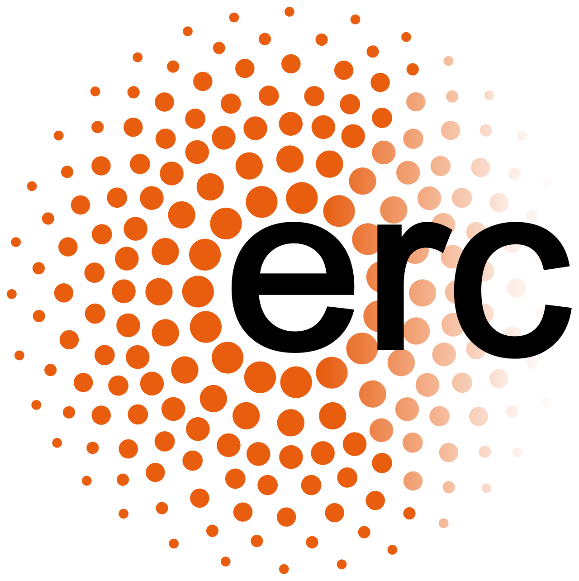
Transliteration Guide

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 0](#_Toc201134747)

[1.1. Version History 0](#_Toc201134748)

[1.2. Coverage 0](#_Toc201134749)

[1.3. Abbreviations 0](#_Toc201134750)

[1.4. Brackets for linguistic notation 0](#_Toc201134751)

[1.5. Terms and definitions 1](#_Toc201134752)

[1.6. Working with Unicode 2](#_Toc201134753)

[1.6.1. Fonts and supported characters 2](#_Toc201134754)

[1.6.2. Entering Unicode characters 3](#_Toc201134755)

[1.6.3. Precomposed characters 3](#_Toc201134756)

[2. Theoretical framework 4](#_Toc201134757)

[2.1. Scripts and writing systems 4](#_Toc201134758)

[2.1.1. Writing system typology 4](#_Toc201134759)

[2.2. Conversion between writing systems 5](#_Toc201134760)

[2.2.1. Transliteration and transcription 5](#_Toc201134761)

[2.3. The elusive grapheme 6](#_Toc201134762)

[2.3.1. A sketch of grapholinguistic concepts 6](#_Toc201134763)

[2.3.2. Refining the concept of the grapheme 7](#_Toc201134764)

[2.3.2.1. The linguistic aspect of the grapheme 8](#_Toc201134765)

[2.3.2.2. The graphic aspect of the grapheme 9](#_Toc201134766)

[2.4. Graphic structures and their elements 10](#_Toc201134767)

[2.4.1. Characters and glyphs 10](#_Toc201134768)

[2.4.2. Polygraphy 11](#_Toc201134769)

[2.4.3. Glyph complexity 11](#_Toc201134770)

[2.4.3.1. Glyph components 12](#_Toc201134771)

[2.4.3.2. Graphic elements 12](#_Toc201134772)

[2.5. Problematic cases: what is a grapheme, and what is not? 13](#_Toc201134773)

[2.5.1. *Diacritical marks* 13](#_Toc201134774)

[2.5.2. *Virāma* 13](#_Toc201134775)

[2.5.3. *Anusvāra* relatives 14](#_Toc201134776)

[2.5.4. Other signs of vague status 14](#_Toc201134777)

[2.5.5. Non-phonographic signs 15](#_Toc201134778)

[2.5.6. Fuzzy segmentation 15](#_Toc201134779)

[2.6. Revisiting allography 16](#_Toc201134780)

[3. General principles of the DHARMA transliteration scheme 18](#_Toc201134781)

[3.1. DHARMA transliteration, ISO-15919 and IAST 18](#_Toc201134782)

[3.2. Case sensitivity 18](#_Toc201134783)

[3.2.1. A note on the use of uppercase for independent vowels and consonants 18](#_Toc201134784)

[3.3. Transliteration and graphemes 19](#_Toc201134785)

[3.4. Transliteration and allographs 19](#_Toc201134786)

[3.5. Strict transliteration 19](#_Toc201134787)

[3.6. Loose transliteration 20](#_Toc201134788)

[3.7. Disambiguation 20](#_Toc201134789)

[3.8. Transliteration and TEI encoding 21](#_Toc201134790)

[3.9. Shorthand 21](#_Toc201134791)

[3.9.1. Transliteration shorthand 22](#_Toc201134792)

[3.9.2. Private markup shorthand 22](#_Toc201134793)

[3.9.3. Public/optional markup shorthand 22](#_Toc201134794)

[3.10. Low-level editorial markup in transliteration 23](#_Toc201134795)

[3.11. @DESCRIPTIVE: ENCODING PREFERRED 23](#_Toc201134796)

[3.11.1. Representation of *akṣara* composition @i.e. graphetic particulars 23](#_Toc201134797)

[3.11.2. Split *akṣaras* 23](#_Toc201134798)

[3.12. @INTERPRETIVE ADDITIONS 24](#_Toc201134799)

[3.12.1. Editorial spaces for word segmentation 24](#_Toc201134800)

[3.12.2. Editorial hyphenation 24](#_Toc201134801)

[3.12.3. Editorial *avagraha* 26](#_Toc201134802)

[3.12.4. Representation of elided overshort final *u* in Tamil 26](#_Toc201134803)

[3.13. @INTERPRETIVE ALTERATIONS 27](#_Toc201134804)

[3.13.1. Silent identification of homographs 27](#_Toc201134805)

[3.13.2. Distinction of long and short e and o 27](#_Toc201134806)

[3.13.3. Short vowel written where a corresponding long vowel is expected 27](#_Toc201134807)

[4. Transliterating alphabetic graphemes 28](#_Toc201134808)

[4.1. The basic inventory of Indic (Sanskrit) graphemes 28](#_Toc201134809)

[4.1.1. Digraphs in the transliteration 28](#_Toc201134810)

[4.1.2. Vocalic *r* and *l* 28](#_Toc201134811)

[4.1.3. Transliteration of *e* and *o* 28](#_Toc201134812)

[4.1.3.1. Distinction of long *ē* and *ō* from short *e* and *o* 29](#_Toc201134813)

[4.1.4. *Anusvāra* and its relatives 29](#_Toc201134814)

[4.1.4.1. *Anunāsika* or *candrabindu* 29](#_Toc201134815)

[4.1.4.2. Other *anusvāra* variants 30](#_Toc201134816)

[4.1.5. *Visarga* and its relatives 30](#_Toc201134817)

[4.2. Graphemes extending the basic repertoire 31](#_Toc201134818)

[4.2.1. Graphemes of other Indian writing systems 31](#_Toc201134819)

[4.2.2. Graphemes of Southeast Asian writing systems 31](#_Toc201134820)

[4.3. Independent consonants 32](#_Toc201134821)

[4.3.1. Distinguishing final forms from characters with a vowel killer 32](#_Toc201134822)

[4.3.2. Final consonants as simplex characters 32](#_Toc201134823)

[4.3.3. Independent consonants as complex characters involving a vowel killer 33](#_Toc201134824)

[4.3.4. The “implicit *puḷḷi*” of Tamil inscriptions 33](#_Toc201134825)

[4.4. Systemic variation in writing independent vowels 33](#_Toc201134826)

[4.4.1. Independent vowels as simplex characters 33](#_Toc201134827)

[4.4.2. Independent vowels as complex characters involving a “vowel support” 33](#_Toc201134828)

[4.4.3. Independent vowel signs behaving like consonant signs 35](#_Toc201134829)

[4.5. Other kinds of systemic variation 36](#_Toc201134830)

[4.5.1. Repurposed vowel markers 36](#_Toc201134831)

[4.5.2. Characters with alternative or optional phonemic values 36](#_Toc201134832)

[5. @Markup 38](#_Toc201134833)

[5.1. Graphetic oddities and *akṣara* composition 38](#_Toc201134834)

[5.1.1. Conjunct consonants in writing systems where they are not the norm 38](#_Toc201134835)

[5.1.2. Multiple vowel markers within an *akṣara* 38](#_Toc201134836)

[5.1.3. The behaviour of the superscript *r* 39](#_Toc201134837)

[5.1.4. Independent and dependent *upadhmānīya* and *jihvāmūlīya* 39](#_Toc201134838)

[5.2. When the materiality of the support interferes 40](#_Toc201134839)

[5.2.1. Legibility problems 40](#_Toc201134840)

[5.2.2. Complex characters split by an intervening feature 40](#_Toc201134841)

[6. Transliterating non-alphabetic signs 42](#_Toc201134842)

[6.1. Numeric signs 42](#_Toc201134843)

[6.1.1. Numbers denoted by bars 42](#_Toc201134844)

[6.1.2. Fractions 43](#_Toc201134845)

[6.2. Symbols 43](#_Toc201134846)

[6.2.1. Punctuation signs 43](#_Toc201134847)

[6.2.2. Space filler signs 44](#_Toc201134848)

[6.2.3. Generic symbols 45](#_Toc201134849)

[6.2.4. Space 45](#_Toc201134850)

[7. References 46](#_Toc201134851)

# 

# 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 epigraphic texts, which must follow these instructions rigorously We do however hope (and, to some degree, expect) that project members will use the same transliteration method, as far as applicable, in their print publications and other work, 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 the DHARMA Transliteration Guide (the present document)

EGD 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 **numeric 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 for transliteration characters 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 will require you to use a particular 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. In a pinch, we recommend using letters with an underdot instead of an undercircle, and substituting f and x for ḫ (upadhmānīya) and ẖ (jihvāmūlīya) respectively.

