Tamil

Juneki Hong

February 17, 2015

தமிழ்

According to Wikipedia

[1] 2007 edition of Nationalencyklopedin.
[2] With about 8 million as a second language.

According to Wikipedia

- Somewhere around 70-100 million native speakers[1][2].

[1] 2007 edition of Nationalencyklopedin.
[2] With about 8 million as a second language.

According to Wikipedia

- Somewhere around 70-100 million native speakers[1][2].
  - India: 60.8 million (2001 census)
  - Sri Lanka: 4.7 million

[1] 2007 edition of Nationalencyklopedin.
[2] With about 8 million as a second language.

According to Wikipedia

- Somewhere around 70-100 million native speakers[1][2].
  - India: 60.8 million (2001 census)
  - Sri Lanka: 4.7 million
- 1% of all speakers worldwide.
- 5th most spoken language in India

[1] 2007 edition of Nationalencyklopedin.
[2] With about 8 million as a second language.

According to Wikipedia

- Somewhere around 70-100 million native speakers[1][2].
  - India: 60.8 million (2001 census)
  - Sri Lanka: 4.7 million
- 1% of all speakers worldwide.
- 5th most spoken language in India
- A *classical* language that is over 2000 years old.

[1] 2007 edition of Nationalencyklopedin.
[2] With about 8 million as a second language.

