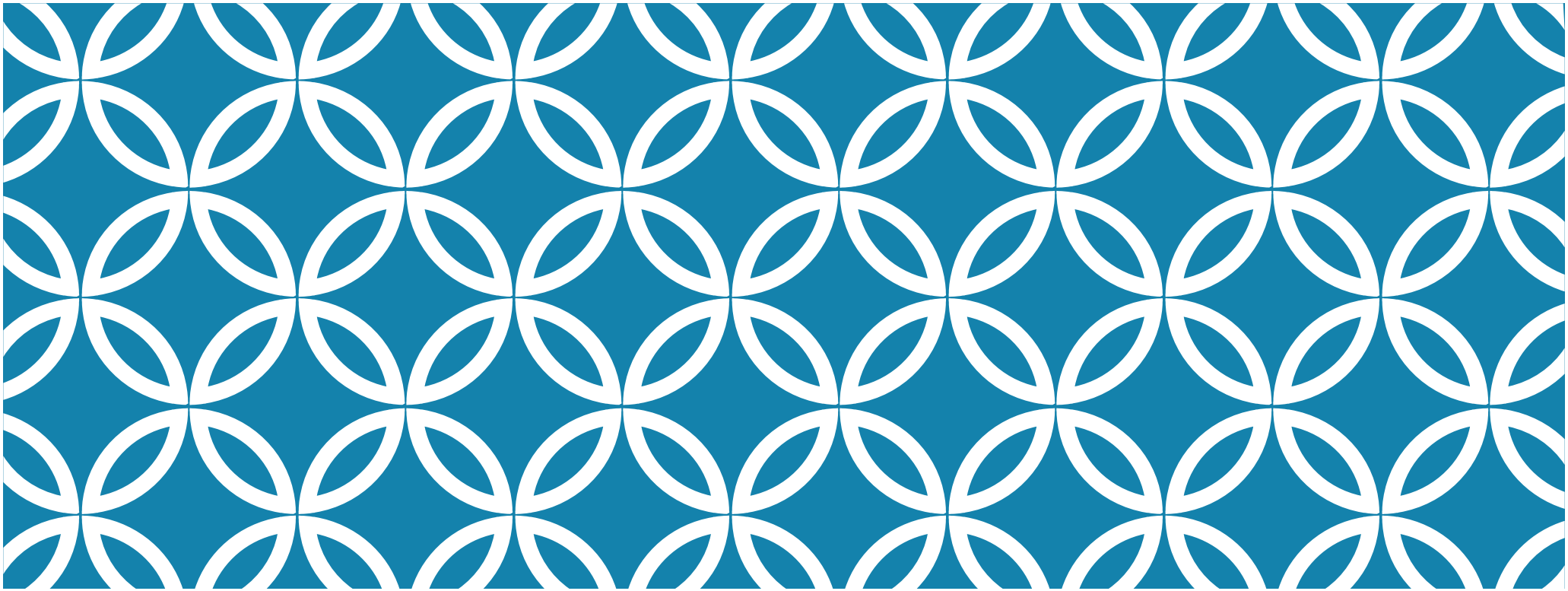
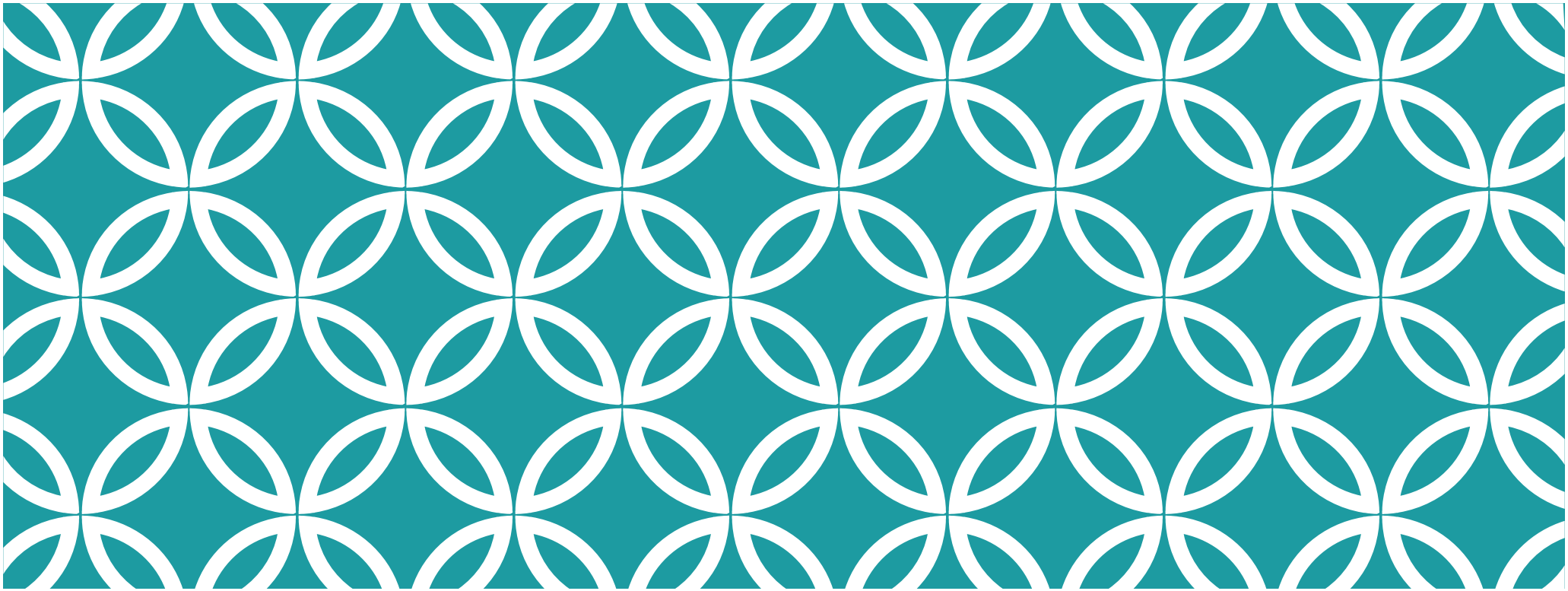


TYPOLOGY, DIACHRONY, AND COMPUTATIONAL GRAMMARS

David Inman

ROADMAP

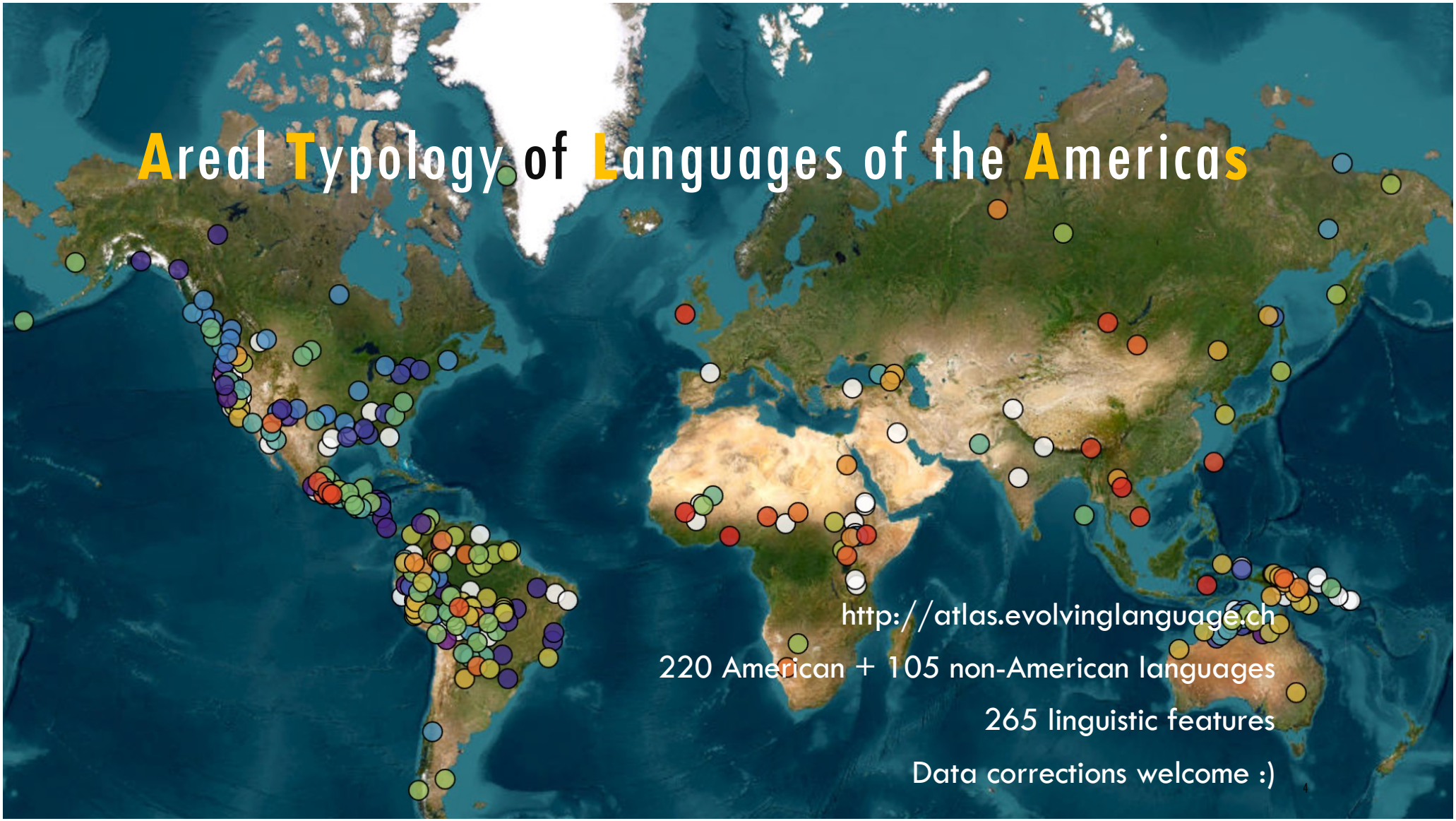
1. Large-scale typology
   1. Areal Typology of Languages of the Americas (ATLAs)
   2. Detecting linguistic areas
2. Language diachrony
   1. Quantifying phonological geometries
   2. Ancestral state reconstruction



