Publishing linguistic typological data at the University of Helsinki

HELDIG Summit 2019: From Text to Knowledge

Kaius Sinnemäki

7 November 2019

UNIVERSITY OF HELSINKI

European Research Council
Established by the European Commission

Linguistic diversity

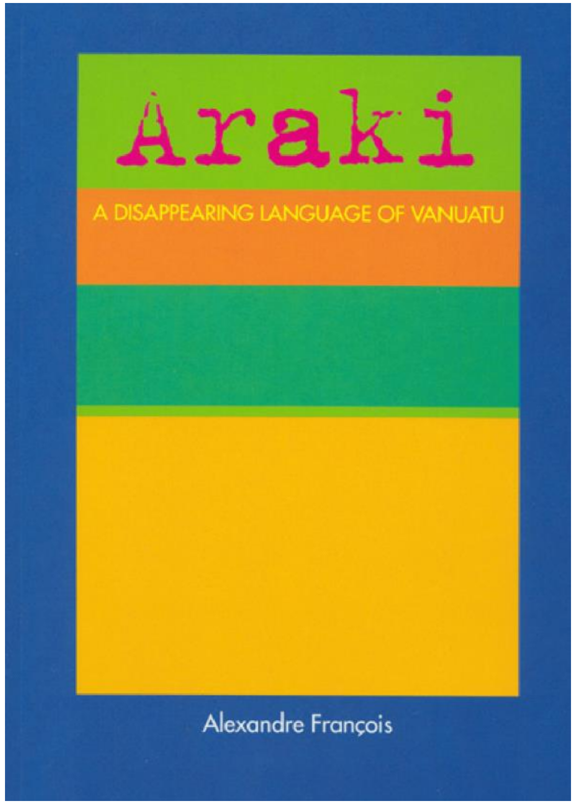
- = about understanding the nature of human language and the systematicities in its variation across different languages.

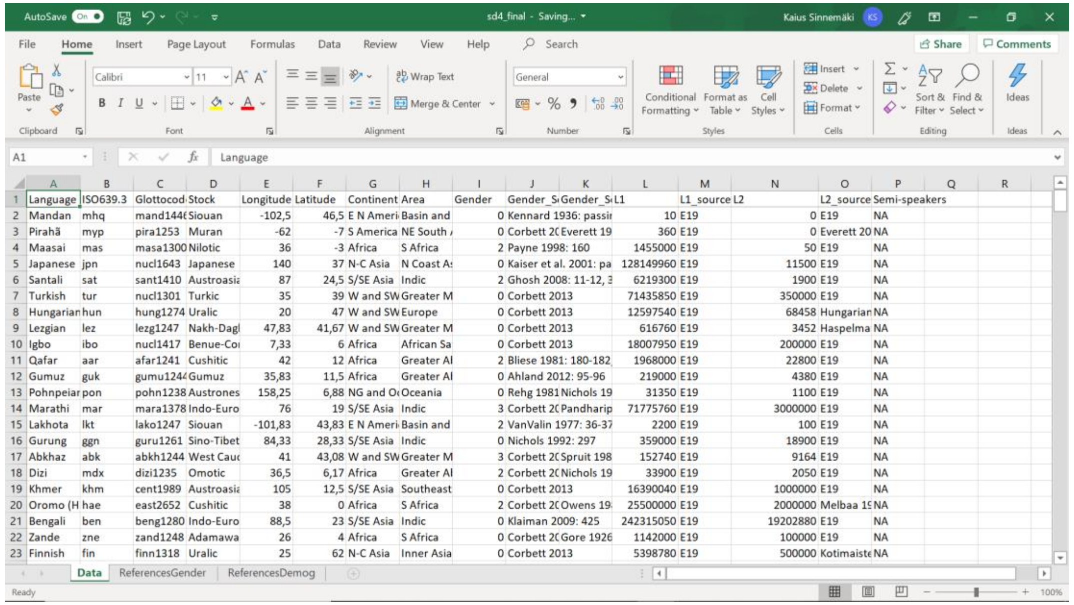
- One of the foci in the humanities research at UH.
  - Long traditions in historical-comparative linguistics.
  - *Helsinki Area and Language Studies* –network.
  - MA programme *Linguistic Diversity and Digital Humanities* (2020-).

