 10.36824/2020-meletis].

———. 2020b. ‘Types of Allography’. *Open Linguistics* 6 (1): 249–66. [DOI: 10.1515/opli-2020-0006].

Meletis, Dimitrios & Christa Dürscheid. 2022. *Writing Systems and Their Use: An Overview of Grapholinguistics*. De Gruyter Mouton. [DOI: 10.1515/9783110757835].

Neef, Martin. 2015. ‘Writing Systems as Modular Objects: Proposals for Theory Design in Grapholinguistics’. *Open Linguistics* 1 (1). [DOI: 10.1515/opli-2015-0026].

Nöth, Winfried. 1990. *Handbook of Semiotics*. Advances in Semiotics. Bloomington: Indiana University Press.

Ollett, Andrew & Sarah Pierce Taylor. forthcoming. ‘Representing Kannada Text’.

Salomon, Richard. 1998. *Indian Epigraphy: A Guide to the Study of Inscriptions in Sanskrit, Prakrit, and the Other Indo-Aryan Languages*. South Asia Research. New York: Oxford University Press. [http://archive.org/details/indianepigraphyg0000salo].

———. 2003. ‘Writing Systems of the Indo-Aryan Languages’. In *The Indo-Aryan Languages*, edited by George Cardona and Dhanesh Jain, 75–114. Routledge Language Family Series. Oxford: Routledge.

Weingarten, Rüdiger. 2013. ‘Comparative Graphematics’. In *Typology of Writing Systems*, edited by Susanne R. Borgwaldt and Terry Joyce, 13–39. Benjamins Current Topics 51. Amsterdam: John Benjamins.

Wellisch, Hans H. 1978. *The Conversion of Scripts—Its Nature, History, and Utilization*. New York: Wiley.

Wikipedia. 2025a. ‘International Alphabet of Sanskrit Transliteration’. Wikipedia. [https://en.wikipedia.org/w/index.php?title=International\_Alphabet\_of\_Sanskrit\_Transliteration&oldid=1270629261].

———. 2025b. ‘ISO 15919’. Wikipedia. [https://en.wikipedia.org/w/index.php?title=ISO\_15919&oldid=1292297131].

1. Version 2 bears the internal version number 1.1, but since the automatic versioning in the HAL-SHS repository assigned it number 2, we have chosen to adopt that numbering to eliminate future inconsistencies. [↑](#footnote-ref-1)
2. References to the EGD in this document pertain to EGD version 2, released simultaneously with TG version 4. The acronym EGD stood for Encoding Guide for Diplomatic Editions in the first version. [↑](#footnote-ref-2)
3. See note 47 about the concept of a Unicode character. [↑](#footnote-ref-3)
4. With Meletis (2020a, 20), we thus exclude *semasiography* — the graphic representation of *meaning* as independent of language — from the scope of writing proper. However, the written texts we are concerned with do include signs which we consider to be semasiographic, q.v. §2.5.5. [↑](#footnote-ref-4)
5. With Weingarten (2013, 17–18), in agreement with Meletis and Dürscheid (2022, 65–66) and essentially reconcilable with Wellisch (1978, 15, 13), and Coulmas (2003, 35). [↑](#footnote-ref-5)
6. With Weingarten (2013, 18) and most recent theorists. [↑](#footnote-ref-6)
7. Such as that for writing twenty-first century Bronx English on a computer in Arial, or that for writing Shakespearean English by hand in insular minuscule. [↑](#footnote-ref-7)
8. The terms ‘Roman’ and ‘Latin’ are near-synonyms in the context of scripts and writing systems. We prefer ‘Latin’ in reference to the specific script and writing system used in ancient Rome for the Latin language (Coulmas 2006, 285–87, s.v. Latin alphabet), and ‘Roman’ in reference to the broad family of scripts and writing systems derived from the former (ibid. 2006, 438–39, s.v. Roman alphabet). [↑](#footnote-ref-8)
9. There exist various and not entirely compatible typologies of writing systems. We broadly follow Meletis (2020a, 142–51), q.v. for a discussion of other typologies. For an in-depth study, see e.g. Daniels (2018). [↑](#footnote-ref-9)
10. In addition to *aksharic* (with varied spellings), technical terms widely used for such writing systems include *abugida* and *alphasyllabary*. The diverse definitions offered for each of these terms are not altogether compatible. See also note 41 below, and see e.g. Gnanadesikan (2017) for a discussion and an attempt at a clearer typology of phonographic writing systems. She endorses the term *āksharik*, which we adopt here apart from the spelling, since it fits the system of related terms better than “akshara script” suggested by Salomon (2003, 78). [↑](#footnote-ref-10)
