



GÖTEBORGS  
UNIVERSITET

**Språk**  
**BANKEN**



UPPSALA  
UNIVERSITET

# 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](mailto:lars.borin@svenska.gu.se)

CLARIN 2017, Budapest

19th September 2017



**SWE-CLARIN**

# 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)

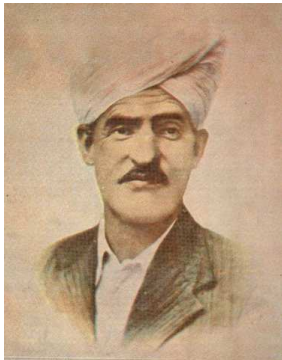
# large-scale comparative linguistics

- ▶ 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

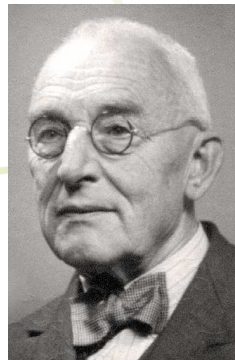
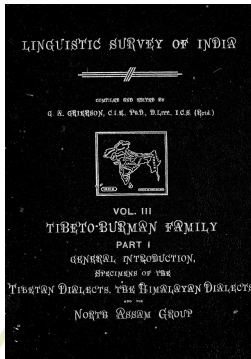
# similarities among languages

- ▶ common ancestry:
  - ▶ IA: Hindi *dō*, Assamese *dui*, Marathi *dōn*, Poguli *dīh* 'two'
  - ▶ TB: Kinnauri *niš*, Tibetan *gñis*, Bodo *nè*, Limbu *nech* 'two'
  - ▶ DR: Tamil *iraṇḍaᵘ*, Telugu *reṇḍaᵘ*, 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](http://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?
Does the language have an ergative adnominal case (affixes and/or postpositions)?
Does the same adnominal case marker function both as the ergative and as the instrumental?
4
5 Does the language mark dual in at least one pronoun?
6 Is the third person pronoun the same as one of the demonstratives?
Is the first person plural pronoun also used for the first person singular referent as a marker?
7
8 Are there reflexive pronouns?

## grammatical features

Language	latitude	longitude	<ADJ>	<NUM>
CENTRAL_PAHARI_KUMAU	29,6	79,7	yes	yes
EASTERN_PAHARI_OR_KH	27,42	85,2	yes	yes
SIRMAURI	31,1	77,17	yes	yes
KAGATE	27,3	88,16	yes	yes
SHARPA	27,3	88,16	yes	yes
BAGHATI	30,7972	76,9172	yes	yes
KIUTHALI	31,0088	77,5317	yes	yes
BHADRAWAHI	32,98	75,72	yes	yes
CHAMEALI	32,33	76,07	yes	yes
GUJURI_OF_HAZARA	34,3768	73,1461	yes	yes
WESTERN_PAHARI_JAUN	30,5587	77,9084	yes	yes
KULUI	31,58	77,06	yes	yes
MANDEALI	31,42	76,55	yes	yes
PANGWALI	31,58932	78,278275	yes	yes
YAKHA	27,3	87,32	yes	yes
KANAURI	31,35	78,25	yes	yes
KANASHI	32,0626008	77,2603548	yes	yes
RAI_OR_JMDAR	26,8794	87,3296	yes	yes
MAGARI	27,4135	87,0617	yes	yes
NEWARI	27,2614	84,9577	yes	yes
KHAMBU	27,1606	87,5717	yes	yes
SUNWAR	27,5201	86,2476	yes	yes
ABOR	27,6281	94,3538	yes	yes

## ... using (information extraction from) the LSI grammar sketches

**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 maṛshang-ka-di*, with a man.

**Nouns.**—**Gender** is distinguished in the common way, by using different words or adding terms denoting the sex; thus, *maṛshang*, man; *beṭrī*, woman: *chho*, son; *chāme*, daughter; *rāng*, horse; *māch rāng*, mare; *kui*, dog; *māch kutī*, 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; *chanditso maṛshang-gan-dits*, from good men.

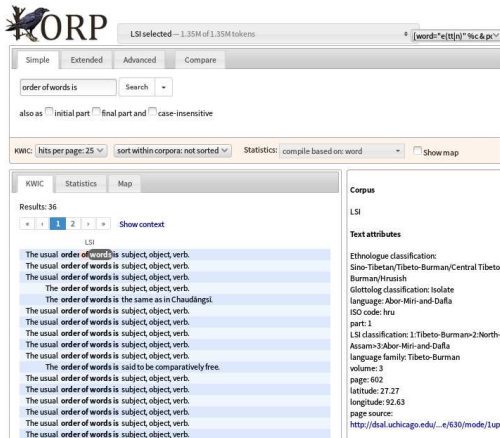
**Case.**—The subject of intransitive verbs is not distinguished by means of any suffix. The final *i* in *du-i tot-kō*, he was; *duga-i tot-ke*, they were, is probably an emphasizing particle.

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-phata-p*, property; *sūra-p rwang-m*, swine-to feed; *ba-p lon-mo*, father-to said.



# existing own infrastructure components for “rapid prototyping” of exploration tools ...



LSI Explorer interface showing search results for the query "order of words". The interface includes a search bar, filters, and a list of results.

