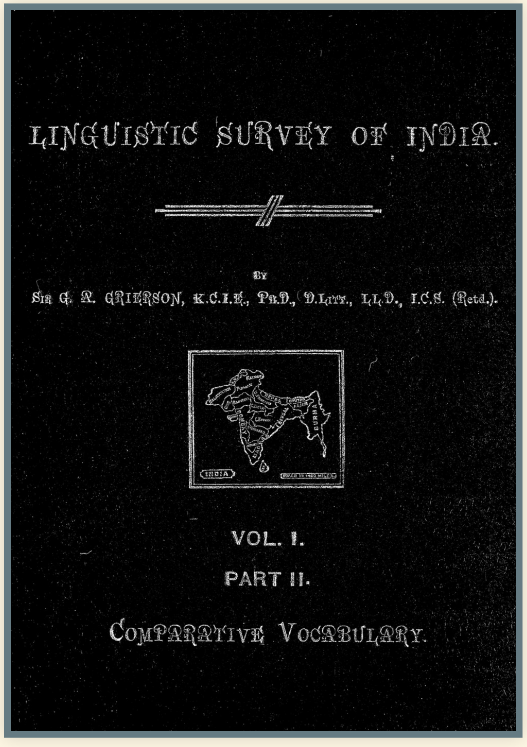
Linguistic Survey of India and Polyglotta Africana: Two Retrostandardized Digital Editions of Large Historical Collections of Multilingual Wordlists

Robert Forkel, Johann-Mattis List, Christoph Rzymski, Guillaume Segerer

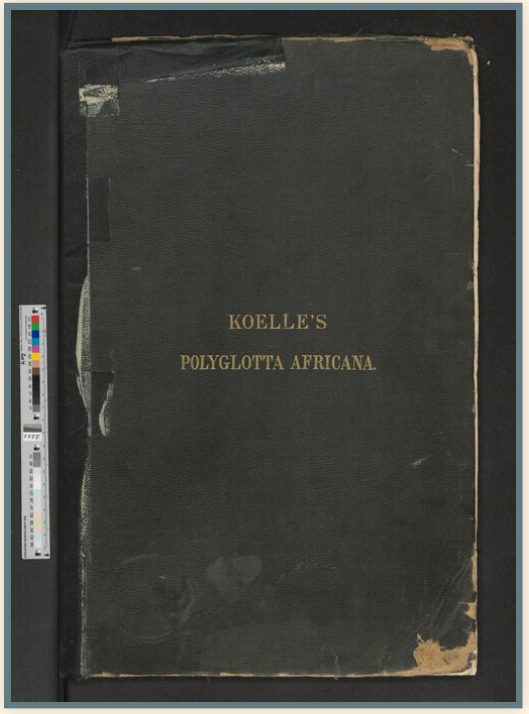
Historical Data for Historical Linguistics

- comparative wordlist have been collected for more than 230 years
- importance of comparative lexical data has increased with automated analysis methods like cognate detection or phylogenetic methods
- automated re-use requires standardized data

The Linguistic Survey of India



Polyglotta Africana



CLDF - an interoperable data format

- tabular data in CSV files
- typed data thanks to schema description according to CSVW
- domain-specific semantics via an ontology
- shared domain data via reference catalogs

Language Identification

- Glottolog - the reference catalog for languoids - is referenced via
- Glottocodes specified using the CLDF `glottocode` property in a
- CLDF `LanguageTable` component.

Concept Mapping

- Concepticon - the reference catalog for concepts - is referenced via
- conceptset IDs specified using the CLDF `concepticonReference` property in a
- CLDF `ParameterTable` component.