11. The inherent vowel may be absent (or optional) in some specific Indic writing systems, notably early Tamil Brāhmī (Salomon 1998, 36; 2003, 104). The only truly essential feature of an aksharic writing system is that it has dependent signs for postconsonantal vowels. [↑](#footnote-ref-11)
12. The *graphic syllables* predominantly represented by the signs of syllabic writing systems do not necessarily correspond to speech syllables; for further discussion, see e.g. Coulmas (2003, 62–66); Meletis and Dürscheid (2022, 240–42). [↑](#footnote-ref-12)
13. See Meletis and Dürscheid (2022, 243–49) for a discussion. [↑](#footnote-ref-13)
14. Expressions such as ‘script conversion’ are often used because of their facility, but conversion is not merely the replacement of the signs of one script (sign inventory) with those of another. [↑](#footnote-ref-14)
15. With Wellisch (1978, 18) and Coulmas (2003, 36). [↑](#footnote-ref-15)
16. Paraphrased at the end of §2.2.1; see also §3. [↑](#footnote-ref-16)
17. See Meletis (2019, 27–34) for an overview. [↑](#footnote-ref-17)
18. Notably Daniels (2018, 164–71). [↑](#footnote-ref-18)
19. Following Meletis (e.g. 2020a, 20–28), who in turn builds on the work of Neef (e.g. 2015). [↑](#footnote-ref-19)
20. Note that this dichotomy is not quite the same as the anthropological distinction of “emic” and “etic” as an insider and outsider perspective; the point of connection is that “etic” refers to an objective and impartial description of characteristics, while “emic” is concerned with subjective meaning within a system. [↑](#footnote-ref-20)
21. On the few occasions where we mention phones, we represent them in the IPA phonetic alphabet. The accurate pronunciation of the IPA signs is not relevant to our discussion. [↑](#footnote-ref-21)
22. The term ‘phonology’ is sometimes used in this specific sense, but this leaves us without a technical term encompassing both phonetics and phonemics. [↑](#footnote-ref-22)
23. We use our transliteration system for the phonemes of source languages where the transliteration intuitively suggests a good approximation of the original phoneme (e.g. /a/). For other source language phonemes and for English, we use IPA (e.g. /ɔ/). [↑](#footnote-ref-23)
24. When the illustration of actual graphic appearance is desired, we use an approximation of the shape concerned. When discussing graphs and graphic units of Indic scripts without needing to (or being able to) illustrate their form, we use transliteration (e.g. |rddhe|, |r|). [↑](#footnote-ref-24)
25. We would intuitively prefer the term ‘graphemics’, but recent theorists overwhelmingly use ‘graphematics’, so we follow suit. [↑](#footnote-ref-25)
26. Since graphemes are emic concepts and not etic written signs, the symbols used for their notation are arbitrary (cf. Meletis 2020a, 252–53, n. 9). In this document, we always represent the graphemes of Indic writing systems in transliterated form. [↑](#footnote-ref-26)
27. Homography can also occur at intermediate or mixed levels, for instance between handwritten Roman |*k*| and |*lc*| or Devanagari |ख| <kha> and |रव| <rava>. [↑](#footnote-ref-27)
28. Including complex characters, introduced in §2.4.1. [↑](#footnote-ref-28)
29. Generally with Coulmas (2006, 379–80, s.v. orthography), Neef (2015, 718) and Meletis (2020a, 28). Sometimes, especially in less recent literature, ‘orthography’ is largely synonymous with ‘writing system’ as defined here. [↑](#footnote-ref-29)
30. Most importantly, we — with Fedorova (2013, 50) and Weingarten (2013), contra Meletis (e.g. 2019, 35–36) — do not require the grapheme to be semantically distinctive. Meletis’s stipulation of semantic distinctiveness seems to be a nod to the so-called analogical or autonomous view according to which writing is a modality of language that is interpretable in itself, without recourse to spoken language (compare note 39 below). We find the requirement problematic to begin with, and since it is entirely irreconcilable with the synchronic and diachronic scope of our subject matter, we ignore it hereafter without further discussion. [↑](#footnote-ref-30)
