COD891: Minor Project

Transliteration amongst Indian languages

Nikhil Chaturvedi
2013CS50291

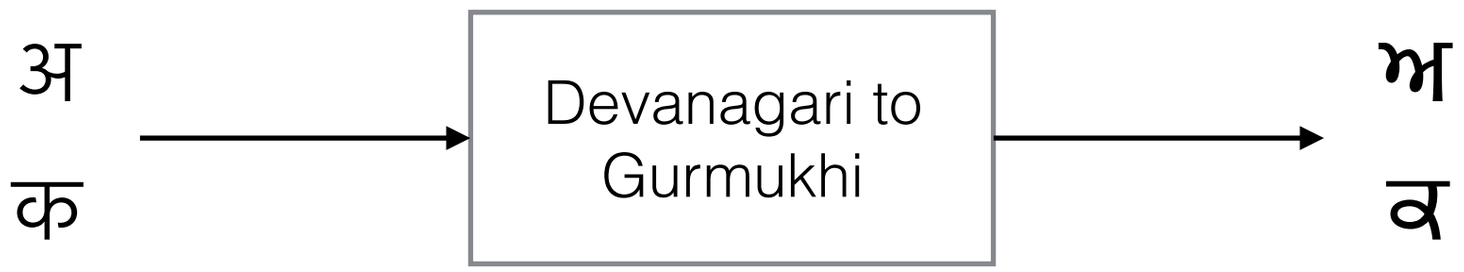
Supervisor: Prof. Rahul Garg

Written Text

- Any computerised written text has 3 characteristics:
  - **Language**: Signifies the semantical meaning of words. Eg. Hindi, Punjabi, Sanskrit etc.
  - **Script**: Signifies the symbolic representation used for depicting the alphabet of the language. Eg. Devanagari, Gurmukhi, SLP1 etc.
  - **Encoding**: The binary encoding schema through which the symbols of the script are stored or presented. Eg. ASCII, UTF-8, UTF-16 etc.

Transliteration

- Transliteration is a type of conversion of a text from one script to another that involves swapping letters (thus trans- + liter-) in predictable ways.
- It makes use of the alphabet correspondence across languages of similar origins.



Motivation

- The Brahmi-derived writing systems of Indian languages are mostly rather similar in structure, but have different letter shapes.

- These scripts are based on similar phonetic values which allows for easy transliteration.

- The phonetic sound [ki] will be rendered as कि in Devanagari, as ਕਿ in Gurmukhi, and as கி in Tamil. Each having different code-points in Unicode and ISCII.

- *This correspondence forms the basis of the motivation behind deriving a uniform encoding schema that is based on the underlying phonetic value rather than the symbolic representation.*

Existing Work: Unicode

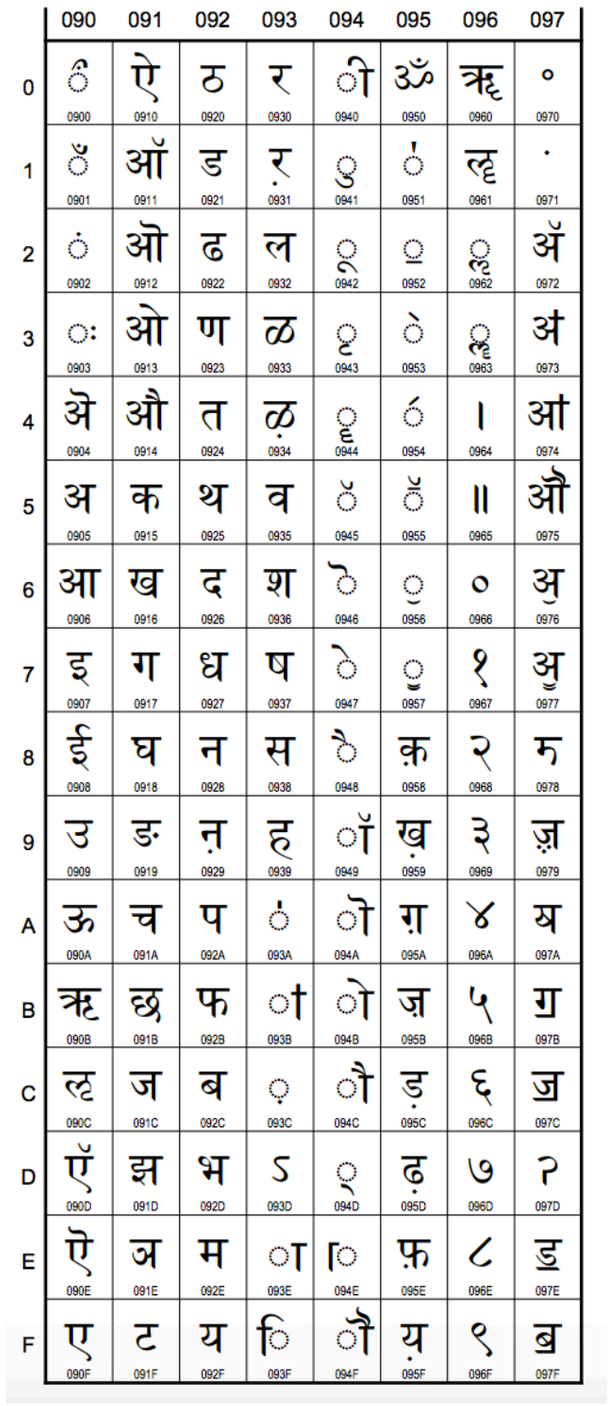
- Unicode has designated code blocks for almost all major Indian scripts.

