Challenges of Processing South Asian Languages (CPSAL)

Kengatharaiyer **Sarves**waran
University of Konstanz, Germany
University of Jaffna, Sri Lanka.

Tafseer Ahmed
Senior NLP Scientist
Alexa Translations

ESSLLI 2024 EUROPEAN SUMMER SCHOOL OF LOGIC, LANGUAGE AND INFORMATION LEUVEN

Course outline

- **Topics (Tentative):**
  - Day 01: Languages, Scripts, and Encoding of South Asian Languages.
  - Day 02: Phonology, Transliteration and Morphology of South Asian Languages.
  - Day 03: Part of Speech and Multiword tokenisation
  - Day 04: Syntax, Morphosyntax, and Semantics of South Asian Languages.
  - Day 05: Deep Learning for South Asian Languages and winding up the course.

Introductions

- Name
- Where are you from
- Are you already working in SALs?
- Other

Challenges of
Processing South Asian Languages
(CPSAL)

Day 01: Languages and Scripts of South Asian Languages

Topics

- South Asia
- Languages
- Scripts
- Encodings
- Challenges

Some Background

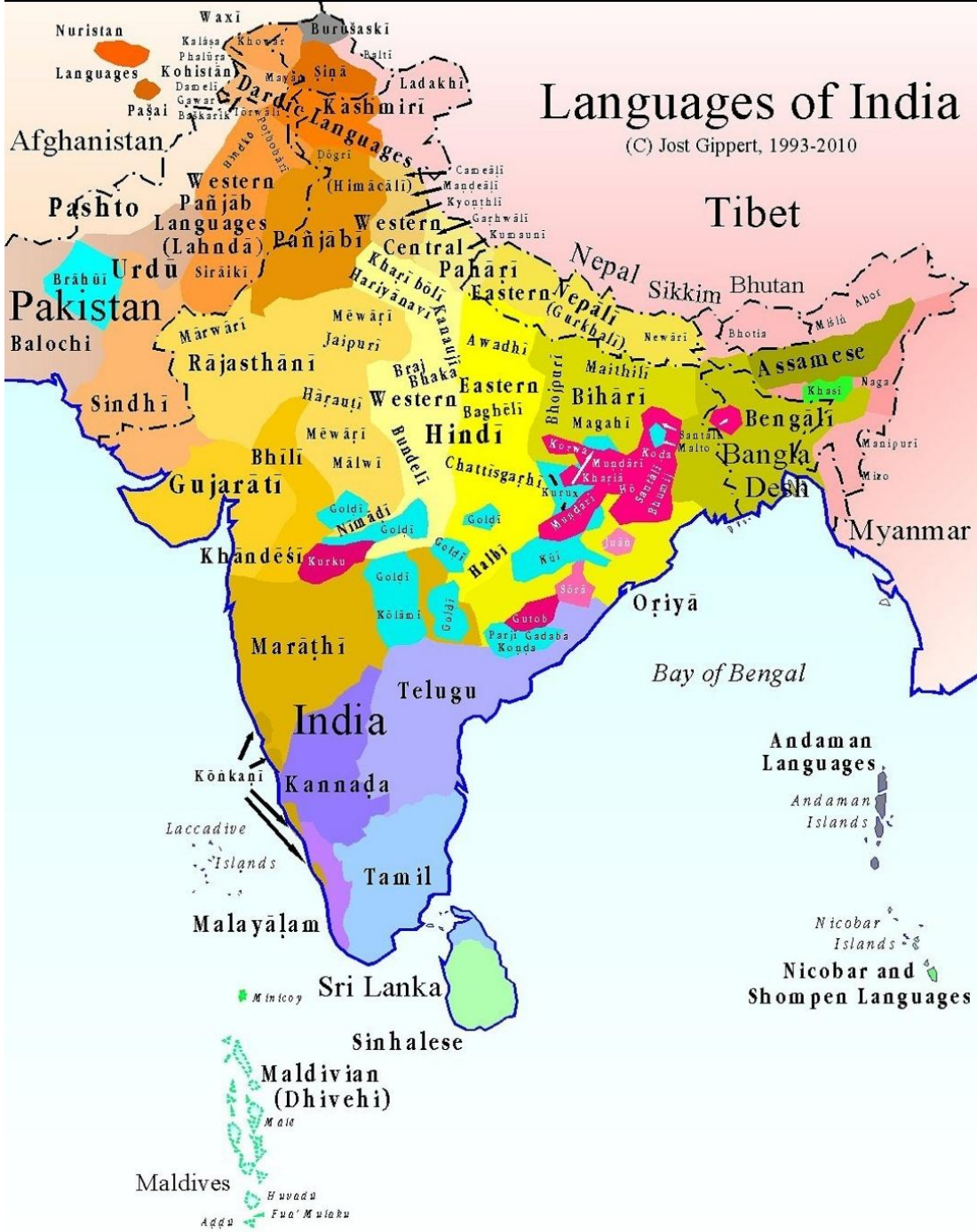
South Asia


Afghanistan
Pakistan
Bhutan
Nepal
India
Bangladesh
Sri Lanka
Maldives

South Asia

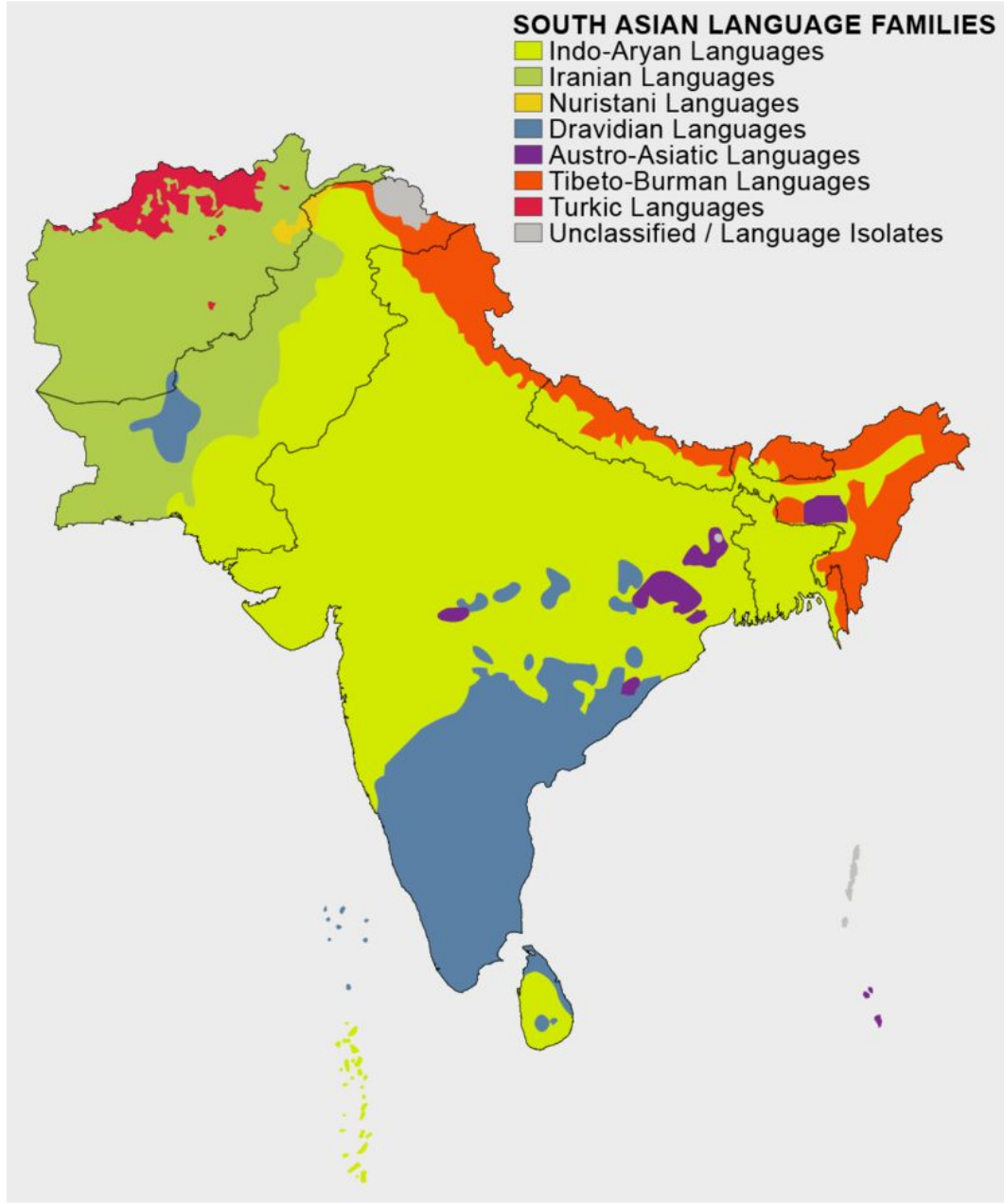
Country	Population	Languages (Primary)
Afghanistan	42+ millions	Pashto, Dari
Bangladesh	172+ millions	Bangla
Bhutan	787 thousands	Dzongkha
India	1.4+ billions	Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Odia, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu, Urdu
Maldives	521 thousands	Divehi
Nepal	30+ millions	Nepali
Pakistan	240+ millions	Urdu, Punjabi, Saraiki, Sindhi, Baluchi and Pashto
Sri Lanka	21+ millions	Sinhala and Tamil
(3% of world's land area)	**1.9+ billions** *(24+% of the world's population)*	**650 living languages**
Migrants, Diaspora	?	New dialects?