31. With Fedorova (2013, 50) and Meletis (e.g. 2019, 35), contra Weingarten (2013). [↑](#footnote-ref-31)
32. For the concept of dyadic signs in semiotics, see e.g. Nöth (1990, 59–60). [↑](#footnote-ref-32)
33. With, among others, Fedorova (2013, 50), Weingarten (2013, 19) and Meletis (e.g. 2019, 35). [↑](#footnote-ref-33)
34. With Meletis (e.g. 2019, 36; Meletis and Dürscheid 2022, 127). [↑](#footnote-ref-34)
35. See Meletis (2020a, 147–51 and Table 5) for a discussion of such information represented in various writing systems. [↑](#footnote-ref-35)
36. With Meletis (2020a, 148; Meletis and Dürscheid 2022, 132), contra Weingarten (2013, 20), whose concept of the grapheme includes all of these graphematic units. [↑](#footnote-ref-36)
37. Graphic signs which operate at a different level of representational mapping than the dominant one (such as punctuation marks and numeral signs) — are a special case to which we return in §2.5.5. Groups of graphemes which together conventionally represent a phoneme (polygraphs) will be discussed in §2.4.2, and graph elements with a graphematic function (diacritical marks) will be discussed in §2.5.1. [↑](#footnote-ref-37)
38. With Meletis (e.g. 2020a, 65; Meletis and Dürscheid 2022, 121). [↑](#footnote-ref-38)
39. Strangely, Meletis balks from this conclusion while we see it as a logical extension of his thoughts about the separateness of grapheme-phoneme correspondence rules. He considers it crucial for graphemes to be direct (if often imprecise) representations of actual phonemes. This seems to be a nod to the referential or heteronomous view of writing, which considers written language to be entirely secondary to and dependent on spoken language (compare note 30 above). His insistence entails complicated (and, frankly, absurd) consequences which we prefer to avoid, such as that in the German writing system, <c> is not a grapheme while <ch> is one, though <sch> is not (Meletis 2019, 36–38). [↑](#footnote-ref-39)
40. With Meletis (e.g. 2019, 41) and Weingarten (2013, 18), contra Fedorova (2013, 50). [↑](#footnote-ref-40)
41. Incidentally, denying grapheme status to these components would obscure the pivotal difference between aksharic writing systems and syllabic ones, as reflected in the term *alphasyllabary*, which we consider inappropriate. [↑](#footnote-ref-41)
42. With Weingarten (2013, 20), contra Meletis (e.g. 2019, 35). [↑](#footnote-ref-42)
43. Largely in agreement with Meletis (2019, 41, 45–46 n. 32), who recognises such “graphetically non-segmentable clusters” as a challenge, but, in spite of his blanket requirement of visual isolability (ibid*.*, 35), his last word on them is that they correspond to sequences of individual graphemes in the same way as the more transparent complex akṣaras. [↑](#footnote-ref-43)
44. We deem this to be the key difference between aksharic systems and abjads. A consonantal grapheme in an abjad can normally signify either the consonant alone or the consonant and an unspecified vowel, so whether a vowel is present can only be determined by referring to other modules of the language system. In a typical aksharic system, the writing alone is sufficient for determining whether the default vowel is present or not. (Note that the grapheme corresponding to the inherent vowel is of course present even when — in particular contexts in particular languages — the inherent vowel is not pronounced, or pronounced differently than in other contexts. That is a matter of grapheme to phoneme mapping, not essentially different from cases such as the grapheme <e> of English, which may be silent or may be pronounced in various ways.) [↑](#footnote-ref-44)