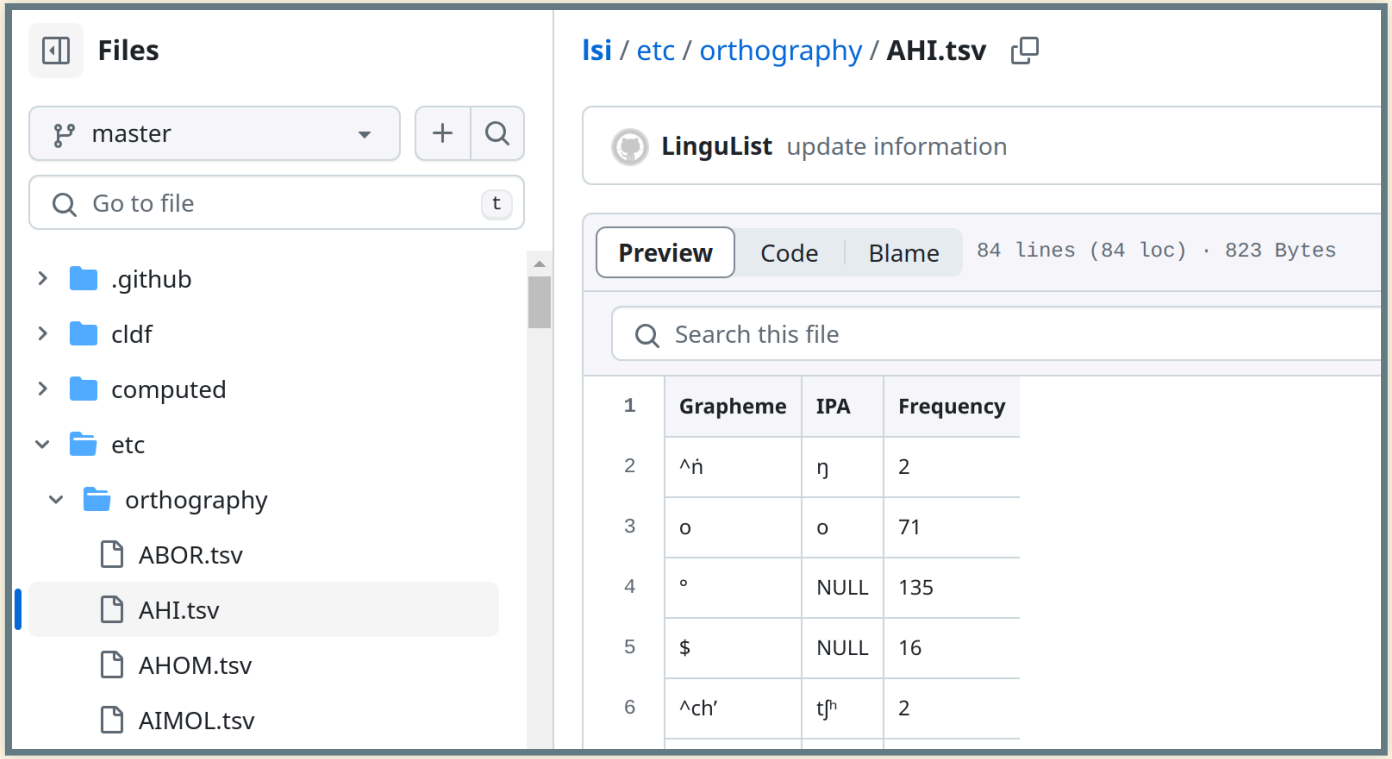
Orthography Conversion

- Original orthography is converted via orthography profiles to
- segmented phonetic transcriptions linked to
- CLTS - the reference catalog for transcription systems.

