Beyond cognacy:

Building the Lexibank World Tree

Gerhard Jäger, University of Tübingen

Leipzig, June 11, 2025

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

phylogenetic linguistics

- main goal: infer **phylogenetic trees** from **lexical data**

input

Glottocode	Glottolog_Name	Concepticon_Gloss	Segments	Cognateset_ID
homs1234	Western Armenian	EAR	ɑ g ɑ n dʒ	208
nucl1235	Eastern Armenian	EAR	ɑ k ɑ n dʒ	208
beng1280	Bengali	EAR	k a n	768
bulg1262	Bulgarian	EAR	u x ɔ	208
bela1254	Belarusian	EAR	v u x a	208
stan1289	Catalan	EAR	u r ɛ ʎ ə	208
czec1258	Czech	EAR	ʔ u x ɔ	208
dani1285	Danish	EAR	øː ʌ	208
dutc1256	Dutch	EAR	o r	208
faro1244	Faroese	EAR	ɔ iː ɹ a	208

phylogenetic linguistics

intermediate step

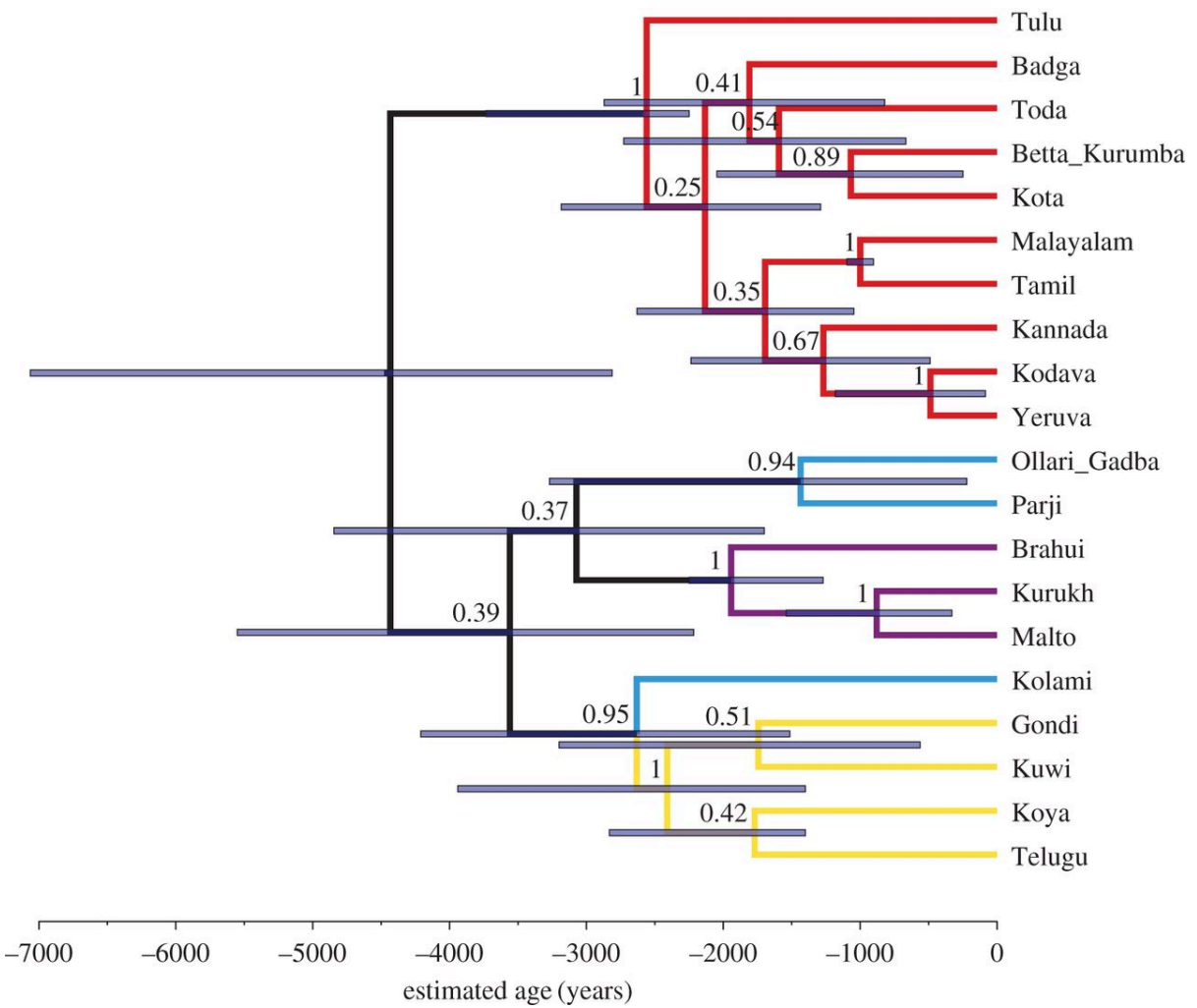
convert cognate classification into binary character matrix

```
TNG.ENGAN.MAIBI                 1000000000000000000000000000000000000000000000
TNG.ENGAN.POLE                  0000000000000000000000000000000000000010000000
TNG.ENGAN.SAU                   0000000000000000000000000000000000000000000000
TNG.ENGAN.YARIBA                1000000000000000000000000000000000000000000000
TNG.FASU.FASU                   0000000000000000000000000000000000000010000000
TNG.FASU.NAMUMI                 0000000000000000000000000000000000000000001000
TNG.FINISTERRE-HUON.AWARA       0000000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.BORONG      0000000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.BURUM       0000000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.BURUM_MIND  0000000000000000000000000000000000001010000000
TNG.FINISTERRE-HUON.DEDUA       0000000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.HUBE        0000000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.KATE        0000000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.KOMBA       0000000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.KOSORONG    0000000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.MAPE        0000000000000000000000000000000000001000000000
TNG.FINISTERRE-HUON.MAPE_2      0000000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.MIGABAC     0000000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.MINDIK      0000000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.MOMOLILI    0000000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.NABAK       0000000000000000000000000000000000001000000000
TNG.FINISTERRE-HUON.NANKINA     0000000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.NEK         0000000000000000000000000000000000000001000000
TNG.FINISTERRE-HUON.NUKNA       0000000000000000000000000000000000000001000000
TNG.FINISTERRE-HUON.ONO         0000000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.SELEPET     0000000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.TIMBE       0000000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.TOBO        0000000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.WANTOAT     0000000000000000000000000000000000000110000000
TNG.FINISTERRE-HUON.YOPNO       0000000000000000000000000000000000000001000000
TNG.GOILALAN.AFOA               0000000000000000000000000000000000000110000000
TNG.GOILALAN.KUNIMAIPA          0000000000000000000000000000000000000010000000
TNG.GOILALAN.MAFULU             0000000000000000000000000000000000000010000000
```

