

# TALKS of the PAST

open  
seminars by CHP

4/12-24  
seminar speaker

*Harald Hammarström*

THE LANGUAGE FAMILIES OF THE WORLD:  
CURRENT STATE AND FUTURE  
**PERSPECTIVES**



# The Language Families of the World: Current State and Future Perspectives

Harald Hammarström  
Uppsala University  
`harald.hammarstrom@lingfil.uu.se`

4 Dec Uppsala

# Language Families

- There are some 7,000 attested (spoken L1) languages in the world
- Some languages **resemble** each other (much) **more than** expected by **chance**
- Such far-reaching similarities can be explained if the languages in question **stem from a common ancestor**
- A set of languages deemed to stem from a common ancestor constitutes a **language family**
- A language which resembles no other known language this way is a **language isolate**

## Example Lexical Data

wʌn	bir	yek	yek	wæ:hed	ek	en
tu	iki	do	dû	etne:n	do:	tvo:
θri	ytʃ	se	sê	tælæ:tæ	ti:n	tre:
neim	isim/ad	esm	naw	ʔesm	na:m	namn
nous	burun	dama:gh	lût	mænæxi:r	na:k	nɛ:sa
watər	su	a:b	aw	ma:ʃa	pa:ni:	vaten
stoun	taʃ	sang	berd	ħagara	patthar	ste:n
hed	baʃ/kafa	sar	ser	ra:s	sar	huvud
nat	gedze	ʃab	ʃev	le:læ	ra:tri:	nat
boun	kemik	ostokha:n	hestî	ʃadm	hadḏi:	be:n
fiʃ	balık	ma:hi	masi	sæmækæ	machli:	fisk
hɔ:n	boynuz	ʃa:x	ʃax	ʔarn	si:ng	huŋ
li:f	yaparak	barg	valge	wæræʔæ	patti:	löv
nu:	yeni	naw/ta:ze	nwê	ge <sup>1</sup> di:d	naya:	ny
wi:	biz	ma:	ême	ehnae	ham	vi:

# Example Lexical Data

wʌn	bir	yek	yek	wæ:hed	ek	en
tu	iki	do	dû	etne:n	do:	tvo:
θri	ytʃ	se	sê	tælæ:tæ	ti:n	tre:
nem	isim/ad	esm	naw	ʔesm	na:m	namn
nous	burun	dama:gh	lût	mænæxi:r	na:k	nɛ:sa
watər	su	a:b	aw	majja	pa:ni:	vaten
stoun	taʃ	sang	berd	hagara	patthar	ste:n
hed	baʃ/kafa	sar	ser	ra:s	sar	hu:vud
nart	gedze	ʃab	ʃev	le:læ	ra:tri:	nat
boun	kemik	ostokha:n	hestî	ʃadm	haḍdi:	be:n
fiʃ	balık	ma:hi	masi	sæmækæ	machli:	fisk
hɔ:n	boynuz	ʃa:x	ʃax	ʔarn	si:ng	hu:ŋ
li:f	yaparak	barg	valge	wæræʔæ	patti:	löv
nu:	yeni	naw/ta:ze	nwê	gedi:d	naya:	ny
wi:	biz	ma:	ême	ehnæ	ham	vi:

# Example Lexical Data

English	Turkish	Persian	Kurdish (Sorani)	Arabic (Egyptian)	Hindi	Swedish
wʌn	bir	yek	yek	wæ:hed	ek	en
tu	iki	do	dû	etne:n	do:	tvo:
θri	ytʃ	se	sê	tælæ:tæ	ti:n	tre:
neim	isim/ad	esm	naw	ʔesm	na:m	namn
nous	burun	dama:gh	lût	mænæxi:r	na:k	næ:sa
watər	su	a:b	aw	majja	pa:ni:	vaten
stoun	taʃ	sang	berd	ħagara	patthar	ste:n
hed	baʃ/kafa	sar	ser	ra:s	sar	hu:vud
nart	gedʒe	ʃab	ʃev	le:læ	ra:tri:	nat
boun	kemik	ostokha:n	hestî	ʔadm	ħaḍḍi:	be:n
fiʃ	balik	ma:hi	masi	sæmækæ	machli:	fisk
hə:n	boynuz	ʃa:x	ʃax	ʔarn	si:ng	hu:ŋ
li:f	yaprak	barg	valge	wæræʔæ	patti:	löv
nu:	yeni	naw/ta:ze	nwê	gedi:d	naya:	ny
wi:	biz	ma:	ême	ehnæ	ham	vi:

# Demonstrating Language Families

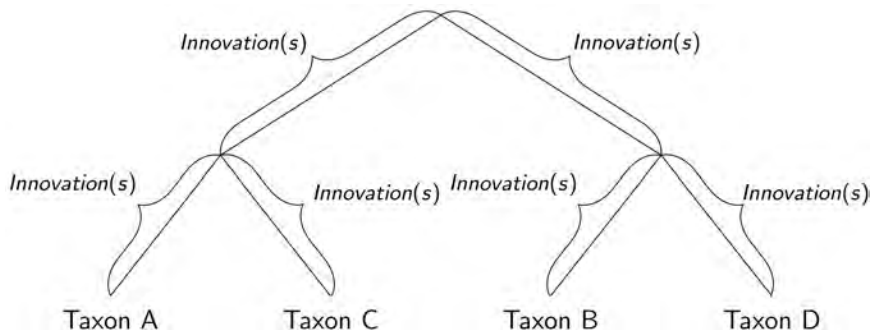
*Languages have to have similarities “beyond chance” to count as related*

- Typically vocabulary
  - ▶ Linked by *regular sound correspondences*
  - ▶ Regular = affects all items in the lexicon
  - ▶ Chance correspondences won't be regular!
- Morphological paradigms
- Any other “quirky” grammatical property
- ...

*A very exact characterization of how many/much/types of similarity is enough to count as “beyond chance” is difficult to do*

# Language Families are Trees

- The basic model for language diversification is Split + Extinction + Descent-with-modification
- In other words, the tree model

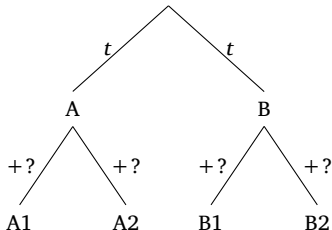
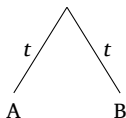


- (There is also language contact with is real and significant, but let us ignore that for today)



# Temporal Limits

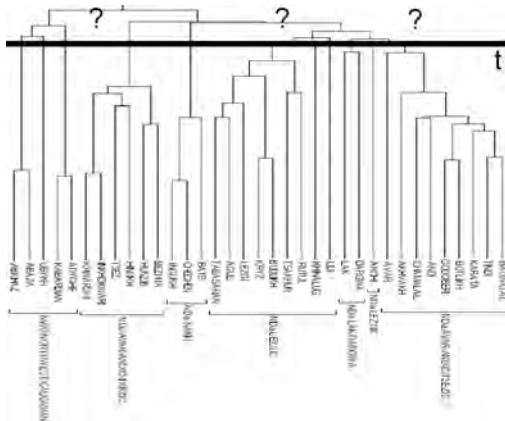
- Language changes at such a pace that after some limit  $t$  there won't be any similarities left from a common ancestor (i.e., there will not be enough to distinguish the similarities from chance)
- Informed speculation says  $t$  is about 10,000 years
- There is no good theoretical understanding (what stochastic process(es)?) or extensive empirical data underlying this estimate
- Should  $t$  depend on the # of languages in each branch?



- For a small number of (famous) languages there are written records, then of course the limit is  $t$  from date of attestation

# Scratching Treetops

- Human language arguably dates back at least to anatomically modern humans (~ 300,000 years, Hublin et al. 2017)
- Presumably all extant languages are related in one world tree



- The language families we (can) discover are thus the treetops of the underlying world tree cut at  $t$

# How Many Language Treetops/Lineages?

*Lineage = Family + Isolate*

- a) 26?
- b) 251?
- c) 250 + ?
- d) 250-300?
- e) 398?
- f) 422?

*Although most authors like to pretend otherwise, clearly there is subjectivity involved!*

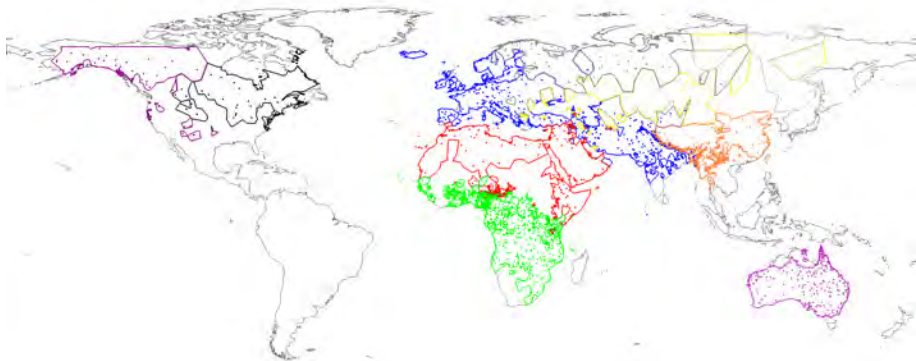
# How Many Language Lineages?