45. Meletis (e.g. 2020a, 100) explicitly denies grapheme status to the inherent vowel of aksharic systems, even though it seems to us to be a logical extension of his approach to graphetically non-segmentable grapheme clusters (q.v. note 43 above). Moreover, Meletis and Dürscheid’s (2022, 235–36) discussion of the Cree writing system explicitly awards grapheme status to all vowels of this system, which are represented through the orientation of consonant signs, without any graphic addition whatsoever. [↑](#footnote-ref-45)
46. Ollett and Taylor (forthcoming) do define a character in the same way as we do: as “an element of the writing system that can be used independently according to the logic of that writing system”. [↑](#footnote-ref-46)
47. In Unicode, a 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 2020, 2, §3.5). That is to say, a Unicode character can be either a “graphic character” (which is much like the grapheme as we define it; ibid. 5, §3.28), a “control character”, or a “format character”; both of the latter influence the processing and rendition of adjacent characters. Accordingly, many specimens of what we call a character are composed of several Unicode characters. The Devanagari character (actually, glyph, q.v. below) |क्त्र| <ktra> is composed of six: one each for the graphemes <k>, <t>, <r> and <a>, plus two instances of the Unicode virāma control character (one each after the first two consonants). [↑](#footnote-ref-47)
48. Neither Coulmas (2003), nor Meletis and Dürscheid (2022) define ‘character’ or employ it in a technical sense. Wellisch (1978, 16) defines it in a way vaguely resembling our definition of the grapheme. Neef’s (2015, 711) treatment of the character seems to encompass both our grapheme (a term Neef avoids) and our character. Iyengar (2024), while rejecting the notion of the grapheme, argues at length that Indic akṣaras are segments determined on a graphetic, and not phonological, basis. This is wholly compatible with our approach, in which ‘character’ is a graphetically determined segment in any writing system, while akṣara is the specific term for a character of an aksharic system. [↑](#footnote-ref-48)
49. The typographic ligatures of Roman writing systems (e.g. |ﬁ| and |ﬃ|) are in our opinion better understood as character sequences which manifest in a particular form (see also note 81), but the boundary between character sequences and complex characters is not always clear (§2.5.6). [↑](#footnote-ref-49)
50. In true syllabic systems, graphemes and characters again coincide. Our notion of the character seems difficult to reconcile with cursive abjadic systems such as Arabic, but it could be applied productively to non-cursive abjads such as Hebrew, to typologically problematic writing systems such as Korean Hangul and Mayan hieroglyphics, and probably also to morphographic systems. [↑](#footnote-ref-50)
51. The Unicode definition of a ‘graphic symbol’ as the “visual representation of a graphic character or of a composite sequence” (ISO/IEC 10646 2020, 5, §3.29) appears to be identical to our ‘graphic sign’. [↑](#footnote-ref-51)
52. Iyengar (2024, 420) defines a graph as “the written counterpart to a phonological segment,” which is in effect identical to our more elaborate definition, the difference being only that he rejects the concept of the grapheme (which we find nonetheless compatible with his graphetically focused discussion) and does not use the term ‘glyph’ (referring only to the akṣara instead, since he is concerned only with aksharic systems). According to Meletis (2020a, 46, n. 57), ‘graph’ and ‘glyph’ can be considered synonyms. Neef (2015, 711) uses ‘glyph’ in a sense that seems to cover both our graph and our glyph, and does not use the former term, while Coulmas (2006, 173–74, s.v.) defines a graph much as we define a glyph, and all he has to say about ‘glyph’ (2006, 168, s.v.) is that the term has been “used in the description of writing systems whose units were not well understood”, such as Mayan, where it is a “collective designation that can refer to a logogram, a phonetic sign or a compound sign”. [↑](#footnote-ref-52)