1 LARGE-SCALE TYPOLOGY

1.1 ATLAs

1.2 Detecting areas


Areal Typology of Languages of the Americas
http://atlas.evolvinglanguage.ch
220 American + 105 non-American languages
265 linguistic features
Data corrections welcome :)

1.1 Areal Typology of Languages of the Americas

- 4(+) years of development at the University of Zurich, ~21 people
- Designed to maximize areal strength within the Americas
- Typological depth > breadth
- CLDF data structure[1]
- Independent modules for alignment,[2] possession,[3] and Sg-Pl alternation.[4]

[1] Forkel, Robert, et al. 2018. "Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics." *Scientific data* 5.1 (2018): 1–10

[2] Inman, David, Alena Witzlack-Makarevich, Natalia Chousou-Polydouri, and Melvin Steiger. 2024. "Alignment everywhere all at once: Applying the late aggregation principle to a typological database of argument marking." *Journal of Language Modelling* 12(2). 287–347.

[3] Chousou-Polydouri, Natalia, David Inman, Thomas C. Huber, and Balthasar Bickel. 2023. "Multi-variate coding for possession: methodology and preliminary results." *Linguistics* 61(6). 1365–1402.

[4] Inman, David, and Marine Vuillermet. 2024. "Singular-plural verb stem alternation: uncovering global and local drivers of typological variation." *Linguistic Typology*.

1.1 ATLAs

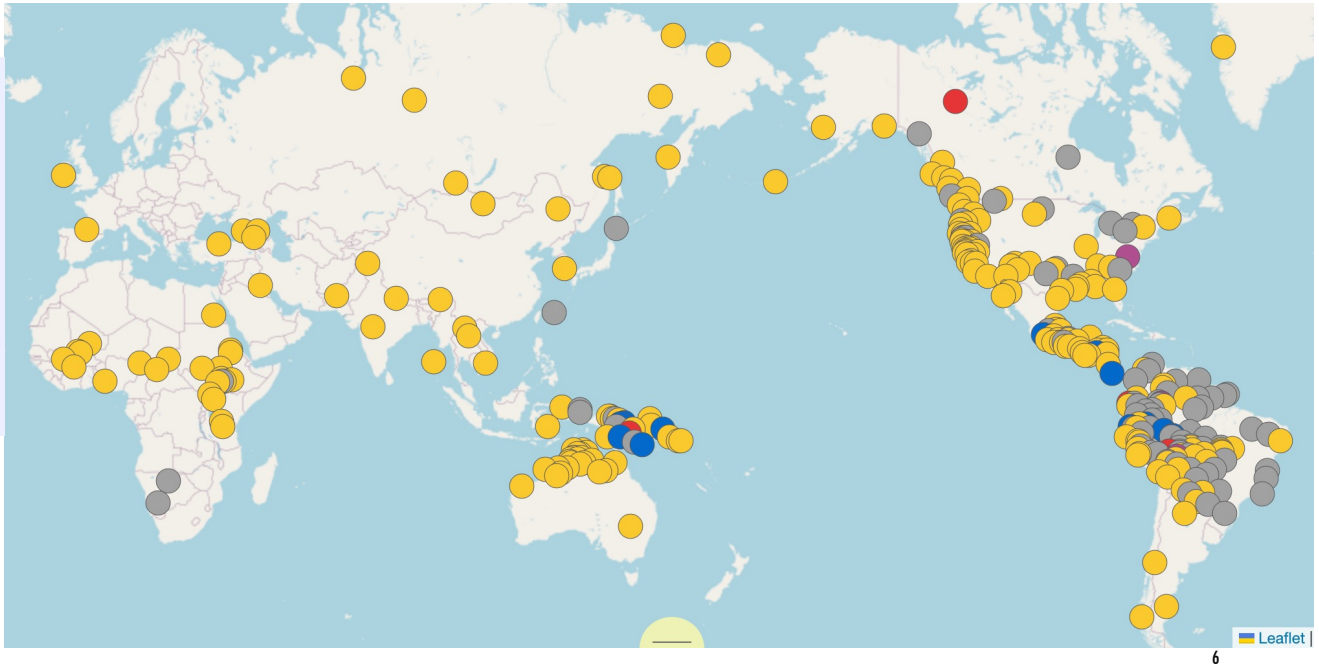
Laterals

Lat-01: Does the language have a lateral approximant phoneme /l/ or an allophone [l] and if so, with what other allophones does this /l/ or [l] alternate?

This feature is described in the feature set Lateral consonants.

Values

	exclusive	partial	all
laterals or glides	217	0	217
rhotic	12	0	12
n or d	4	0	4
n or d and rhotic	3	0	3
no	86	0	86
Total Languages:			322



1.1 ATLAs

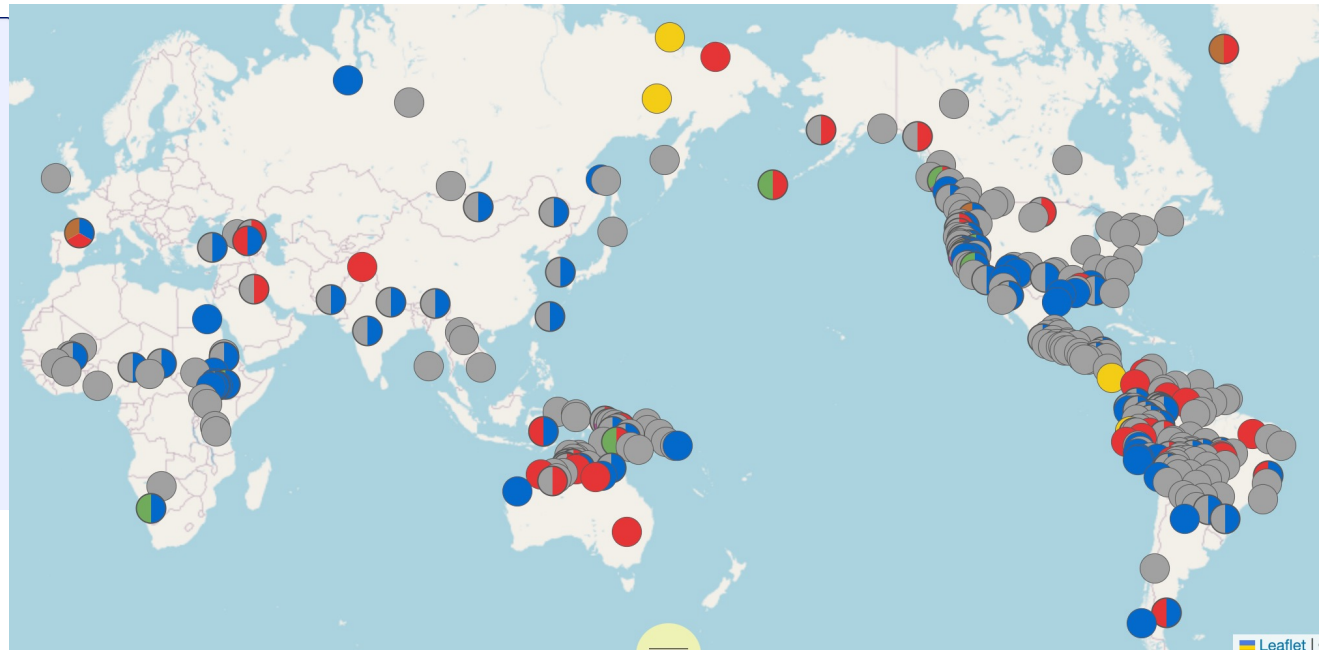
Noun all alignments

Align-02: What are all the alignments of nouns?