Languages of South Asia

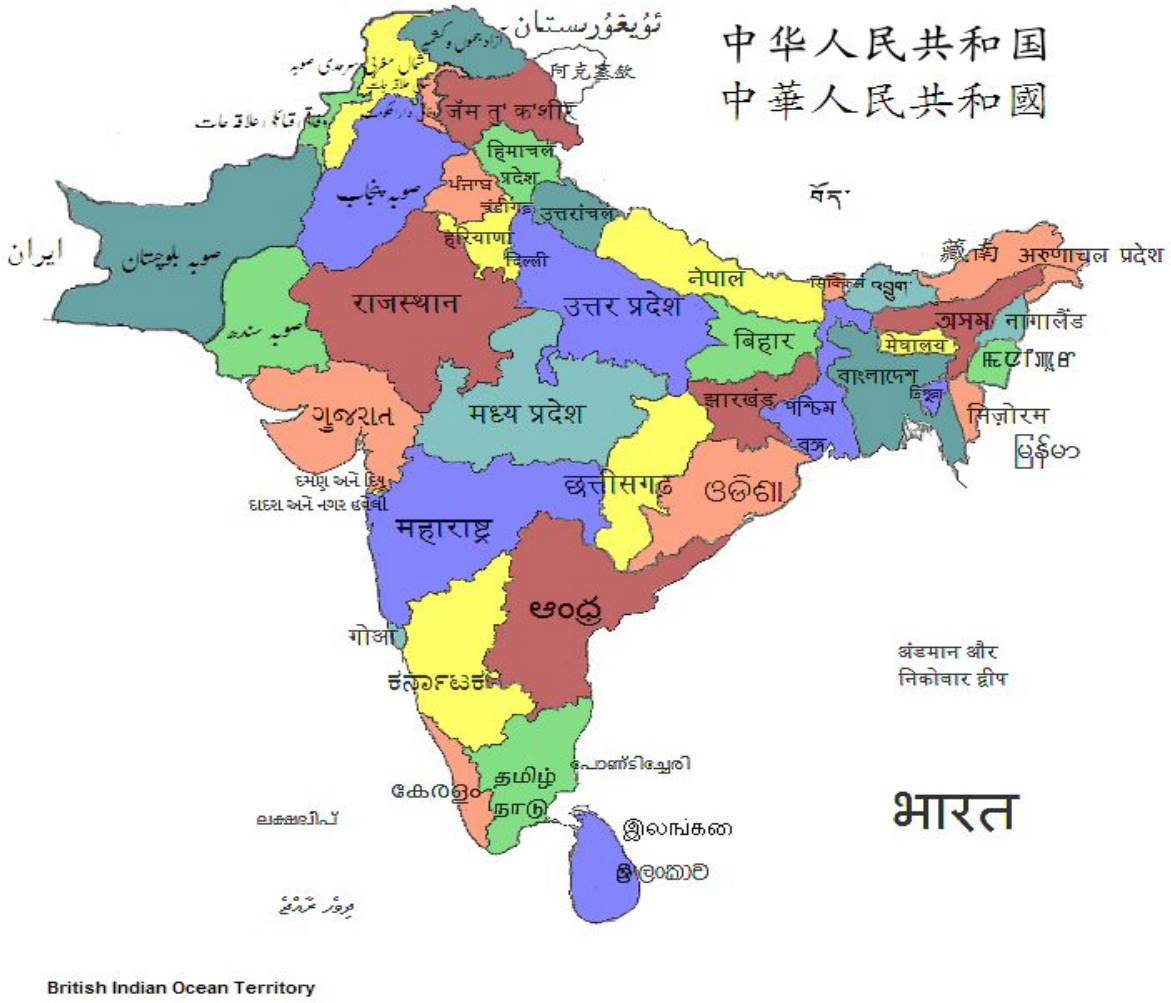


Language families

- **Indo-Aryan**
  - Urdu, Hindi, Rajasthani, Nepalese…
- **Iranian**
  - Pashto, Balochi, Dari…
- **Nuristani**
  - Kamkata-viri, Ashkun, Tregami…
- **Dravidian**
  - Tamil, Malayalam, Telugu, Kannada, Brahui…
- **Austro-Asiatic**
  - Munda, Santali, Khasi, Mundari…
- **Tibeto-Burman**
  - Bodo, Manipuri (Meitei)...
- **Turkic**
  - Uzbek, Turkmen…



Scripts of South Asia



Scripts of South Asia: Examples of scripts used to write Indo-Aryan languages

- कोन्स्टांज़ - Devanagari (Hindi)
- কোনস্টান্জ - Bengali
- ਕੋਨਸਟਾਂਜ਼ - Gurmukhi (Punjabi)