phylogenetic linguistics

- main goal: infer **phylogenetic trees** from **lexical data**

output



(source: Kolipakam et al. 2018, https://doi.org/10.1098/rsos.171504)

applications

- control for common ancestry in statistical models *(Dunn et al. 2011, Jäger & Wahle 2021, Skirgård et al. 2023, …)*
- estimate time depth and geographic location of ancestral populations *(Bouckaert et al 2012)*
- track cultural change using language phylogeny as scaffold *(Watts et al 2015, …)*
- reconstruct properties of ancestral populations *(Cathcart et al 2021, Carling & Cathcart 2021a,b, …)*
- statistic identification of patterns of language change *(Blasi et al. 2019)*
- …

from word lists to trees

1. perform cognate classification (manual or automatic)
2. construct binary *character matrix*
3. let computer search the tree(s) that best explain(s) the distribution of 0s and 1s in the character matrix

manual cognate annotation

manual cognate annotation

- labor intensive
- available data are geographically skewed
- requires tons of prior classical historical linguistics work

automatic cognate detection

- lot of computational research over the past years to automate the process
- some relevant papers: Hauer & Kondrak (2011), List (2014), Rama (2015), Jäger, List & Sofroniev (2017)
- results are usable but far from perfect

phylogenetic signal below cognacy

- sound change and morpological change contains relevant phylogenetic information