- The supported scripts are: Assamese, Bengali (Bangla), Devanagari, Gujarati, Gurmukhi, Kannada, Malayalam, Oriya, Tamil, and Telugu among others.

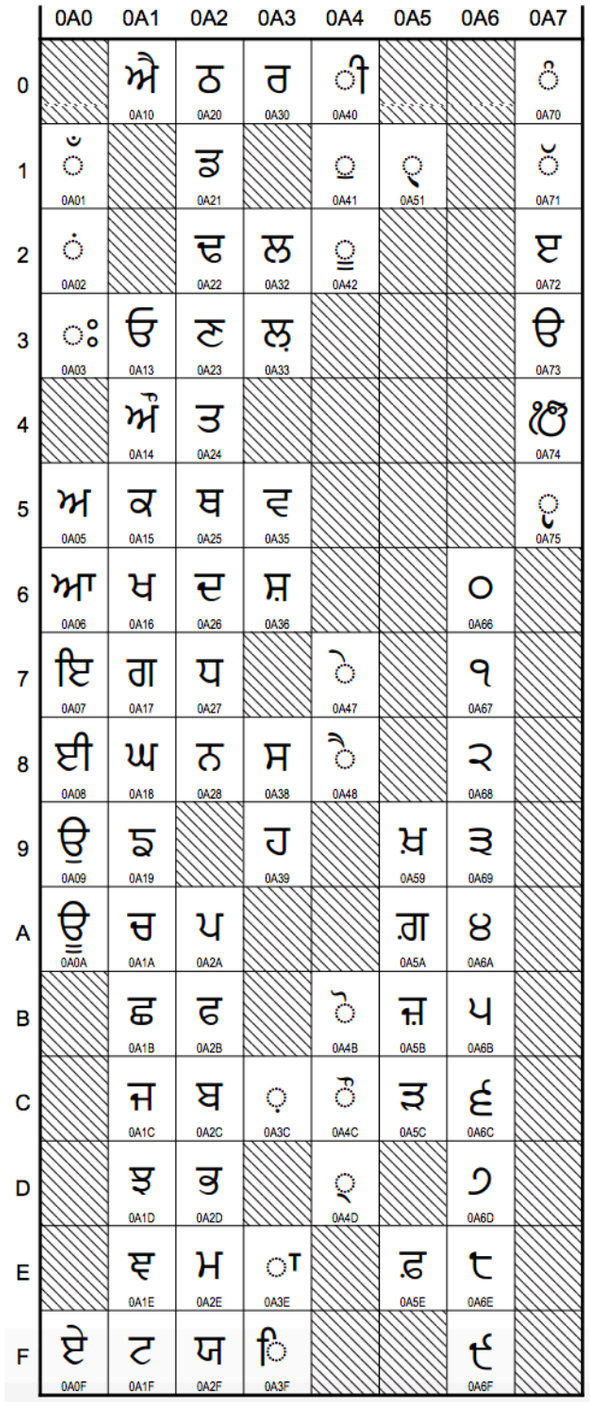
- Across scripts, Unicode respects alphabet correspondence and letters with similar phonetic values are assigned the same codepoints.

- As a result, transliteration can be done easily with a mere offsetting of the alphabet code. For example, अ is U+0905 while ਅ is U+0A05. क is U+0915 while ਕ is U+0A15.