53. Coulmas (2006, 129, s.v. digraph 1) asserts that polygraphs are graphemes in their own right. Meletis and Dürscheid (2022, 128–29) classify certain polygraphs as “complex graphemes” (cf. note 39), but deny grapheme status to most. By our definitions, a polygraph always involves more than one individual grapheme, which may or may not manifest as separate characters. [↑](#footnote-ref-53)
54. Examples can, however, be found even in the Indic system, such as the use of the combination <ys> in some North Indian scripts of the first millennium CE to represent the phoneme /z/, which does not occur in Sanskrit; the use of |ळ्ह| for [ḷh], a Vedic allophone of /ḍh/; or, arguably, the simultaneous use of the dependent vowels <ui> in Khmer, Burmese and Mon to represent a vowel phoneme alien to Sanskrit. [↑](#footnote-ref-54)
55. If the anusvāra and visarga are perceived as dependent (cf. §2.5.6.1), then complex glyphs also include combinations such as |तं| and |अः|. [↑](#footnote-ref-55)
56. Keep in mind that in the context of Roman typography, ‘ligature’ has a different meaning (q.v. note 49). [↑](#footnote-ref-56)
57. The concept of ‘glyph component’ is thus not applicable to glyphs such as |क्ष|, where no discernible components correspond to the graphemes constituting the character. [↑](#footnote-ref-57)
58. Meletis (2020a, 100) simply describes the corresponding graphemes as graphetically subsegmental and dependent or bound, while Weingarten (2013, 18) calls them affigated graphemes. [↑](#footnote-ref-58)
59. See also §2.5.6.1 about the anusvāra and visarga. [↑](#footnote-ref-59)
60. See also §2.5.6.2 for further intricacies. [↑](#footnote-ref-60)
61. For instance, the diaeresis (Umlaut) distinguishes |ö| from |o| in the Roman writing system for German, the macron distinguishes |ū| from |u| in the DHARMA transliteration system, and the nuqta (dot) distinguishes |ड़| from |ड| in Devanagari for Hindi. For our purposes, these elements are in the same class as the horizontal stroke that distinguishes Brāhmī |𑀓| <ka> from |𑀭| <ra>. We do, however, treat the length markers used in some Southeast Asian writing systems (see note 61 above) differently, as per (§###). [↑](#footnote-ref-61)
62. Thus, Devanagari |ड| is transliterated ḍa, while |ड़| is transliterated ṛa. In the source graphemes, a diacritical mark in |ड़| indicates that the grapheme stands for a flap allophone of the phoneme represented by the base graph |ड|; incidentally, an identical-looking diacritical mark in the transliteration of both indicates the retroflex quality of both phonemes, but the distinction of the graphemes takes place in the base graph. [↑](#footnote-ref-62)
63. Although the virāma is almost always conjoined to basic consonant glyphs, there is nothing inherent in the Indic writing system that would forbid adding it to a conjunct glyph, and practical examples of the latter do exist (e.g. Devanagari |र्द्| <rd·>, Tamil |க்ஷ்| <kṣ·>). [↑](#footnote-ref-63)
64. We are not aware of any linguistic publication that discusses the status of the virāma vis-à-vis graphemes. Theoretical frameworks which equate akṣaras to graphemes can unproblematically handle the virāma as a diacritical mark which changes one grapheme to another, and this implicit understanding may be in the background of the fact that the virāma is often referred to as a diacritic (e.g. Gnanadesikan 2017, 18). However, such approaches implicitly treat the Indic script as syllabic rather than aksharic (cf. note 41), which we find unacceptable. Weingarten (2013, 22) definitely implies that he considers the virāma to be a grapheme, but since his broad grapheme concept includes all diacritical marks (cf. note 36), this does not support our stance. Iyengar (2024, 427–28) treats the virāma as a graph (defined much as we define it, cf. note 52) . Meletis does not address the status of the virāma as far as we are aware, but Meletis and Dürscheid (2022, 233) state that it “appears as the diacritic <◌्>” which, unless the angle brackets (indicating graphemes) are a mistake for vertical bars (denoting graphs), implies that they are willing to recognise it as at least some sort of grapheme. [↑](#footnote-ref-64)