This feature is described in the feature set Morphological alignment.

Values

	exclusive	partial	all
no marking	184	68	252
accusative	35	61	96
ergative	18	34	52
tripartite	0	3	3
horizontal	0	4	4
overt neutral	0	7	7
sensitive	4	3	7
	Total Languages:		325



1.1 ATLAs

- Lots of ancillary work has come out of this project:
  - Method for cross-linguistic comparison of alignment[2]
  - Cross-linguistic comparison of possession + semantic grounding of possession classes[3]
  - A typology of singular-plural verb stem alternations[4]
  - Sociative causative as a grammatical phenomenon[5]
  - Are isolates typologically different from non-isolates?[6]

[2] Inman, David, Alena Witzlack-Makarevich, Natalia Chousou-Polydouri, and Melvin Steiger. 2024. "Alignment everywhere all at once: Applying the late aggregation principle to a typological database of argument marking." *Journal of Language Modelling*, 12(2):287–347.

[3] Chousou-Polydouri, Natalia, David Inman, Thomas C. Huber, and Balthasar Bickel. 2023. "Multi-variate coding for possession: methodology and preliminary results." *Linguistics*, 61(6):1365–1402.

[5] Rose, Françoise and Marine Vuillermet. *accepted.* "Reassessing the areality of sociative causative markers: a South American feature."

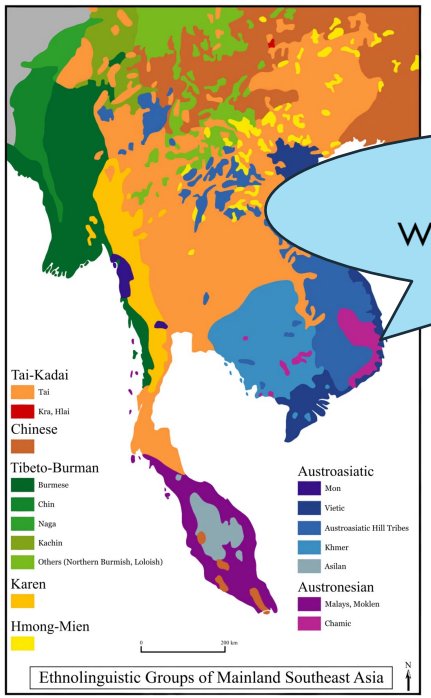
[6] Vuillermet, Marine, David Inman, Natalia Chousou-Polydouri, Kellen Parker van Dam, Shelece, Easterday, and Françoise Rose. 2025. "Is there a typological profile of isolates?" In Iker Salaberri et al., editors, *Investigating Language Isolates: Typological and diachronic perspectives*, 22–47. John Benjamins.

1.1 ATLAs

- How can this resource be used?
  - Extraction of linguistic features for encoded languages
  - Evaluation of typological claims or universals
  - Empirical grounding of linguistic areas

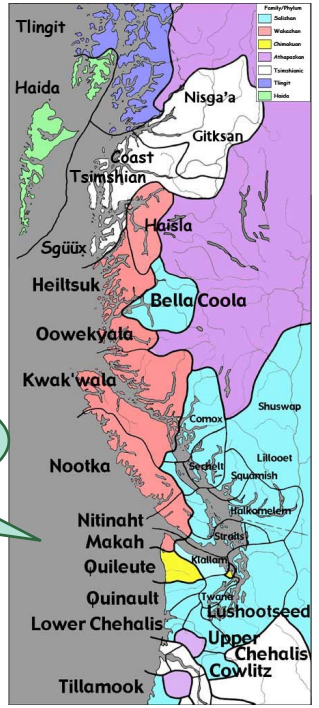
1.2 DETECTING LINGUISTIC AREAS

A geographic region in which languages share properties from prolonged contact.



We love tone!
We love numeral classifiers!

We love consonant clusters!
We love VSO syntax!



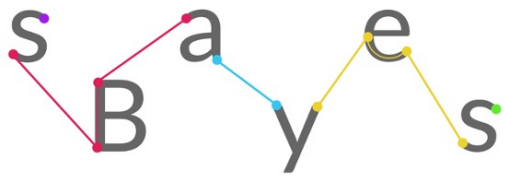
1.2 DETECTING LINGUISTIC AREAS

- Traditionally, areas are "discovered" through expert analysis
  - Are there lots of similar and unusual features?
  - Do enough people agree that this establishes an area?
- But sometimes it's hard to tell
  - Is the Amazon an area?[7]
  - Is Anatolia an area?[8]

[7] Campbell, Lyle, Thiago Chacon, and John Elliott. 2020. "Contact and South American languages." *The Handbook of Language Contact* 625–648.

[8] Haig, Geoffrey. 2014. "East Anatolia as a linguistic area? Conceptual and empirical issues." *Bamberger Orientstudien* 1.

1.2 DETECTING LINGUISTIC AREAS



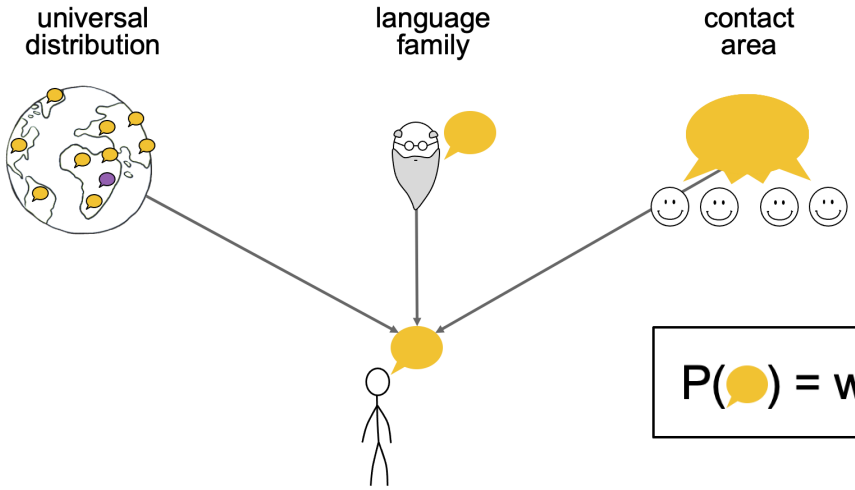
✧sBayes: A Bayesian algorithm for area detection in the presence of confounders.[9]

[9] Ranacher, Peter, Nico Neureiter, Rik Van Gijn, Barbara Sonnenhauser, Anastasia Escher, Robert Weibel, Pieter Muysken, and Balthasar Bickel. 2021. "Contact-tracing in cultural evolution: a Bayesian mixture model to detect geographic areas of language contact." *Journal of The Royal Society Interface* 18, no. 181.

1.2 DETECTING LINGUISTIC AREAS

sBayes assumptions:

✧3 underlying distributions determine each feature:
(1) universal distribution; (2) family distribution; (3) contact area distribution



$$P(\text{💬}) = w_u P_u(\text{💬}) + w_i P_i(\text{💬}) + w_c P_c(\text{💬})$$

1.2 DETECTING LINGUISTIC AREAS

sBayes assumptions:

- 3 underlying distributions determine each feature $P(\text{🟡}) = w_u P_u(\text{🟡}) + w_i P_i(\text{🟡}) + w_c P_c(\text{🟡})$
- Features are *independent*
- Contact regions are more likely if they are geographically close