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, then use shorthand (§**Hiba! A hivatkozási forrás nem található.**)

### 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 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”.

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), which have a very similar or fully identical graphic manifestation, but carry different linguistic information, and are thus different graphemes.

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.[[27]](#footnote-27) 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.[[28]](#footnote-28) 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,[[29]](#footnote-29) 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[[30]](#footnote-30) comprised of a linguistic aspect as its signified and a graphic aspect as its signifier.[[31]](#footnote-31) 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.[[32]](#footnote-32) 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).[[33]](#footnote-33) These other graphematic units can be relegated to a separate stage of analysis,[[34]](#footnote-34) and are generally not relevant to our topic.[[35]](#footnote-35)

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.[[36]](#footnote-36) 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.[[37]](#footnote-37)

#### 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).[[38]](#footnote-38) 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.[[39]](#footnote-39) 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,[[40]](#footnote-40) 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[[41]](#footnote-41) 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.[[42]](#footnote-42) 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.[[43]](#footnote-43)

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.[[44]](#footnote-44) Importantly, ‘character’ as defined in information technology, specifically in the Unicode standard,[[45]](#footnote-45) is *not* what we define as a character. The grapholinguistic literature rarely uses this term,[[46]](#footnote-46) 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,[[47]](#footnote-47) but this is not necessarily so in other types of writing systems.[[48]](#footnote-48)

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**.[[49]](#footnote-49) 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’.[[50]](#footnote-50)

### 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.[[51]](#footnote-51) 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.[[52]](#footnote-52) The DHARMA transliteration scheme employs polygraphs for certain source graphemes, the practical consequences of which will be discussed in §3.5.

### 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 graphतic 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>.[[53]](#footnote-53) 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,[[54]](#footnote-54) 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.[[55]](#footnote-55) 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,[[56]](#footnote-56) 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.[[57]](#footnote-57)

#### 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|).[[58]](#footnote-58)

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 in our terms, 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.[[59]](#footnote-59) 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.[[60]](#footnote-60)

### 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[[61]](#footnote-61) 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 §3.4).[[62]](#footnote-62) 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.[[63]](#footnote-63)

### 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.[[64]](#footnote-64)

### 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.[[65]](#footnote-65) 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.[[66]](#footnote-66) 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 **numeric sign** or **cipher** is one that denotes a number, transliterated as per #§. Alphabetic and numeric signs together are referred to as **alphanumeric signs**, and all others are non-alphanumeric. Among the latter, we give differential treatment to the avagraha (#§), and distinguish the category of **punctuation sign** (#§). All other non-alphanumeric signs are referred to as **symbols** or **miscellaneous signs** (#§), with the caveat that the boundary between punctuation signs and miscellaneous signs is permeable and cannot always be established objectively. 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,[[67]](#footnote-67) 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.[[68]](#footnote-68) 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”,[[69]](#footnote-69) 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 merely an element when it combines with other elements to constitute a different graph (such as Devanagari |ओ| <O> or |ो| <o>). 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.[[70]](#footnote-70) Analogous dual patterning is present in other aspects of language as well: some morphemes (i.e. meaningful[[71]](#footnote-71) 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, and even with consonants.[[72]](#footnote-72) 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).[[73]](#footnote-73) 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.[[74]](#footnote-74) **Graphetic allographs** are alternative graphs 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|.[[75]](#footnote-75) **Graphotactic allographs** are alternative graphs of which the rules of the writing system permit only one in a given graphematic context, as in the <r> of |र्क| and |क्र| or the <u> of |कु| and |रु|.[[76]](#footnote-76) 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 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|.[[77]](#footnote-77) 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>).[[78]](#footnote-78) 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.[[79]](#footnote-79) 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

## DHARMA transliteration, ISO-15919 and IAST

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).[[80]](#footnote-80) 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 small details in order to allow for the representation of certain graphetic details and editorial alterations. The key difference between IAST and ISO-15919 (as well as the DMARMA scheme) is 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.

## 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 (§3.7) 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 ° character before transliterated vowels and/or after transliterated consonants to denote their special independent forms. The principal investigators have agreed 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 (e.g. when cutting words to be cited in a critical apparatus)

## Transliteration and graphemes

Transliteration systems in actual practice include a number of transcription-like features, at least for the sake of pronounceability and ease of acquisition.[[81]](#footnote-81) 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.[[82]](#footnote-82)

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.

## Transliteration and allographs

In order for transliteration to be truly reversible, we also need to give distinguished treatment to any graphematic allographs (as defined in §2.6) that may be present in the source text. If this distinction disappears in transliteration, then we lose potentially important linguistic information represented in the source. 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 (§3.2.1). Conversely, attempting to reflect graphotactic and graphetic allography in transliteration would complicate the transliteration scheme by an order of magnitude, without much gain in return. Therefore, all allography that is not graphematic must be ignored in our transliteration. There are, however, solutions for indicating cases of graphotactic and graphetic allography in the encoding of digital editions, for use in individual cases where such alternation is deemed to be important.

## 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 (§3.7) 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
* disambiguation where a language uses one feature of a writing system to represent more than one phonological feature, e.g.
  + Old Sundanese sastra, rahiyaṅ and ku nu reya (even when written as sasṭā, ku nu rye and rahyiṁ as in the examples under §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

## Disambiguation

ISO-15919 (International Organization for Standardization 2001, 17, §8.1.15) employs the colon : as a disambiguation sign in situations where the target graphemes constituting digraphs in the transliteration scheme appear in their own role rather than as a digraph, as well as in sundry situations 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 (§#EQUALSIGN). As for the former, the need to distinguish vowels in hiatus from diphthongs — as in Devanagari प्रउग, चउत्थो and दइआ — is eliminated by our use of uppercase for independent vowels in strict transliteration: instead of ISO-15919 pra:uga, ca:uttho and da:iā, we transliterate praUga, caUttho and daIĀ. The disambiguation colon is therefore only needed in the very 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 (e.g. p:h as opposed to ph).[[83]](#footnote-83) In loose transliteration, the colon may also be employed for the disambiguation of the diphthongs <ai> and <au>, but we recommend instead that you follow the established convention of using a diaeresis (pair of dots) above the second vowel, thus e.g. caüttho and daïā. Moreover, no disambiguation sign is required in cases where an editorial space (§3.9) or hyphen (§3.12.2) intervenes between the target graphemes that might otherwise be read as a digraph.

* @BUT THIS IS really integral to the transliteration scheme itself, while the akṣara composition thing is descriptive, graphetic

Since the disambiguation colon is thus extremely rare 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).

## Transliteration and TEI encoding

When digitally representing the text of inscriptions and manuscripts for preservation and for computer-aided research, we strive to keep recorded content (i.e. what text is written on a certain support) separate, or at least separable, from our annotations *describing 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 in the form of TEI-compliant markup as detailed in the Encoding Guides. The terms ‘markup’ and ‘encoding’ are fully synonymous in the context of digital editions, but ‘markup’ can also mean a non-electronic annotation system, such as the various brackets employed in printed epigraphic editions. 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.

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 text content should be omitted from the transliterated text and recorded only in markup. In practice, there are a number of borderline cases that could arguably belong to either of these domains, 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; @@@

while others (such as the possibility of interpreting an ambiguous glyph as either of two or more characters) are dealt with in markup. We moreover expect our transliteration scheme to be used in situations where TEI markup is not applicable, and therefore suggest some text-based markup alternatives (shorthand, §**Hiba! A hivatkozási forrás nem található.**) for certain phenomena.

## Shorthand

In the context of this Guide, ‘shorthand’ refers to using an accessible substitute for a special character you cannot easily produce on your keyboard, and to using simple text-based markup in place of more complicated TEI encoding. Shorthand speeds up work because it takes less time to produce and because markup shorthand can be scanned more easily than computer markup by the human reader. Indeed, we have chosen globally to represent certain kinds of common and low-level editorial addition and disambiguation on the level of transliteration. These, to which we refer as ‘essential shorthand’, will be discussed in §3.10 below.

