Extensions to Brahmic script processing within the Nisaba library: new scripts, languages and utilities

Alexander Gutkin, Cibu Johny, Raiomond Doctor, Lawrence Wolf-Sonkin, Brian Roark

Google Research; United Kingdom, India, United States.

Overview

Motivation: Brahmic Scripts

The Brahmic family of scripts is very important:

- It is used to record some of the most spoken languages in the world.
- Is arguably the most diverse family of writing systems.

Observation

The suite of linguistically well-motivated compact & efficient *low-level script* processing utilities beyond what Unicode offers is somewhat scarce, especially for *smaller* scripts and languages.

Nisaba Finite-state Script Processing Library

Nisaba is an open-source library for processing Brahmic scripts [1].

Design

- Formal models of akṣara "orthographic" syllable [2].
- Using Pynini finite-state grammars [3, 4].
- Compile offline into finite-state transducers (FSTs) using OpenFst [5].

The aim of this work: Extend Nisaba to further scripts, languages and operations.

Background

Original Set: The "Big" Scripts

Bengali, Devanagari, Gujarati, Gurmukhi, Kannada, Malayalam, Oriya (Odia), Sinhala, Tamil, and Telugu [1].

NFC (Devanagari)

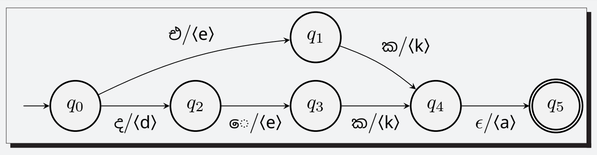
Visual	Legacy sequence	NFC normalized
न	NA NUKTA (U+0928 U+093C)	NNNA (U+0929)
क़	QA (U+0958)	KA NUKTA (U+0915 U+093C)

Visual Normalization (Devanagari)

Standard sequence	Do-Not-Use sequence
अे (U+0904)	अ (U+0905) े (U+0946)
ख (U+0916)	ख (U+0916) ् (U+094D) ा (U+093E)

... *also*, sequences found in the wild.

Reversible Translit (Sinhala)



Sinhala එක ("one") and දෙක ("two").

Extensions

Our implementation extends ISO 15919, e.g.:

- Bengali *letter khanda ta* (ৎ) ⟺ ⟨t$^{\text{ʂ}}$⟩,
- Tamil *visarga* + *letter pa* (ஃப) ⟺ ⟨f⟩.

Well-Formedness Automata

Finite state automata that accept well-formed (according to the akṣara principles) strings in the native script as well as documented exceptions. E.g.,

Malayalam Exceptions

- *letter a* (U+0D05) + *virama* (U+0D4D),
- *vowel sign u* (U+0D41) + *virama* (U+0D4D).

Brahmic Script Extensions

Supported Scripts

Name	Id	Sample	Name	Id	Sample
Bengali	Beng	বাংলা	Prachalit	Newa	𑐣𑐵𑐫𑐮𑐫
Lontara	Bugi	ᨒᨚᨈᨗ	Oriya	Orya	ଓଡ଼ିଆ
Devanagari	Deva	देवनागरी	Sinhala	Sinh	සිංහල
Gujarati	Gujr	ગુજરાતી	Sylheti Nagri	Sylo	ꠍꠤꠟꠐꠤ
Gurmukhi	Guru	ਗੁਰਮੁਖੀ	Takri	Takr	𑚛𑚀𑚭𑚦
Kannada	Knda	ಕನ್ನಡ	Tamil	Taml	தமிழ்
Lepcha	Lepc	ᰛᰩᰵ	Telugu	Telu	తెలుగు
Limbu	Limb	ᤕᤠᤰᤌᤠᤱ	Baybayin	Tglg	ᜊᜌᜊᜌᜒᜈ᜔
Malayalam	Mlym	മലയാളം	Thaana	Thaa	ތާނަ
Meetei Mayek	Mtei	ꯃꯤꯇꯩ ꯃꯅꯥ	Tirhuta	Tirh	𑒞𑒪𑒯𑒳𑒞𑒲

New Scripts

South and Maritime South-East Asia: Baybayin (Tagalog), Lepcha, Limbu, Lontara (Bugis), Maithili (Tirhuta), Meetei Mayek, Prachalit (Newa), Sylheti Nagri, Takri and Thaana.

Minimal Requirements & Observations

- Support NFC and visual norms, reversible translit and well-formedness.
- Some scripts are relatively simple (e.g., Baybayin).
- Others are challenging (e.g., Lepcha).