  - ERC-funded project *Linguistic Adaptation: Typological and Sociolinguistic Perspectives to Language Variation* (2019–2023). https://www.helsinki.fi/en/researchgroups/linguistic-adaptation

Nature of cross-linguistic data

- From linguistic fieldwork to language description to data matrices.





Language	ISO639.3	Glottocod	Stock	Longitude	Latitude	Continent	Area	Gender	Gender_S	Gender_S	L1	L1_source	L2	L2_source	Semi-speakers
Mandan	mhq	mand1446	Siouan	-102,5	46,5	E N Ameri	Basin and	0	Kennard 1936: passi		10	E19	0	E19	NA
Pirahã	myp	pira1253	Muran	-62	-7	S America	NE South	0	Corbett 20	Everett 19	360	E19	0	Everett 20	NA
Maasai	mas	masa1300	Nilotic	36	-3	Africa	S Africa	2	Payne 1998: 160		1455000	E19	50	E19	NA
Japanese	jpn	nucl1643	Japanese	140	37	N-C Asia	N Coast As	0	Kaiser et al. 2001: pa		128149960	E19	11500	E19	NA
Santali	sat	sant1410	Austroasia	87	24,5	S/SE Asia	Indic	2	Ghosh 2008: 11-12, 3		6219300	E19	1900	E19	NA
Turkish	tur	nucl1301	Turkic	35	39	W and SW	Greater M	0	Corbett 2013		71435850	E19	350000	E19	NA
Hungarian	hun	hung1274	Uralic	20	47	W and SW	Europe	0	Corbett 2013		12597540	E19	68458	Hungarian	NA
Lezgian	lez	lezg1247	Nakh-Dagl	47,83	41,67	W and SW	Greater M	0	Corbett 2013		616760	E19	3452	Haspelma	NA
Igbo	ibo	nucl1417	Benue-Co	7,33	6	Africa	African Sa	0	Corbett 2013		18007950	E19	200000	E19	NA
Qafar	aar	afar1241	Cushitic	42	12	Africa	Greater Al	2	Bliese 1981: 180-182		1968000	E19	22800	E19	NA
Gumuz	guk	gumu1244	Gumuz	35,83	11,5	Africa	Greater Al	0	Ahland 2012: 95-96		219000	E19	4380	E19	NA
Pohnpeiar	pon	pohn1238	Austrones	158,25	6,88	NG and O	Oceania	0	Rehg 1981	Nichols 19	31350	E19	1100	E19	NA
Marathi	mar	mara1378	Indo-Euro	76	19	S/SE Asia	Indic	3	Corbett 20	Pandharip	71775760	E19	3000000	E19	NA
Lakhota	lkt	lako1247	Siouan	-101,83	43,83	E N Ameri	Basin and	2	VanValin 1977: 36-37		2200	E19	100	E19	NA
Gurung	ggn	guru1261	Sino-Tibet	84,33	28,33	S/SE Asia	Indic	0	Nichols 1992: 297		359000	E19	18900	E19	NA
Abkhaz	abk	abkh1244	West Cauc	41	43,08	W and SW	Greater M	3	Corbett 20	Spruit 198	152740	E19	9164	E19	NA
Dizi	mdx	dizi1235	Omotic	36,5	6,17	Africa	Greater Al	2	Corbett 20	Nichols 19	33900	E19	2050	E19	NA
Khmer	khm	cent1989	Austroasia	105	12,5	S/SE Asia	Southeast	0	Corbett 2013		16390040	E19	1000000	E19	NA
Oromo (H	hae	east2652	Cushitic	38	0	Africa	S Africa	2	Corbett 20	Owens 19	25500000	E19	2000000	Melbaa 19	NA
Bengali	ben	beng1280	Indo-Euro	88,5	23	S/SE Asia	Indic	0	Klaiman 2009: 425		242315050	E19	19202880	E19	NA
Zande	zne	zand1248	Adamawa	26	4	Africa	S Africa	0	Corbett 20	Gore 1926	1142000	E19	100000	E19	NA
Finnish	fin	finn1318	Uralic	25	62	N-C Asia	Inner Asia	0	Corbett 2013		5398780	E19	500000	Kotimaiste	NA

