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

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.

Challenges of Processing South Asian Languages (CPSAL)

Day 02: Transliteration/Phonology and Morphology of South Asian Languages.

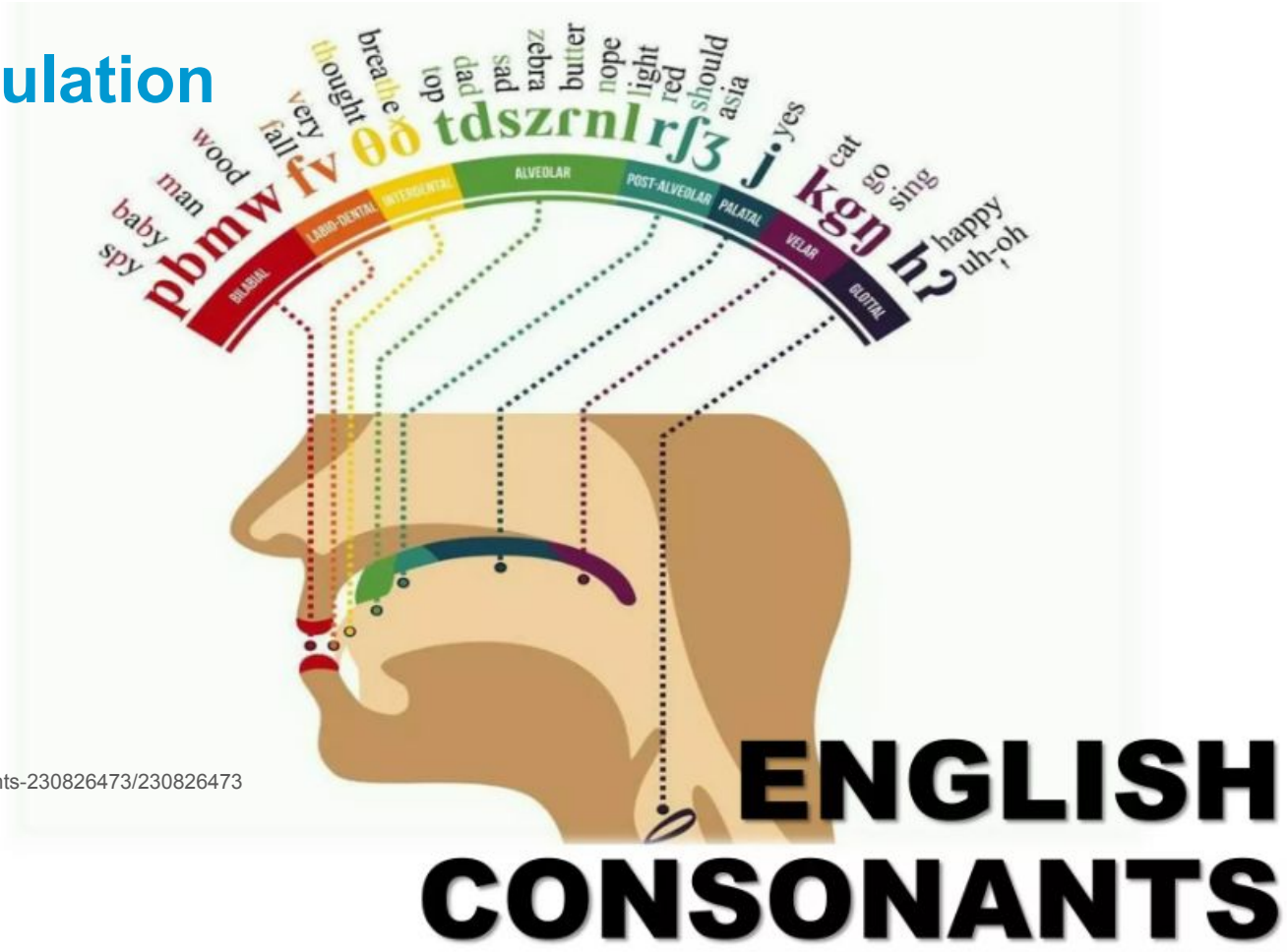
Kengatharaiyer Sarveswaran
University of Jaffna, Sri Lanka.
University of Konstanz, Germany

Tafseer Ahmed
Senior NLP Scientist
Alexa Translations

ESSLLI 2024 European Summer School of Logic, Language and Information Leuven

Position of Articulation

- Bilabial
- Labio-Dental
- InterDental
- Alveolar
- Palatal
- Velar
- Glottal



https://www.slideshare.net/slideshow/english-consonants-230826473/230826473

Position and manner of articulation

	bilabial	labio-dental	dental	alveolar	post-alveolar	palatal	velar	glottal
stop	p b			t d			k g	ʔ
nasal	m			n			ŋ	
flap				ɾ				
fricative		f v	θ ð	s z	ʃ ʒ			h
approximant				ɹ		j		
lateral approximant				l				

https://open.maricopa.edu/essentialsoflinguistics/chapter/3-3-need-to-add-tables-ipa-for-canadian-english/

Consonants in Sindhi

PLACE / MANNER		Bilabial	Labio-dental	Dental	Alveolar	Retroflex	Alveo-Palatal	Palatal	Velar	Glottal
Stop	Plosive	p b pʰ bʰ		t̪ d̪ t̪ʰ d̪ʰ		ʈ ɖ ʈ ʰ ɖʰ			k g kʰ gʰ	ʔ
	implosive	ɓ				ɗ		ʃ	ɠ	
Nasal		m			n nʰ	ɳ ɳʰ		ɲ	ŋ	
Fricative			f v		s z			ʃ	x ɣ	h
Affricate							tʃ dʒ tʃʰ dʒʰ			
Trill					r rʰ					
Flap						ɾ ɾʰ				
Lateral					l lʰ					
Approximant								j		