South Asian Brahmic

- Prachalit (Newa): Primarily used to write Sino-Tibetan Newar (Nepal Bhasa), endangered (Nepal & Sikkim; ∼860K speakers). Similar to Bengali and Devanagari.
- Takri: Primarily used to write Indo-Aryan Dogri (Jammu and Kashmir, etc.; ∼5M speakers). Structurally simple.
- Sylheti Nagri: Used to write Indo-Aryan Sylheti closely related to Bangla (Bangladesh, Assam & Tripura; ∼11M speakers). Salient features:
  - Virama (*hasanta*) rarely used. Pronunciations determined from context.
  - Special sign *dvisvara* attaches to consonants, vowels and vowel signs to form diphthongs.
- Lepcha: Records Tibeto-Burman severely endangered Lepcha (Sikkim; ∼30K speakers). Salient features:
  - Lacks virama. Instead: explicit akṣara-final silent consonant signs.
  - Tibetan subjoined consonant model.
  - Sign *ran* marks vowel length and accent. Combines with vowels, vowel signs, final consonants.
- Limbu: Used to write endangered Sino-Tibetan Limbu (eastern Nepal, Sikkim, Darjeeling; ∼380K speakers). Salient features:
  - Subjoined consonants.
  - Silent syllable-final consonant signs for native words, *but* sign *sa-i* can act as virama or vowel length marker.
- Maithili (Tirhuta): Used to write Indo-Aryan Maithili (Bihar; ∼33M speakers). Salient feature:
  - Vowel signs (for short vowels, only word-internal) with no full-form equivalents.
- Meetei Mayek: Used to write vulnerable Tibeto-Burman Meitei (Manipuri) (Manipur; ∼1.8M L1 speakers). Salient features:
  - Silent syllable-final consonants, *but* these are *full* letters.
  - Has virama (*apun iyek*), but not for final consonants.

Maritime South-East Asian Brahmic

- Baybayin: Used to write Malayo-Polynesian Tagalog (Indonesia).
  - Simple script: No consonant conjuncts are formed.
  - Virama marks silent consonants.
- Lontara: Used to write Malayo-Polynesian Buginese, Makassarese and Mandar (Indonesia). Structurally similar to Baybayin, *but* ...
  - ... no virama to mute final consonants. Use context for pronunciation.

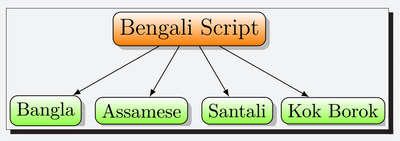
Thaana: Between *abugida*, *abjad* and alphabet

Thaana is used to write Indo-Aryan Dhivehi (Maldivian), closely related to Sinhala (Maldives; ∼340K speakers). Salient features:

- From Perso-Arabic: Rendered right-to-left, *sukun* as virama.
- From Brahmic and Perso-Arabic: Vowels are *only* diacritics, subordinate to consonants.
- Standalone vowels: *alifu* + vowel diacritic.
- Alphabetic: Vowels are *always* recorded, *no* inherent vowel.
- Sukun also marks gemination.

Language-specific Transducers: Deriving from Bengali

Assamese & Bangla



4 out of 41 languages using the script.

- Assamese has two extra consonant letters not found in Bangla, the Assamese *ra* (U+09F0) and *wa* (U+09F1).
- Assamese *wa* is unique, but Assamese *ra* should override Bengali *ra* (U+09B0) in Assamese transducer.
- Given a Bengali script-specific transducer $\mathcal{B}$, construct by FST composition: $\mathcal{B} \circ \mathcal{T}_1^l \circ \ldots \circ \mathcal{T}_N^l$, where $\mathcal{T}_i^l$ is language-specific.

Santali & Kok Borok

Santali – Austroasiatic language that uses (official) Ol Chiki alphabet, but is also recorded in Bengali. The Santali adapter for Bengali makes sure that romanizations of Santali in Bengali and in Ol Chiki are equivalent.

Kok Borok – threatened Sino-Tibetan language:

- Two added diphthongs *vowel letter aw* and *vowel letter ua* encoded using two code-points in Unicode. Require adaptation.

... similar transformations to support Meitei (Manipuri) in Bengali.

Reading Normalization

Motivation

NFC and visual normalization are *visually invariant*. Reading normalization produces visually distinct forms.

Malayam and Hindi Examples

In Malayalam, virama (*candrakkala*) sign has dual function:

1. suppresses the inherent vowel,
2. replaces it with a neutral vowel sound (*samvruthokaram*).

Resolution

For (2), rewrite traditional use *vowel sign u* + virama with modern form (virama only), i.e.: {U+0D41, U+0D4D } (ു ്) → U+0D4D (്).

In (Standard) Hindi:

- *anusvra* sign is traditionally defined as representing a nasal consonant homorganic to a following plosive,
- ... in contrast to *candrabindu*, which marks vowel nasalization.

Pronunciation Ambiguity

In modern writing, the two are used interchangeably.

Sometimes this can be resolved partially at a script level:

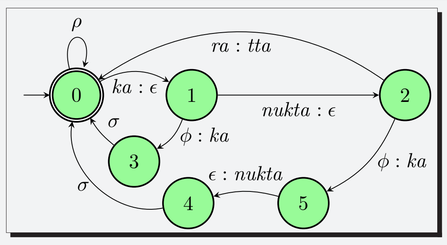
Example: Bilabial Plosive & Explicit Marking

{ *ma* (U+092E), *virama*, *bha* (U+092D) } (म्फ) → { *anusvara*, *bha* (U+092D) } (ंफ).