Devanagari and Gurmukhi Codeblocks



	090	091	092	093	094	095	096	097
0	ऀ 0900	ऐ 0910	ठ 0920	र 0930	ी 0940	ॐ 0950	ॠ 0960	॰ 0970
1	ँ 0901	ऑ 0911	ड 0921	ऱ 0931	ु 0941	॑ 0951	ॡ 0961	ॱ 0971
2	ं 0902	ओ 0912	ढ 0922	ल 0932	ू 0942	॒ 0952	ॢ 0962	ॲ 0972
3	ः 0903	ओ 0913	ण 0923	ळ 0933	ृ 0943	॓ 0953	ॣ 0963	ॳ 0973
4	ऄ 0904	औ 0914	त 0924	ऴ 0934	ॄ 0944	॔ 0954	। 0964	ॴ 0974
5	अ 0905	क 0915	थ 0925	व 0935	ॅ 0945	ॕ 0955	॥ 0965	ॵ 0975
6	आ 0906	ख 0916	द 0926	श 0936	ॆ 0946	ॖ 0956	० 0966	ॶ 0976
7	इ 0907	ग 0917	ध 0927	ष 0937	े 0947	ॗ 0957	१ 0967	ॷ 0977
8	ई 0908	घ 0918	न 0928	स 0938	ै 0948	क़ 0958	२ 0968	ॸ 0978
9	उ 0909	ङ 0919	ऩ 0929	ह 0939	ॉ 0949	ख़ 0959	३ 0969	ॹ 0979
A	ऊ 090A	च 091A	प 092A	ऺ 093A	ॊ 094A	ग़ 095A	४ 096A	ॺ 097A
B	ऋ 090B	छ 091B	फ 092B	ऻ 093B	ो 094B	ज़ 095B	५ 096B	ॻ 097B
C	ऌ 090C	ज 091C	ब 092C	़ 093C	ौ 094C	ड़ 095C	६ 096C	ॼ 097C
D	ऍ 090D	झ 091D	भ 092D	ऽ 093D	् 094D	ढ़ 095D	७ 096D	ॽ 097D
E	ऎ 090E	ञ 091E	म 092E	ा 093E	ि 094E	फ़ 095E	८ 096E	ॾ 097E
F	ए 090F	ट 091F	य 092F	ी 093F	ॏ 094F	य़ 095F	९ 096F	ॿ 097F



	0A0	0A1	0A2	0A3	0A4	0A5	0A6	0A7
0		ਅ 0A05	ਠ 0A20	ਰ 0A30	ੀ 0A40			ੰ 0A70
1	ੁ 0A01		ਡ 0A21		ੂ 0A41	ੑ 0A51		ੱ 0A71
2	ੂ 0A02		ਢ 0A22	ਲ 0A32	੃ 0A42			ੲ 0A72
3	੃ 0A03	ਉ 0A13	ਣ 0A23	ਲ਼ 0A33				ੳ 0A73
4		ਅੰ 0A14	ਤ 0A24					ੴ 0A74
5	ਅ 0A05	ਕ 0A15	ਥ 0A25	ਵ 0A35				ੵ 0A75
6	ਆ 0A06	ਖ 0A16	ਦ 0A26	ਸ਼ 0A36			੦ 0A66	
7	ਇ 0A07	ਗ 0A17	ਧ 0A27		ੇ 0A47		੧ 0A67	
8	ਈ 0A08	ਘ 0A18	ਨ 0A28	ਸ 0A38	ੈ 0A48		੨ 0A68	
9	ਉ 0A09	ਙ 0A19		ਹ 0A39		ਖ਼ 0A59	੩ 0A69	
A	ਊ 0A0A	ਚ 0A1A	ਪ 0A2A			ਗ਼ 0A5A	੪ 0A6A	
B		ਛ 0A1B	ਫ 0A2B		ੋ 0A4B	ਜ਼ 0A5B	੫ 0A6B	
C		ਜ 0A1C	ਬ 0A2C	਼ 0A3C	ੌ 0A4C	ੜ 0A5C	੬ 0A6C	
D		ਝ 0A1D	ਭ 0A2D		੍ 0A4D		੭ 0A6D	
E		ਞ 0A1E	ਮ 0A2E	ਾ 0A3E		ਫ਼ 0A5E	੮ 0A6E	
F	ਏ 0A0F	ਟ 0A1F	ਯ 0A2F	ਿ 0A3F			੯ 0A6F	

Problems with Unicode

- The encoding doesn't represent the language in its true essence.
- Hindi, Sanskrit and most other Indian languages are centred around phonetic values. Hence the encoded token should ideally represent the entire sound rather than it being split into different symbols for vyanjana and maatra.
- We cannot figure out anything about the letter from its corresponding encoding. Which section of vyanjana it belongs to, whether it has a sweet or a bitter sound etc.
- The vyanjana symbols have a pre-added 'अ' (क् + अ = क). It is this conjoined sound which gets representation in Unicode rather than the plain क्.

Existing Work: Romanization

- There are several methods of transliteration from Devanagari to the Roman script (a process known as romanization) which share similarities, although no single system of transliteration has emerged as the standard. Eg. SLP1, Velthius, Harvard-Kyoto etc.

- These can represent not only the basic Devanagari letters, but also phonetic segments, phonetic features and punctuation. SLP1 also describes how to encode classical and Vedic Sanskrit.

