Tamil

Juneki Hong

February 17, 2015
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\(^1\).  
  - 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
The Universal Declaration of Human Rights that everyone else put in their presentations

Sample Text³

Mañitap piraviyinar čakalarum čutantiramākavē pirakkiṉṟaṇar; avarkaḷ matippilum urimaikaḷilum čamamāṇavarkaḷ. Avarkaḷ niyāyattaiyum maņačāṭčiyaiyum iyarpanpākap peṟṟavarkaḷ. Avarkaḷ oruvaruṭaṇoruvar 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

³www.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 \((12 + 18 + 1 + (12 \times 18))\)
Tamil Script

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

<table>
<thead>
<tr>
<th>Formation</th>
<th>Compound form</th>
<th>ISO 15919</th>
<th>IPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ஆ + உ</td>
<td>க</td>
<td>ka</td>
<td>[kʌ]</td>
</tr>
<tr>
<td>ஆ + ஃ</td>
<td>கா</td>
<td>kā</td>
<td>[kɑː]</td>
</tr>
<tr>
<td>ஆ + இ</td>
<td>கி</td>
<td>ki</td>
<td>[ki]</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google (2011)</td>
<td>13.51</td>
</tr>
<tr>
<td>Post et al. 2012</td>
<td>9.81</td>
</tr>
</tbody>
</table>

JHU, University of Edinburgh

<table>
<thead>
<tr>
<th></th>
<th>$\text{BLEU}_{\text{suff_sep}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ramasamy et al. 2012</td>
<td>15.12</td>
</tr>
</tbody>
</table>

Charles University in Prague

---

4 joshua-decoder.org/data/indian-parallel-corpora/
5 Reference and Hypothesis translations suffix-separated before evaluation