- કોન્સ્ટાંઝ - Gujarati

- කොන්ස්ටාන්ස් - Sinhala (Sinhala)

Scripts of South Asia: Examples of scripts used to write **Iranian** languages

- (Urdu) Naskh - کونسٹانس
- (Urdu) Nastaliq - کونسٹانس
- (Sindhi) Naskh - کونسٹانس
- (Pashto) Naskh - کونسټانس،

Scripts of South Asia: Examples of scripts used to write **Dravidian** languages

- கோன்ஸ்டான்ஸ் - Tamil
- కొన్స్టాంజ్ - Telugu
- ಕೊನ್ಸ್ಟನ್ಸ್ - Kannada

- കോൺസ്റ്റാൻസ് - Malayalam

Writing systems

- **Brahmi scripts: Tamil, Sanskrit, Punjabi and others.**
  - **Left to right** writing system
  - **Abugida/alphasyllabary** writing system: vowels, consonants, and composites.
  - Composite = Consonant+Vowel modifier = One Unit/Grapheme
    - க் (k) + அ (a) = க (ka)
    - க் (*k*) + உ (*u*) = கு (*ku*)
    - க் (*k*) + ஆ (*aa*) = கா (*kaa*)
    - க் (*k*) + ஐ (*ai*) = கை (*kai*)
    - க் (*k*) + ஓ (*oo*) = கோ (*koo*)

Writing systems

- **Arabic scripts: Urdu, Punjabi, Sindhi, Pastho**
  - **Abjad** writing system
  - **Right to left**
  - Consonants and long vowels are written – short vowels exist, but **optional**.
  - Different shapes for the character

1st character	2nd character	3rd character	Result
س	ب	ب	سبب

Language vs Script

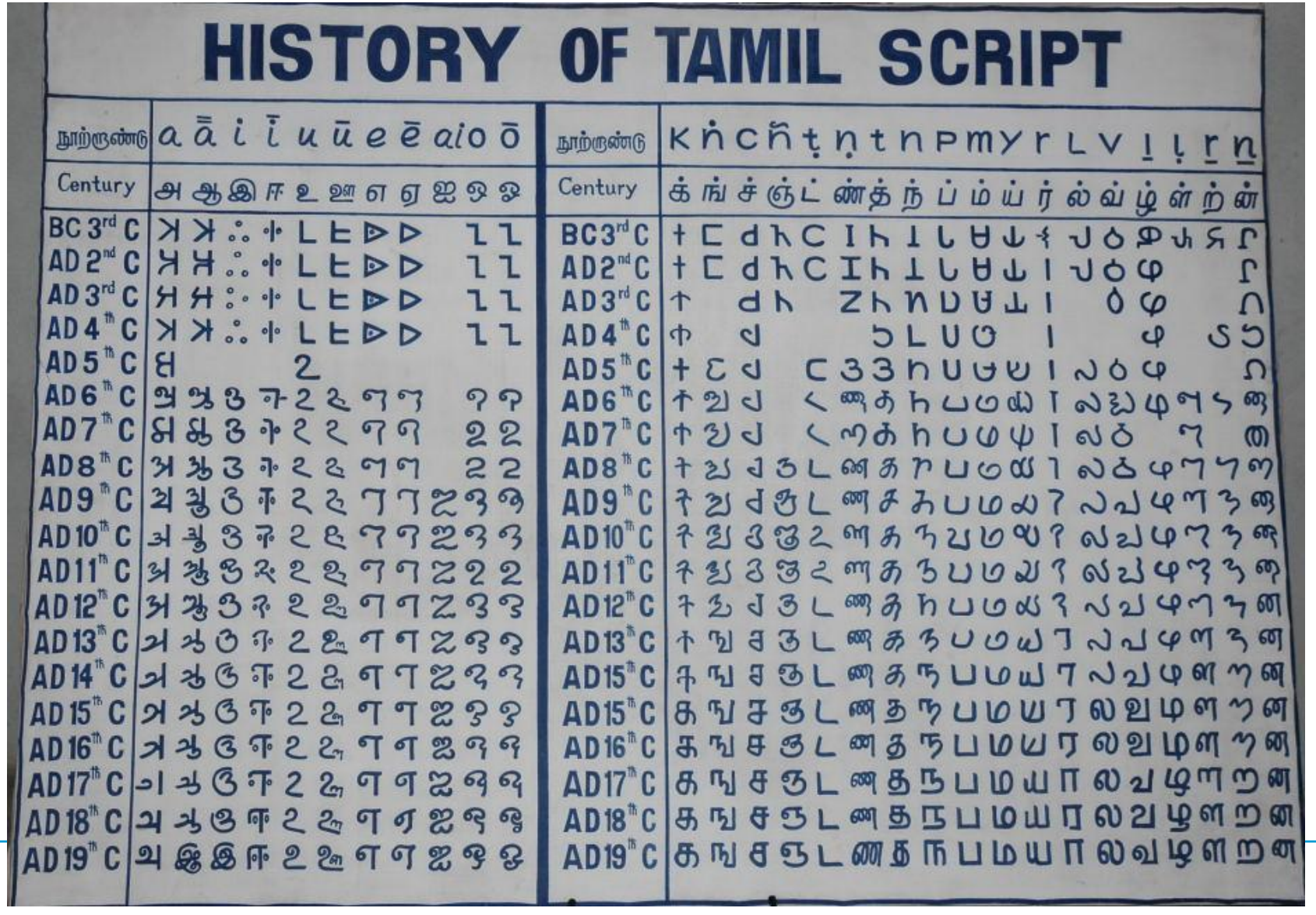
- **One script -> Many languages**
  - आप कैसे हैं? (*Āp kaise hain?*) - Hindi
  - आप केम बा? (*Āp kem bā?*) - Bhojpuri
  - तुम्ही कसे आहात? (*Tumhī kase āhāt?*) - Marathi
  - कथं असि? (*Katham asi?*) - Sanskrit
  - तपाइंलाई कस्तो छ? (*Tapā'īnlā'ī kasto cha?*) - Nepali
  - …