Nature of cross-linguistic data

- From linguistic fieldwork to language description to data matrices.
- Data types:
  - Descriptions of grammatical and lexical phenomena
  - Linguistic examples (transcribed and glossed)

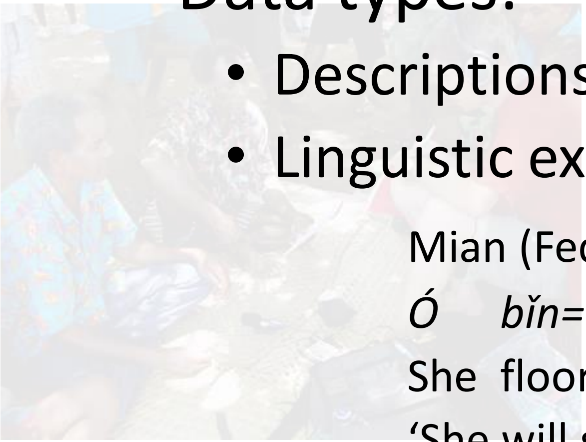
    Mian (Fedden 2004)

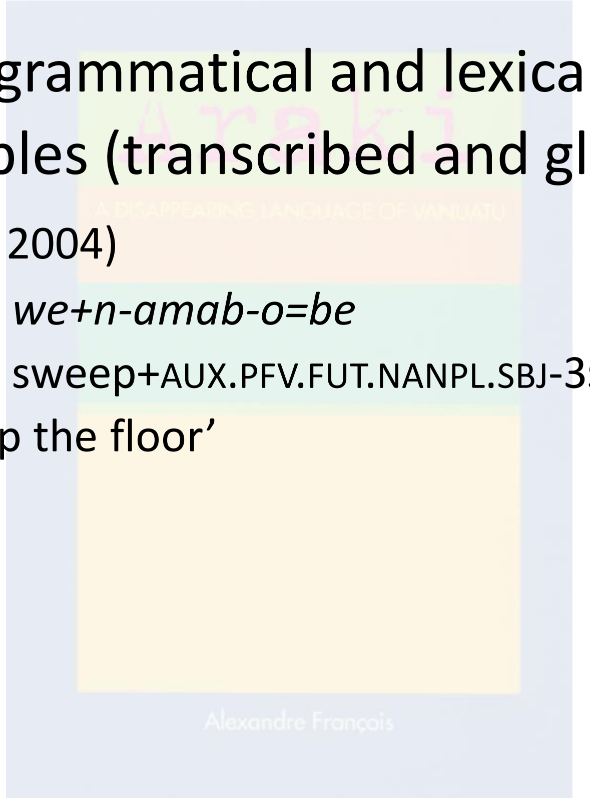
    Ó bǐn=o we+n-amab-o=be

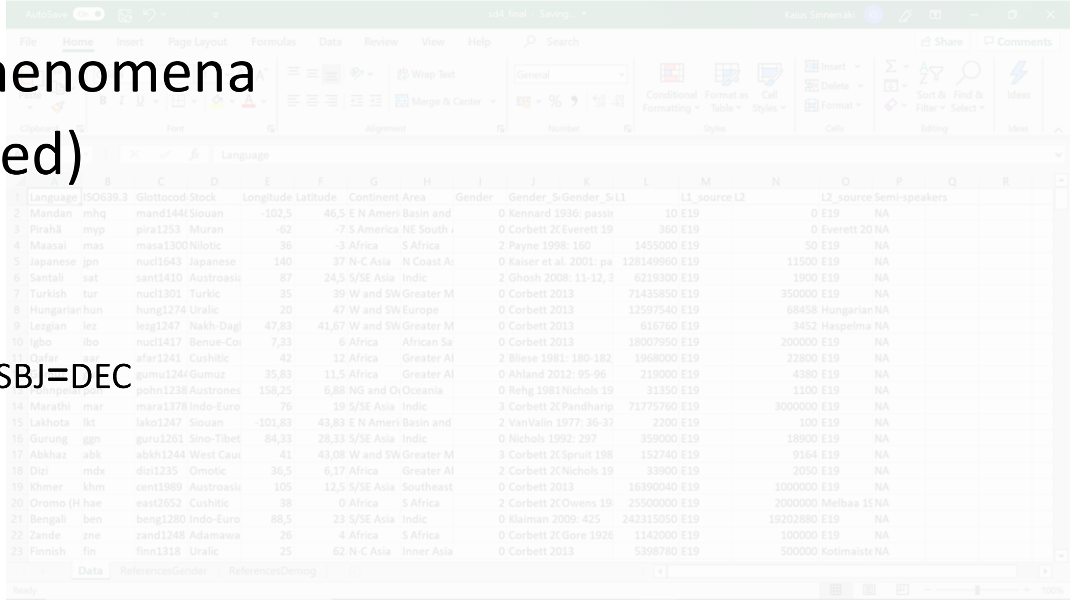
    She floor=N2 sweep+AUX.PFV.FUT.NANPL.SBJ-3SG.F.SBJ=DEC