1.2 DETECTING LINGUISTIC AREAS

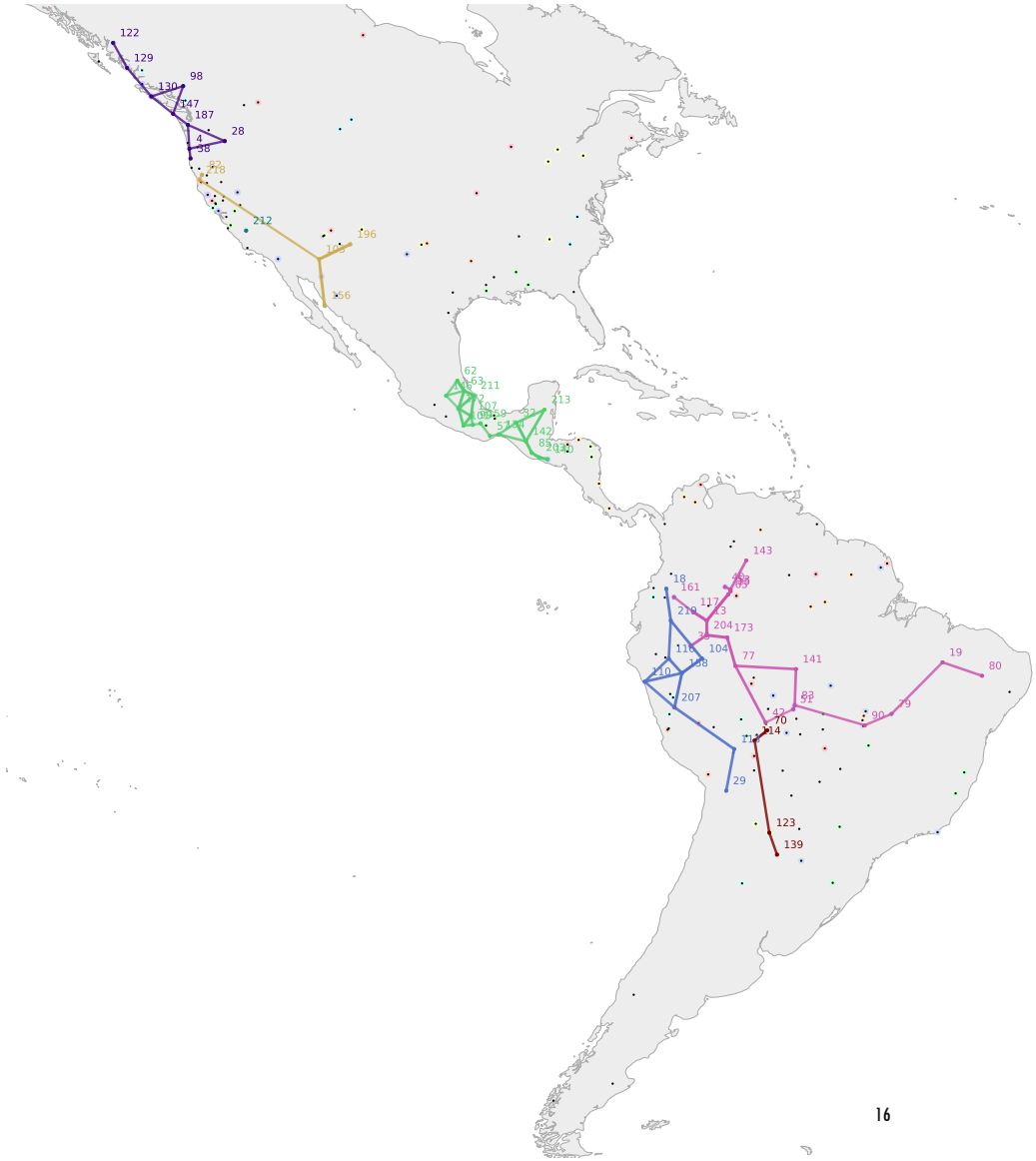
Prior distributions for sBayes:

- Geographic prior
  how to penalize large areas
- Universal prior
  can be empirical from worldwide data
- Family priors
  can be empirical from family data outside the studied area

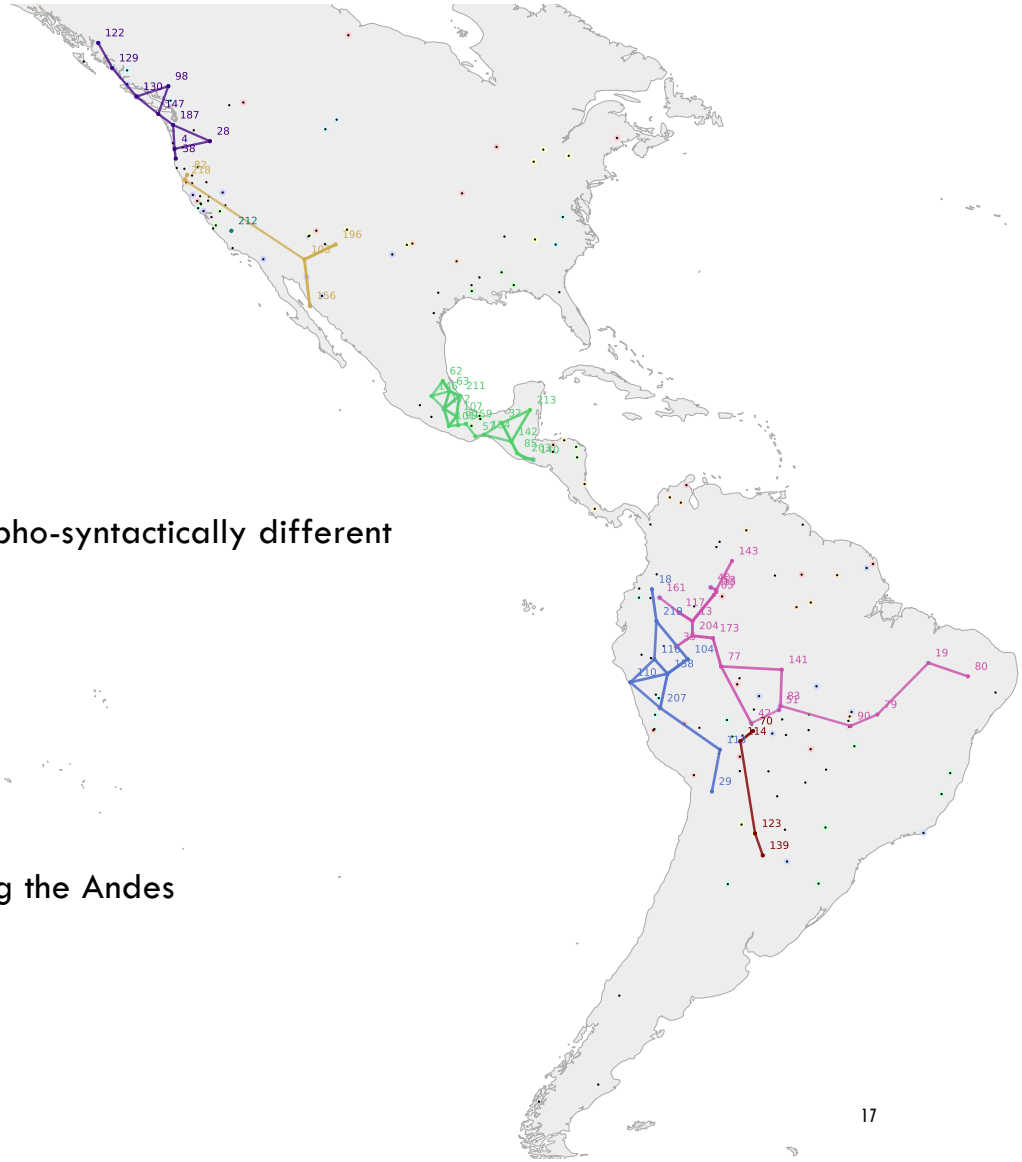
1.2 AMERICAN AREAS

<u>Preliminary sBayes results</u>

- ✧Pacific Northwest
- ✧Southwest(–California)
- ✧Mesoamerica
- ✧Amazon
- ✧Andean corridor
- ✧Mamoré–Guaporé



1.2 AMERICAN AREAS



✧Pacific Northwest

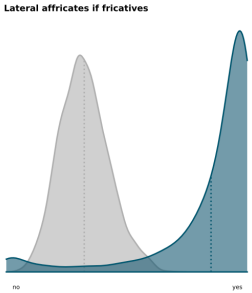
- **Excludes** Na-Dene languages
- Despite phonological similarities, Na-Dene is morpho-syntactically different

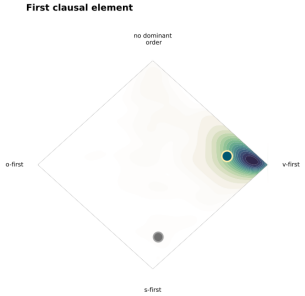
✧Andean corridor

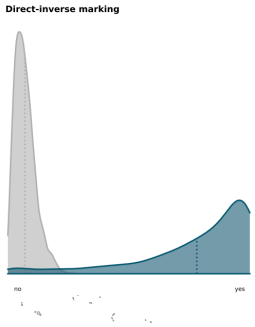
- **Excludes** Quechuan languages
- Likely reflects a pre-Quechuan contact zone along the Andes