Language vs Script

अ	a	
इ	i	
उ	u	
ऋ	ṛ	ॠ
ऌ [4]	ḷ	ॡ
ए	e	ē
ओ	o	ō
अं / ◌ं [1,2]	ṃ	ṁ
आ	ā	
ई	ī	
ऊ	ū	
ॠ [4]	ṝ	ॣ
ऌ [4, 5]	ḹ	ॢ
ऐ	ai	
औ	au	
अः / ◌ः [1]	ḥ	

- **Many scripts -> One language**
  - Sanskrit: Devanagari and Grantha
  - Punjabi and Sindhi are written using different writing systems
    - Punjabi: Shahmukhi and Gurmukhi
      - **Shahmukhi - تسی کویں ہو (tusI kiveN ho)**
      - **Gurmukhi - ਤੁਸੀ ਕਿਵੇਂ ਹੋ (tusī kivēṁ hō)**
    - Sindhi: Perso-Arabic, Devanagari, Khudabadi, and Khojki
      - **Naskh - تون ڪيئن آهين**
      - **Devanagari - तूं कीअं आहीनि (tūṁ kī'aṁ āhīni)**
  - All the languages are also written in the Roman script.

Scripts have evolved: Example - Tamil



Numerals and other symbols - Arabic scripts

- **Different shapes of some digits**

Each numeral in the Persian variant has a different Unicode point even if it looks identical to the Eastern Arabic numeral counterpart. However, the variants used with Urdu, Sindhi, and other Languages of South Asia are not encoded separately from the Persian variants.

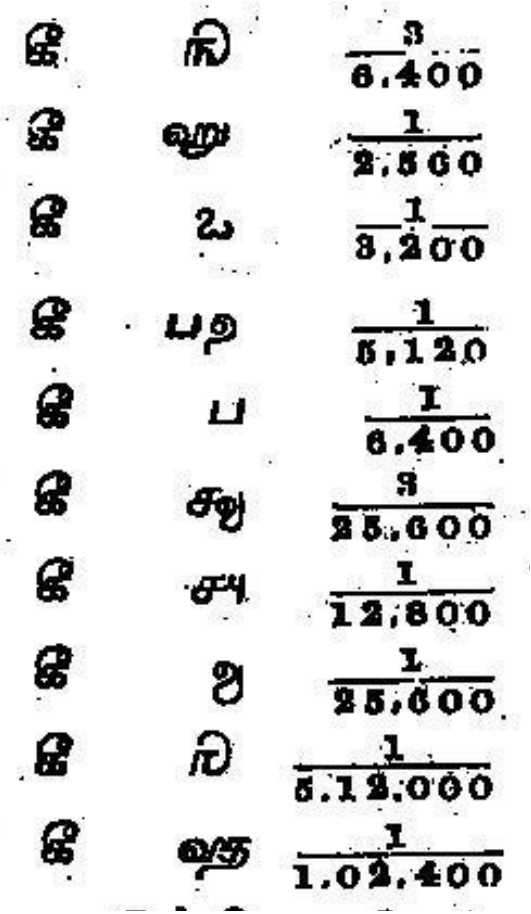
Western Arabic	0	1	2	3	4	5	6	7	8	9	10
Eastern Arabic[a]	٠	١	٢	٣	٤	٥	٦	٧	٨	٩	١٠
Persian[b]	۰	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
Urdu[c]	۰	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰

Numerals and other symbols - the Tamil script

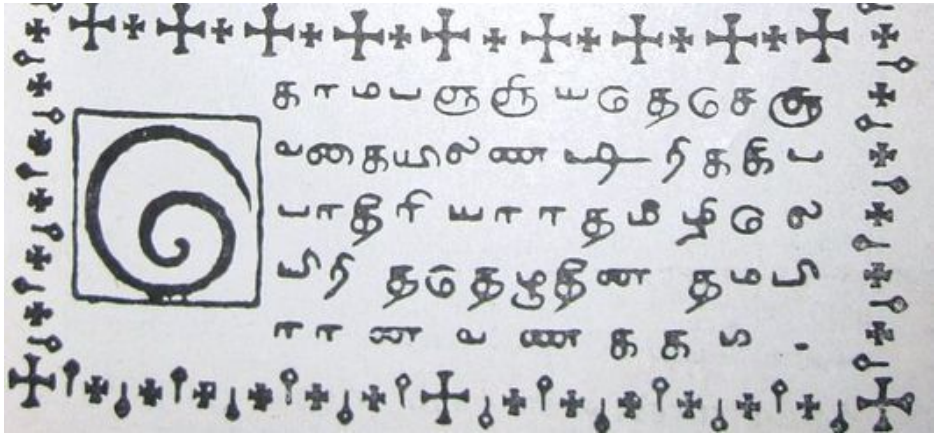
- **Examples**

Comparing Ninal Decimal and Tamil Decimal Counting

0	1	2	3	4	5	6	7	8	9	10	100	1000
	க	உ	ங	ச	ஞ	ண	எ	அ	க	ய	ன	த
	11	12	13	14	15	16	17	18	19	20		
	கக	கஉ	கங	கச	கஞ	கண	கஎ	கஅ	ககி	உய		
	21	22	23	24	25	26	27	28	29	30		
	உக	உஉ	உங	உச	உஞ	உண	உஎ	உஅ	உக	ஙய		
	31	32	33	34	35	36	37	38	39	40		
	ஙக	ஙஉ	ஙங	ஙச	ஙஞ	ஙண	ஙஎ	ஙஅ	ஙக	சய		
									49	50		
									சக	ஞய		
									59	60		
									ஙக	ணய		
									69	70		
									ணக	எய		



Writing system vary based on medium/technology



தமிழ், உலகில் உள்ள முதன்மையான மொழிகளில் ஒன்றும் செம்மொழியும் ஆகும். இந்தியா, இலங்கை, மலேசியா, சிங்கப்பூர் ஆகிய நாடுகளில் அதிக அளவிலும், ஐக்கிய அரபு அமீரகம், தென்னாப்பிரிக்கா, மொரிசியசு, பிசி, இரீயூனியன், திரினிடாடு போன்ற நாடுகளில் சிறிய அளவிலும் தமிழ் பேசப்படுகிறது.

- Written together without spaces, different shapes for the same character in different medium
- Consonants are written without *pulli* (a diacritic)

Character Encoding, Fonts, Input methods of South Asian Scripts

ASCII & Unicode

- **ASCII**
  - American Standard Code for Information Interchange - 128 code points.
  - 1 byte per character.
- **Unicode**
  - Version 15.1 of the standard defines 149,813 code points and 161 scripts.
  - https://unicode.org/charts/
  - 3 (UTF-8) or 2 (UTF-16) bytes per Unicode (by default).

Fonts

- Understanding fonts is important for South Asian languages.
- More than just providing different shapes, some of the styles and ligatures are managed at the font rendering level.

Shakespeare	Naskh شیکسپیئر	Nastaliq 

Unicode - Ligatures

- **Multiple characters -> single ligature**
  - By rendering - using fonts
  - By using special Unicode characters
- **Ligatures by rendering characters**
  - **ஸ்ரீ (*shri*) - ஸ் (sh) + ரீ (ri)**

1st character	2nd character	3rd character	Result
س	ب	ب	سبب

- **Ligatures by using special Unicode characters**

ಠ

ಠ + ◌್ + ZWJ + ಠ --> ಞ

ಠ + ◌್ + ಠ --> ಠ್ಠ

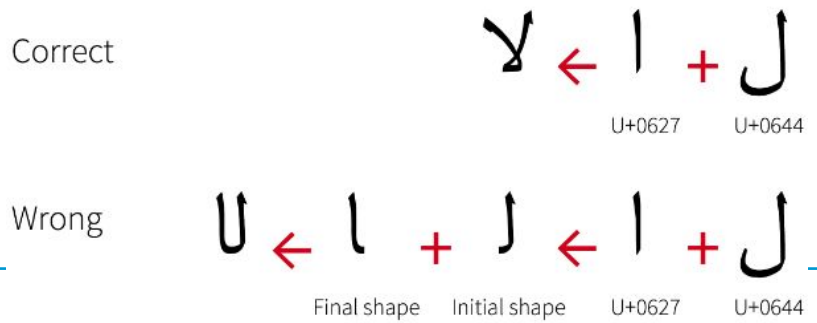
https://r12a.github.io/app-conversion/

Unicode Character, Glyphs, and Ligatures

- **Unicode Character**
  - U+0627 ا , U+0628 ب , U+062A ت , U+062B ث
- **Glyphs**