*Lineage = Family + Isolate*

- a)  $16 = 12 + 4$  (Ruhlen 1991:390)
- b)  $251 = 137 + 114$  Ethnologue (E27, Eberhard et al. 2024)
- c)  $250 +$  (Campbell 1999:163-165)
- d)  $250-300$  (Campbell 2004:184-186)
- e)  $398 = 235 + 163$  (Campbell 2020:220-229)
- f)  $422 = 239 + 183$  Glottolog (G51, Hammarström et al. 2024)

*If linguists can't resolve the matter internally are there "outside checks" that can shed light?*

# Geospatial sizes: 10 Largest



# The Largest (# lgs) Language Families

Family	# languages	Continent
Atlantic-Congo	1436	Africa
Austronesian	1274	Greater New Guinea
Indo-European	581	Eurasia
Sino-Tibetan	486	Eurasia
Afro-Asiatic	373	Africa/Eurasia
Nuclear Trans New Guinea	316	Greater New Guinea
Pama-Nyungan	242	Australia
Otomanguean	179	North America
Austroasiatic	162	Eurasia
Tai-Kadai	96	Eurasia
Dravidian	81	Eurasia
Arawakan	77	South America
Mande	75	Africa
Tupian	71	South America
Uto-Aztecan	69	North America
Central Sudanic	63	Africa
Nuclear Torricelli	55	Greater New Guinea
...	...	...

# Ruhlen (1991)'s Sizes

## Lineage

AUSTRIC  
NIGER-KORDOFANIAI  
AMERIND  
INDO-PACIFIC  
EURASIATIC  
DENE-CAUCASIAN  
AFRO-ASIATIC  
AUSTRALIAN  
NILO-SAHARAN  
ELAMO-DRAVIDIAN  
KHOISAN  
KARTVELIAN  
HURRIAN  
SUMERIAN  
MEROITIC  
ETRUSCAN



Language Families of the World (after Greenberg)



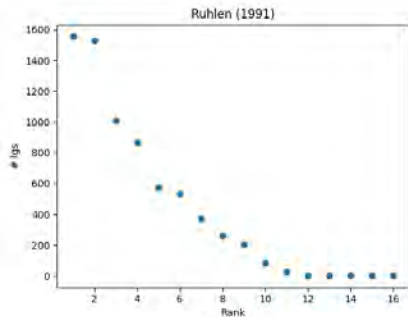
# Family Size Distributions

- If a classification really discovers treetops within a time range  $t$  we should see a size-distribution reflecting the underlying (stochastic) birth-death process
- Typically, e.g., with a Galton-Watson process (Athreya and Ney 1972, Chu and Adami 1999, Watson and Galton 1875), this implies a power-law (aka Zipfian) distribution on the distribution of language family sizes
- If the subgrouping is complete, this can be used to gauge the parameters of the process
- The same power-law should hold in sub-areas which are sufficiently enclosed, old in occupation, and large (e.g., Africa)
- Is this what we find? (cf. Arnold and Bauer 2006, Wichmann 2005)

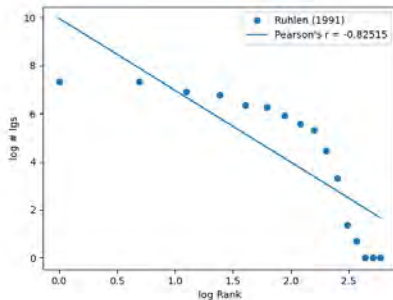


# Ruhlen (1991) Rank-Size Distribution

## Rank-Size



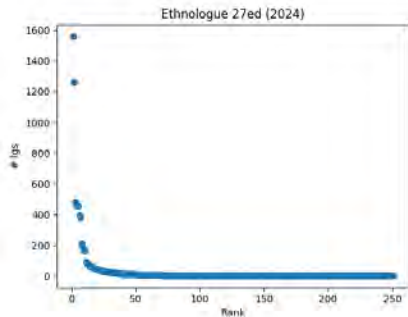
## log Rank-log Size



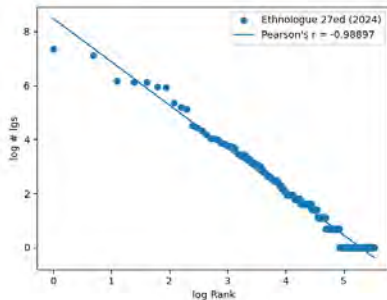
*Does not seem to fit a power-law distribution very well*

# Ethnologue 27ed (2024) Rank-Size Distribution

## Rank-Size



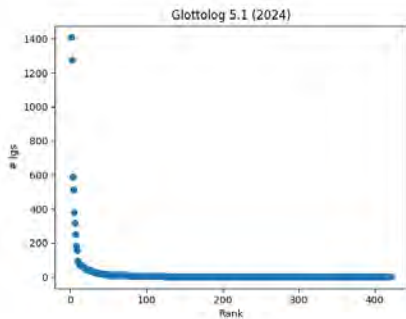
## log Rank-log Size



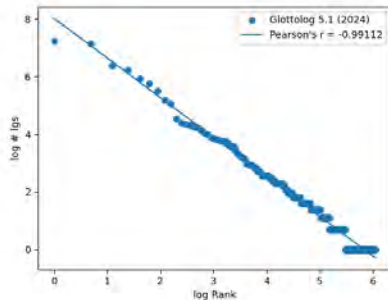
*Fits a power-law distribution very well*