LSI selected — 1.35M of 1.35M tokens

Simple Extended Advanced Compare

order of words is Search

also as ☐ initial part ☐ final part and ☐ case-insensitive

KWIC: hits per page: 25 sort within corpora: not sorted Statistics: compile based on: word Show map

Results: 36

1 2 3 Show context

LSI

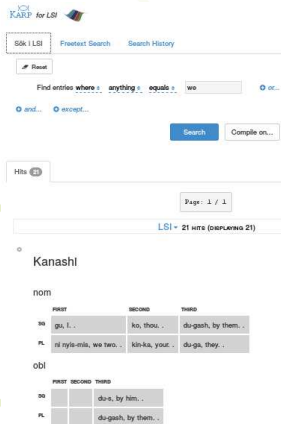
The usual order of words is subject, object, verb.  
 The usual order of words is subject, object, verb.  
 The usual order of words is subject, object, verb.  
 The order of words is subject, object, verb.  
 The order of words is the same as in Chaudāngsi.  
 The usual order of words is subject, object, verb.  
 The usual order of words is subject, object, verb.  
 The usual order of words is subject, object, verb.  
 The usual order of words is subject, object, verb.  
 The usual order of words is subject, object, verb.  
 The order of words is said to be comparatively free.  
 The usual order of words is subject, object, verb.  
 The usual order of words is subject, object, verb.  
 The usual order of words is subject, object, verb.  
 The usual order of words is subject, object, verb.  
 The usual order of words is subject, object, verb.  
 The usual order of words is subject, object, verb.

Corpus

LSI

Text attributes

Ethnologue classification:  
 Sino-Tibetan/Tibeto-Burman/Central Tibeto-Burman/Hrusish  
 Glottolog classification: Isolate  
 language: Abor-Miri-and-Dafla  
 ISO code: hru  
 part: 1  
 LSI classification: 1:Tibeto-Burman-2:North-Assam-3:Abor-Miri-and-Dafla  
 language family: Tibeto-Burman  
 volume: 3  
 page: 602  
 latitude: 27.27  
 longitude: 92.63  
 page source:  
<http://dsal.uchicago.edu/...e/630/mode/1up>



LSI Explorer interface showing search results for the query "Kanashi". The interface includes a search bar, filters, and a list of results.

LSI Explorer for LSI

LSI Freetext Search Search History

Raw

Find entries where: anything equals we

and... except...

Search Compile on...

Hits 21

Page: 1 / 1

LSI - 21 hits (displaying 21)

Kanashi

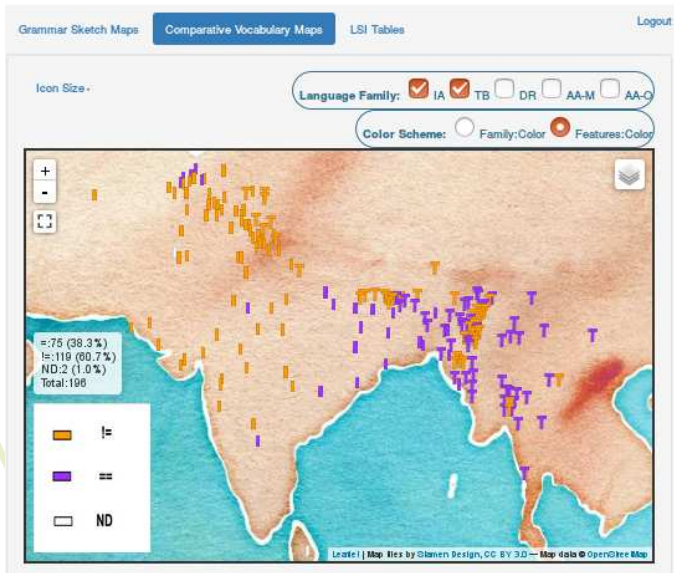
nom

	FIRST	SECOND	THIRD
SG	gu, I. .	ko, thou. .	du-gash, by them. .
PL	ni nyis-mis, we two. .	kin-ka, you. .	du-ga, they. .

obl

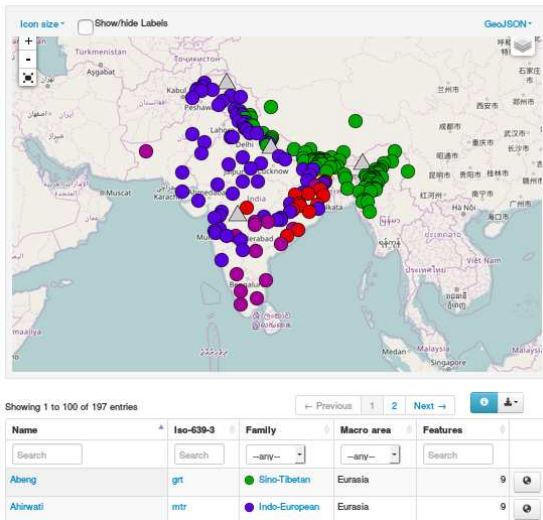
	FIRST	SECOND	THIRD
SG		du-a, by him. .	
PL		du-gash, by them. .	

# South Asia as a Linguistic Area



# ... and external standard solutions

## Languages



## 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 accomodate 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-Clarín

Thank you!



(Photo: Boyd Michalovsky)