beh-arab	beh-arab.isol	beh-arab.init	beh-arab.medi	beh-arab.fina
ب	ب	بـ	ـبـ	ـب
teh-arab	**teh-arab.isol**	**teh-arab.init**	**teh-arab.medi**	**teh-arab.fina**
ت	ت	تـ	ـتـ	ـت

- **Ligatures**



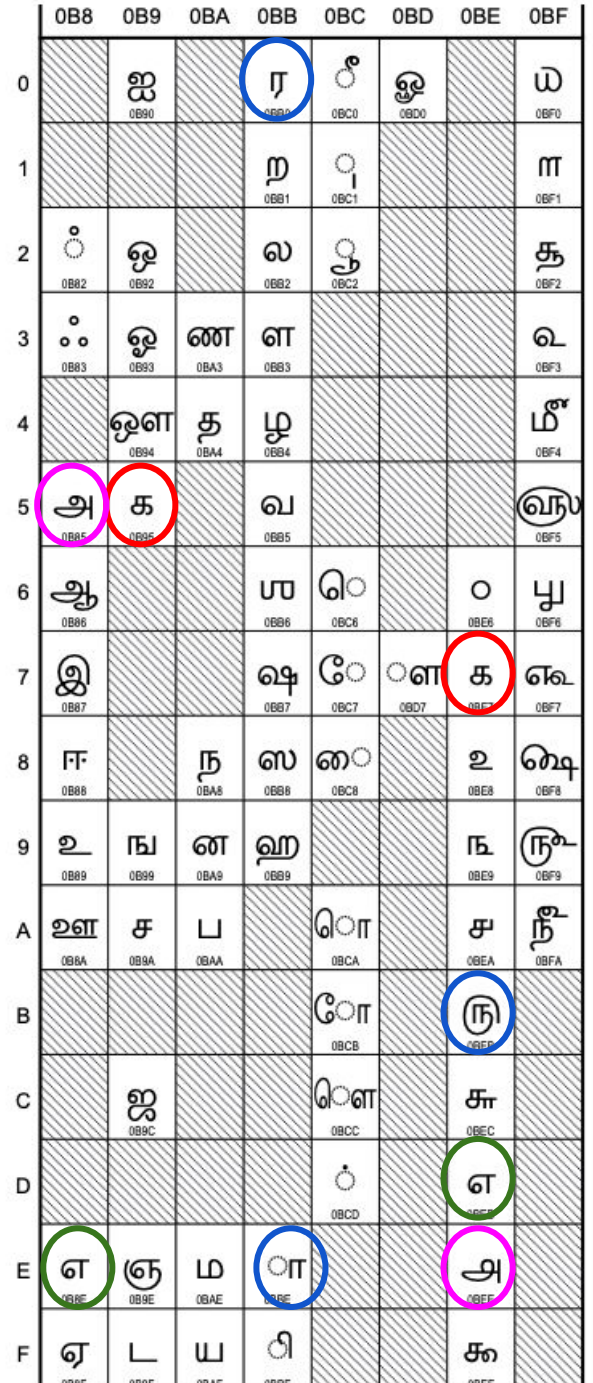
Challenges

Identification

- **How many characters in South Asian Languages?**
- **Language identification:**
  - Several languages are being written using a single script.
  - Single language is being written using multiple scripts.
  - Dialectal variations.

Confusing shapes

- **Visually confusing graphemes/numbers/symbols:**
  - கண் (*kan* - Eye) - கண் (*1n(?)* - ?)
  - அரசு - அரசு (*arasu* - Government) -
- **Optical Character Recognition:**
  - Character, Word, and Sentence segmentations.
  - Evolution of scripts.
  - Similar shapes.