Conversely, the shorthand practices introduced in this section are for private and informal use. They may be useful for your work process and for collaboration with other project members, but they should not be present in completed digital editions and @@@(generally) in printed publications. It is therefore **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.

### Transliteration shorthand

If any special character required for transliterating your texts is not easily accessible in your editing environment, feel free to substitute a character (or character sequence) that you can simply enter on your keyboard. The substitute character or sequence should never occur in your documents in any other function. If it does, then you will have to pay attention to its context when replacing it with the correct character. The following is a list of transliteration shorthand solutions that have been useful in some task forces of the project.

* for the right single quote ’ (representing the avagraha, §3.12.3 or elision, 3.12.4) — substitute the plain apostrophe '
  + @possibly essential instead
  + this shorthand is counter-recommended in documents that contain text in a modern language
* for characters involving an undercircle (r̥, r̥̄ and l̥, §0) — substitute underdotted characters (ṛ, ṝ and ḷ)
  + this shorthand must not be used in documents that also employ the underdotted characters in their proper function (ṛ for the retroflex flap and/or ḷ for the retroflex glide)
* for the schwa ə (pepet, §––) — substitute ĕ
* for the middle dot · (virāma, §0) — substitute the asterisk \*
  + this shorthand should only be used in documents that do not employ an asterisk for any other purpose, including the legitimate transliteration of certain variant signs
* **ALSO**, move the bit with “Some publishers will require you to use a particular font” over to this section, and refer from there; that is actually a form of public transliteration shorthand

### Private markup shorthand

editorial stuff that had better be encoded but may need simple representation in loose transliteration and has probably been used by project members with the expectation that it will be auto-converted to markup

* avagraha: still need decision whether public or private, and whether
  + ’ is shorthand for <supplied reason="subaudible">’</supplied> and original avagrahas get no markup; or
  + ’ always means an editorial avagraha, and original avagrahas get <orig>
* §6.2.2 §abc for space fillers of a particular shape
* §6.2.3 $abc for symbols

### Public/optional markup shorthand

keep only a list here, and the detailed descriptions in the referenced sections

* the Tamil sandhi thing
  + Tamilex: <c type="elided">u</c>
    - doesn’t work for us because we work with the witnessed text, in which there is nothing here
* Tamil short and long vowels (Manu on GitHub: option c: ē and ō in these languages are always understood to be editorial when not marked up, and any original instances must be wrapped in <orig>)
* ??Indonesian short vowels where long expected
* §6.1 + for numeric signs
  + multi-digit
  + Cambodian bars
  + fraction signs
* §6.2.1 easy characters for punctuation signs @these could be viewed as transliteration graphemes like the § sign, except that they have a markup alternative (but probably so should the §)
  + |
  + ||
  + /
  + //
  + ,
  + ~
  + @
  + dingbats for symbols
* §6.2.4 \_ for space (but this may be a dedicated character like standalone § and the ceilings)
* . for punctuation
  + (add , for low-level punctuation and something for high-level punctuation?)
* § for space fillers

## Low-level editorial markup in transliteration

, are in fact also a kind of shorthand replacing more complex computer markup, and may in the future be automatically replaced with such

* @@@as a general rule, do not add anything to your transliteration that is not already present in the original text
* in order to handle editorial additions and alterations, you must rely on markup; see EGD §6
* however, this general rule comes with the following exceptions, which serve as a low-level editorial markup to facilitate the analysis and segmentation of a text for human readers, and which may at a later stage be converted to machine-readable XML markup
* @add some intro here or in subsection about
  + how transliteration always involves some subjective interpretation
  + and the treatment of homography, with benefit of the doubt for the scribe or editorial normalisation
    - e.g. short or long i; subscript ṇ looks like n, whatever
* integral parts of the transliteration system:
  + the disambiguation colon, which is not really editorial markup, so best keep it separate as now
  + the ceilings, which are markup but don’t have a feasible encoding alternative
* basic things which we don’t want to do by encoding but could
  + word spacing
  + hyphen

## @DESCRIPTIVE: ENCODING PREFERRED

this is a skeleton for listing these features, but I now think it will be better to treat them in the appropriate sections, and only refer to them from here

### Representation of akṣara composition @i.e. graphetic particulars

* @at least mention here, a general thing on the = sign

### Split akṣaras

* the ceilings belong here

## @INTERPRETIVE ADDITIONS

### Editorial spaces for word segmentation

* **words** should be **separated** from one another with a space wherever Romanised transliteration allows, notwithstanding that the original inscription or a published edition, whether in Indic or Latin script, does not do so
* emphatically, **do add spaces**
  + where the end of one word and the beginning of the next word constitute a single akṣara in the original
    - even if such an akṣara involves a sandhi modification, e.g.
      * Sanskrit tad dhi (for tat + hi – space goes between d and dh)
      * Sanskrit gacchaty eva (space goes after the y)
      * Sanskrit putrām̐l lakṣmīḥ (space goes between the two l-s)
      * Old Javanese tann inaku (space goes between the -nn and the i-)
      * Tamil arit’ eṉṟu (for aritu + eṉṟu; see also §3.12.4 for elision of overshort u in Tamil)
    - including non-standard sandhi and orthographic practice, e.g.
      * nasals used where standard orthography would employ an anusvāra, e.g. Sanskrit uktañ ca or śaraṇaṅ gataḥ
      * Sanskrit dīnārair ddaśabhiḥ
      * Old Javanese darpaṇa ryy avakta
  + before an avagraha, unless it occurs within a compound
    - thus use, e.g. so ’bhūt and not so’bhūt for separate words, but use e.g. saro’nte or saro-’nte in compound, depending on whether or not you use hyphens for compound segmentation
  + in close-knit structures such as atha vā, kiṁ ca and kiṁ tu (even if spelt kiñ ca and kin tu), tad yathā; including grammaticalised structures such as
    - Sanskrit periphrastic perfects, e.g. varayāṁ cakāra (especially since other words may intrude inside such a construction, e.g. saṁraṁjayāṁ ca prakr̥tīr babhūva)
    - Sanskrit past tense formed with imperfect and sma, e.g. samādiśati sma
    - Sanskrit formations with -sāt prefixed to a verb such as brāhmaṇasād gatāḥ
    - Sanskrit prepositions such as ā samudrāt, anu gaṅgām
  + in repetitions of Sanskrit inflected pronouns and nouns (āmreḍita) expressing a generalised or distributive meaning, e.g. yasya yasya, dine dine
* **do not**, however, use spaces (nor hyphens) to separate
  + successive words where the final vowel of the first and the initial vowel of the second are fused in vowel sandhi, e.g.
    - tasyāyam stays as is, though so yam is separated
    - gacchatīva stays as is, though gacchaty eva is separated
  + Tamil enclitic particles (e.g. ē, ō) and forms of the verb āku-tal (e.g. āṉa, āy, āka) when used adverbially
  + Old Javanese enclitic pronominal suffixes (-(ṅ)ku etc.), possessive constructions built with the linker -ni (-nikaṅ, etc.); definite article -*ṅ*; conjunction -*n*
  + for Sanskrit close-knit structures borrowed into other languages, follow the spelling with or without space (generally the latter) of the relevant dictionaries, if there are any
    - e.g. Old Javanese kimuta, Old Cam kintu
* in sub-standard Sanskrit, strings of words without case endings but apparently intended as nominatives should preferably be spaced instead of being treated as compounds (e.g. *dvandva*), unless the latter in fact facilitates interpretation, e.g.
  + lamvoṣṭha dedamita mahādeva guṇṭhaka ity evam-ādibhyo
  + samrāṬ vākātakānāṁ mahārāja śrī-pravarasenasya