1.2 AMERICAN AREAS

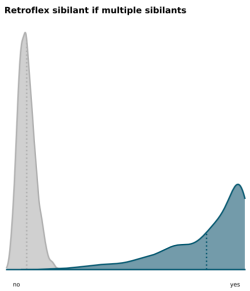
✧Pacific Northwest

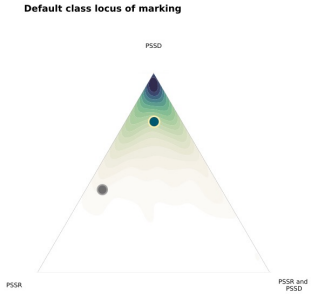


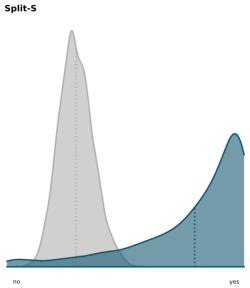




✧Andean corridor



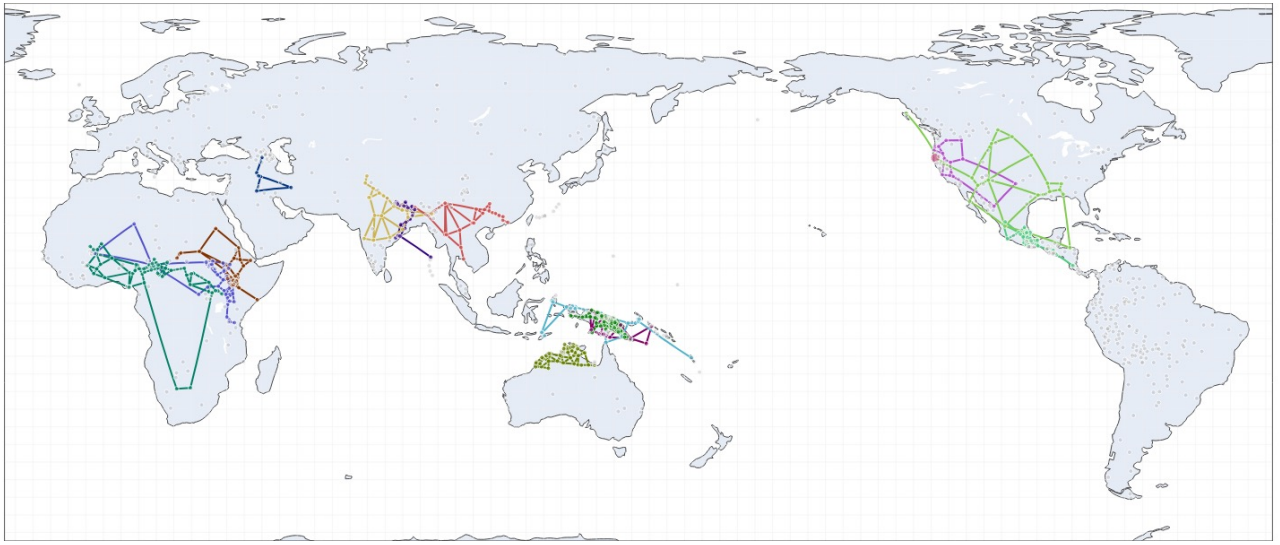




1.2 DETECTING LINGUISTIC AREAS

Ongoing work applying sBayes to independent global datasets.[10]

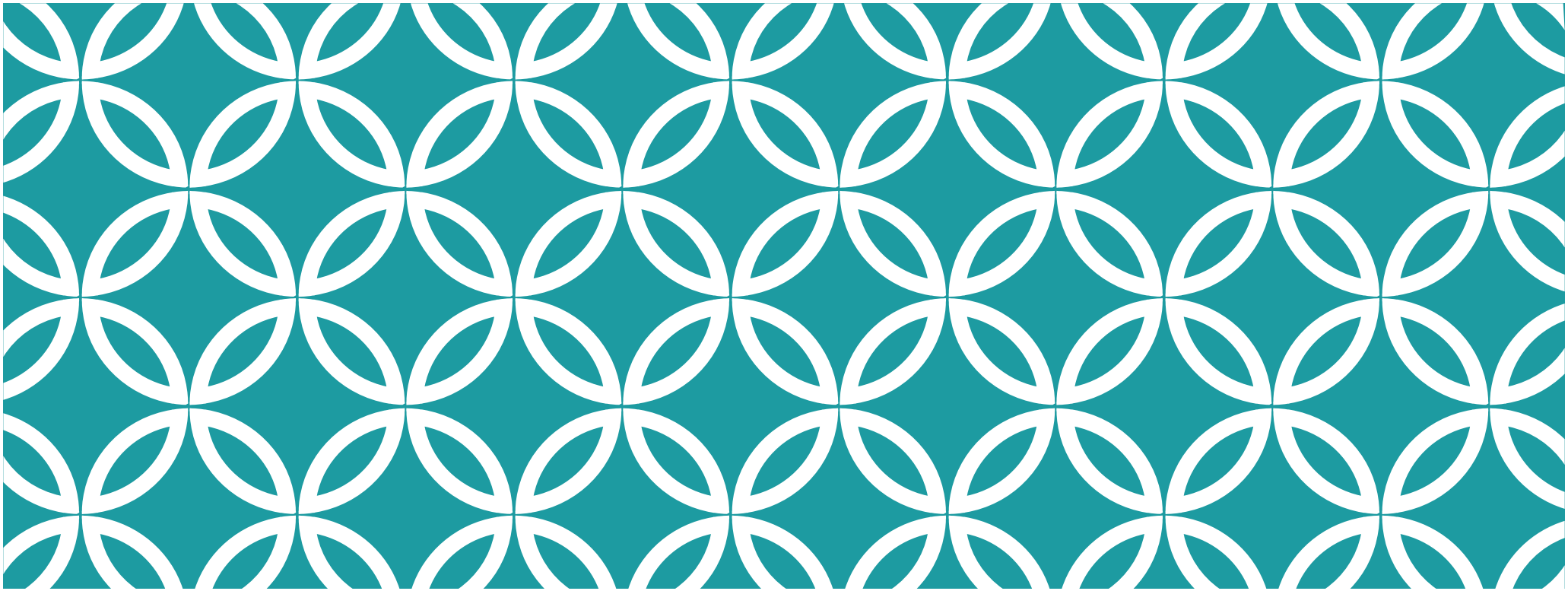
- Fails to find detailed structure in the Americas!



[10] Graff, Anna, Natalia Chousou-Polydouri, David Inman, et al. 2025. "Curating global datasets of structural linguistic features for independence." *Scientific Data* 12, 106.

1.2 DETECTING LINGUISTIC AREAS

- Feature selection and well-motivated priors are ***very important***
- This can be applied to variables derived from implemented grammars



2 LANGUAGE DIACHRONY

2.1 Phonological geometries

2.2 Ancestral state reconstruction

2.1 PHONOLOGICAL GEOMETRIES

Pilot study on language diachrony:

What is the stability of oral obstruent series over time?
(= rows in a phonological chart)

- Why is this an interesting question?
  - Lots of data available from PHOIBLE[11]
  - This feature is *continuous*
  - Reason to suspect this changes slowly

[11] Moran, Steven & McCloy, Daniel (eds.) 2019. PHOIBLE 2.0. *Jena: Max Planck Institute for the Science of Human History.*

2.1 PHONOLOGICAL GEOMETRIES

Proto-Indo-European[12]

3 stop series

1 fricative series - /ɸ, x^j, x, x^w/

	labial	alveolar/dental	palatalized velar	velar	labialized velar	laryngeal
stop (-voice)	*p	*t	*k^j	*k	*k^w	
stop (+voice, -aspiration)	(*b)	*d	*g^j	*g	*g^w	
stop (+voice, +aspiration)	*b^h	*d^h	*g^{jh}	*g^h	*g^{wh}	
fricative		*s				*h_1, *h_2, *h_3
nasal	*m	*n				
liquid		*r, *l				
glide			*j		*w	

3 stop series + /s/ ≈ 3.3

[12] Mallory, James P., and Douglas Q. Adams. 2006. *The Oxford introduction to proto-Indo-European and the proto-Indo-European world.* Oxford University Press, USA, 2006.

2.1 PHONOLOGICAL GEOMETRIES

Gheg Albanian:

2 stop series

2 fricative series

	labial	dental	alveolar	palatal	velar
stop (-voice)	p	t	ts	tʃ	k
stop (+voice)	b	d	dz	dʒ	g
fricative (-voice)	f	θ	s	ʃ	
fricative(+voice)	v	ð	z	ʒ	
liquid		ɾ, r, l, ɬ			
nasal	m	n			ŋ
glide				j	

≈ 3.5

2.1 PHONOLOGICAL GEOMETRIES

Calculating partial obstruent series:

- Estimate a "markedness" value for each phoneme:
  - Assign each language family to one macroarea
  - Count occurrences of each phoneme, weighted by language

$$language\ weight = \frac{avg\ number\ of\ languages\ per\ macroarea}{number\ of\ stocks\ in\ macroarea \times number\ of\ languages\ in\ family}$$

- Each cell has a markedness score, which is the total markedness value of all phonemes that *could* go in the cell.

$$row\ score = \frac{\sum_{filled\ cells} markedness(cell)}{\sum_{all\ cells} markedness(cell)}$$

2.1 PHONOLOGICAL GEOMETRIES

Gheg Albanian:

	labial	dental	alveolar	palatal	velar
stop (-voice)	p	t	ts	tʃ	k
stop (+voice)	b	d	dz	dʒ	g
fricative (-voice)	f	θ	s	ʃ	**-0.25**
fricative(+voice)	v	ð	z	ʒ	**-0.21**
liquid		ɾ, r, l, ɬ			
nasal	m	n			ŋ
glide				j	

≈ 3.5

2.2 ANCESTRAL STATE RECONSTRUCTION

Reconstruct ancestral states on linguistic phylogenies

Multiple goals (together with Gerhard Jäger):

