Degradation of Indic scripts in PanLex

Yadav Gowda

August 8th, 2013

Degradation

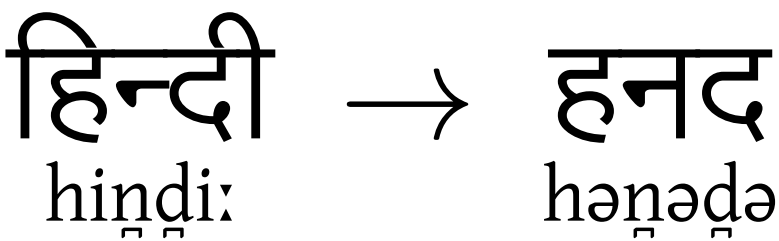
- To help with search as well as normalization of entries, PanLex stores all of its expressions in a full and ‘degraded’ form.
- The degraded form is intended to increase collisions between search terms, by neutralizing orthographic differences.

Champs-Élysées $\rightarrow$ champselysees

Degradation: Issues

What's wrong with the script?

- It doesn't work properly for some non-Latin scripts, including Indic scripts.
- Degradation removes too much information from expressions written in Indic scripts.



- To see why this happens, we need to look at Indic scripts and how they're encoded in Unicode.

Indic scripts

- Indic scripts are part of a larger family of scripts known as Brahmic scripts, which are all descendants of the Brahmi script, from the 3rd century BC, which may itself be descended from Aramaic script.

Indic scripts

- Indic scripts are used to write almost all of the languages of India, as well as some in Bangladesh and Pakistan.

Bengali
বাংলা

Devanagari
देवनागरी

Gujarati
ગુજરાતી

Gurmukhi
ਗੁਰਮੁਖੀ

Kannada
ಕನ್ನಡ

Malayalam
മലയാളം

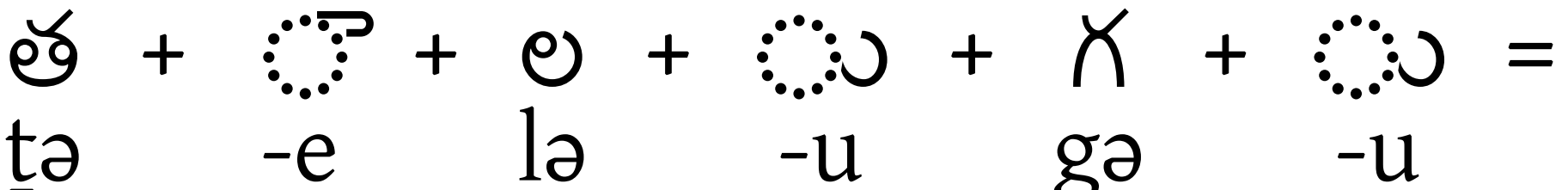
Tamil
தமிழ்

Telugu
తెలుగు

Oriya
ଓଡ଼ିଆ

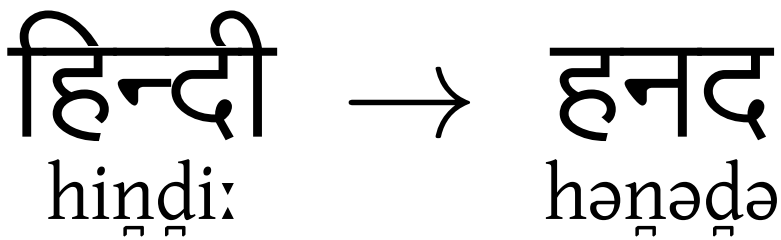
Indic scripts

- Indic scripts are all abugidas, which means that each character represents a syllable, with the main body of the character representing the consonant part of the syllable, and diacritics used to indicate vowels.
- Consonant clusters are represented using ‘conjunct consonants.’
- The unmarked consonants have an implicit vowel, usually /ə/.