Glottocode	Glottolog_Name	Concepticon_Gloss	Segments	Cognateset_ID
nucl1235	Eastern Armenian	TOOTH	ɑ t ɑ m	328
beng1280	Bengali	TOOTH	d̪ ã t̪	328
stan1289	Catalan	TOOTH	d e n	328
dani1285	Danish	TOOTH	t a nˀ	328
dutc1256	Dutch	TOOTH	t ɑ n t	328
faro1244	Faroese	TOOTH	tʰ ɔ nː	328
stan1290	French	TOOTH	d ɑ̃	328
stan1295	German	TOOTH	tsʰ aː n	328
mode1248	Modern Greek	TOOTH	ð o̱ n d i	328
hind1269	Hindi	TOOTH	d̪ ɑ̃ t̪	328

Multiple Sequence Alignment for phylogenetic inference

- Akavarapu and Bhattacharya (2024):
  - apply Multiple Sequence Alignment (MSA) to all reflexes for a given concept, regardless of cognacy
  - treat MSA columns as characters for phylogenetic inference

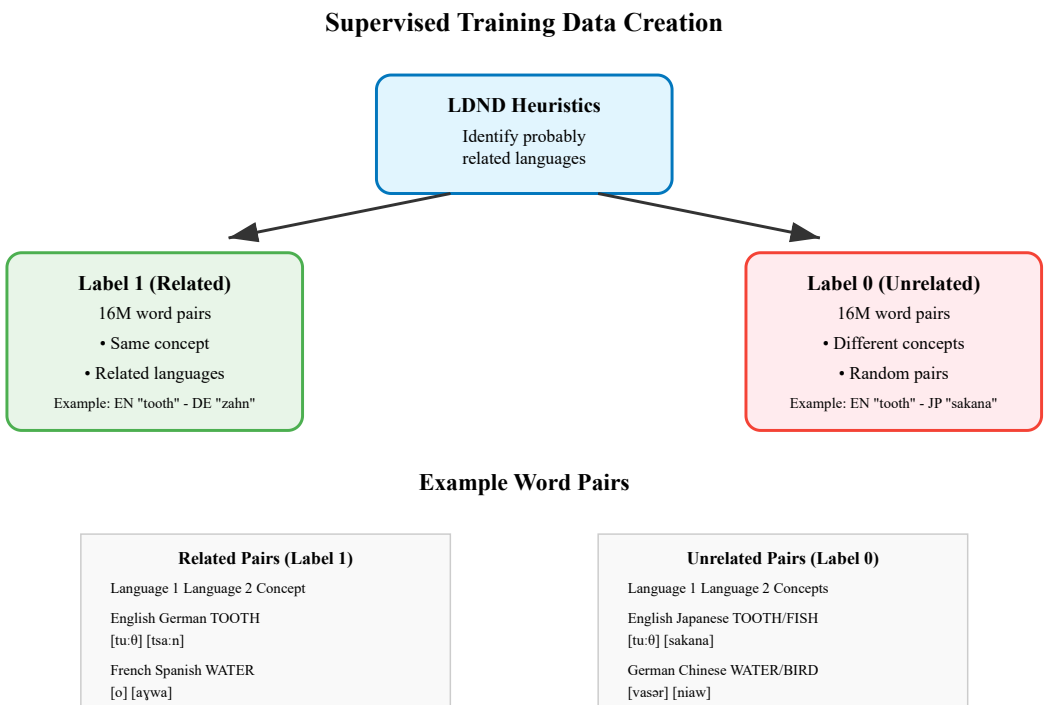
Name	Cognateset_ID	1	2	3	4	5	6	7	8
Even	16_lousen-37	k	-	u	-	m	-	k	e
Kilen	16_lousen-37	q	h	u	-	m	I	k	I
Negidal	16_lousen-37	k	-	u	-	m	-	k	I
Oroch	16_lousen-37	k	-	u	-	m	-	-	I
Udihe	16_lousen-37	k	-	u	-	m	u	x	I
Nanai	16_lousen-38	t	-	i	k	t	-	-	I
Orok	16_lousen-38	t	-	i	k	t	-	-	I
Ulch	16_lousen-38	t	-	i	q	t	-	-	I

Name	sound class	k	q
Even	0	0	0
Kilen	0	0	0
Negidal	0	0	0
Oroch	0	0	0
Udihe	0	0	0
Nanai	1	1	0
Orok	1	1	0
Ulch	1	0	1

Table 1: Example of a multiple sequence alignment. Alignment cells are shaded to indicate different cognate sets. (left) Binarized version of column 4. (right)