### Editorial hyphenation

* editorial hyphens may be optionally added for the following purposes
  + **segmentation of compounds** in Sanskrit and other compound-heavy languages
    - such segmentation need not be exhaustive
      * feel free to hyphenate only long or difficult compounds and leave others intact
    - as a rule, do not use hyphens within proper names (e.g. śrī-viṣṇuvarddhana-mahārājasya, not śrī-viṣṇu-varddhana-mahārājasya)
      * but do feel free to use hyphens when part of a compound name may be perceived as an epithet rather than an essential part of the name (e.g. śrī-jayasiṁha-vallabha-mahārājasya, where the name Jayasiṁha also occurs without the epithet vallabha)
    - in Sanskrit, keep in mind that secondary derivatives of compound nouns are not themselves compounds and should therefore not be hyphenated, e.g.
      * although derived from the compound catur-varṇa, cāturvarṇya is not a compound and should not be divided up into the members \*cātur and \*varṇya
    - Sanskrit compounds where a verb is combined with a nominal stem with the ending -ī (or -ū) should for this purpose be regarded as compounds and may be hyphenated when this is deemed helpful for the reader; specifically, we recommend that you
      * preferably hyphenate such compounds when they are not themselves members of compounds and/or when their nominal member is itself a compound (in which case preferably avoid hyphenation within that member), e.g. brahmadeyī-kr̥tya
      * preferably avoid hyphenating such compounds when they are themselves members of longer compounds and/or when their nominal member is a simple word, e.g. aśvamedhāvabhr̥tha-snāna-pavitrīkr̥ta-śirasāṁ
    - in the case of Old Javanese, consider that reduplicated expressions are always compounds, whereas close-knit structures consisting of two different elements only become compounds if any morphological derivation takes place
  + **sandhi analysis** when hyphens are conventionally used for this purpose in your field, specifically:
    - epenthesis in Tamil may be indicated by joining the added letter to the preceding word with a hyphen (see the examples below)
    - non-standard Sanskrit sandhi involving an epenthetic m, s or r may be indicated in the same way, e.g. mleccha-rājye-m apūjitaḥ; paṁca-s-triṁśottaratame
* as with editorial spacing, feel free to add hyphens between transliterated characters that belong to a single akṣara of the original, but do not use a hyphen at points where the final and initial vowels of two successive compound members are fused in sandhi
* some examples of Tamil hyphenation:
  + - tiru-makaḷ (திருமகள் tiru+makaḷ)
    - koṇṭ-āṭu (கொண்டாடு koṇṭu+āṭu)
    - I-p-peruṅ-kōyil (இப்பெருங்கோயில் i+perum+kōyil)
    - tiru-mēṉi-y āṭa (திருமேனியாட tiru+mēṉi āṭa)
* some examples of Old Javanese hyphenation:
  + vulu-vulu
  + tahi tikus > manahi-tikusa
  + no fully satisfactory hyphenation point can be identified when an h-initial word is the second member of a compound whose first member ends in an unvoiced stop, and an aspirated consonant akṣara is employed to represent the final phoneme of the initial member of the compound plus the phoneme /h/ of the second member, and consonant gemination at morpheme boundaries occurs as well: e.g., bvat haji > buAtthajyanya, to be hyphenated, for lack of a better solution, as buAt-thajyanya
* **do not use hyphens** for any other purpose, e.g. to show that a word has been broken into two parts by the end of an inscribed line
  + this should be noted in markup (see EGD §3.2.4)
  + if you are not adding any markup, please use the character ¬ (U+00AC Not Sign; do not use a hyphen), which will be auto-converted into the proper markup
* if you use hyphens for editorial compound analysis, and
  + **a physical line break** coincides with such a hyphen, then
    - first encode the physical line break as one inside a word (as per EGD §3.2.4 or with the shorthand ¬)
    - then put the editorial hyphen at the beginning of the new line
  + **a verse line break** coincides with such a hyphen, then
    - first encode the verse line break as one inside a word (as per EGD §2.3.6)
    - then put the editorial hyphen at the beginning of the new line

### Editorial avagraha

* @this should probably go to the non-alphabetic section, not the editorial
  + @no: separate editorial avagraha from original avagraha
* use the transliteration ’ (right single quote) to represent any original avagraha in your text
  + alternatively, you may use the shorthand ' (plain apostrophe) if this is easier for you to type
* the inscriptions we work with very rarely use an actual avagraha sign, but avagrahas may be optionally supplied by the editor
  + this is recommended especially in cases where the text would be meaningful (and even contradictory in meaning) both with and without an avagraha
    - e.g. the inscribed sequence sohataḥ may stand for so hataḥ or so ’hataḥ, so if you interpret the text as the latter, then supply an avagraha to make this clear
    - if you feel that ambiguity of this kind 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 avagrahas applies likewise to words in compound (and regardless of whether you use hyphens for compound segmentation or not), e.g. yaśo’mr̥tam or yaśo-’mr̥tam; saro’nte or saro-’nte
* since most of the avagrahas in our texts will be supplied, any and all avagrahas in an electronic edition will be assumed to be supplied by the editor, and the required markup signifying this (for which see EGD §6.3.7) will be added automatically
  + in the exceptional cases where there is an original avagraha in your texts, use the shorthand ’! or '! (right single quote or plain apostrophe followed by an exclamation mark)
    - the automatic conversion routine will not add supplied markup to these, but will remove the exclamation marks after adding XML tags to all supplied avagrahas
* in order for automatic conversion to be possible, an apostrophe representing a supplied avagraha must never be followed by a space, in order to distinguish it from the apostrophe used to represent elision in Tamil (q.v. §3.12.4), which is always followed by a space
  + such a distinction is important because whereas the supplying of the Sanskrit avagraha is an act of normalisation, namely projecting modern orthographic practice (e.g. सो ऽहतः) onto an original text in which this practice was not (or not widely) used, the use of an apostrophe in the transliteration of Tamil is an act of linguistic analysis, considered an integral part of our transliteration scheme
  + a Sanskrit avagraha must normally be preceded by a space (§3.9), but it will be preceded by an alphabetic sign or a hyphen when it appears within a compound; only the following space is relevant to automatic conversion

### Representation of elided overshort final u in Tamil

@need to provide for similar apostrophes in elided final vowels in Kannaḍa. See Ollett’s Kannada guide under Elision (lōpasandhi)

* in the transliteration of Tamil text, use an apostrophe followed by a space to represent the elided overshort u at the end of an independent word, e.g.
  + arit’ eṉṟu (அரிதென்று for aritu + eṉṟu)
* but do not use an apostrophe for the elided overshort u inside a lexicalised compound, e.g.
  + koṇṭ-āṭu (for koṇṭāṭu)
* note that an apostrophe used for this purpose must always be followed by a space (and not be preceded by one), in order to distinguish it from the apostrophe used to represent Sanskrit avagraha (q.v. §3.12.3)
  + such a distinction is important because these apostrophes are used in the transliteration of Tamil as a feature of linguistic analysis (in a way similar to our use of editorial hyphens, §3.12.2): they are understood to be integral parts of our transliteration system which do not correspond to anything graphically present in the original, whereas avagrahas in Sanskrit could conceivably be present in the original, and when they are not, they are supplied by the editor for the sake of normalisation

## @INTERPRETIVE ALTERATIONS

### Silent identification of homographs

dsdfdf

choice of using markup for ambiguity when warranted

### Distinction of long and short e and o

* now written up as §4.1.3.1

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

* where a short vowel is written in place of an otherwise identical long vowel, optionally add a breve to the transliterated short vowel in order to highlight the fact that the short vowel is not an editorial mistake
  + i.e. use ă, ĭ or ŭ when a, i or u is used for expected ā, ī or ū
* this option is especially recommended for Sanskrit loanwords in Indonesian vernacular documents, following Damais (1955, 15), but need not be applied in such documents if notation of vowel length is absent altogether from their orthography
* this shorthand notation will be converted to XML markup involving the tag <orig> as per EGD §6.3.7

# Transliterating alphabetic graphemes

## The basic inventory of Indic (Sanskrit) graphemes

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.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |  |  |  |  |

Figure .. The basic inventory of Indic graphemes

### 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 §3.5, 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 (§3.7).

### 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 (§3.9).

public shorthand for vocalic r and l

* use underdots instead of undercircles, 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.

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

* @pending decision on github shorthand issue
* **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/, it is permitted to transliterate (actually, transcribe) the long phonemes using a macron, i.e. as ē or ō
  + whether you do so or not is dependent on the conventions of your specific field
  + since most of the source texts in our purview do not make a graphic distinction between short and long /e/ and /o/, the transliterations ē and ō will always be **understood as editorial**
* **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 transliteration ē and ō must be used for the long graphemes
  + any instances of graphically distinct long ē or ō in the source will have to be marked up in XML as original (EGD §###) in order to distinguish them from the editorial transcription ē and ō used where a graphic distinction is not present
    - since such cases will be rare in our practice, we choose to mark up the few original instances rather than to mark up each of the many editorial instances

### 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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
| anusvāra | anunāsika/ candrabindu | ulu ricem | anusvāra-candra | Bengali variant anusvāra |
|  | §0 | @ | @ | @ |
|  | ँ |  |  |  |
| saṁ | m̐ | nāśaṁ\* | m̃ | m\* |

Figure .. Anusvāra relatives

#### Anunāsika or candrabindu

If a candrabindu (crescent-and-dot) graph (Figure 4.1.4.A/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 (§––)
  + @can’t we use m̃ for ALL special anusvāras? we’d get rid of the confusion with the Burmese abbreviation marker and have more consistency

### 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 4.1.5.A). 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)
* **jihvāmūlīya** → ẖ
  + (U+1E96 Latin Small Letter H with Line Below)

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 or behave like a consonant graph and form a conjunct with the following consonant. This is irrelevant to transliteration, since all members of the visarga family are recognised as graphemes and are thus transliterated with a separate character. However, the character joiner sign = may, when expedient, be used to indicate which adjacent character is more closely associated with jihvāmūlīya or upādhmānīya; see §5.1.4.