# Glottolog 5.1 (2024) Rank-Size Distribution

## Rank-Size



## log Rank-log Size



*Fits a power-law distribution very well (a slight bit better than E27)*

# Demonstrated Families and Documentation?

- # families should increase as more languages are discovered to the scientific world

*This is a logical necessity, needs no empirical demonstration!*

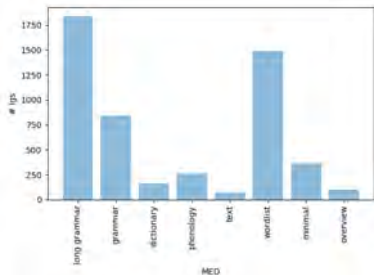
- # families should decrease as we get richer documentation for the already known languages

*This hypothesis can be investigated empirically*

# State of Description of the World's Languages

- MED = Most Extensive Description for each language

	MED type	# lgs	
5	long grammar	1 759	<b>23.1%</b>
4	grammar	859	<b>11.3%</b>
3	grammar sketch	1 969	<b>25.9%</b>
2	specific feature	443	<b>5.8%</b>
2	phonology	266	<b>3.5%</b>
2	dictionary	164	<b>2.1%</b>
2	text	82	<b>1.0%</b>
1	wordlist	1 526	<b>20.0%</b>
0	minimal	410	<b>5.3%</b>
0	overview	120	<b>1.5%</b>
		7 598	



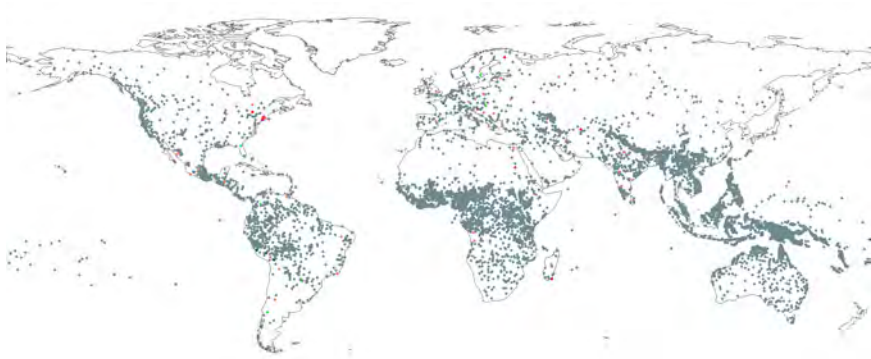
Source: Glottolog 5.1

- The numerical average (0-5) is just above half: 2.84

# Can One Researcher Read It All?

- The MED:s for all languages sum up to 1 692 549 pages
- The average MED is thus 222.8 pages
- At the reading rate of
  - ▶ 250 words per minute
  - ▶ 60 pages per hour
  - ▶ 8 hours a day
- It will take you 3526 days  $\approx$  9.65 years to read all MED:s
- At least if you can access them
- So it is probably possible to obtain single-individual consistency in a world-wide classification