1. Evaluate how consonant series change in general
2. Evaluate the phylogenetic signal of this feature

2.2 ANCESTRAL STATE RECONSTRUCTION

Methods:

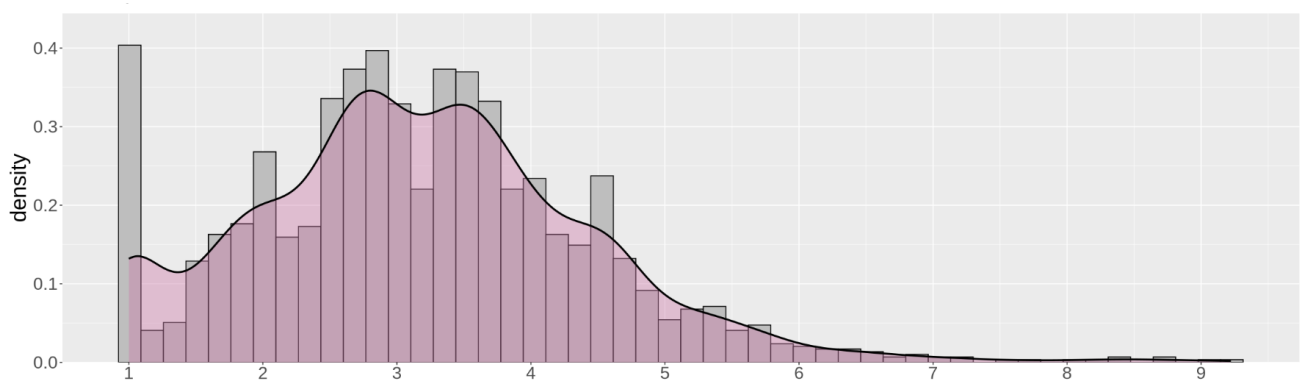
- Modified Ornstein-Uhlenbeck process (FPK)[13]
- Using a world tree[14] and individual phylogenies

[13] Boucher, Florian C., Vincent Démery, Elena Conti, Luke J. Harmon, and Josef Uyeda. 2018. "A general model for estimating macroevolutionary landscapes." *Systematic biology* 67, no. 2 304–319.

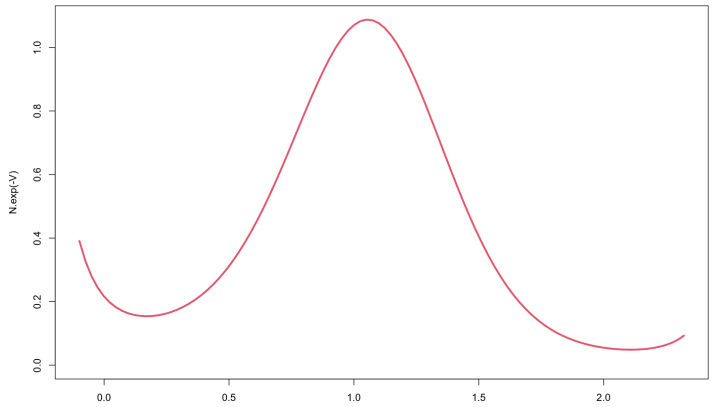
[14] Bouckaert, Remco, David Redding, Oliver Sheehan, Thanos Kyritsis, Russell Gray, Kate E. Jones, and Quentin Atkinson. 2022. "Global language diversification is linked to socio-ecology and threat status." *Preprint at https://doi. org/10.31235/osf. io/f8tr6*.

2.2 ANCESTRAL STATE RECONSTRUCTION

Raw data:



Fitness landscape (log scale):

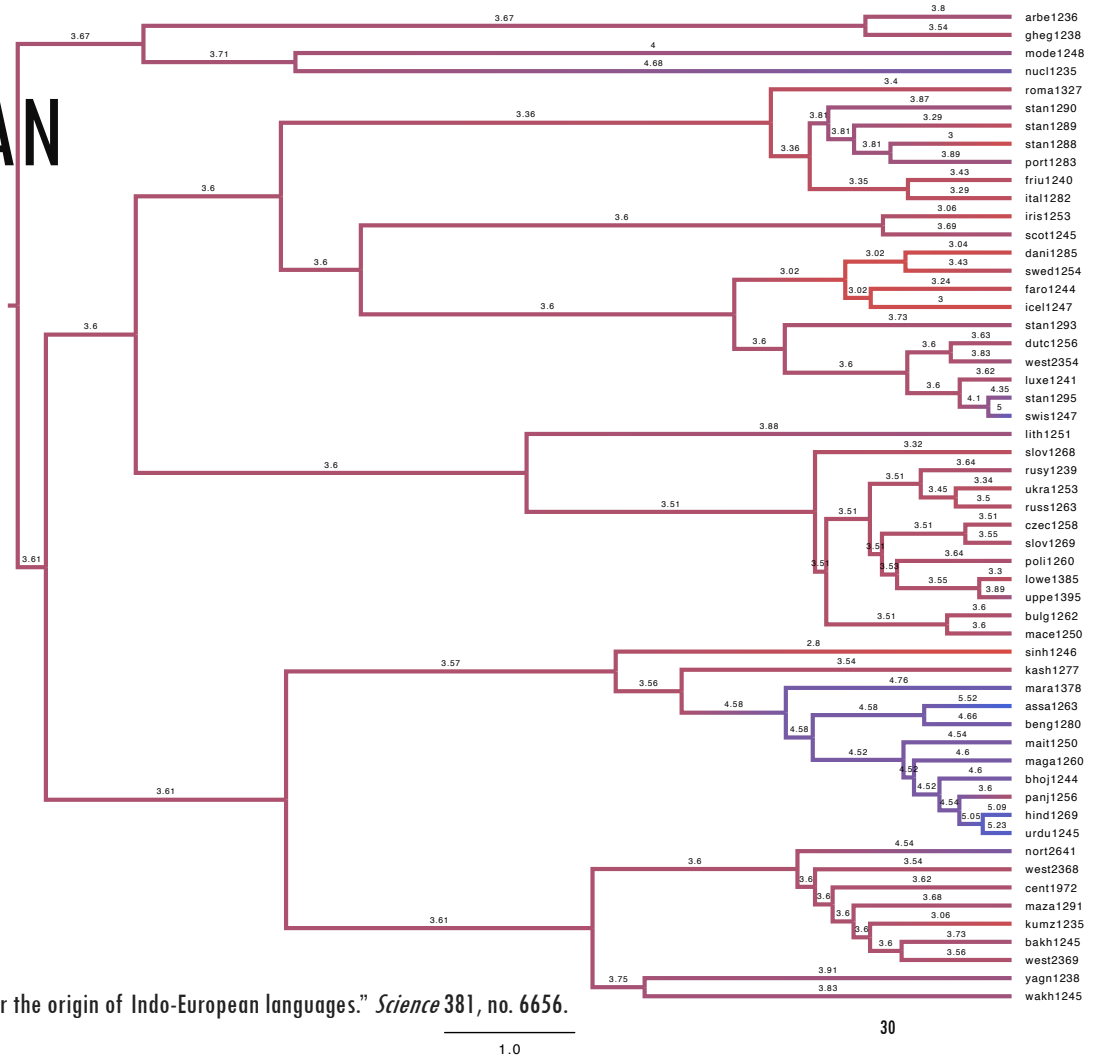


Characteristic time: ~6,000 years

2.2 ASR: INDO-EUROPEAN

Ancestral state reconstruction on Heggarty et al's Indo-European tree[18]