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

Figure .. Visarga relatives

## Graphemes extending the basic repertoire

In addition to the transliteration signs listed below, which are relevant to the DHARMA project, the inventory of transliteration equivalents 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

* **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

* @consider: move the abbreviation marker and underdot to another section?
  + join up with anusvāra and visarga? or what else belongs in such a section?
* **short mid-central vowel, “schwa”**, Javanese/Balinese pepet |ᭂ| → ə
  + U+0259 Latin Small Letter Schwa
  + @private shorthand: ĕ
* **long mid-central vowel**, Javanese/Balinese pepet tedung |ᭃ| → ə:
  + U+0259 followed by a colon[[84]](#footnote-84)
  + in loose transliteration, ə̄ may be used instead
    - not available as a precomposed glyph: enter U+0259 and ◌̄ (U+0304 Combining Macron)
  + not covered by ISO-15919
* **Khmer (and Mon-Burmese) glottal stop** → q
  + the Latin letter q
  + see also §### about the representation of independent vowels involving this character component
  + not covered by ISO-15919
* **barred/dotted variant of b in Mon and Pyu** → ḅ
  + U+1E05 Latin Small Letter B with Dot Below
  + not covered by ISO-15919
* **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
  + not covered by ISO-15919
* ***akṣaras with abbreviation markers in Burmese***
  + use an asterisk to represent the abbreviation marker, e.g. n\* to transliterate ၌
  + note that you should avoid using an asterisk as shorthand for the vowel killer (§###) if the abbreviation marker occurs in your corpus, especially within the same text
  + not covered by ISO-15919

## 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 4.3.2.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 (§3.4), 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.[[85]](#footnote-85) 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 4.3.2.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 Latin 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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
|  |  |  |  |  |
| yaN | dattaM | ṇāM|| | yaT | dr̥K |

Figure .. Final consonant graphs

### 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·
  + @private shorthand: use an asterisk \* instead of the middle dot if you need to transliterate vowel killers frequently but have difficulty entering that sign
    - @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

### The “implicit puḷḷi” of Tamil inscriptions

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. 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,[[86]](#footnote-86) 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, simply transliterate the consonant cluster without any additional characters
  + e.g. |தத| (understood as த்த /tta/) → tta
* when puḷḷi is present in the source, transliterate as per §4.3.3
* when an actual ligature occurs in Tamil script, this must be treated as unusual akṣara composition, for which see §###

## Independent vowels

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 (§4.3), 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 (§3.4), as described in §4.4.1.

In some of the writing systems in our purview, the independent vowel graphs of the Indic writing system, especially the graph normally corresponding to the independent <A>, have come to be used in manners that would not be expected in India and that are not foreseen by ISO-15919. The character used in such cases as a “vowel support” may represent a glottal stop, or be only a zero consonant sign, serving no other function than to permit the notation of a vowel

* + in some writing systems, we see over time that a consonant character that is not needed to express any consonant phoneme gets redeployed as vowel support (e.g., the h in Modern Javanese script where the Sanskrit loanword eka is spelt heka)
  + in others, inherited “independent vowel” characters gain the ability to behave as consonant characters, i.e., to bear vowel markers and to enter into ligatures
    - this systemic change is complete in Khmer, where អ functions fully as a consonant akṣara, but has been carried through to varying degrees in the writing traditions of Java, Bali and Lombok[[87]](#footnote-87)
* the fact that “vowel support” characters can enter into combinations not foreseen by ISO-15919 may cause problems of ambiguity in transliteration in ISO-15919, or it may simply be impossible to transliterate them without adapting the system

### Independent vowels as simplex characters

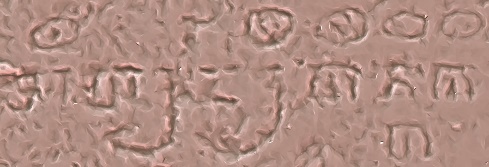
* if the original script employs a distinct character for vowel-only akṣaras (initial vowels and vowels in hiatus), these shall be mandatorily transliterated as follows
  + type a corresponding uppercase Latin consonant, e.g. A
    - thus, इति becomes Iti, whereas कृतमिति becomes kr̥tam iti
    - for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

@my take on this

* **vowel support**: seems to be a combination of two evolutionary processes
  + a (graphetic?) development, where the fact that some vowel graphs are formed by adding markers to other vowel graphs (e.g. आ from अ) is generalised to deriving all other vowel graphs from the same base
    - this alone is evident in Balinese, so here we ignore the graphetic complexity (just as we ignore the complexity of आ) and transliterate as we do for the entire writing system
  + a (graphemic?) development, where there was a need to give written representation to a minimal consonant (the glottal stop or h), and the |A| base was available for this function
    - in a given writing system, this is evidenced by cases where the “support” (actually, minimal consonant) graph enters ligatures with consonants, as it does in Khmer
    - when this evidence is present, we treat the graph as representing the consonantal grapheme q
  + plus an extra complication where other independent vowel signs could then also enter ligatures, how does that work?
    - Arlo distinguished it from vowel support cases, so these are regular old-school independent vowels
    - they seem always to involve either R̥ or a repha plus an independent vowel
      * R̥ was presumably already understood as a syllabic sign (consonant r plus schwa)
      * repha plus vowel constitutes syllabic signs for other rV syllables
      * so this is an adaptation whose probable purpose is to reduce the use of subscript consonants

@original

* in the remainder of this section, we will be concerned only with “independent vowel” characters that are used as “vowel supports”, and whose functional equivalence to consonant characters will be made transparent in our system either by their position (immediately after a consonant character, without intervening inherent vowel a) or by our use of the dedicated character q
* in short,
  + **when the “independent vowel” character that we would otherwise transliterate as A is employed in the original script as a “vowel support”**, i.e.,
    - (1) when it enters into a ligature with another consonant,
    - (2) when a vowel marker is attached to it, or
    - (3) when it enters into a ligature with another consonant and a vowel marker is attached to the ligature,
    - **then mandatorily transliterate the “vowel support” with the letter q followed by the applicable (lowercase) Latin vowel**
    - thus, the text in the image to the right is to be transliterated as qət r̥ṅyəkən tikiṁ
  + in all other situations, even if it is clear that they express a consonant phoneme, transliterate “independent vowel” characters in the manner prescribed in §**Hiba! A hivatkozási forrás nem található.**
    - see also 4.4.3 about other situations where independent vowel signs form a ligature with consonants
* here is an overview of how to transliterate relevant cases from Khmer script
  + 1. Simplex “independent vowel” characters
    - អ A, អា Ā, ឥ I, ឦ Ī, ឧ U, ឪ Ū, ឯ E, ឰ Ai, ឱ O, ឳ Au
  + 2.  អ used as vowel support in syllable-initial position
    - អិ qi, អី qī, អុ qu, អូ qū, អេ qe, អៃ qai, អោ qo, អៅ qau
  + 3.  អ in the function of vowel support (2) and simplex equivalents (1)
    - អេត qeta — ឯត្ត Etta
  + 4. អ entering into a ligature as vowel support in syllable-initial position
    - អ្នក qnaka
  + 5. អ entering into a ligature as vowel support in post-consonantal position
    - 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[[88]](#footnote-88)

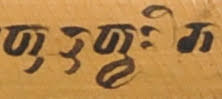
The following table repeats some of the same examples but adds potential equivalents in Balinese glyphs, and shows that the same identically transliterated sequence may not lead to the same transcribed sequence between one writing system and another, because at the level of transcription, we must start to take into account the sounds that are being expressed. The last column shows the romanization that you may use, for instance, in commentary and translation, when you wish to cite words as words, not as transliterated strings of characters.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| combination | Khmer glyph | Balinese glyph | DHARMA  transliteration | IPA  transcription | DHARMA transcription  (lower and upper case) |
| A | អ | ᬅ | A | Khmer /ʔa/  Bal. /a/ | qa ’a  or qA ’A  a A |
| A with e marker | េអ | ᬅᬾ | qe | Khmer /ʔe/  Bal. /e/ | qe ’e or qE  ’E  e E |
| A with u marker | អុ | ᬅᬸ | qu | Khmer/ʔu/  Bal. /u/ | qu ’u or qU ’U  u U |
| A with i marker | អិ | ᬅᬶ | qi | Khmer /ʔi/  Bal. /i/ | qu ’u or qU ’U  u U |
| A with e and ā markers | េអា | ᬅᭀ | qo | Khmer /ʔo/  Bal. /o/ | qo ’o or qO ’o  o O |
| A with ə marker |  | ᬅᭂ | qə |  | ə Ə |
| A with ə and ā markers |  | ᬅᭃ | qə: [[89]](#footnote-89) |  | ə̄ Ə̄ |