# State of Description in 1700

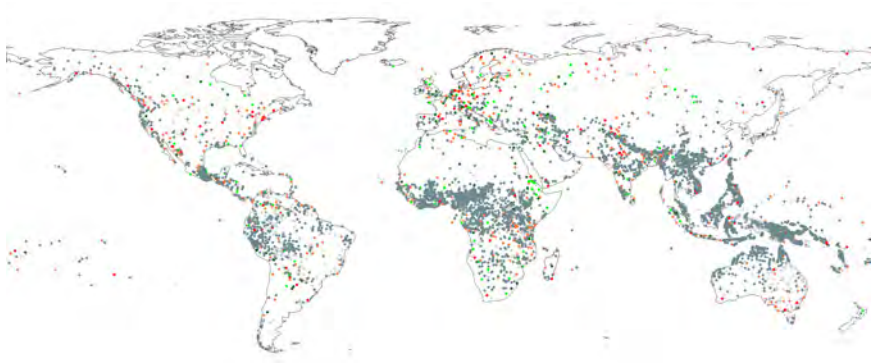


# State of Description in 1850

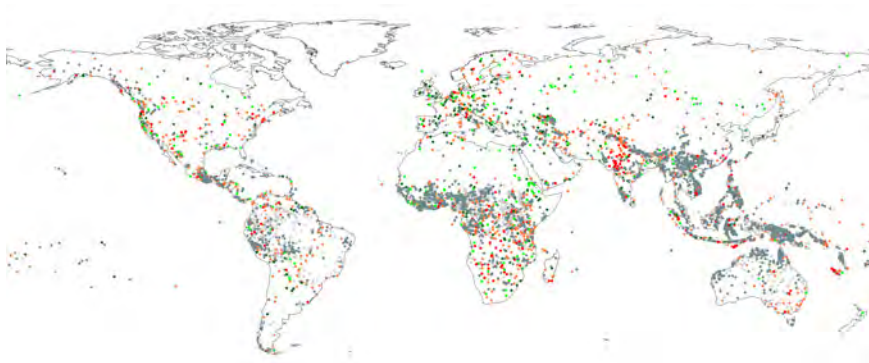




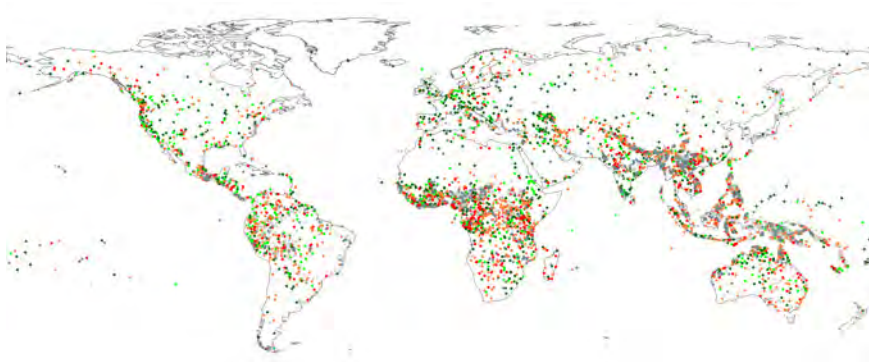
# State of Description in 1900



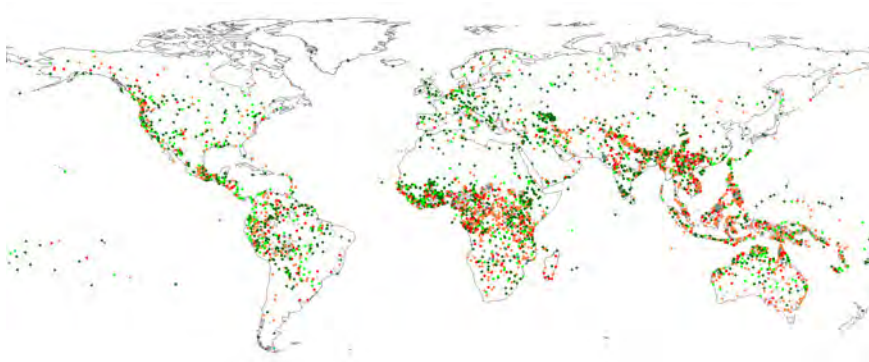
# State of Description in 1950



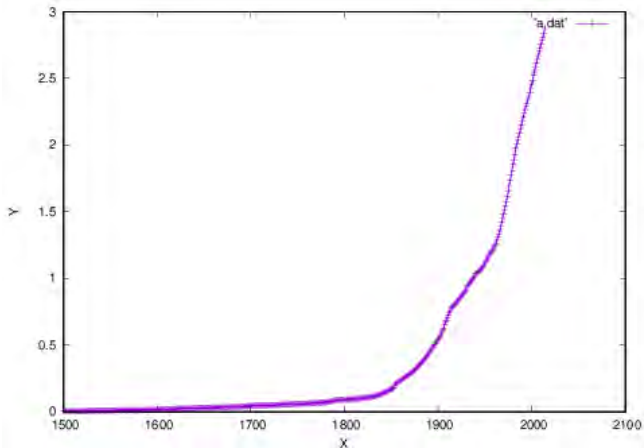
# State of Description in 2000



# State of Description in 2020



# State of Description Across Time



- Since the 50s the rate of new documentation has been essentially constant
  - ▶ A yearly increase in 0.027 “description points” per language
  - ▶  $\approx 40$  long grammar **equivalents** per year
- At this rate, the maximum desc level (4.68) will be reached in 2084

# Poor vs Rich Documentation?

## ● South America

- ▶ Loukotka (1968:15, 29) relied **only on basic vocabulary** inspection for practical reasons: the time limitations of one single human and general lack of more extensive data
- ▶ Loukotka (1968)'s classification is nearly identical to that of Campbell (2012) (as well as G51) despite the appearance of hundreds of South American grammars in the meantime