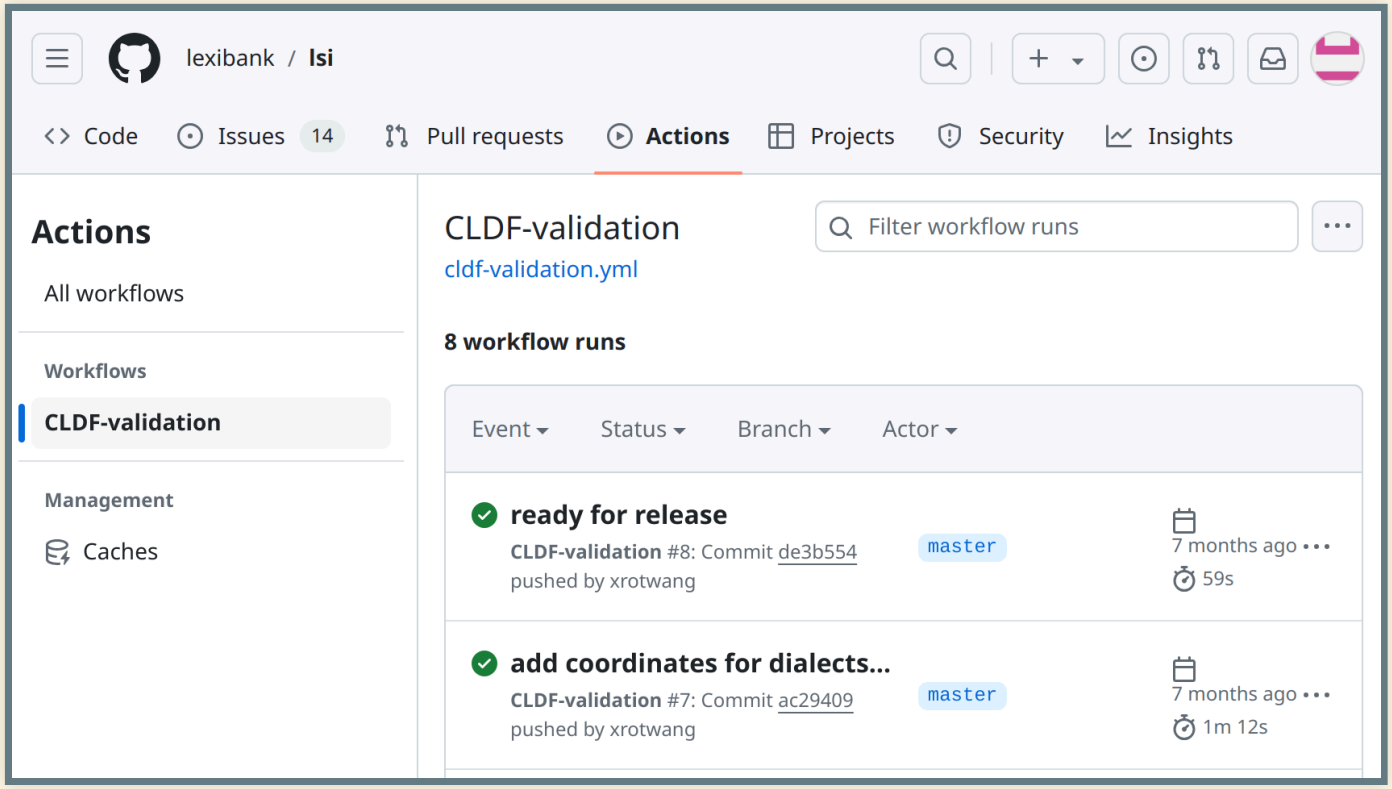
CLDF Creation

- “Raw” digital data is fed into
- `cldfbench` where it is enriched with
    - language and concept mappings and segmented via
    - orthography profiles
- resulting in CLDF datasets.

Orthograhy Profiles



Validation


lexibank / lsi
Code
Issues 14
Pull requests
Actions
Projects
Security
Insights
Actions
All workflows
Workflows
CLDF-validation
Management
Caches
CLDF-validation
cldf-validation.yml
Filter workflow runs
8 workflow runs
Event
Status
Branch
Actor
ready for release
CLDF-validation #8: Commit de3b554
pushed by xrotwang
master
7 months ago
59s
add coordinates for dialects...
CLDF-validation #7: Commit ac29409
pushed by xrotwang
master
7 months ago
1m 12s