* finally, note that when you are representing a Sanskrit name or loanword from an Old Khmer text in transcription, you should transcribe it as you would if the context were fully Sanskrit
  + e.g., Amoghapura can remain Amoghapura, notwithstanding the facts
    - (1) that this name would have been pronounced by Khmer speakers with an initial glottal stop and
    - (2) that you would transcribe the same initial syllable as ’a/qa or  ’A/qA if it stood inside or at the beginning of a Khmer word or name

## Other kinds of systemic variation

### Repurposed vowel markers

* this subsection is about the deliberate and consistent use of vowel markers for a purpose other than their ancestral function; see also §5.1.1 about other circumstances in which an akṣara may have more than one vowel marker
* for the **ā marker** (Javanese tarung, Balinese tedong) used as a marker of vowel length or consonant doubling in Indonesian texts:
  + @
  + generalisation of the original function: it lengthens the inherent a to an ā, so let it lengthen other vowels too
  + special treatment: it is strictly speaking a diacritic, yet with regard to its evolutionary background we transliterate it separately as an honorary grapheme, but distinguish it from ā proper
  + when used in conjunction with another vowel marker to transform the latter into a long vowel, mandatorily type a colon (:) after the short vowel to transliterate the length marker
    - e.g. qə:bni pilaṁ for the image on the right (see also §0 about the vowel support)
  + when representing a doubling of the consonant component of the akṣara to which it is attached, mandatorily transliterate this by typing a colon (:) after the transliterated consonant to which the doubling pertains, even if it is not adjacent to the length marker in the original script, e.g. Old Sundanese
    - Macintosh HD:Users:username:Documents:PHILOLOGY PROJECT:Edisi Bhīmaswarga:BS Cahier d'Archipel:Figure4.jpg (pronounce /gәnәp pipitu/, “fully seven”) is to be transliterated as gnәp:ipitu
    - the image on the right (/turut tvaḥna/, “You should follow his behaviour”) is to be transliterated as turut:vaḥna
  + @refer to §3.7 about the disambiguation colon and mention:
    - this is a slight inconsistency in the system which we find tolerable 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
* for the **vowel markers u/ū and i** used together to represent a particular phoneme in Khmer, Burmese and Mon (as in the Khmer character shown in the image):
  + mandatorily transliterate the vocalisation as ui or ūi
  + however, the deliberate use of u and i markers in conjunction to signify deletion belongs in the domain of markup (see EGD §4.5.1), not that of transliteration

### Characters with alternative or optional phonemic values

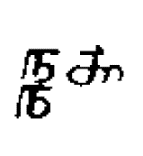
* some writing systems may use certain glyphs to represent more than one phoneme or sequence of phonemes, or may use a non-alphabetic sign in an alphabetic function[[90]](#footnote-90)
  + **in strict transliteration**, always prioritise the primary value of such glyphs
  + **in loose transliteration**, however, it is preferable to transcribe the phonemic value intended in the context
  + **in an EpiDoc edition**, you may add XML markup to normalise the transliterated primary value to a transcription of the intended value (see EGD §6.3 on editorial normalisation)
* some specific examples:
  + 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, but in loose transliteration transcribe it as tra, e.g.
    - Macintosh HD:Users:username:Pictures:Naskah Nusantara:Perpusnas:Bima Swarga:Bhīma Swarga 623 P 16:sastra.png – strict: sasṭā; loose: sastra
  + when a vowel marker added to a ligature with subscript y is intended to be pronounced before the y (and after the primary consonant) in Old Sundanese, write the transliterated vowel after the y in strict transliteration, but transcribe in the intended order when using loose transliteration, e.g.
    - Macintosh HD:Users:username:Pictures:Naskah Nusantara:Perpusnas:L 621 P15:ku nu rye.png – strict: ku nu rye; loose: ku nu reya (“by many [people]”)
    - strict: rahyiṁ; loose: rahiyaṅ (with anusvāra normalised to ṅ)
  + when the numeral 2 is used in Old Sundanese to represent the phonemes /ro/, represent it as 2 in strict transliteration (as per EGD §4.2.7, without adding the XML markup prescribed for numerals proper in EGD §7.1), but use ro in loose transliteration
    - e.g. Macintosh HD:Users:username:Pictures:Naskah Nusantara:Perpusnas:69 L 626:di jeroning wawangunan.png – strict: di jә2niṁ vavaṁṅun:an·; loose: di jәroniṅ vavaṅunan (“in the interior of the building”)[[91]](#footnote-91)

# @Low-level markup in the transliteration of alphabetic graphemes

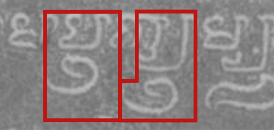
## Graphetic oddities and akṣara composition

* ideally, transliteration would be concerned with graphemes alone, and disregard the choice of allograph used in a particular instance to represent a particular grapheme
  + however, we find that it may be important for our research interests to preserve in the transliterated text some alternative ways of representing the same character or character combination
* for this reason, in strict transliteration we shall employ some mandatory distinctions set out in the following subsections
* certain further distinctions set out in the following subsections may be optionally made using markup or a shorthand notation that will be auto-converted to markup
* other potentially interesting allographs – for instance the use of two alternative glyphs within the same inscription for the same simplex character, or different ways in which a vowel marker is attached to a consonant – will need to be described in your metadata, and will not be directly represented in the transliteration or the markup[[92]](#footnote-92)
* in order to highlight certain formations that deviate from the standard glyph composition for any particular language and writing system, our transliteration scheme permits the use of the dedicated character = (equals sign), in the specific cases set out below
  + this notation is optional, but if you do employ it anywhere within an edition, please attempt to use it consistently throughout that edition wherever applicable
* the = sign will be ignored by search and processing software, but serve as a marker that something strange is going on in the text here, and may be used as a starting point for future analysis or harvesting of such cases
  + @let’s call it the character joiner sign
* should you need to add an editorial space or hyphen between such characters, always put it *after* the = sign
  + e.g. …k= ka… (for a single akṣara at a word boundary) and …k=-ka… (for a single akṣara at the boundary between two elements of a compound word)
* this notation will be auto-converted to markup (EGD §4.1.1)

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

* **where a Tamil text written in Tamil script employs a ligature** such as nna and kka in the image
  + use an = sign between the transliterated consonants to distinguish the ligature from the script’s default method of writing conjunct consonants as two glyphs with an explicit or implicit zero vowel marker (for which see also ––), e.g.
    - n=na as distinct from n·na and nna
    - *k=ka* as distinct from k·ka and kka
  + it is **strongly recommended** that whenever feasible, you should make Tamil ligatures explicit in this way
  + however, never add an = sign where ligatures are a writing system’s default method of representing conjunct consonants (including Tamil written in Grantha)

### Independent vowel signs behaving like consonant signs

* this subsection is about vowel signs forming conjuncts with consonant signs, apart from the special case where the glyph for independent A is repurposed as a vowel support, for which see §0
* **where an independent vowel sign is conjoined to a consonant sign to form a ligature, as the second R̥ in the image on the right**
  + transliterate the vowel in uppercase to indicate that it is an independent vowel grapheme
  + optionally use the = sign between the consonant and the vowel sign to indicate that the two are conjoined in an akṣara
  + thus, the text in the image is transliterated as maR̥k= R̥mpva
* **where an independent vowel sign is combined with a superscript *r* marker**, as in the Javanese example on the right
  + use an = sign between the transliterated characters to make it explicit that the transliteration is not a mistake (for an original text involving a final consonant, a zero vowel marker or a dependent vowel)
  + thus, the text in the image is Umiṅsor= I