	Grammar Sketches	(Long) Grammars
1964	94	46
2012	119	233

## ● North America

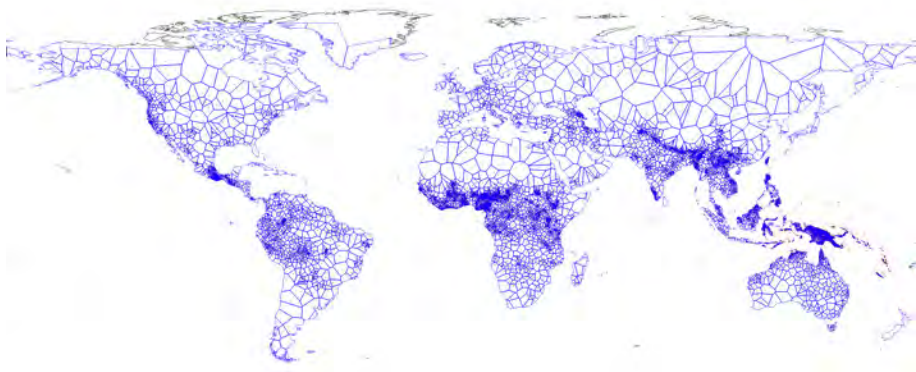
- ▶ Powell (1891:11) relied **only on basic vocabulary** inspection on theoretical grounds (would not have used other data even if available)
- ▶ Powell (1891)'s classification is nearly identical to that of Goddard (1996) (as well as G51) despite a century of additional data and intensive study of historical relationships.

● ...

# Going Beyond?

- Geography?
  - ▶ There were no helicopters in prehistory, so one's closest relatives are likely to be geographically close
- Genetics?
  - ▶ If genes and languages travel together, one should be able to use genetics to gauge deeper
- Archaeology?
  - ▶ Associating proto-languages with archaeological cultures sometimes seems plausible

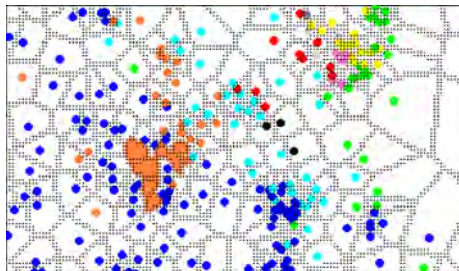
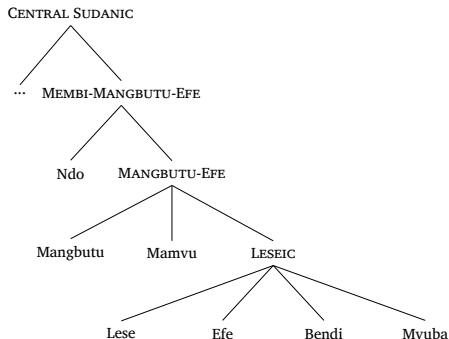
# Geography



- Voronoi Regions from language centre coordinates from Glottolog



# Example



- Probability that one's immediate sister language is a Voronoi neighbour? E.g., is Lese a neighbour of Efe, Bendi or Mvuba?
- Probability that a language from one's immediate clade is a Voronoi neighbour? E.g., is Ndo a neighbour of any of the other languages?
- Probability that a language from one's family is a Voronoi neighbour? E.g., is Ndo a neighbour of any Central Sudanic (orange) language?

## Related Languages as Geographical Neighbours

- Probability that one's immediate sister language is a Voronoi neighbour?

$$\frac{4155}{6225} \approx 54\%$$

- Probability that a language from one's immediate clade is a Voronoi neighbour? E.g., is Ndo a neighbour of any of Mangbutu-Efe languages?

$$\frac{5419}{7665} \approx 71\%$$

- Probability that a language from one's family is a Voronoi neighbour? E.g., is Ndo a neighbour of any Central Sudanic (yellow) language?

$$\frac{7141}{7665} \approx 93\%$$

# Use Geographical Prior?

- The geographical prior seems very strong
- However:
  - ▶ Is the sister from  $t \approx 10,000$  years ago still around?
    - ★ Only (approximately) 422 languages from  $t \approx 10,000$  years ago survived to be reflected today
    - ★ Let's say 7665 languages were spoken at  $t \approx 10,000 \rightarrow$  only  $\frac{422}{7665} \approx 5.5\%$  chance of survival
  - ▶ We usually do not know the location of a (proto-)language  $t \approx 10,000$  years ago, but have to estimate with uncertainties