Workflow for this paper

pairwise sequence alignment

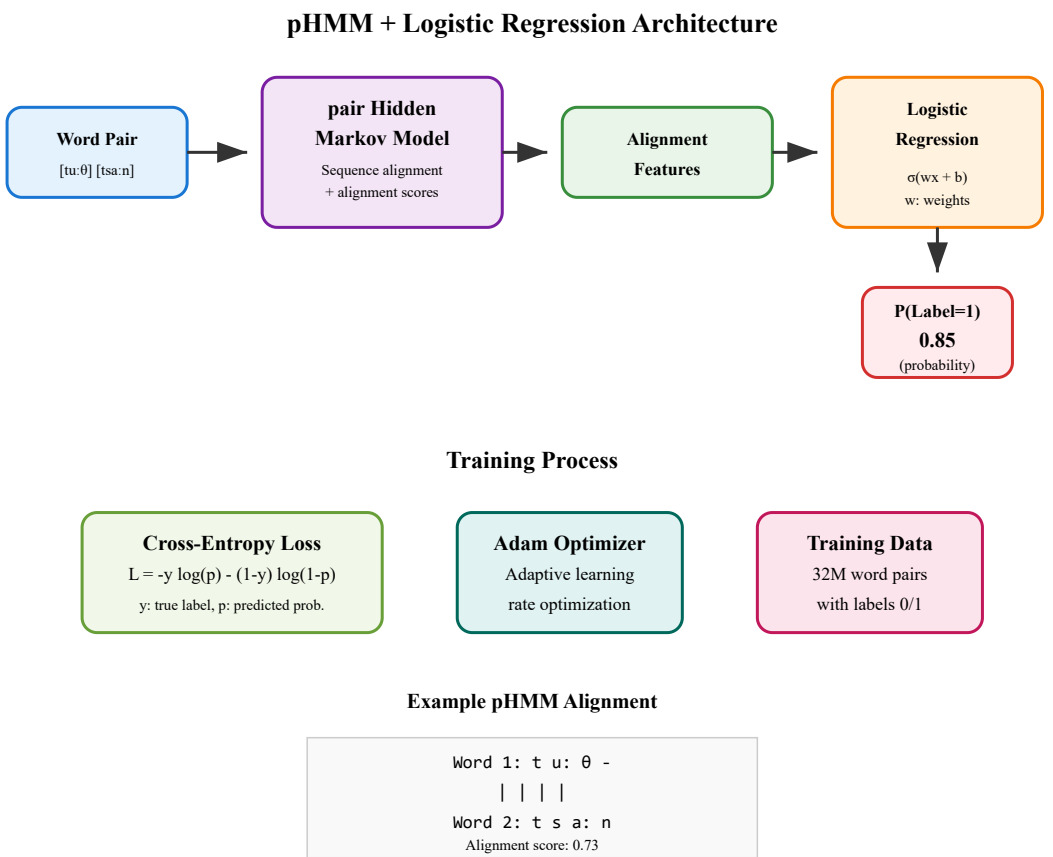


- convert all segments to ASJP
- pairwise sequence alignment via *pair Hidden Markov Model* (pHMM)
- supervised training:
  - use simple heuristics (LDND) to identify *probably related* languages
  - extract 16,000,000 word pairs that
    - share a Concepticon gloss
    - come from probably related languages.
      These pairs receive a label of 1.
  - extract the same number of random word pairs *not* sharing a Concepticon gloss.

    These pairs receive a label of 0.

