Google Research

Criteria for Useful Automatic Romanization in South Asian Languages

Işın Demirşahin, Cibu Johny, Alexander Gutkin, Brian Roark {isin, cibu, agutkin, roark}@google.com

LREC 2022

Introduction

Transliteration is the conversion of language in one script to the same language in another script.

Romanization is transliteration where the target script is the Latin script.

Native	Transliteration	Translation
гласность	glasnost	transparency
идет снег	idet sneg	It is snowing

Romanization is common in South Asia due to:

- lack of types for native scripts in 16th c,
- poor encoding and font support,
- challenges in mobile text entry.

There is no common standard romanization.

Automatic transliteration is useful for:

- Multiscript training for statistical MT
- Multilingual NLP tasks (morphological analysis, NER, POS tagging)

Nisaba library:

Low-level script processing for South Asia, particularly Brahmic scripts.

- Visual normalization
- Well-formedness
- Reversible transliteration

Romanization Criteria

Invertibility

The result can be transliterated back to the exact original input string, eg. ISO 15919

- अस्पताल -> aspatāla

Pronunciation transparency

Increased pronunciation transparency conflicts with invertibility and can introduce ambiguity.

- अस्पताल -> aspatāl

Naturalness

The way speakers of the language tend to spontaneously romanize words.

- अस्पताल -> aspatāla -> aspatal or hospital

Conventionality

Words with a standard spelling, eg. names or loanwords.

- अस्पताल -> aspatāla -> hospital

Ease of Input

Romanizations restricted to the basic Latin script without diacritics.

- अस्पताल -> aspatāla -> aspataal

Stability

The same word in different scripts.

- अस्पताल -> aspatāla -> hospital
- হাসপাতাল -> hāsapātāla -> hospital

Linguistic Challenges

Inherent vowels

Marked vs. unmarked deletion. Eg. "Farsi".

- ఫార్సీ -> pʰārsī (Telugu - virama ్)
- ഫാർസി -> pʰārˀsi (Malayalam - chillu r ർ)
- फ़ारसी -> fārasī (Hindi - unmarked deletion)

Nasal assimilation

Place of articulation assimilation.

- चंदा -> caṁdā -> chanda
- चंबा -> caṁbā -> chamba

Voicing

Intervocalic / post nasal voicing.

- കടൽ -> kaṭal -> kadal

Pronunciation adjustments

Context-independent pronunciation changes.

- कृष्ण -> kr̥ṣṇa -> krishna
- अज्ञान -> ajñāna -> agyaan
- കുറ്റം -> kurraṁ -> kuttam

Dental marking

Natural transliteration conventions for dentals.

- താൾ -> tāḷˀ -> thal
- தூங்க -> tūṅka -> thoonga

Long vowel transliteration

Different strategies to represent long vowels.

- फ़ारसी -> fārasī -> Farsi
- వీధి -> vīdʰi -> veedhi

Implementation

Input

- Visually-normalized ISO 15919 romanization

Intermediate steps

- iso2typ: Ease-of-input mapping.
- typ2txn: Default phoneme mapping.
- phon_ops: Phonological operations.

Output

- txn2nat: Multilingual output formats
- <lang>_e2e: Language specific combinations of modules.

Output Formats

Natural transliteration

Naturalness and conventionality over pronunciation transparency and stability across scripts.

- കേൾക്കാത്ത -> kēḷˀkkātta -> kelkkaatha
- उदात्त -> udātta -> udatt

ISO pronunciation

Invertability over all else, while increasing pronunciation transparency

- फ़ारसी -> fārasī -> fārasī
- हिंदी -> hiṁdī -> hiⁿdī

Pan-South Asian transliteration

Consistency and stability above all else.
Fine grained (PSAF):

- കേൾക്കാത്ത -> kēḷˀkkātta -> keelkkaatta

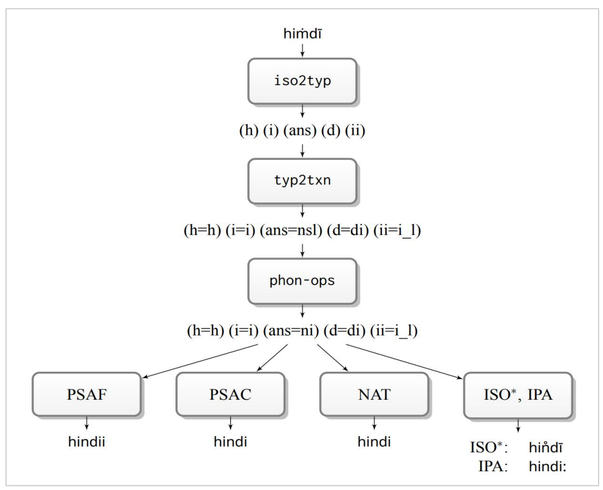
Coarse grained (PSAC):

- കേൾക്കാത്ത -> kēḷˀkkātta -> kelkata

IPA pronunciation

Conventionality and pronunciation transparency.

- വെളുത്ത -> veḷutta -> /ʋeɭutta/




```
phon_ops.py
letter = vowel_letter | consonant_letter
nasal = "ni"
approximant = "y"
sonorant = vowel_letter | nasal | approximant

exporter["VOICING"] = p.cdrewrite(
  p.cross("(tt=tt)", "(dd=dd)"),
  "(" + letter.star + "=" + sonorant + ")",
  "(" + letter.star + "=" + sonorant + ")",
  sigma_star_txn).optimize()
```

Sample snippet for voicing.

- കടൽ -> kaṭal -> kadal

Original word:	फ़ारसी	अस्पताल	വെളുത്ത
Language:	Hindi	Hindi	Malayalam
English gloss:	*Farsi*	*hospital*	*white*
ISO 15919:	fārasī	aspatāla	veḷutta
uconv:	farasi	aspatala	velutta
uroman:	phaarasii	aspataal	vellutta
NAT:	farsi	aspatal	velutha
PSAF:	faarsii	aspataal	velutta
PSAC:	farsi	aspatal	veluta
ISO-pron:	fārᵃsī	aspatālᵃ	veḷutta
IPA:	/fa:rsi:/	/əspəta:l/	/ʋeɭuta/

Comparison of the outputs and the well known romanization utilities *uconv* and *uroman*.

Libraries

- Nisaba: https://github.com/google-research/nisaba
- Natural transliteration: https://github.com/google-research/nisaba/tree/main/nisaba/scripts/brahmic/natural_translit

References

- Demirşahin, Jansche, Gutkin (2018). A unified phonological representation of South Asian languages for multilingual text-to-speech
- Gutkin, Johny, Doctor, Wolf-Sonkin, Roark. (2022). Extensions to Brahmic script processing within the Nisaba library: new scripts, languages and utilities.
- Johny, Wolf-Sonkin, Gutkin, Roark.(2021). Finite-state script normalization and processing utilities: The nisaba Brahmic library
- Roark, Wolf-Sonkin, Kirov, Mielke, Johny, Demirsahin, Hall. (2020). Processing South Asian languages written in the Latin script: the Dakshina dataset.