Visualization: HTML

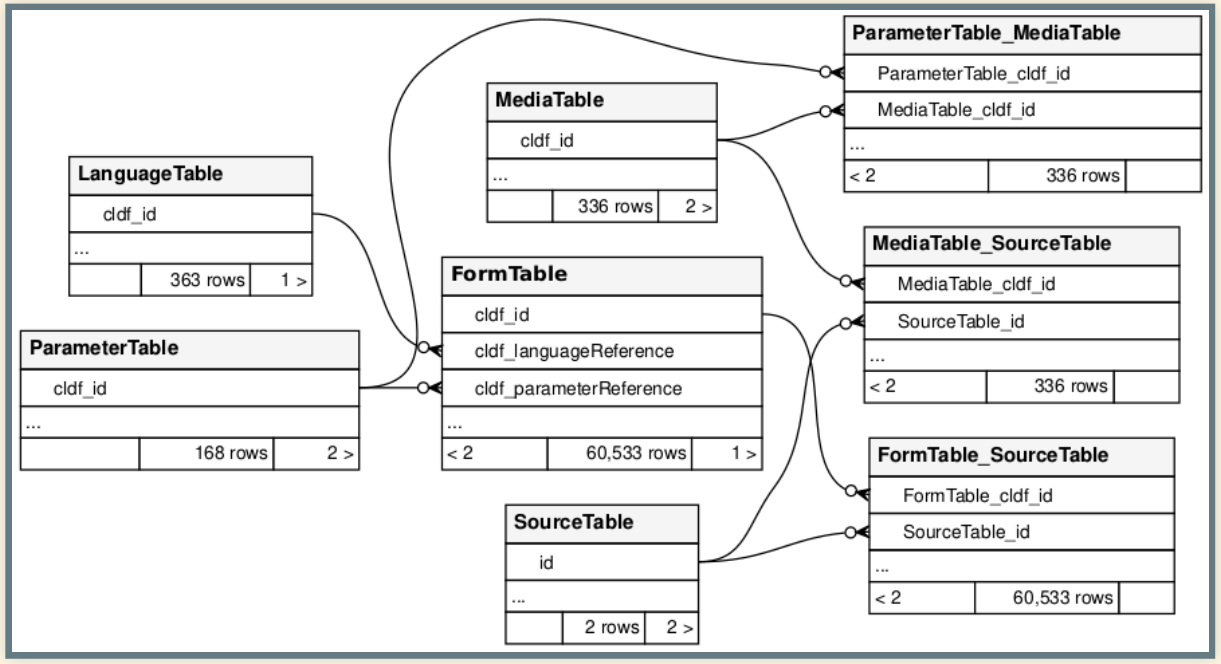
Wordlist CLDF dataset derived from Grierson's "Linguistic Survey of India" from 1928