తెలుగు
ṭelugu

Indic scripts



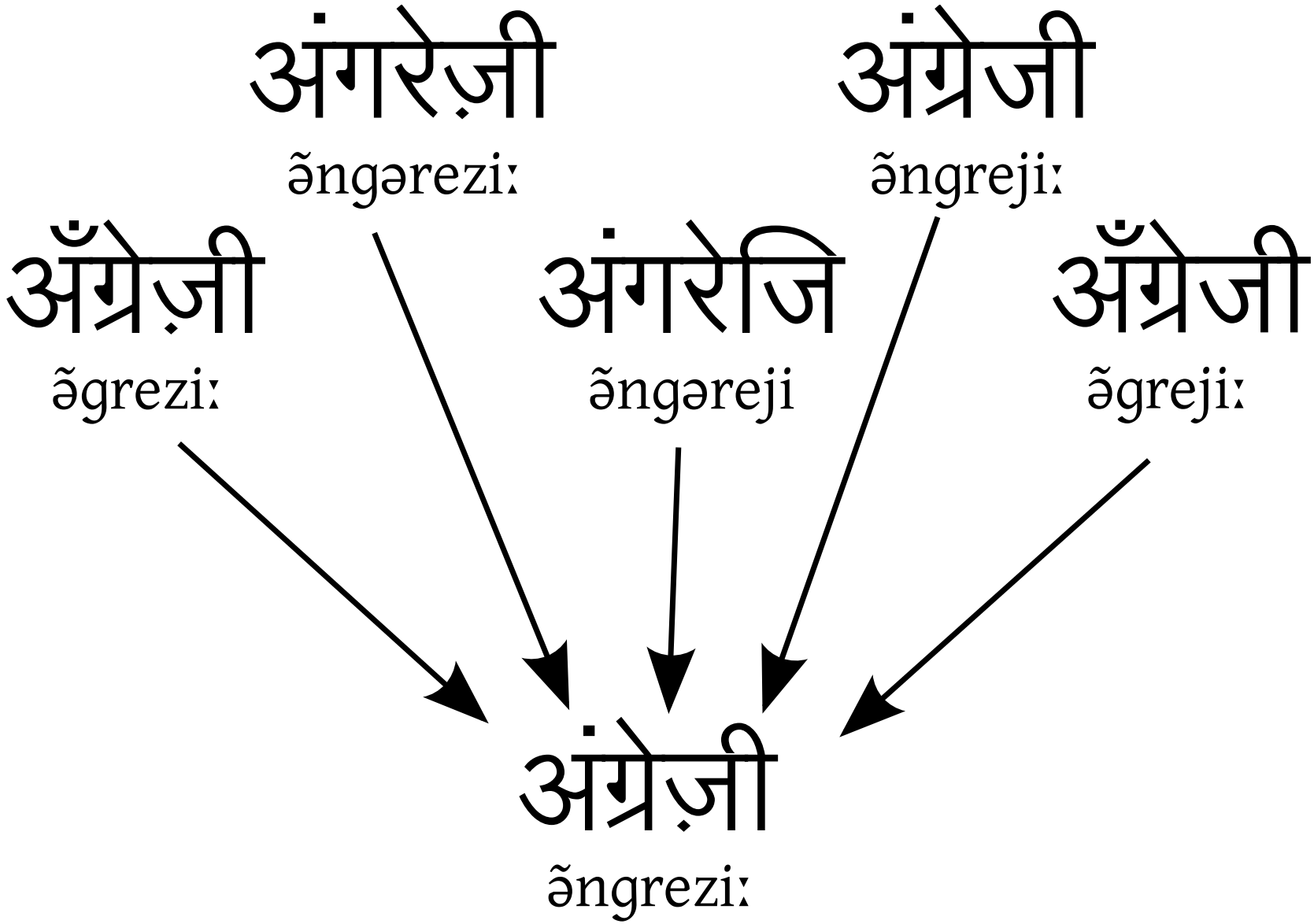
- Now we know part of the reason that degradation doesn't work for Indic scripts: Degradation removes vowels because they are combining characters in Unicode, and our current degradation script removes all combining characters:

```
$td =~ s/[^\p{Ll}\p{Lo}\p{Nd}]//g;
```