Here is a Map

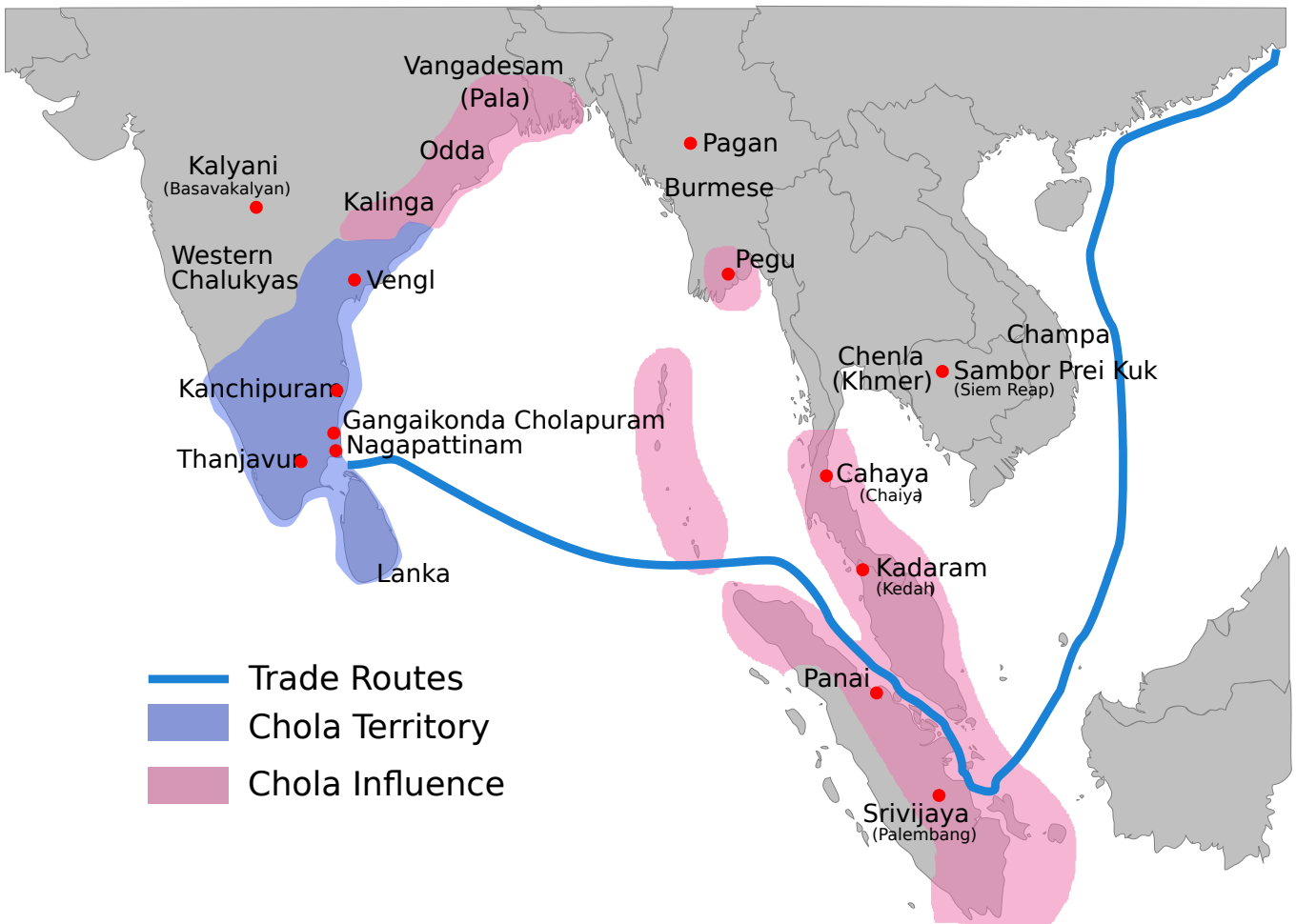

Figure : The Chola Dynasty: ~300BC - 1279CE

Sample Text[3]

மனிதப் பிறவியினர் சகலரும் சுதந்திரமாகவே பிறக்கின்றனர்; அவர்கள் மதிப்பிலும் உரிமைகளிலும் சமமானவர்கள். அவர்கள் நியாயத்தையும் மனசாட்சியையும் இயற்பண்பாகப் பெற்றவர்கள். அவர்கள் ஒருவருடனொருவர் சகோதர உணர்வுப் பாங்கில் நடந்துகொள்ளல் வேண்டும்.

Tamil Text

Maṉitap piṛaviyiṉar čakalarum čutantiramākavē piṛakkiṉṛaṉar; avarkaḷ matippilum urimaikaḷilum čamamāṉavarkaḷ. Avarkaḷ niyāyattaiyum maṉačāṭčiyaiyum iyaṛpaṇpākap peṛṛavarkaḷ. Avarkaḷ oruvaruṭaṉoruvar čakōtara uṇarvup pāṅkil naṭantukoḷḷal vēṇṭum.

Transliterated Text

All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

Translated Text

[3] www.omniglot.com/writing/tamil.htm

Tamil Script

- 12 vowels, 18 consonants, 1 special character.

Tamil Script

- 12 vowels, 18 consonants, 1 special character.
  - Script is syllabic, with 216 combinations of characters.
  - Total of 247 characters (12 + 18 + 1 + (12*18))

Tamil Script

- 12 vowels, 18 consonants, 1 special character.
  - Script is syllabic, with 216 combinations of characters.
  - Total of 247 characters (12 + 18 + 1 + (12*18))

Formation	Compound form	ISO 15919	IPA
க் + அ	க	ka	[kʌ]
க் + ஆ	கா	kā	[kɑ:]
க் + இ	கி	ki	[ki]

Syntax

- SOV
  - But with flexible word orders denoting different meanings.
- Null-Subject language.
  - A verb by itself can be a valid sentence.
  - Or the subject and object without a verb.

Parallel Corpora

- Corpus taken from `www.tamilnet.com` manually translated into English
  - 1,300 sentences
  - USC, 2001
- Joshua Decoder offers Indian parallel corpora including English-Tamil.
  - 36,252 sentences
  - JHU and University of Edinburgh, 2012
- EnTam: An English-Tamil Parallel Corpus taken from bible, cinema, and news domains.
  - 169,871 sentences
  - Charles University in Prague, 2014

MT Systems

	BLEU
Google (2011)	13.51
Post et al. 2012	9.81

JHU, University of Edinburgh [4]

	$BLEU_{suff\_sep}$ [5]
Ramasamy et al. 2012	15.12

Charles University in Prague

[4] `joshua-decoder.org/data/indian-parallel-corpora/`

[5] Reference and Hypothesis translations suffix-seperated before evaluation