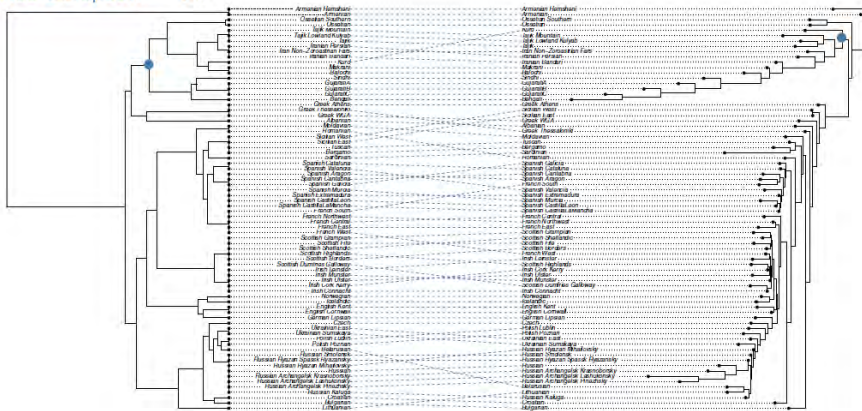
# Genes vs Languages

- Sometimes genes and languages travel together, sometimes not
- To what degree?
- GeLaTo (Barbieri et al. 2022) dataset has
  - ▶ 4,000 individuals representing
  - ▶ 397 genetic populations speaking
  - ▶ 295 languages
- Does closest genetic relative match closest linguistic relative (within the dataset?)

# Genetic vs Linguistic Relatedness: Indo-European

Indo-European

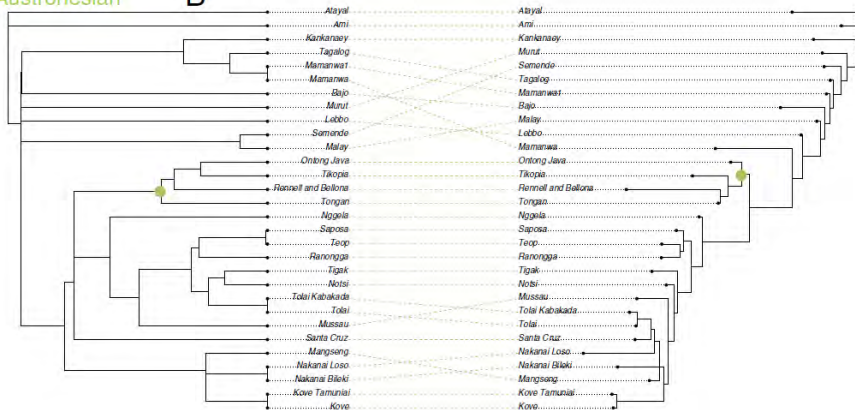
A



# Genetic vs Linguistic Relatedness: Austronesian

Austronesian

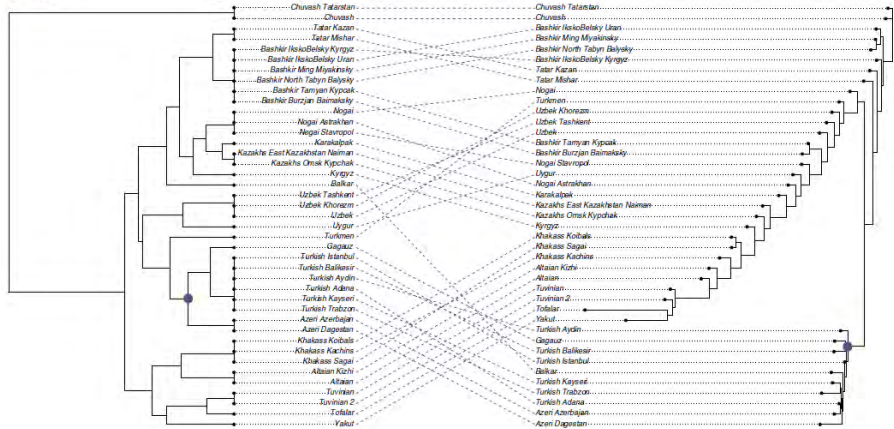
B



# Genetic vs Linguistic Relatedness: Turkic

Turkic

C



# Genes vs Languages

- Sometimes genes and languages travel together, sometimes not
- To what degree?
- GeLaTo (Barbieri et al. 2022) dataset has
  - ▶ 4,000 individuals representing
  - ▶ 397 genetic populations speaking
  - ▶ 295 languages
- Does closest genetic relative match closest linguistic relative (within the dataset?)