65. Anyone who is utterly disinclined to accept that the virāma’s contribution of a zero-vowel qualifies it for grapheme status may instead choose to regard it as a non-phonographic grapheme (§2.5.5) in an otherwise phonographic writing system, but this is not a position we agree with. [↑](#footnote-ref-65)
66. With Iyengar (2024, 430–31). [↑](#footnote-ref-66)
67. However, in non-diplomatic contexts (i.e. loose transliteration, §3.4.2), primarily when transliterating a modern language with a known orthography and pronunciation, we have no objection to using diacritical marks indicating vowel nasality, e.g. to transliterating Hindi हाँ as hā̃. [↑](#footnote-ref-67)
68. The addak |ੱ| usually indicates the gemination of the following consonant, but may also function as a stress marker. [↑](#footnote-ref-68)
69. Largely in agreement with e.g. Coulmas (2006, 86, 292, 421; s.vv. cipher 1, letter, punctuation 1) and Neef (2015, 711). [↑](#footnote-ref-69)
70. Examples include spectra such as |oe| - |œ| - |oͤ| - |ö|, from sequence to complex character to simplex character with diacritical mark; and [sz] - |ſʒ| - |ß|, from sequence to simplex character. A similar series culminating in a non-phonographic sign is |et| - |🙰| - |&|. [↑](#footnote-ref-70)
71. The same might be said of the final consonants of Pyu, which are dwarfed by and graphetically subordinate to the regular consonant signs, and may therefore be perceived as components of the preceding akṣara in a system tweaked for the representation of closed syllables. [↑](#footnote-ref-71)
72. We refer not to semantic meaning but to the fact that each graph is the signifier of linguistic information. [↑](#footnote-ref-72)
73. For contrast, compare the digraphs mentioned in note 54 above. [↑](#footnote-ref-73)
74. We thus emphatically disagree with Meletis (e.g. 2020a, 101), in whose view Tamil |ொ| <o> corresponds to two separate graphemes. His reasoning is that the graphs |ெ| <e> and |ா| <ā> can occur separately on their own, hence they correspond to separate graphemes, hence their combination is not minimal, and the association of |ொ| with the phoneme /o/ is a matter of higher-level grapheme-phoneme correspondence. We assert that it is simply the strokes comprising Tamil |ொ| that happen to look identical (be homographic) to those comprising the graphs |ெ| and |ா| <ā>. This is just as irrelevant to the grapheme status of Tamil <o> as the fact that Latin |d| or |Y| look like combinations of |c| and |l| or |V| and |I| is irrelevant to the grapheme status of <d> and <y>. (An added twist is that |Y| was in fact deliberately created from the primary graphs for |U| and |I|, originally to represent a vowel phoneme “in between” these two). [↑](#footnote-ref-74)
75. Here, actual semantic meaning is present. [↑](#footnote-ref-75)
76. This appears to be a straightforward generalisation of the phenomenon that in most Indic writing systems the graph for independent <Ā> is distinguished from that for independent <A> by the addition of a stroke similar or identical to that which normally represents dependent <ā>, as in Devanagari |आ| comprised of the elements |अ| and |ा|. [↑](#footnote-ref-76)
77. We suspect two evolutionary processes here, which may have worked synergistically. On the graphetic side, certain independent vowel graphs in Indic writing systems serve as bases from which graphs for related vowels are derived by the addition of a stroke similar or identical to the corresponding dependent vowel graph. Thus, in many specific systems, the graph for <Ā> is derived from that for <A>, <Ai> from <E> and <Au> from <O>; in some, such as modern Devanagari, <O> is also derived from <A>. In yet others, this generalisation has been carried to the extreme, whereby all independent vowel graphs are derived in this manner from the basic graph for the neutral vowel. On the linguistic side, some languages which adopted an Indic writing system had a minimal consonant phoneme for which the original writing system had no provision, and therefore adapted the writing system to accommodate that phoneme. [↑](#footnote-ref-77)
