Many a little makes a mickle
Infrastructure component reuse for a massively multilingual linguistic study

Lars Borin, Shafqat Mumtaz Virk, Anju Saxena
- U. of Gothenburg & Uppsala U. / Sweden
lars.borin@svenska.gu.se

CLARIN 2017, Budapest
19th September 2017
language in South Asia

- **South Asia** (SA), i.e., the seven countries
  1. Pakistan
  2. India
  3. Nepal
  4. Bhutan
  5. Bangladesh
  6. Sri Lanka
  7. The Maldives

- (and adjacent parts of neighboring countries)

- is the home of ~600 languages (according to the *Ethnologue*) belonging to four major language families
  1. Indo-Aryan (<Indo-European) – **IA**
  2. Tibeto-Burman (<Sino-Tibetan) – **TB**
  3. Dravidian – **DR**
  4. Austroasiatic (>Munda, Mon-Khmer) – **AA**

- (+ some small families and language isolates)
large-scale comparative linguistics

• the traditional narrower sense:
  • (historical-)comparative linguistics
    • finding out how (if) language varieties are related through descent from a common ancestor (a proto-language)

• the wider sense of this project:
  • the investigation of similarities among language varieties in order to find out about their causes
    • common ancestry
    • language contact (borrowing, areal linguistics)
    • structural-typological tendencies
    • some combination of the above
similarities among languages

- common ancestry:
  - IA: Hindi $dō$, Assamese $dui$, Marathi $dōn$, Poguli $dīh$ ‘two’
  - TB: Kinnauri $niš$, Tibetan $gñis$, Bodo $nè$, Limbu nech’ī ‘two’
  - DR: Tamil $iraṇḍu$, Telugu $reṇḍu$, Gondi $raṇḍ$, Kurukh $eṇḍ$ ‘two’

- language contact/areality
  - retroflex consonants
    (80/20 in SA vs. 20/80 in the world)
  - dative experiencer/subject

- structural-typological tendencies
  - OV constituent order ↔ postpositions
Linguistic Survey of India (LSI) – 1903–1927

Sir George Abraham Grierson

Sten Konow
infrastructure reuse: from the LSI . . .

- Grierson’s (and Konow’s) *Linguistic Survey of India* (1903–1927) remains the most complete source on SA lgs
- 19 tomes (9500 pages) w. 723 linguistic varieties (two tomes not used in our project)
- Comparable lexical and grammatical information on 267 varieties (141 TB; 95 IA; 18 DR; 13 AA)
- Most of the tomes now digitized (using double keying)
- This is **big data**, (approximately) in the sense of

  “data that is too diverse (…) for conventional technologies, skills and infra-structure to address efficiently”

  <www.mongodb.com/big-data-explained>

  or

  “data that was previously ignored because of technology limitations” (Matt Aslett)
to linguistic database

feature questionnaire

1. Does this language have retroflex consonant(s) as phoneme(s)?
2. Does this language mark plural for nominals?
3. Does the language have an ergative absolutive case (all cases and/or postpostional)?
4. Does the same adnominal case marker function both as the ergative and as the instrumental?
5. Does the language mark dual in at least one pronoun?
6. Is the third person pronoun the same as one of the demonstratives?
7. Is the first person plural pronoun also used for the first person singular relative as a marker?
8. Are there reflexive pronouns?

grammatical features

<table>
<thead>
<tr>
<th>Language</th>
<th>Latitude</th>
<th>Longitude</th>
<th>&lt;ADJ&gt;</th>
<th>&lt;NUM&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>CENTRAL_PAHARI_KUMAL</td>
<td>29.6</td>
<td>79.7</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>EASTERN_PAHARI_OR_KH</td>
<td>27.42</td>
<td>85.2</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>SIRMAURI</td>
<td>31.1</td>
<td>77.17</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>KAGATE</td>
<td>27.3</td>
<td>88.16</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>SHARPA</td>
<td>27.3</td>
<td>88.16</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>BAGHATI</td>
<td>30.7972</td>
<td>76.9172</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>KUHTALI</td>
<td>31.0088</td>
<td>77.5317</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>BHADRAWAHI</td>
<td>32.98</td>
<td>75.72</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>CHAMEALI</td>
<td>32.3</td>
<td>76.07</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>GUJURU_OF_HAZARA</td>
<td>34.3768</td>
<td>73.1461</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>WESTERN_PAHARI_JAUN</td>
<td>30.5587</td>
<td>77.9084</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>KULUI</td>
<td>31.58</td>
<td>77.06</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>MANDALI</td>
<td>31.42</td>
<td>76.55</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>PANGWALI</td>
<td>31.5832</td>
<td>78.278275</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>YAKHA</td>
<td>27.3</td>
<td>87.32</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>KANUARI</td>
<td>31.35</td>
<td>78.29</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>KANASHI</td>
<td>32.0626008</td>
<td>77.26039548</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>RAI_OR_JIMDAR</td>
<td>26.8794</td>
<td>87.3296</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>MAGARI</td>
<td>27.4135</td>
<td>87.0617</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>NEWARI</td>
<td>27.2614</td>
<td>84.9577</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>KHAMBURI</td>
<td>27.1606</td>
<td>87.5717</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>SUNWAR</td>
<td>27.5201</td>
<td>86.2476</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>ABOR</td>
<td>27.6281</td>
<td>94.3538</td>
<td>yes</td>
<td></td>
</tr>
</tbody>
</table>
Articles.—There are no articles, but i, the shortest form of the first numeral, is often used as a kind of indefinite article; thus, i marshang-ka-di, with a man.

Nouns.—Gender is distinguished in the common way, by using different words or adding terms denoting the sex; thus, marshang, man; betri, woman; chime, daughter; räng, horse; mich räng, mare; kui, dog; mich kuit, bitch.

Number.—There are two numbers, the singular and the plural. The latter is not necessarily marked, when it appears from the context; thus, shum rhad, three bulls. There is, however, a separate plural suffix ga, which usually takes the fuller form gan before suffixes; thus, nyish bā-ga, two fathers; bā-gan-ka, of fathers; chanditsō marshang-gan-dits, from good men.

Case.—The subject of intransitive verbs is not distinguished by means of any suffix. The final i in du-i tot-kō, he was; duga-i tot-ke, they were, is probably an emphasizing particle.

The subject of transitive verbs is usually distinguished by means of a suffix sh or s; thus, bā-sh tang-mo, father-by saw, the father saw; jang-s sanemuk, God-by slew, the God slew. The two first personal pronouns do not appear to possess any such case.

The object is often distinguished by adding a p; thus, laṭa-phāṭa-p, property; sūra-p rwan-m, swine-to feed; ba-p lon-mo, father-to said.
existing own infrastructure components for “rapid prototyping” of exploration tools . . .
South Asia as a Linguistic Area
... and external standard solutions
current project status

- manual extraction of selected features ongoing
- IE experiments being conducted (also on other, non-LSI, grammars)
- some modifications of our infrastructure are being implemented to accommodate the needs of the project (with a view to their wider usefulness)

- we wish to acknowledge the support of
  - The Swedish Research Council (main funder)
  - University of Gothenburg
  - Swe-Clarin
Thank you!

(Photo: Boyd Michalovsky)