**Across the whole dataset the closest genetic relative belong to the same language family in 82% of the cases!**



# Proto-Languages and Archaeological Correlates

Node	Timespan	Source
Indo-Iranian	2000-1900 BC	Lubotsky 2023:259-262
Romance	1850-1650 BP	Chang et al. 2015:223, 226; Embleton 1991:381
Eastern Romance	900-1200 AD	Sala 2010:855
Permian	1100 BP	Maurits et al. 2020:15-16
Sinitic	2700 BP	Zhang et al. 2019:S2:16
Middle Old Tibetan	1150 BP	Zhang et al. 2019:S2:16
Aceh-Chamic	200 BC-0 AD	Brunelle 2019
Cholan	1600 BP	Holman et al. 2011:6
East Polynesian	1050 BP	Holman et al. 2011:6, 23
Southeast Barito	1300-1400 BP	Adelaar 1989
Mongolic	1100-1300 AD	Robbeets et al. 2020:759
Romani	1200 AD	Benisek 2020:18, Matras 2002:43-48
Northern Songhay	650-1300 AD	Souag 2012:204-208
Greenlandic Inuit	1350 AD	Bergsland and Vogt 1962:127
Japonic	300 BC	Miyake 2020:11
Tungusic	600 BC-200 AD	Robbeets et al. 2020:763-763
NE Coastal Bantu	100 AD	Walsh 2017:122
Quechuan	2000-1500 BP	Beresford-Jones and Heggarty 2013
Nikio	700-800 AD	Horton 1998:233
Dhivehi-Sinhala	100 BC - 0 BC	Cain 2000
Oriya-Gauda-Kamrupa	700 AD	Toulmin 2006:292
Bukatanic	250-350 BP	Smith and Rama 2022:5
Ha-Ya	400 BP	Bradley 2022:191
Tanalaric	1580-1707 AD	Bischoff et al. 2016:275-276



# Proto-Languages and Archaeological Correlates

- I curate a “decent” list, currently some 126 cases
  - ▶ Not known how “complete” this is
- Association typically arguable
- Pinpointing the association to a specific node less so, especially if we take the extinct sisters issue seriously
- What can we do with such lists?
  - ▶ No trees over archaeological cultures
  - ▶ No full-ish global database on archaeological sites

# Conclusions

- There are reasons to believe any consistent language classification into family a should exhibit power-law properties
- The number and extent of language families is less dependent on rich documentation than what one might believe
- Geography indeed appears to be a good prior for linguistic relatedness
- Genes indeed appears to be a good prior for linguistic relatedness  
→ bigger databases needed!
- How can we use archeology to guess deeper?

- Arnold, R. and Bauer, L. (2006). A note regarding 'on the power-law distribution of language family sizes'. *Journal of Linguistics*, 42:373–376.
- Athreya, K. B. and Ney, P. E. (1972). *Branching Processes*. Berlin: Springer.
- Barbieri, C., Blasi, D. E., Arango-Isaza, E., Sotiropoulos, A. G., Hammarström, H., Wichmann, S., Greenhill, S. J., Gray, R. D., Forkel, R., Bickel, B., and Shimizu, K. K. (2022). A global analysis of matches and mismatches between human genetic and linguistic histories. *Proceedings of the National Academy of Sciences*, 119(47(e2122084119)):1–9, S1–S42.
- Campbell, L. (1999). *Historical Linguistics: An Introduction*. Cambridge, Massachusetts: MIT Press.
- Campbell, L. (2004). *Historical Linguistics: An Introduction*. Edinburgh: Edinburgh University Press, 2 edition.
- Campbell, L. (2012). Classification of the indigenous languages of south america. In Campbell, L. and Grondona, V., editors, *The Indigenous Languages of South America: A Comprehensive Guide*, volume 2 of *The World of Linguistics*, pages 59–166. Berlin: Mouton.

- Campbell, L. (2020). *Historical Linguistics: An Introduction*. Edinburgh: Edinburgh University Press, 4 edition.
- Chu, J. and Adami, C. (1999). A simple explanation for taxon abundance patterns. *Proceedings of the National Academy of Sciences*, 96:15017–15019.
- Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2024). *Ethnologue: Languages of the World*. Dallas: SIL International, 27 edition.
- Goddard, I. (1996). The classification of the native languages of north america. In Goddard, I., editor, *Languages*, volume 17 of *Handbook of North American Indians*, pages 290–324. Washinton, D.C.: Smithsonian Institution, Washinton, D.C.
- Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2024). Glottolog 5.1. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at <http://glottolog.org>. Accessed on 2024-11-15.
- Hublin, J.-J., Ben-Ncer, A., Bailey, S. E., Freidline, S. E., Neubauer, S., Skinner, M. M., Bergmann, I., Cabec, A. L., Benazzi, S., Harvati, K., and Gunz, P. (2017). New fossils from jebel irhoud, morocco and the pan-african origin of homo sapiens. *Nature*, 546:289–292.  

- Powell, J. W. (1891). Indian linguistic families. In *Seventh Annual Report of the Bureau of Ethnology to the Secretary of the Smithsonian Institution, 1885-1886*, pages 1–142. Washington: Government Printing Office, Washington.
- Ruhlen, M. (1991). *A guide to the world's languages. Vol. 1, Classification with a postscript on recent developments*. Stanford: Stanford University Press, Stanford.
- Watson, H. W. and Galton, F. (1875). On the probability of extinction of families. *Journal of the Anthropological Institute of Great Britain and Ireland*, 4:138–144.
- Wichmann, S. (2005). On the power-law distribution of language family sizes. *Journal of Linguistics*, 41:117–131.
- Čestmír Loukotka (1968). *Classification of the South American Indian Languages*, volume 7 of *Reference Series*. Los Angeles: Latin American Center, University of California.