Table 1: Consonantal Inventory of Sindhi

https://uogenglish.wordpress.com/wp-content/uploads/2011/07/sindhi-phonemic-inventory.pdf

Tamil consonants

	Labial	Dental	Alveolar	Retroflex	(Alveolo-) palatal	Velar	Glottal
Nasal	m ம்	(n̪) ந்	n ன்	ɳ ண்	ɲ ஞ்	(ŋ) ங்	
Plosive/ Affricate	p ப்	t̪ த்	(tːr ற்ற)	ʈ ட்	t͡ɕ ~ t͡ʃ ச்[5]	k க்	
Fricative	(f)[1]		s[5] ஸ் (z)[1]	(ʂ)[1] ஷ்	(ɕ)[1] ஶ்	(x)[2]	(h)[2] ஹ்
Tap			ɾ ர்				
Trill			r ற்				
Approximant	ʋ வ்			ɻ ழ்	j ய்		
Lateral approximant			l ல்	ɭ ள்			

https://www.cambridge.org/core/services/aop-cambridge-core/content/view/5BE3F9D1C8291CDA6B3610F0E1079C02/S0025100304001549a.pdf/tamil.pdf
https://en.wikipedia.org/wiki/Tamil_phonology

Interesting/Uncommon Features - Some examples

- **Aspirated Consonants**

		Sindhi	Hindi	Urdu
unaspirated	b	ب	ब	ب
aspirated	b^h	ڀ	भ	ب ھ

- **Retroflex Consonants**

		Sindhi	Hindi	Urdu
Dental	t̪	ت	त	ت
Retroflex	ʈ	ٹ	ट	ٹ

Interesting/Uncommon Features

Implosive Consonants:

		Sindhi
Plosive	b	ب
Implosive	ɓ	ٻ

Foreign Sounds

Consonants:

Foreign	Origin	Nearest Native	Sindhi	Hindi
f	Persian/Arabic, English	pʰ	ف	फ़
x	Persian/Arabic	kʰ	خ	ख़

Vowels:

Urdu/Hindi: bɛ:l (bull), be:l (creeper plant)

People know how to pronounce the borrowed English words *bell (bɛl)*, and *cell (sɛl)* etc. However, there is no symbol to represent these in traditional orthography.

Sandhi - Example Tamil

- Internal Sandhi / External Sandhi
  - Phonological change that occurs when two morphs are concatenated together
- External Sandhi (Gemination of stops)
  - Word initial stops are geminated in Tamil when preceded by either a short vowel or a glide - this rule does not always apply even when these phonological conditions are met!
  - The phonological phenomenon is shown to be 'directly syntax-driven??
    - If word initial k, t, p is preceded by a noun in the accusative case, a noun in the dative case, a demonstrative or interrogative adjective (inta, anta,enta), an infinitive or one of the adverbs ippaDi 'like this', appaDi 'like that', eppaDi 'how', the realisation is with a voiceless plosive….

Nagarajan, H. (1995). *Gemination of stops in Tamil: implications for the phonology-syntax interface* (Vol. 7, pp. 495-509). UCL Working Papers in Linguistics.

Sandhi - A few example in Tamil

anta = ppustakattay = kkuDu
that book (acc) give imp.
Give (me) that book.
(Dem=N,Obj.NP=V)

avanukku = ppaNam kuDutteen
him (dat) money give past 1s.
(I) gave him money.
(I.O=D.O)

raamanukku = ppacikkiradu
ram (dat) hunger pres 3p.
Rama is hungry.
(Dative subj.=V)

kaar-ootti centran
kaar-driver go past 1sm
Car driver went
(kaarootti - the one who drives car)

kaar-ootti =(c) centran
Car-drive.Inf =(c) go past 1sm
(he) drove a car.

Transliteration standards

- ISO_15919: Transliteration of Devanagari and related Indic scripts into Latin characters
  - https://en.wikipedia.org/wiki/ISO_15919
- ISO 233-3:2023: Transliteration of Arabic characters into Latin characters
  - https://en.wikipedia.org/wiki/ISO_233

Transliteration standards

- Not all sounds have an orthographic character

p	paṯʉ	*pattu*	‘ten’
[b]	ṯambi	*tampi*	‘younger brother’
(b)	baḏıl	*patil*	‘answer’
ṯ	ṯapʉ	*tappu*	‘mistake’
[ḏ]	paṉḏʉ	*pantu*	‘ball’
(ḏ)	ḏinəm	*tinam*	‘day’
ʈ			
[ɖ]	ʋanɳɖi	*vanʈi*	‘cart’
k	kaːl	*kaal*	‘leg’
[g]	aŋgeː	*aŋkee*	‘there’
(g)	ganəm	*kanam*	‘heaviness’
tʃ	tʃınːə	*cinna*	‘small’

Morphology

What can you find out about morphology?

1	அவன் அழுதான்	avaṉ aḻutāṉ
	He cried	
2	அவள் அழுதாள்	avaḷ aḻutāḷ
	She Cried	
3	அவன் அழுகிறான்	avaṉ aḻukiṟāṉ
	He is crying	
4	அவள் அழுகிறாள்	avaḷ aḻukiṟāḷ
	She is crying	
5	அவன் அழுவான்	avaṉ aḻuvāṉ
	He will cry	
6	அவள் அழுவாள்	avaḷ aḻuvāḷ

7	அவர்கள் அழுதனர்	avarkaḷ aḻutaṉar
	They cried	
8	அவர்கள் அழுகின்றனர்	avarkaḷ aḻukiṉṟaṉar
	They are crying	
9	அவர்கள் அழுவார்கள்	avarkaḷ aḻuvārkaḷ
	They will cry	