- Note increase in Indic, decrease in North Germanic

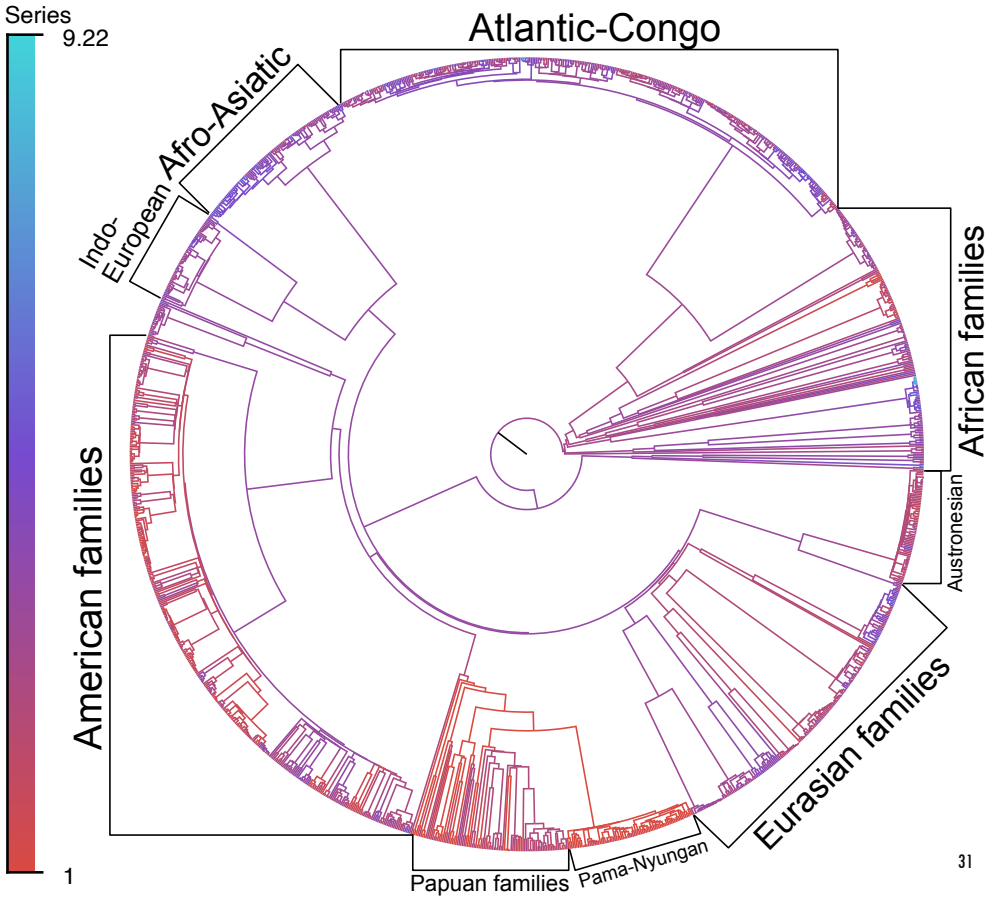


[18] Heggarty et al. 2023. "Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages." *Science* 381, no. 6656.

2.2 ASR: WORLD TREE

Ancestral state reconstruction on the world tree[14]

- Note ~1 series in Papua/Australia
  - Could be inherited, could be an (extremely ancient) area

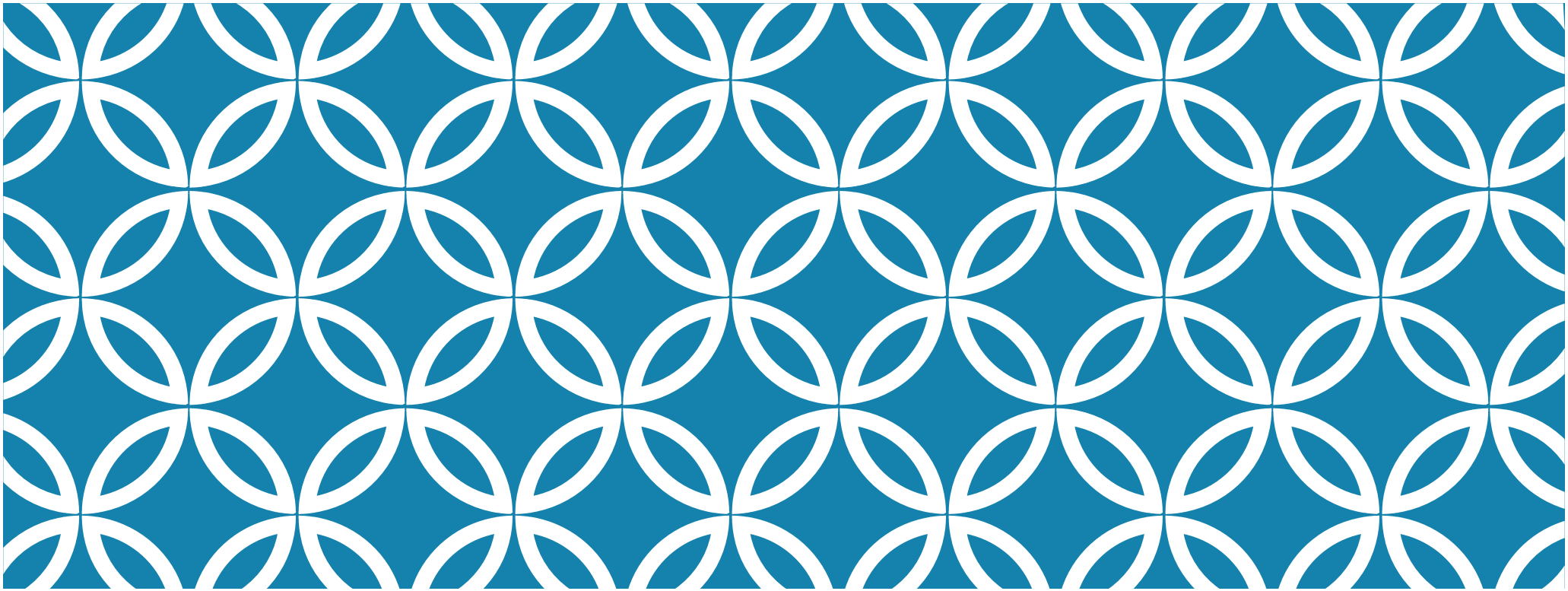


[14] Bouckaert, Remco, David Redding, Oliver Sheehan, Thanos Kyritsis, Russell Gray, Kate E. Jones, and Quentin Atkinson. 2022. "Global language diversification is linked to socio-ecology and threat status." *Preprint at https://doi. org/10.31235/osf. io/f8tr6.*

2.2 ANCESTRAL STATE RECONSTRUCTION

What does this show?

- Some linguistic features (a) change slowly; (b) hold deep-time phylogenetic signal
- Such features are not pre-existing: consonant geometry isn't a WALS variable!
- This methodology can be expanded to grammar:
  - Feature space must be carefully designed
  - Large grammatical artifacts might be a good bet



THANK YOU FOR YOUR ATTENTION!

Bedankt

EXTRA SLIDES

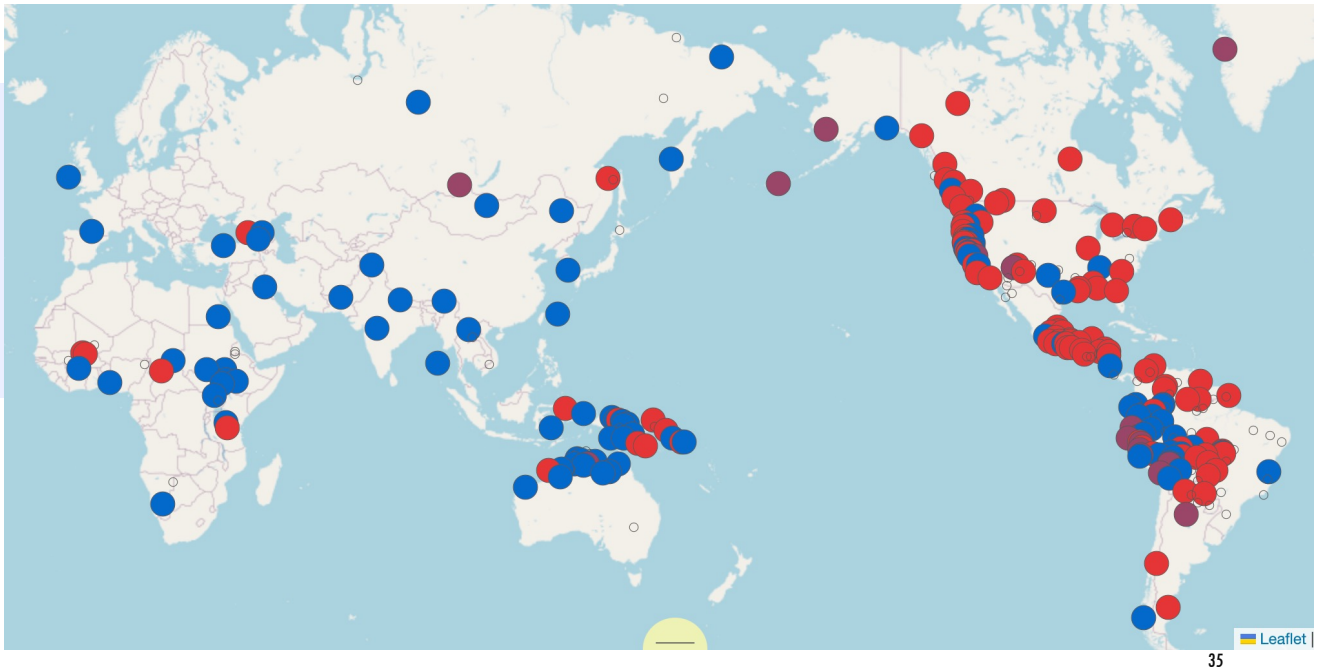
1.1 ATLAs

Default class locus of marking

NounPoss-08c: If the default class is possessed with some type of MARKER construction that is not a LINKER construction, what is the locus of this marker or markers?