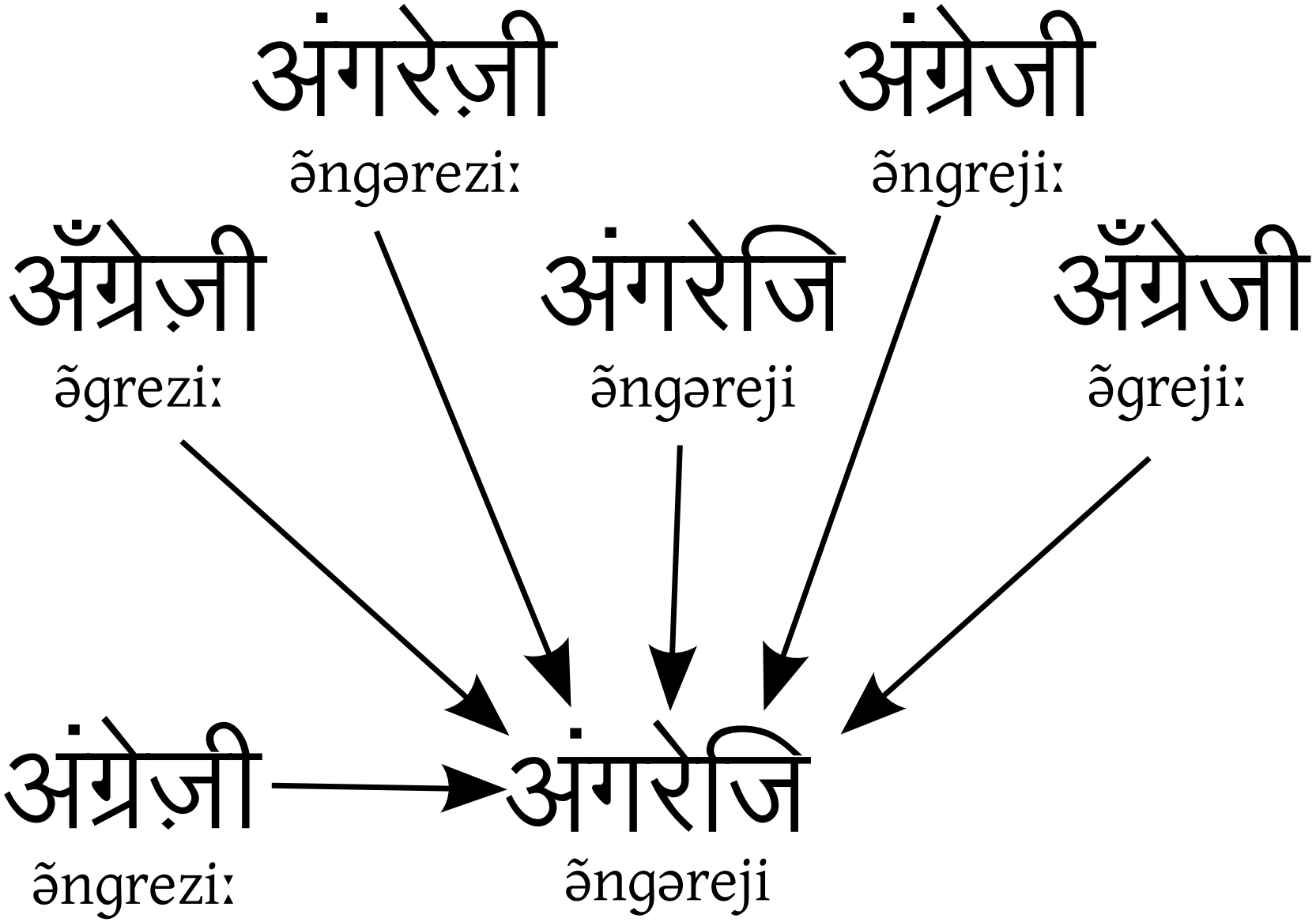
Degradation: Improvements

- The minimal solution would be to prevent the degradation script from removing the vowel diacritics.
- However, we might as well make the degradation more robust, as there are other issues with Indic scripts that this solution doesn't fix.