Also resolving orthographic ambiguities:

- Syloti Nagri: *dvisvara* sign vs. *letter i* for diphthongs ending with /i/.
- Lepcha: Traditional → modern orthography for /ʈ/, /ʈʰ/, /ɖ/.

Example: Lepcha Retroflex Clusters



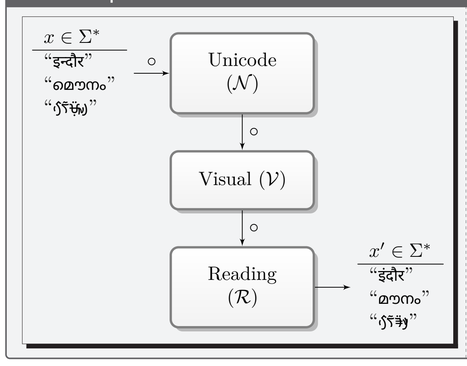
FST semantics:

- Non-consuming transitions:
  - ϵ-transitions: always taken,
  - ϕ-transitions (*failure* transitions) are traversed only when trying to match with a symbol that does not label any arc leaving that state, an "otherwise" arc [6].
- Symbol consuming:
  - σ-transitions are like ϵ,
  - ρ-transitions are like ϕ.

Contextless rewrite

"ൎ" (*ka*, *nukta*, *subj. ra*) → "ഺ" (*tta*)

Full Script Normalization Flow



Transform strings x into x' over a particular script Σ using FST composition ($\circ$): $\mathcal{N} \circ \mathcal{V} \circ \mathcal{R}$

Fixed-Input Transliteration

Input Methods: Challenges & Examples

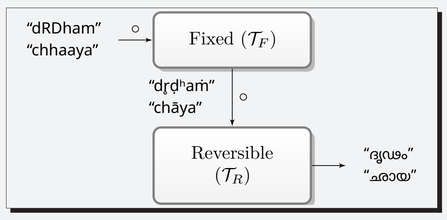
Example: ISO 15919 vs. ITrans

Devanagari	Reversible	ITrans
उ	u	u
ऊ	ū	U
अं	aṁ	aM
द	da	da
ड	ḍa	Da
ढ	ḍʰa	Dha
श	śa	sha
ष	ṣa	Sha
स	sa	sa

Transliteration types: pros and cons on an input side:

- Model-based: Accurate, but ...
  - requires training data. Hard to obtain for low-resource languages,
  - arbitrary input alphabet requires output candidate selection: Malayalam: "padaṁ" → { "പടം", "പാടം", "പാടാം", "പാഠം", "പദം", "പാദം", "പതം", "പഥം" }.
- ISO 15919 & variants: 8-bit Latin is visually compact & keystroke savings, but hard to arrange the keys.
- ITrans, Mozhi & variants: Compact ASCII input space - simple keyboard layout.

Fixed-Input Transliteration Workflow



Transliteration pipeline: $\mathcal{T} = \mathcal{T}_F \circ \mathcal{T}_R$

- $\mathcal{T}_F$: Lightweight (small) adapter FST from ASCII to 8-bit Latin (extended ISO 15919).
- $\mathcal{T}_R$: 8-bit Latin (extended ISO 15919) to native script.

Note: $\mathcal{T}_F$ is *non-reversible*.

Bibliographical References

[1] C. Johny, L. Wolf-Sonkin, A. Gutkin, and B. Roark, "Finite-state script normalization and processing utilities: The Nisaba Brahmic library," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, April 2021, pp. 14–23". [Online]. Available: https://aclanthology.org/2021.eacl-demos.3

[2] F. Coulmas, *The Blackwell Encyclopedia of Writing Systems*. Oxford: John Wiley & Sons, 1999.

[3] K. Gorman, "Pynini: A Python library for weighted finite-state grammar compilation," in *Proceedings of the SIGFSM Workshop on Statistical NLP and Weighted Automata*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 75–80. [Online]. Available: https://www.aclweb.org/anthology/W16-2409

[4] K. Gorman and R. Sproat, *Finite-State Text Processing*, ser. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2021, vol. 14.

[5] M. Riley, C. Allauzen, and M. Jansche, "OpenFst: An open-source, weighted finite-state transducer library and its applications to speech and language," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*. Boulder, Colorado: Association for Computational Linguistics, May 2009, pp. 9–10. [Online]. Available: https://aclanthology.org/N09-4005

[6] C. Allauzen, M. Mohri, and B. Roark, "Generalized algorithms for constructing statistical language models," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan: Association for Computational Linguistics, Jul. 2003, pp. 40–47. [Online]. Available: https://aclanthology.org/P03-1006