CLDF Metadata: cldf-metadata.json

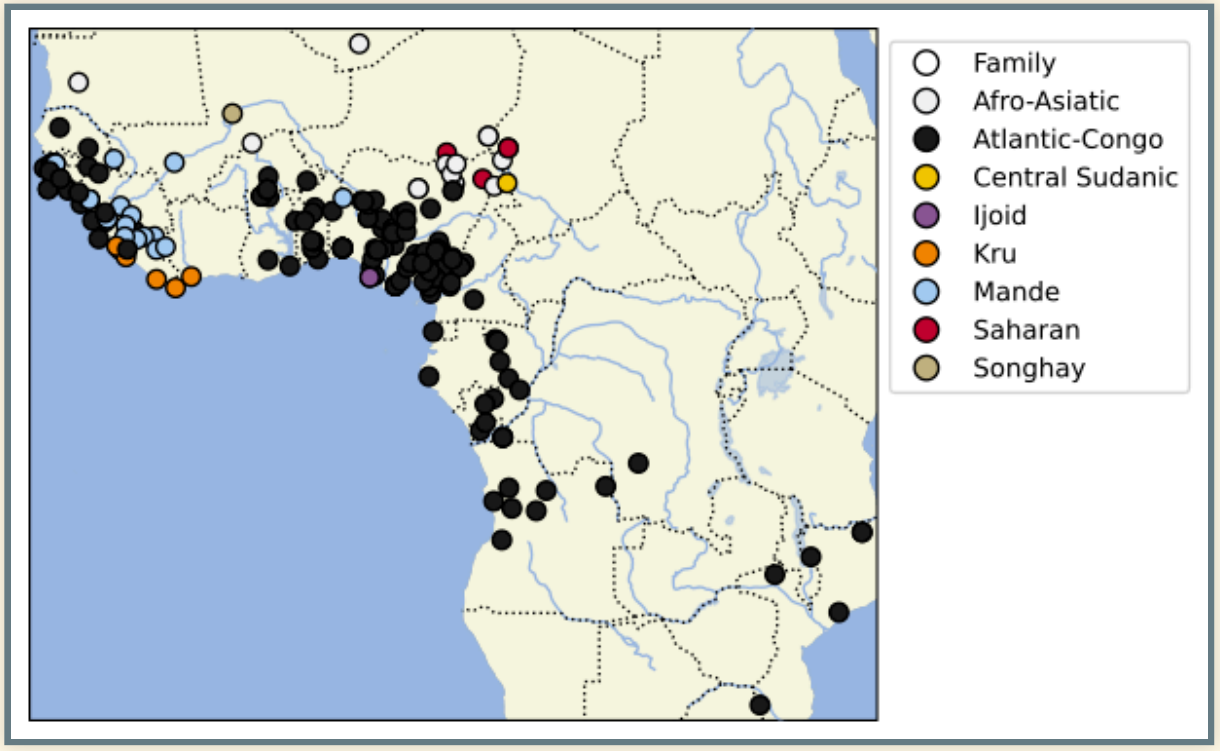
Sources: sources.bib

property	value
dc:bibliographicCitation	Grierson, George Abraham (1928): Linguistic Survey of India. Comparative Vocabulary. Calcutta: Government of India Central Publication Branch.
dc:conformsTo	CLDF Wordlist
dc:format	1. http://concepticon.clld.org/contributions/Grierson-1928-168
dc:identifier	https://lsi.clld.org
dc:license	https://creativecommons.org/licenses/by/4.0/
dcat:accessURL	https://github.com/lexibank/lsi
prov:wasDerivedFrom	1. lexibank/lsi ac29409 2. Glottolog v4.8 3. Concepticon v3.1.0 4. CLTS v2.2.0