- *This motivates the second part of the project which involves auto-detection of the script through Machine Learning. As figuring out the difference between script schemas which use latin letters is a non-trivial problem.*

Universal Encoding

- In this project, we set out to design a universal encoding for all Indian languages.
- The encoding is 16-bit unlike Unicode which requires 24-bits to represent most Indian characters.
- Initial 5 bits are for specifying the script (hence we can support 32). Next 6 bits are for the vyanjana (consonant). Last 5 bits are for the swar/maatra (vowel).
- In most cases, the vyanjana as well as the mantra both require 24-bits each in Unicode. We fit both in 16 bits.
- The encoding respects the structure of the language as described by Panini and we can figure out important characteristics about the letter just by seeing the encoding.
- The first 3 bits of vyanjana represent the source of origin (tongue, throat, teeth etc). The last bit of vyanjana represents prayatna. The second last bit of vyanjana represents sweet or pungent sound. The last bit of star represents deergh (long) or laghu (short).

Encoding for Devanagari

Vyanjana

	000	001	010	011	100	101	110	111
000	क	ख	ग	घ		ह	ङ	
001	च	छ	ज	झ		श	ञ	य
010	ट	ठ	ड	ढ		ष	ण	र
011	त	थ	द	ध		स	न	ल
100	प	फ	ब	भ			म	व
101	क़	ख़	ग़	ज़	ड़	ढ़	फ़	
110	Uddat	Anuddat	ः	ं	ँ	ऽ		
111	Latin	Space	Punc	Numbers	Vaidika			Null

Encoding for Devanagari

Swara

	00	01	10	11
000	अ	आ	अे	
001	इ	ई	ऐ	ए
010	ऋ	ॠ		ऐ
011	ऌ	ॡ	ओ	ओ
100	उ	ऊ	औ	औ
101	अं	आं	अु	अू
110	अॅ	एॅ	ऑ	ॐ
111				Null

Transliteration Pipeline

- **Fragmentation**: Splitting the given text into smaller fragments (words, sentences, paragraphs etc). The assumption shall be that the script and encoding remain same through these fragments if not through the entire text.

- **Script Detection**: Figuring out the scripts and encodings for the various fragments through a Naive Bayes model.

- **Tokenisation**: Splitting the fragment further into tokens, each of which represent a single sound. Similar to the concept of English syllables. कि will be seen as one single token under this model.

- **Universalisation**: Conversion of the token to the universal 16-bit encoding designed by us. Done through pre-populated hash maps.

- **Specification**: Conversion of the universal encoding to the specified script using pre-populated hash maps.

Work Till Now

- Built a lexer for Devanagari and SLP1 using Flex.
- Populated code maps for the above two languages.
- Built an encoder and decoder using the above maps on C++.
- Can perform transliteration between the above two languages.
- Script addition to the above framework is easy and any new script can be added to the list in 5-10 mins.

Results

- **Devanagari**: पक्षि इस पृष्ठ पर इन्टरनेट पर उपलब्ध विभिन्न हिन्दी एवं देवनागरी सम्बंधित साधनों की कड़ियों की सूची है। इसमें उपकरण टूल्स शामिल हैं।

- **Encoded**: 0400 001f 02a4 073f 07e4 03a0 073f 0408 02bf 0220 073f 0400 02e0 073f 07e4 03df 0200 02e0 03c7 0200 073f 0400 02e0 073f 07f0 0400 03e0 045f 0360 073f 04e4 0464 03df 03c0 073f 00a4 03df 0345 073f 07e7 04e0 067f 073f 0347 04e0 03c1 0040 02e5 073f 03a0 04df 0440 067f 0364 0300 073f 03a1 0360 03cf 067f 073f 0005 073f 0000 0584 01ef 067f 073f 0005 073f 03b1 0105 073f 00ab 075f 073f 07e4 03a0 04c7 067f 073f 07f0 0400 0000 02e0 02c0 073f 0211 03ff 03a0 073f 01a1 04c4 03e0 073f 00ab 067f 075f

- **SLP1**: pakzi isa pfzWa para inwaranewa para upalabDa viBinna hindI evaM devanAgarI sambaMDita sADanoM kI kaN*iyoM kI sUcI hE. isameM upakaraRa wUlsa SAmila hEM.

Work Ahead

- Script addition to the current framework.

- Planned scripts are: Assamese, Bengali (Bangla), Devanagari, Gujarati, Gurmukhi, Kannada, Malayalam, Oriya, Tamil, Telugu, Velthius, ITRANS, Harvard-Kyoto.

- Building corpus for English, SLP1, Velthius, ITRANS, French, Italian and German.

- Building a naive bayes model on substring frequencies to differentiate the above scripts.

References

- http://unicode.org/charts/

- https://en.wikipedia.org/wiki/Indian_Script_Code_for_Information_Interchange

- https://en.wikipedia.org/wiki/Devanagari_transliteration

- https://betterexplained.com/articles/unicode/

- https://en.wikipedia.org/wiki/SLP1