11-737 Multilingual NLP

Lang in 10: Hindi

Example of 10 minute presentation on a language
Hindi हिन्दी

- Hindi spoken in Northern India
  - 320mn native speakers
  - 270 L2 speakers
  - 3rd or 4th most spoken language in world (after Mandarin, English, Spanish)
- Lingua franca for Northern India
  - Taught in schools throughout India
- Hindi != Urdu
  - But mutually intelligible
  - Writing script is different (but writing style is different too)
  - Hindi more Sanskrit words, Urdu more Persian/Arabic words
  - Sometimes both are called “Hindustani”
    - Share phonology, grammar
Hindi

- Indo-European →
  - Indo-Iranian → Indo-Aryan → Western Hindi → Hindustani → Hindi
- Has common cognates with European languages
  - महा रजा (maha raja) → magnus royal (great king)
  - Numbers ek do teen chaar paanch
- Script is Devnagari देवनागरी
  - Brahmi script (as many Indian subcontinent languages are)
  - Also used for nearby languages (e.g. Marathi, Nepali)
  - Urdu uses Perso-Arabic (Nasta’liq)
  - Often in social media contexts Hindi is written in latin script
Grammar

- Default order: S O V
- But often pro-dropped, subject
- Gender for all nouns
- Inflection morphology for agreement
- Ergative marking
  - In transitive verb (sometimes) agreement is object/verb not subj/verb
Phonology

• Vowels: schwa, vowel length (plus borrowed vowel ae from English)

From Ohala 1999 via Wikipedia
# Phonology – consonants

<table>
<thead>
<tr>
<th></th>
<th>Labial</th>
<th>Dental/Alveolar</th>
<th>Retroflex</th>
<th>Palatal</th>
<th>Velar</th>
<th>Uvular</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nasal</strong></td>
<td>m</td>
<td>n</td>
<td>η</td>
<td>(ŋ)</td>
<td>η</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Plosive/Affricate</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>voiceless</td>
<td>p</td>
<td>t</td>
<td>tʃ</td>
<td>k</td>
<td>(q)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>voiceless aspirated</td>
<td>pʰ</td>
<td>tʰ</td>
<td>tʰ</td>
<td>kʰ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>voiced</td>
<td>b</td>
<td>d</td>
<td>dʒ</td>
<td>dʒʰ</td>
<td>g</td>
<td></td>
<td></td>
</tr>
<tr>
<td>voiced aspirated</td>
<td>bʰ</td>
<td>dʰ</td>
<td>dʰ</td>
<td>dʒʰ</td>
<td>gʰ</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Flap and trill</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>plain</td>
<td>r</td>
<td>l</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>voiced aspirated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Fricative</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>voiceless</td>
<td>f</td>
<td>s</td>
<td>ʃ</td>
<td>(x)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>voiced</td>
<td>v</td>
<td>z</td>
<td>(ʒ)</td>
<td>(ɣ)</td>
<td>ɦ</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Approximant</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>l</td>
<td>j</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From wikipedia: Hindi Phonology
Hindi vs Hinglish

- Most educated Hindi speakers are fluent in English
- Code-switching: mixing of two languages in one utterance
  - Common in multilingual environments
  - Typically more common in casual speech (online text)
  - Borrows phonology, morphology, grammar from both languages
- In last 50 years more and more English borrowing in Hindi
  - Hinglish may be the new Hindi.
Lang in 10

- History, geography, social position
- Linguistic: morphology, grammar, phonology
- Examples of something (linguistically) interesting about the language
- Status with respect to resources
- Influences, social use, issues that may affect collection/access