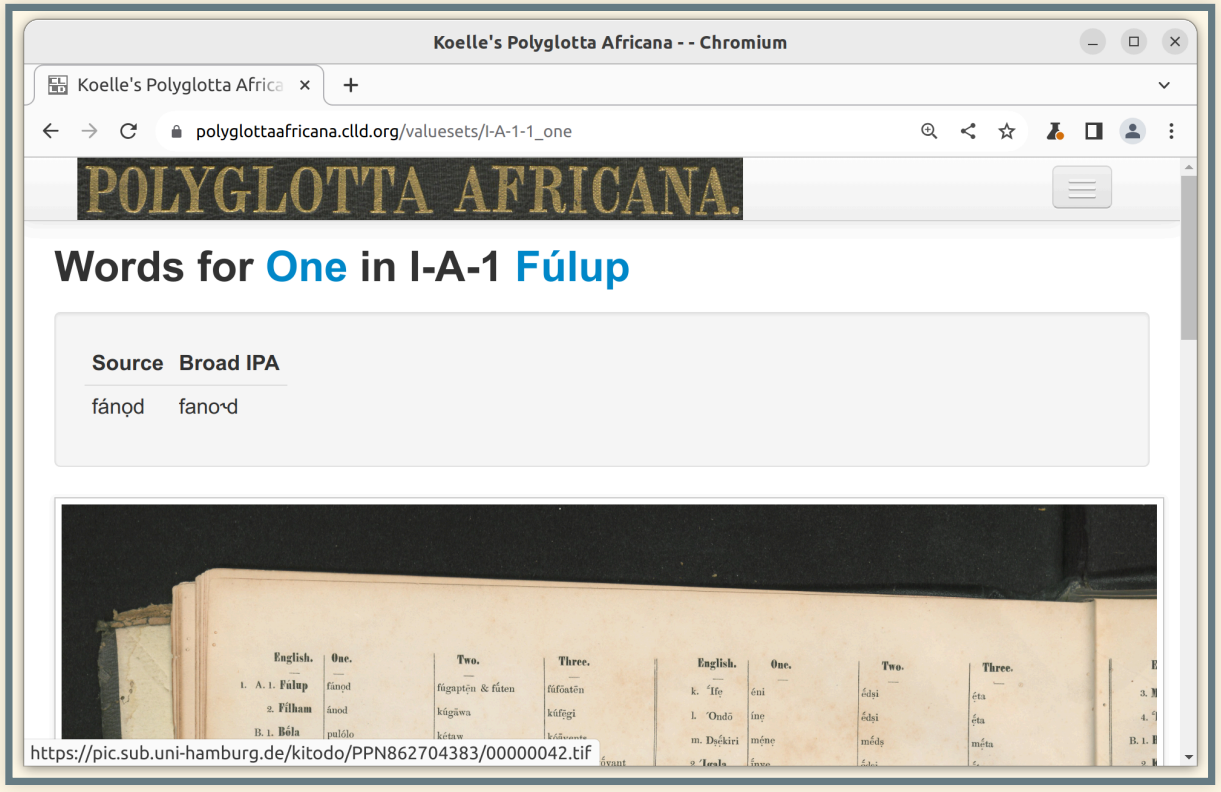
Visualization: ERD



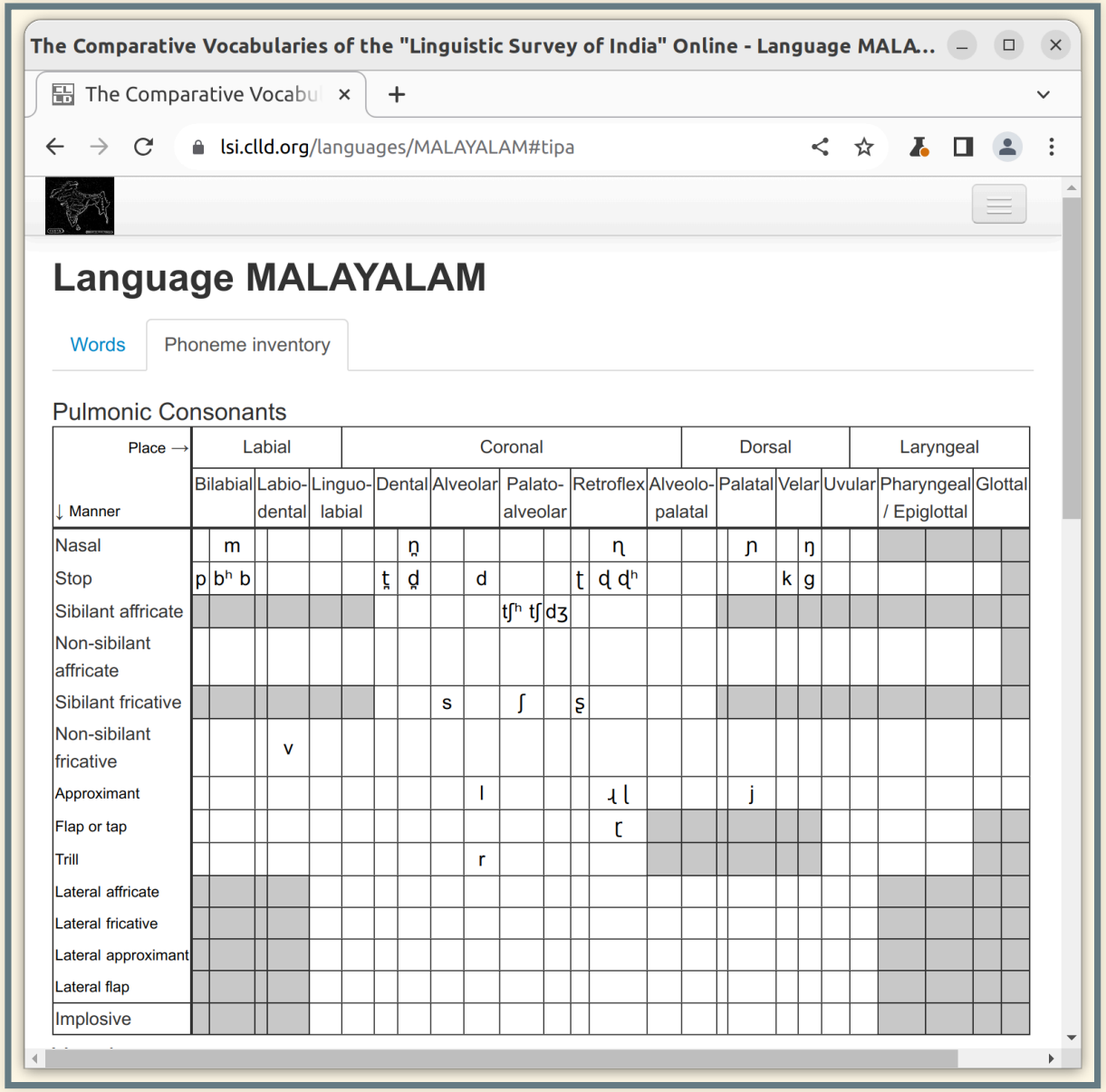
Visualization: Coverage map



Visualization: clld



Visualization: IPA chart



Place → / ↓ Manner	Labial			Coronal						Dorsal				Laryngeal		
	Bilabial	Labio-dental	Linguo-labial	Dental	Alveolar	Palato-alveolar	Retroflex	Alveolo-palatal	Palatal	Velar	Uvular	Pharyngeal / Epiglottal	Glottal			
Nasal	m			n̪			ɳ		ɲ	ŋ						
Stop	p bʰ b			t̪ d̪		d		ʈ	ɖ ɖʰ			k g				
Sibilant affricate							tʃʰ tʃ	dʒ								
Non-sibilant affricate																
Sibilant fricative					s		ʃ		ʂ							
Non-sibilant fricative		v														
Approximant						l			ɻ l			j				
Flap or tap								ɾ								
Trill						r										
Lateral affricate																
Lateral fricative																
Lateral approximant																
Lateral flap																
Implosive																

Aggregation: CLDF SQL

```
ATTACH DATABASE "phoible.sqlite" AS phoible;
ATTACH DATABASE "clts.sqlite" AS clts;
ATTACH DATABASE "lsi.sqlite" AS lsi;
```

```
CREATE TEMP VIEW lsigraphemes AS
SELECT
  DISTINCT grapheme
FROM
(
  WITH split(grapheme, segments) AS (
    SELECT
      '',
      f.cldf_segments || ' '
      FROM lsi.formtable AS f, lsi.languagetable AS l
      WHERE f.cldf_languagereference = l.cldf_id AND l.cldf_glottocode = 'mala1464'
    UNION ALL SELECT
      substr(segments, 0, instr(segments, ' ')),
      substr(segments, instr(segments, ' ') + 1)
    FROM split
    WHERE segments != ''
  ) SELECT grapheme FROM split
  WHERE grapheme!=''
);
```

```
25  CREATE TEMP VIEW phoiblegraphemes AS
26  SELECT