Values

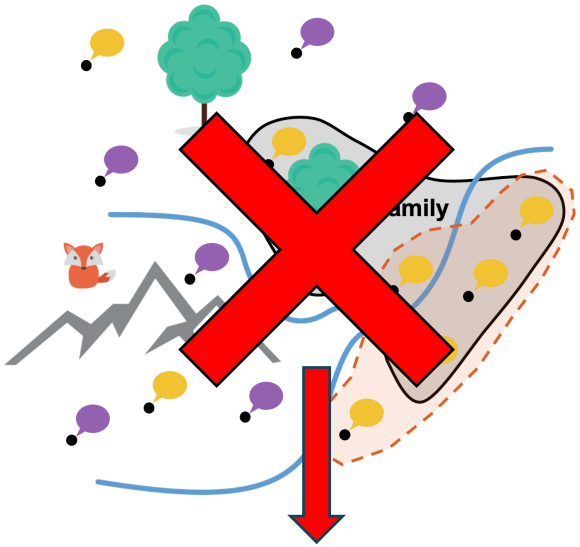
	exclusive	partial	all
PSSR	96	0	96
PSSD	103	0	103
PSSR and PSSD	16	0	16
Total Languages:			215



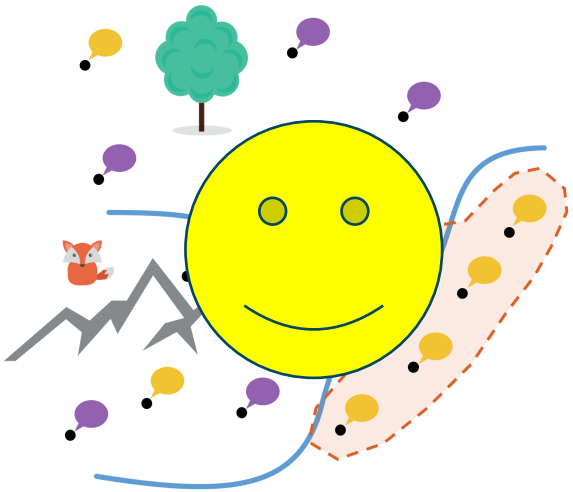
1.2 DETECTING LINGUISTIC AREAS

How to find an area?

Universal preference



Same family



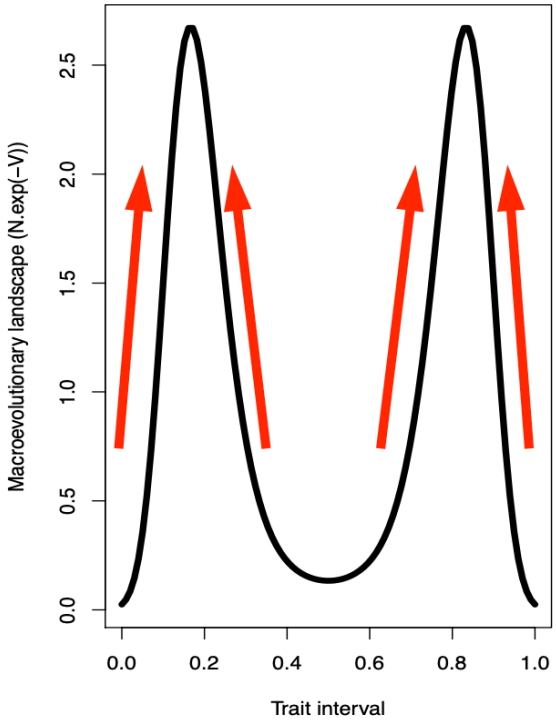
2.1 PHONOLOGICAL GEOMETRIES

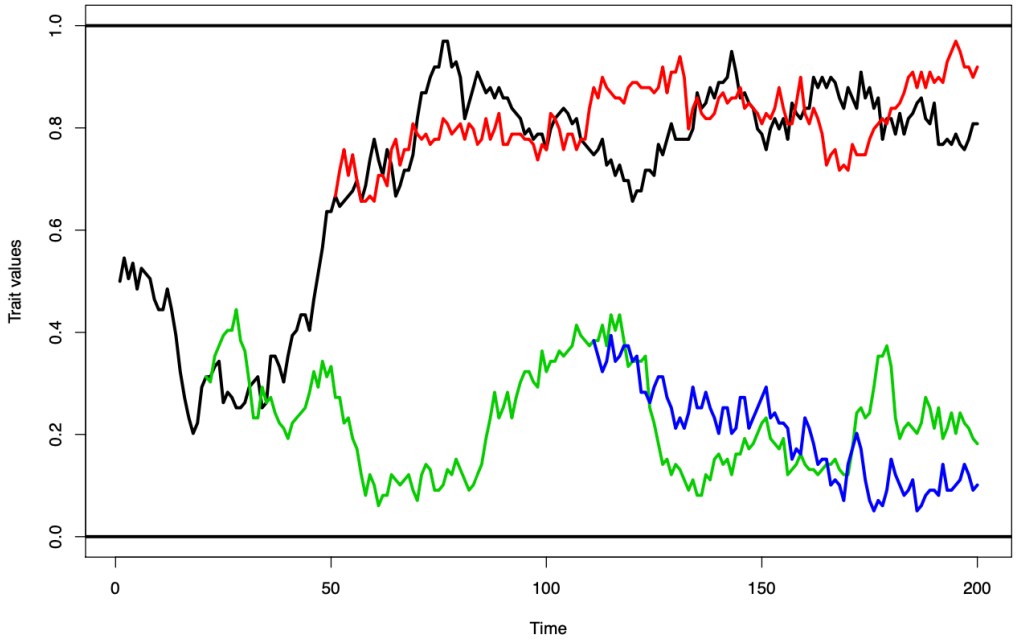
Khakas:

	labial	dental-alveolar	alveolar-sibilant	palatal	velar-uvular
stop (-voice)	p	t	ts	tʃ	k
stop (+voice)	b	d	**-0.04**	dʒ	g
fricative (-voice)	f	**-0.1**	s	ʃ	χ
fricative(+voice)	v	**-0.08**	z	ʒ	ʁ
liquid		r, l			
nasal	m	n			ŋ
glide				j	

≈ 3.8

2.2 ANCESTRAL STATE RECONSTRUCTION

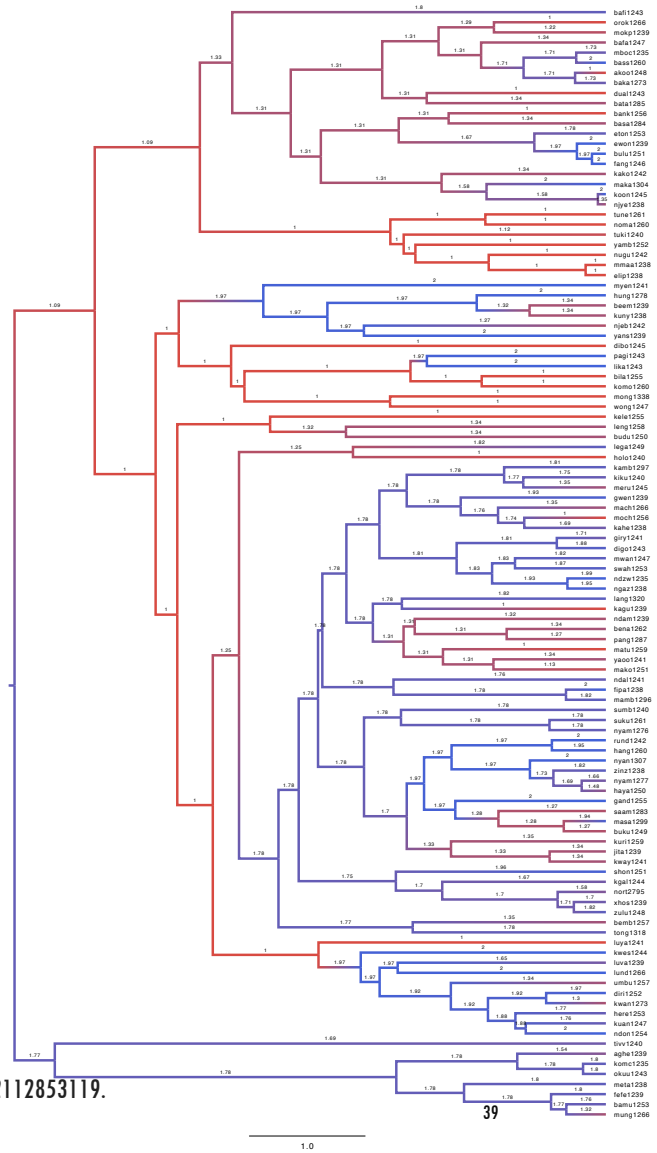




2.2 ASR: BANTU

Restriction to particular natural categories reveals further structure

✧E.g. fricatives in Bantu[15]



[15] Koile, Ezequiel, et al. 2022. "Phylogeographic analysis of the Bantu language expansion supports a rainforest route." *PNAS* 119.32: e2112853119.