### Multiple vowel markers within an akṣara

* multiple vowel markers may be used deliberately in the original
  + to represent a particular phoneme or modification (§4.5.1)
  + to mark segments as deleted (this belongs in the domain of markup, not that of transliteration; see EGD §4.5.1)
  + if you encounter multiple vowel markers that appear to be used deliberately for a purpose other than the above, please contact us to discuss how best to represent them
* other appearances of multiple vowel marks are likely to be cases where the scribe erroneously engraved more than one explicit vowel mark, neither of which appears to be deliberately cancelled
  + if one of these vowels is expected in the context and the other is not, it is acceptable and preferable to encode this as a premodern correction as per EGD §4.5.3
  + in other cases, which we expect to be very rare, transliterate all vowels in an order you deem suitable
    - the fact that the transliterated vowels are lowercase indicates in our system that none of them are independent vowel akṣaras (cf. §3.7)
    - nonetheless, the unusual fact that multiple vowel markers are present in a single akṣara may optionally be made explicit using an = (equals) sign between the transliterated vowels belonging to a single akṣara, e.g. du=ā for the glyph shown on the right (which is probably an engraving mistake for ddhā)
* as a special case, when the vowel markers r̥ and i are used in conjunction to represent the sonant /r̥/ phoneme
  + if the r̥ component is clearly distinguishable from a subscript consonantal r (as in the character mr̥i on the right), transliterate the combination as r̥i (optionally, r̥=i)
  + otherwise, transliterate as ri
  + in your digital edition, either r̥i or ri should be emended to r̥ (by means of normalisation or correction as applicable, see EGD §6.1.1)

### The behaviour of the superscript r

* **where an Indonesian text employs the superscript *r* marker** (repha, layar, surang) **in two modes**,
  + namely
    - the “Indian” mode, i.e. to be read before the consonant it is attached to, as in ᬲᬯᬃ *sarva*; and
    - the “Indonesian” mode, i.e. to be read after the rest of the *akṣara* it is attached to, as in ᬲᬫᬃ *samar*
  + then this fact must be noted in your commentary to the text, including a specification of which mode is the default (dominant) one for that text
  + in addition, you may optionally use an = sign between the transliterated *r* and the other characters transliterating the same akṣara in instances of the non-default mode, i.e.
* the representation of the “Indonesian” (versus “Indian”) positioning of the r marker is handled via markup
  + thus, even in strict transliteration, transcribe the text in the order the script components were meant to be pronounced, i.e.
    - if the “Indonesian” mode is dominant, transliterate ᬲᬯᬃ as sar=va
    - if the “Indian” mode is dominant, transliterate ᬲᬫᬃ as sama=r
    - e.g. Ina=rpaṇakan· for the example on the right
  + for the sake of explicitness you may also choose to use the above notation redundantly for all occurrences of a superscript *r*, not only for ones that deviate from the default

### Independent and dependent upadhmānīya and jihvāmūlīya

Within a specific writing system, upadhmānīya and jihvāmūlīya may be practically independent inline graphs or combining graphs that form a conjunct with the following consonant (Figure 4.1.5.A). Which of these behaviours is present in a text need not and cannot be represented on the level of transliteration, but may be noted in your palaeographic description. 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. In the edition itself, you can optionally use the = sign 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 4.1.5.A/2: ḫ= pu; column 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 4.1.5.A/5: yo=ẖ ka

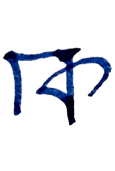
## When the materiality of the support interferes

### Legibility problems

* TODO: perhaps in this section, add instructions on C and V or another solution for partially extant akṣaras

### Complex characters split by an intervening feature

@possibly add <https://github.com/erc-dharma/project-documentation/issues/284> about the consonant body itself broken across lines

* 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 §6.2.3)
    - 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

# Transliterating non-alphabetic signs

## Numeric signs

* numbers written in **decimal place-value notation** in the original shall be transliterated straightforwardly (e.g. 876), since each transliterated character corresponds here to one original character
* however, original ciphers which cannot be represented in transliteration by a single character (i.e. an Arabic digit for integers between 0 and 9, or a common vulgar fraction sign as per §6.1.2 below) require markup to indicate the fact that several transliterated characters together represent a single original character
  + this markup is ideally in the form of XML tags as per EGD §4.2.2
  + but to simplify your work, especially if you are creating an e-text that will not (or not immediately) be marked up in XML, you may take advantage of the following shorthand markup method, which will be automatically converted to the proper XML markup[[93]](#footnote-93)
* for integers 10 and above, written as a single original character,
  + type a + sign after each transliterated number sequence of two or more Arabic numerals that represents a single numeric sign in the original[[94]](#footnote-94)
    - arguably, most Indic numerals in the 100s range could be viewed as combinations of several characters rather than as a single character, but we foresee no useful purpose that such a complex distinction could serve and therefore treat all these Indic numerals as single characters (with distinguishable components)
  + for example:
    - * 10+ means “10” written as a character for 10, e.g. Brahmi 𑁛
      * 10+ 2 means “12” written as a character for 10 followed by one for 2, e.g. Brahmi 𑁛𑁓
      * 80+ 10+ means “90” written as a character for 80 followed by one for 10, e.g. Brahmi 𑁢𑁛
      * 300+ 50+ 2 means “352” written as a character for 300, one for 50, and one for 2
      * 1000+ 8 100+ 3 10+ means “1830” written in Tamil notation as 1000 (plus) 8 (times) 100 (plus) 3 (times) 10
  + note that though this shorthand notation resembles the notation of older printed publications such as Epigraphia Indica, which used a + sign to indicate numerals inscribed in an additive system, in our shorthand + must mark the end of every sequence of two or more Arabic numerals that transliterate a single numeric sign in the original
    - thus, in our shorthand a final + sign is required in cases such as the following
      * 10+ (rather than 10) – “10” written as a character for 10
      * 300+50+ (rather than 300+50) – “350” written a character for 300 and one for 50

### Numbers denoted by bars

* to transliterate numerals represented in Cambodian inscriptions by bars (daṇḍa) instead of numeric signs (as in the image, showing the number 3):
  + type as many I characters as there are bars in the original (NB: uppercase i characters, not vertical bars |)
  + type a + sign after the last I as shorthand markup
    - note that unlike regular numerals, the + sign must be used in this case even after a single I representing the numeral 1
    - this shorthand will be automatically converted to markup indicating that these characters are not alphabetic and constitute a single meaningful character

### Fractions

* for any fractional numbers represented in the original by a single character, use one of the following Unicode characters:
  + ½ (U+00BD Vulgar Fraction One Half)
  + ⅓ (U+2153 Vulgar Fraction One Third)
  + ⅔ (U+2154 Vulgar Fraction Two Thirds)
  + ¼ (U+00BC Vulgar Fraction One Quarter)
  + ¾ (U+00BE Vulgar Fraction Three Quarters)
* should you encounter fraction signs not covered by the above,[[95]](#footnote-95) transliterate them as a common fraction using a slash and add a + sign after the denominator as shorthand markup
  + e.g. 1/8+ to transliterate a numeral sign meaning “one eighth”
  + this shorthand will be automatically converted to markup as per EGD §4.2.2
* the Khmer fraction sign in the shape of a cross (with a single or a double bar, see the images) shall always be transliterated as ½ (U+00BD Vulgar Fraction One Half)

## Symbols

* @avagraha here?
* @abbreviation markers should definitely be moved here
  + or not, after all, since it seems to be a component, not an akṣara-level thing
* in the terms of this Guide, “**symbol**” is an umbrella term for all non-alphanumeric characters
* we consider the diversity of non-alphanumeric symbols used in inscriptions to deserve preservation and investigation, but acknowledge the challenges of representing them in a machine-actionable scheme of manageable complexity, and therefore provide a combination of methods for handling them at different levels:
  + 1. at the level of transliteration, symbols of a definite function (punctuation signs and space fillers) shall be represented by a dedicated character, but symbols whose function is not so straightforward (“miscellaneous symbols”) shall have no representation on this level, only on that of XML markup
  + 2. at the level of XML encoding, all symbols shall be encoded with a relatively simple classification of their shapes, for which see EGD §4.2.3
  + 3. at the level of human-readable metadata, symbols may (and generally should) be described in any detail that you deem appropriate, in the handDesc section of the TEI header (EGD §11.2.1)
* as an intermediate step between levels 1 and 2 above, the following subsections recommend the use of **shorthand markup** for certain symbols
  + we may opt later on to devise automatic conversion of this shorthand to XML markup, but it seems likely at present that we will not do so
  + thus, if you use the shorthand suggested here, please be prepared that you may have to replace it with XML markup yourself, and keep notes with a more accurate description of your punctuation signs than that afforded by the shorthand scheme

### Punctuation signs

* in the terms of this Guide, “**punctuation sign**” is used in a sense restricted to symbols 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 signs **primarily includes** symbols
    - whose shape is derived from the small and originally simple non-figural marks used widely for punctuation in our scripts, e.g. vertical bars, dots, circles and dashes
    - whose primary function is to segment the text into sentences, clauses, list items or metrical units
    - which occur repeatedly in the body of a single text
  + the scope of punctuation signs **does not normally include** symbols
    - whose shape is figural or complexly ornamental, and is not derived from one of the simple basic shapes used for punctuation
    - whose primary function is to mark the beginning or end of an entire inscription or a major section thereof, and/or to express auspiciousness
    - which occur only once per text or once per major section of text
* **original punctuation** must always be transliterated, but you should **never add punctuation signs** not already present in the original
  + editorial punctuation may, however, be supplied using XML markup, see EGD §6.3.6
* as outlined in the introduction to this section (§6.2), punctuation signs 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 sign 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 sign 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 sign 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