Modifiers

- **Short Vowels as diacritics - usually not transcribed resulting in ambiguity**

- **Non-joiner Characters - ambiguity for word boundaries**

برابر برابر بر ابر	برابر brAbar *'equal/next'*	ابر abar *'cloud'*	بر bar *'land'*

Sorting

- **Cannot sort just by using Unicode points**

1.	க	(U+0B95) - KA
2.	ங	(U+0B99) - NGA
3.	ச	(U+0B9A) - CA
4.	ஞ	(U+0B9E) - NYA
5.	ட	(U+0B9F) - ṬA
6.	ண	(U+0BA3) - ṆA
7.	த	(U+0BA4) - TA
8.	ந	(U+0BA8) - NA
9.	ப	(U+0BAA) - PA
10.	ம	(U+0BAE) - MA
11.	ய	(U+0BAF) - YA
12.	ர	(U+0BB0) - RA
13.	ல	(U+0BB2) - LA
14.	வ	(U+0BB5) - VA
15.	ழ	(U+0BB4) - L̲A
16.	ள	(U+0BB3) - ḶA
17.	ற	(U+0BB1) - r̲A
18.	ன	(U+0BA9) - n̲A

1.	(U+0627) - A
2.	ب (U+0628) - B
3.	پ (U+067E) - P
4.	ت (U+062A) - T
5.	ٹ (U+0679) - Ṭ
6.	ث (U+062B) - S
7.	ج (U+062C) - J
8.	چ (U+0686) - Ch
9.	ح (U+062D) - Ḥ
10.	خ (U+062E) - Kh
11.	د (U+062F) - D
12.	ڈ (U+0688) - Ḍ
13.	ذ (U+0630) - Dh
14.	ر (U+0631) - R
15.	ڑ (U+0691) - Ṛ
16.	ز (U+0632) - Z
17.	ژ (U+0698) - Zh
18.	…

Rendering

- **Application support for rendering characters**
  - Special glyphs are not supported - in some applications these are considered as symbols during the processing.
  - Rendering issues - Glyphs are not stored in the order that we see.
    - தே - த ே
    - https://r12a.github.io/app-conversion/
  - Input methods - keyboard software.
- **Not all the letters in South Asian Languages are encoded**
  - Cannot process old text

Encodings

- **Different Encodings:**
  - Standard Unicode
  - Other regional encoding (Tamil - TACE)
  - ASCII for South Asian Languages (Tamil - TAB/TAM)
- **Conversation challenges**
- **Consume more memory**
  - https://onlinetools.com/unicode/count-unicode-characters

Unicode normalisation

- **Some characters can be typed in multiple ways**
  - Example:
    - கொக்கு (*kokku* - Egret) *vs* கொக்கு
    - க ொ க் கு vs க ொ ◌ ா க் கு
- **https://r12a.github.io/app-conversion/**

Normalisation - more confusions

Display	Unicode Character Sequence C_1	C_2	C_3	C_4	C_5	Transformation
رئيس	*reh* U+0631	*yeh with hamza above* U+0626	*yeh* U+064A	*seen* U+0633		
رئيس	*reh* U+0631	*yeh* U+064A	*hamza above* U+0654	*yeh* U+064A	*seen* U+0633	Unicode NFC
رئيس	*reh* U+0631	*alef maksura* U+0649	*hamza above* U+0654	*yeh* U+064A	*seen* U+0633	*Visual* Normalization
رئيس	*reh* U+0631	*yeh with hamza above* U+0626	*farsi yeh* U+06CC	*seen* U+0633		*Visual* Normalization
رئيس	*reh* U+0631	*farsi yeh* U+06CC	*hamza above* U+0654	*yeh* U+064A	*seen* U+0633	*Reading* Normalization
رئيس	*reh* U+0631	*farsi yeh* U+06CC	*hamza above* U+0654	*farsi yeh* U+06CC	*seen* U+0633	*Reading* Normalization

Table 1: Six different spellings of the Arabic word for "leader" (MSA: /ra.ʔi:s/) rendered in Naskh.

https://arxiv.org/ftp/arxiv/papers/2210/2210.12273.pdf

Normalisation - more confusions

- **Different unicode points for Arabic (U+0660 to U+0669) and other languages (U+06F0 to +06F9)**

Each numeral in the Persian variant has a different Unicode point even if it looks identical to the Eastern Arabic numeral counterpart. However, the variants used with Urdu, Sindhi, and other Languages of South Asia are not encoded separately from the Persian variants.

Western Arabic	0	1	2	3	4	5	6	7	8	9	10
Eastern Arabic[a]	٠	١	٢	٣	٤	٥	٦	٧	٨	٩	١٠
Persian[b]	۰	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
Urdu[c]	۰	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰

Processing related challenges

- **Bidirectional Algorithm**
  - Left to Right (LTR)
  - Right to Left (RTL)
  - Processing needs to be done in two level: Character and String levels

Fig. 3. The same directional runs in a LTR context. See live

Fig. 4. The same directional runs in a RTL context. See live

Thank you

Kengatharaiyer Sarveswaran
University of Jaffna, Sri Lanka.
University of Konstanz, Germany
sarves@univ.jfn.ac.lk
sarves.github.io

Tafseer Ahmed
Senior NLP Scientist, Alexa Translations
tafseer@gmail.com