78. Our distinction between graphematic and graphetic allography is inspired by Meletis (e.g. 2020b), but he defines these concepts in a very different way. For him (e.g. Meletis and Dürscheid 2022, 64), the crux is the physical template (“basic shape”), identified strictly on the basis of phenomenal criteria such as the number and topological arrangement of graphic elements. Iyengar (2024, 427–28) uses the label ‘homophonous heterography’ for some cases of what we call graphematic allography, but is concerned only with the phonological signification of alternate graphic representations, and ignores their potential graphematic role. [↑](#footnote-ref-78)
79. Graphetic allography manifesting on the level of glyphs involves alternative forms of certain complex characters, such as |क्ष| and || <kṣa> Devanagari. [↑](#footnote-ref-79)
80. Meletis (2020b, 257–59) would call |अ| and |अ| graphematic allographs, and recognise only the allography of |अ| and |अ| as graphetic (Meletis 2020b, 255 Fig. 3). A distinction between these two kinds of alternation (involving different or identical basic shapes) may be relevant to grapholinguistics, but is irrelevant to transliteration, so we do not discuss it further. Conversely, the dichotomy we consider crucial is addressed by Meletis only in terms of the rules and constraints applicable to the alternation of forms. Our concept of graphetic allography thus includes Meletis’s graphetic allography and his free graphematic allography. [↑](#footnote-ref-80)
81. In Meletis’s terms (2020b, 259–60), our graphotactic allographs are positional graphematic allographs, but that category also includes many cases of what we consider true graphematic allography. We note here that the notion of graphotactic allography may be fruitful for grapholinguistics at large in addressing phenomena like context-dependent letter shapes in cursive alphabetic writing, the typographic ligatures of alphabetic systems, and even graphetically simplex complex glyphs like Devanagari |क्ष|. [↑](#footnote-ref-81)
82. Meletis seems to consider our graphematic allography no different from his positional graphematic allography (Meletis 2020b, 257–60), except possibly for the matter of upper and lower case in Roman, about which he is undecided (2020b, 260–61). We are not familiar enough with the Arabic and Greek writing systems to venture an opinion on whether their alternation of positionally appropriate forms is purely graphotactic or if it can have a graphematic role. [↑](#footnote-ref-82)
83. Since we have already granted grapheme status to the virāma (§2.5.2), in the Devanagari example allography is only involved in the case of the grapheme <e>. [↑](#footnote-ref-83)
84. Meletis (2020a, 116) in fact makes a distinction between suprasegmental graphetic and graphematic variation, where the latter involves different basic shapes while the former does not. He offers bold and italics as an example of the former and all-caps as an example of the latter. We find this problematic, since the italic alternatives of many Roman letters (e.g. |*g*| and |*a*|) are, in Meletis’s own terms, different basic shapes than their regular counterparts (e.g. |g| and |a|). He also seems to contradict himself by equating suprasegmental graphetic variation with a “switch to a different inventory” (Meletis 2020b, 256), since a different inventory normally comprises different basic shapes, so the switch is by Meletis’s definition not graphetic but graphematic. [↑](#footnote-ref-84)