### Space filler signs

* 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 miscellaneous symbols even if they are identical in visual appearance to symbols used as space fillers elsewhere in the document or the corpus
* as outlined in the introduction to this section (§6.2), space fillers shall be represented
  + 1. at the level of transliteration, by the dedicated character § (U+00A7 Section Sign), which by our convention shall be understood as an abstract space filler 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
* we recommend the use of the following shorthand markup as an intermediate step between levels 1 and 2 above
  + since the use of space fillers is most common within our corpus in Javanese documents, a § character without an XML tag encoding its shape will be understood by default as being of the Javanese form, and will be automatically tagged in XML as having that form
  + fillers for any other shape shall ideally be encoded in XML, but if necessary, you may use the shorthand §abc
    - where “abc” (any sequence of letters, followed by a space) will be converted into a symbol token in the XML tag to be added to the § character

### Generic symbols

* in the terms of this Guide, a “**generic symbol**” is any symbol that does not meet the criteria for punctuation signs and space fillers as defined above
* note that auspicious (maṅgala) symbols should never be transliterated as the words siddham or om̐
* as outlined in the introduction to this section (§6.2), 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 miscellaneous 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
  + as dingbats, using any Unicode character approximating the original glyph (e.g. ◊卐✤)

### Space

* spaces in your text must be encoded in XML as per EGD §4.3
* to simplify your work, especially when you are creating an e-text that will not (yet) be marked up in XML, you may use the shorthand markup \_ (underscore) where an inscription employs an interword space (large enough to be called a space but smaller than the width of two average characters)
  + it is recommended that you also use a regular space before and after the underscore, but this is not required
* any other spaces — such as space left blank for filling later, or because of a defect or feature of the material — can only be handled in XML

# References

@hard

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

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].

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 44 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 39 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 §3.2.1; see also §4.1.1. [↑](#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. 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-27)
28. 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 37 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-28)
29. With Fedorova (2013, 50) and Meletis (e.g. 2019, 35), contra Weingarten (2013). [↑](#footnote-ref-29)
30. For the concept of dyadic signs in semiotics, see e.g. Nöth (1990, 59–60). [↑](#footnote-ref-30)
31. With, among others, Fedorova (2013, 50), Weingarten (2013, 19) and Meletis (e.g. 2019, 35). [↑](#footnote-ref-31)
32. With Meletis (e.g. 2019, 36; Meletis and Dürscheid 2022, 127). [↑](#footnote-ref-32)
33. See Meletis (2020a, 147–51 and Table 5) for a discussion of such information represented in various writing systems. [↑](#footnote-ref-33)
34. 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-34)
35. Graphic signs which operate at a different level of representational mapping than the dominant one (such as punctuation signs 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-35)
36. With Meletis (e.g. 2020a, 65; Meletis and Dürscheid 2022, 121). [↑](#footnote-ref-36)
37. 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 28 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-37)
38. With Meletis (e.g. 2019, 41) and Weingarten (2013, 18), contra Fedorova (2013, 50). [↑](#footnote-ref-38)
39. 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-39)
40. With Weingarten (2013, 20), contra Meletis (e.g. 2019, 35). [↑](#footnote-ref-40)
41. 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-41)
42. 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-42)
43. 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 41 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-43)
44. 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-44)
45. 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-45)
46. 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-46)
47. 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 72), but the boundary between character sequences and complex characters is not always clear (§2.5.6). [↑](#footnote-ref-47)
48. 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-48)
49. 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-49)
50. 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-50)
51. 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 37), 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-51)
52. 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-52)
53. 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-53)
54. Keep in mind that in the context of Roman typography, ‘ligature’ has a different meaning (q.v. note 47). [↑](#footnote-ref-54)
55. 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-55)
56. 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-56)
57. See also §2.5.6.1 about the anusvāra and visarga. [↑](#footnote-ref-57)
58. See also §2.5.6.2 for further intricacies. [↑](#footnote-ref-58)
59. 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 59 above) differently, as per (§###). [↑](#footnote-ref-59)
60. 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-60)
61. 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-61)
62. 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 39), 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 34), this does not support our stance. Iyengar (2024, 427–28) treats the virāma as a graph (defined much as we define it, cf. note 50) . 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-62)
63. 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-63)
64. However, in non-diplomatic contexts (i.e. loose transliteration, §2.7), 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-64)
65. The addak |ੱ| usually indicates the gemination of the following consonant, but may also function as a stress marker. [↑](#footnote-ref-65)
66. Largely in agreement with e.g. Coulmas (2006, 86, 292, 421; s.vv. cipher 1, letter, punctuation 1) and Neef (2015, 711). [↑](#footnote-ref-66)
67. 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-67)
68. 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-68)
69. We refer not to semantic meaning but to the fact that each graph is the signifier of linguistic information. [↑](#footnote-ref-69)
70. 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>, and this is just as irrelevant to the grapheme status of Tamil <o> as the fact that Latin |d| or |Y| look like combinations of |o| and |l| or |V| and |I|. [↑](#footnote-ref-70)
71. Here, actual semantic meaning is present. [↑](#footnote-ref-71)
72. 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-72)
73. We suspect two evolutionary processes here, which may have worked hand in hand. On the graphetic side, certain independent vowel graphs in many 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, 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 some writing systems, this generalisation was 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 that could not be expressed in writing, and therefore adapted the writing system to accommodate that phoneme. [↑](#footnote-ref-73)
74. 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-74)
75. 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-75)
76. 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-76)
77. 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-77)
78. 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-78)
79. 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-79)
80. 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-80)
81. 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-81)
82. As noted in §3.3, Wellisch does not define what a grapheme is. His statements about transliteration are, however, fully compatible with our grapheme definition. [↑](#footnote-ref-82)
83. [↑](#footnote-ref-83)
84. 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-84)
85. An explicitly transliterated virāma can be tagged in XML, for instance as unclear, restored or supplied. [↑](#footnote-ref-85)
86. In future projects, similar exceptions may be made for other languages and writing systems as deemed necessary. [↑](#footnote-ref-86)
87. 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-87)
88. The example is taken from Modern Khmer. The combination ae is not included in the DHARMA system because it does not exist in Old Khmer, but it does not seem to cause any problems with our system. [↑](#footnote-ref-88)
89. See also 5.3.5 about the colon as a length marker. [↑](#footnote-ref-89)
90. It would be possible to view these as different characters that happen to manifest as identical-looking glyphs: this phenomenon is known as synoglyphy and happens e.g. in the case of I (uppercase i) and l (lowercase L) in Latin script printed in a sans-serif font. In this case we would transliterate them according to their function in context and disregard their appearance. However, we find that this approach would cover up palaeographically interesting and important features of the original. For a modern Western parallel, the text “going 2 bed” is not entirely equivalent to “going to bed”, and “2” and “to” are not synoglyphic. [↑](#footnote-ref-90)
91. Note that as per §4.2.2, in the loose transliteration of this example, ṁ and ṁṅ are both represented by ṅ (being the sign selected to represent phoneme /ŋ/). Simultaneously, n: (theoretically denoting /nn/) is simplified to n, and ṁṅ (theoretically /ṅṅ/) is simplified to ṅ, because consonant gemination is not considered to be a phonemic feature of the language, but rather an orthographic particularity. [↑](#footnote-ref-91)
92. This is a conscious decision of the authors of this Guide, who consider that we need to impose a limit on the granularity of our representation of potentially interesting phenomena. However, it is possible to use sub-akṣara markup (EGD §4.1.2) to encode the relative positions of certain character components, if you consider it necessary to do so. [↑](#footnote-ref-92)
93. This shorthand may be used in place of the tagging of numeric signs as per EGD §4.2.2, but it does not replace the semantic markup for the value of numerals (EGD §7.1), which is mandatory in XML files for all numerals including those inscribed in the place-value system. [↑](#footnote-ref-93)
94. Keep in mind, however, numerals 1 to 10 do not need this + sign even if they are used in an additive system (in combination with signs for tens or hundreds), since these are represented by a single transliteration character. [↑](#footnote-ref-94)
95. Although further “vulgar fraction” signs are available in the Unicode code table, they are absent from many fonts and not easily accessible via the keyboard. We have therefore chosen to use single Unicode characters only for an arbitrarily delimited set of common fractions. [↑](#footnote-ref-95)