    ‘She will sweep the floor’

  - Data matrices
  - Maps & graphs







Open publishing in language typology

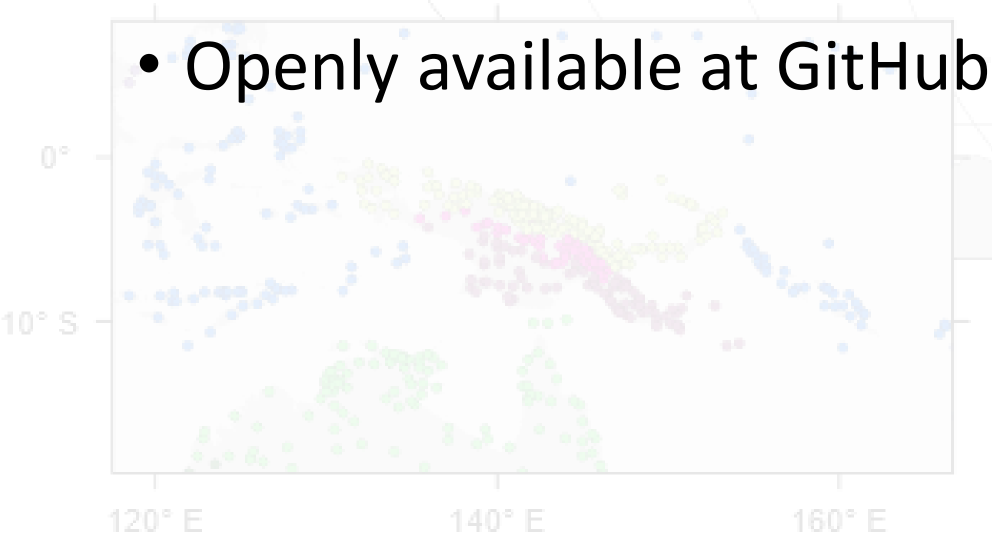
- Cross-Linguistic Linked Data -initiative (https://clld.org/; MPI at Jena).
  - Offers a platform for publishing typological data openly: *WALS, WOLD, APiCS…*
  - Available also at GitHub; archived at Zenodo.

- Standardized data formats
  - CLDF = Cross-Linguistic Data Format (Forkel et al. 2019)
    - Follows the World Wide Web Consortium (W3C) recommendations *Model for Tabular Data and Metadata on the Web* and *Metadata Vocabulary for Tabular Data*.
  - All metadata described by a JSON metadata file according to CSVW for tabular metadata specification.

Open publishing in language typology

- AUTOTYP (University of Zurich & UCB)
  - Data matrices published as csv-files.
  - Metadata published as yaml-files.

- Openly available at GitHub (https://github.com/autotyp/).



Publishing typological data at UH

- Typological data at the University of Helsinki is piling up.
  - Not many good options available for publishing the data at UH.
  - FIN-CLARIN: great for corpus data, but not necessarily for typological data.

- Data published through international initiatives or as supplements.

  →Need for developing a more centralized option at UH.

Publishing typological data at UH

- Aim: bring together typological datasets at UH and publish them openly.

- Data format and platform:
  - Minimally data as csv-files and metadata as yaml-files (following AUTOTYP).
  - Minimally at GitHub, possibly DARIAH-FI (?)

- Timeline:
  - 2020: Bring people together, collect and standardize data, create metadata
  - 2021: continue data standardization and creating metadata
  - 2022: publish data