85. The pamphlet describing the ISO in full (International Organization for Standardization 2001) is accessible to project members in our Sharedocs repository. A summary is freely available on Wikipedia (2025b). IAST is also described on Wikipedia (2025a). The schemes used in many traditional Indic epigraphic publication and those still current in India have much in common with both of these, but are less standardised, and generally employ more polygraphs in order to approximate English phonological transcription better. The transliteration in the Madras Tamil Lexicon is identical to ISO-15919 on all fundamental points. [↑](#footnote-ref-85)
86. There is no hard reason why we should not transliterate धर्म as x%it% (so long as the matching of source signs to target signs is consistent throughout our transliteration scheme), yet most of us prefer transliterations such as dharma. [↑](#footnote-ref-86)
87. As noted in §2.3, Wellisch does not define what a grapheme is. His statements about transliteration are, however, fully compatible with our grapheme definition. [↑](#footnote-ref-87)
88. TEI markup 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. [↑](#footnote-ref-88)
89. Since the disambiguation colon hardly ever occurs in our texts, we have chosen to employ the colon in a different role too, namely for the use of the *ā* marker in Indonesian scripts as an indicator of vowel length or consonant doubling (§4.5.1.1) [↑](#footnote-ref-89)
90. See §### about the transliteration of word breaker signs. [↑](#footnote-ref-90)
91. In the Javanese/Balinese writing system, this phoneme is always written by adding a length marker to the basic grapheme <ə>, so for the sake of consistency we transliterate it as per §###. See also §### about the transliteration of vowel supports. [↑](#footnote-ref-91)
92. See in particular the remark about the Roman graph |Y| in note 74 to that section. [↑](#footnote-ref-92)
93. An explicitly transliterated virāma can be tagged in XML, for instance as unclear, restored or supplied. [↑](#footnote-ref-93)
94. In future projects, similar exceptions may be made for other languages and writing systems as deemed necessary, such as the writing system of the Bhaṭṭiprōḷu inscriptions (Salomon 1998, 35). [↑](#footnote-ref-94)
95. According to Ida Bagus Komang Sudarma (personal communication, 16 Aug. 2019), in Sasak writing ᬅ can be combined with a pasangan consonant, e.g. ᬅ᭄ᬳᬶ qhi and ᬅ᭄ᬳᬸ qhu, but cannot itself become a pasangan, while in Balinese writing neither possibility exists. [↑](#footnote-ref-95)
96. The example is taken from Modern Khmer. The digraph ae is not included in the DHARMA system because it does not exist in Old Khmer, but can function analogously to ui (§4.2.2.1). [↑](#footnote-ref-96)
97. The graph |o| is of course composed of elements which, in themselves, correspond to the graphs |e| and |ā|; cf. §2.5.6.2. [↑](#footnote-ref-97)
98. Unlike the case of |o|, the |ā| marker is used here to indicate length, treated as per §4.5.1.1. [↑](#footnote-ref-98)
99. ISO-15919 already caters for a number of diacritically modified Indic graphemes, for instance Bengali |য়| → ẏ; Devanagari |ज़| → z. Wherever available, Romanisation compliant with ISO-15919 should be preferred. [↑](#footnote-ref-99)
100. In the loose transliteration of this example, ṁ and ṁṅ are both represented by ṅ (since both express the 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 peculiarity. [↑](#footnote-ref-100)
101. In the loose transliteration of this example, anusvāra has been normalised to ṅ. [↑](#footnote-ref-101)
102. Throughout this guide, the term ‘Arabic numeral’ refers to the modern international numeral signs. [↑](#footnote-ref-102)
103. Place-value notation, also called positional notation, is where a digit’s position in the sequence of numbers functions as a multiplier to the digit’s default value. This is distinguished from sign-value notation, where each numeral sign has a fixed value regardless of its position. [↑](#footnote-ref-103)
104. This does not apply to Tamil number notation, where multiplier glyphs can be placed before glyphs for powers of 10, as in Figure 5.1.B/5, but these are visually separate. [↑](#footnote-ref-104)
105. If you have doubts concerning the function of a symbol, encode it as a generic symbol rather than a functional one. [↑](#footnote-ref-105)