Degradation: Improvements



Degradation: Improvements



Degradation: Nasals

अँग्रेज़ी → अंगरेजि

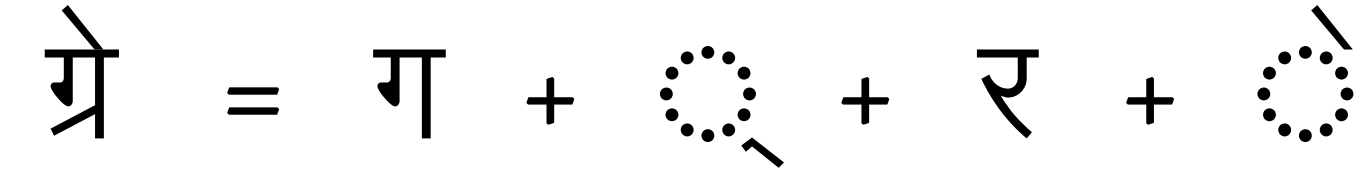
ãgrezi: → ãngəreji

Degradation: Schwa syncope

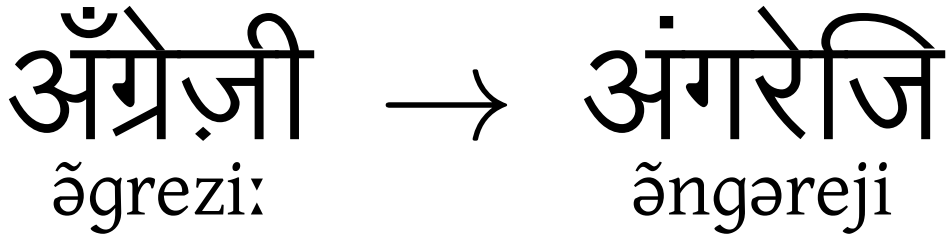
अँग्रेज़ी → अंगरेजि

ãgrezi: → ãngəreji

- Many Indian languages have a process called schwa syncope, in which certain word-internal schwas are deleted during pronunciation.

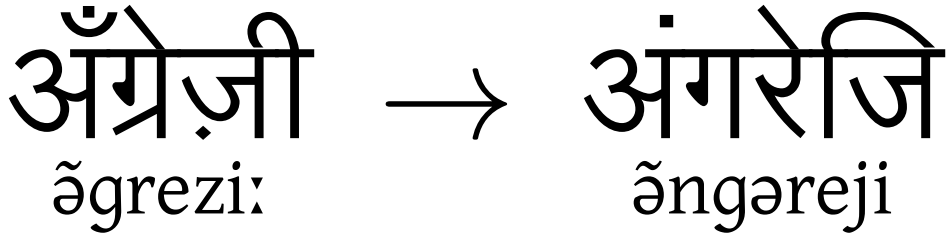


Degradation: Nuqtas and other diacritics



- Indian languages used nuqtas and other diacritics to indicate sounds which were not used when these scripts were invented, such as /f/ and /z/.

Degradation: Vowel length



- Vowel length is phonemic in Indian languages, but it is easy to type a vowel incorrectly.

Degradation: What next?

- Remaining Indic scripts (Saurashtra, Sinhalese, Tibetan)
- SE Asian Scripts (Thai, Sundanese, Khmer)
- Other abugidas (Ge'ez)
- Distinguish by language vs. script?