27    DISTINCT c.cltsgrapheme AS grapheme
28  FROM
29  (
30    SELECT v.cldf_value AS grapheme
31    FROM phoible.valuetable AS v
32    WHERE cldf_languagereference = 'mala1464' AND contribution_id = 1762
33  ) AS p
34  JOIN
35  (
36    SELECT phoible.grapheme AS phoiblegrapheme, clts.grapheme AS cltsgrapheme
37    FROM clts."data/graphemes.tsv" AS phoible, clts."data/sounds.tsv" AS clts
38    WHERE phoible.dataset = 'phoible' AND phoible.name = clts.name
39  ) AS c
40  ON c.phoiblegrapheme = p.grapheme;
```

```
42  SELECT lsi.grapheme, 'LSI', clts.name
43  FROM lsigraphemes AS lsi
44  JOIN
45    clts."data/sounds.tsv" AS clts
46  ON clts.grapheme = lsi.grapheme
47  WHERE clts.name LIKE '%vowel' AND lsi.grapheme NOT IN phoiblegraphemes;
48
49  SELECT phoible.grapheme, 'PHOIBLE', clts.name
50  FROM phoiblegraphemes AS phoible
51  JOIN
52      clts."data/sounds.tsv" AS clts
53  ON clts.grapheme = phoible.grapheme
54  WHERE clts.name LIKE '%vowel' AND phoible.grapheme NOT IN lsigraphemes;
```

Glyph	Dataset	CLTS
ʌ	LSI	unrounded_open_mid_back_vowe
ʌː	LSI	long_unrounded_open_mid_back_
a	PHOIBLE	unrounded_open_front_vowel
aː	PHOIBLE	long_unrounded_open_front_vowe
æ	PHOIBLE	unrounded_near_open_front_vowe
ɨ	PHOIBLE	unrounded_close_central_vowel
ʊ	PHOIBLE	rounded_near_close_near_back_v

Analysis

- LingPy can read CLDF and compute cognates or alignments
- BEASTling can read cognate coded CLDF to compute phylogenetic trees

Conclusion

- We have established a workflow to retro-digitize legacy wordlists.
- Thereby providing another angle from which to attack the data scarcity problem in computational linguistics.