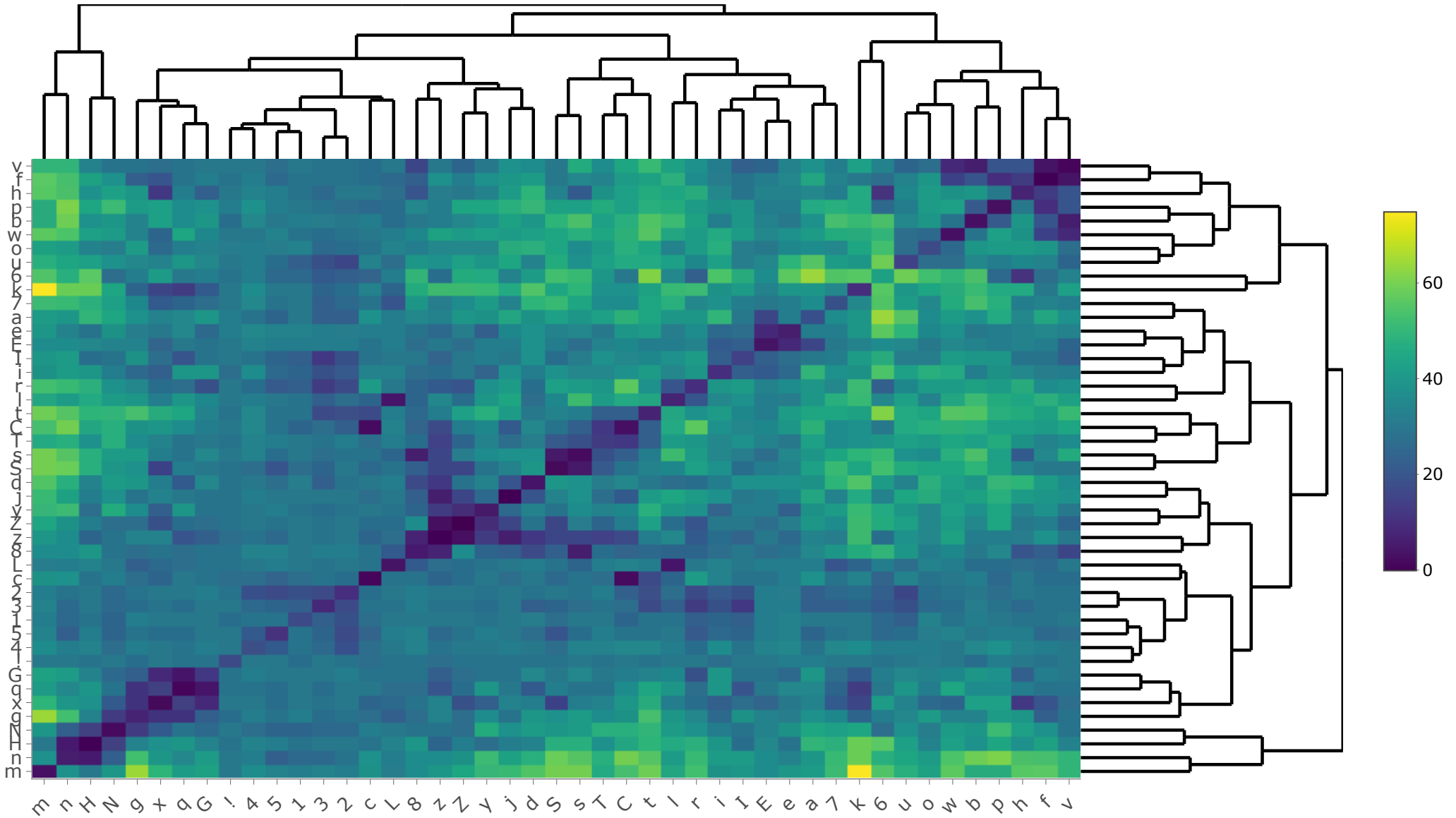
Workflow for this paper

pairwise sequence alignment



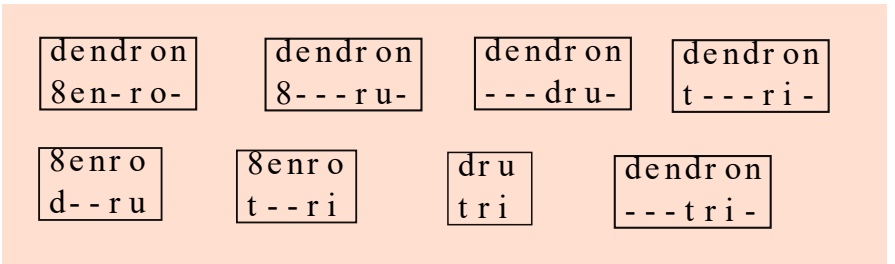
- pHMM + logistic regression
  - input: word pair
  - output: probability of label 1
- loss function: cross-entropy loss
- training with Adam optimizer

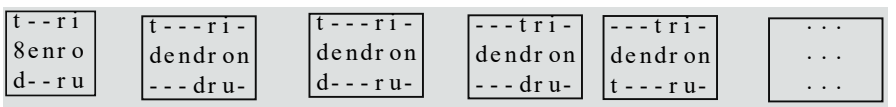
Trained model

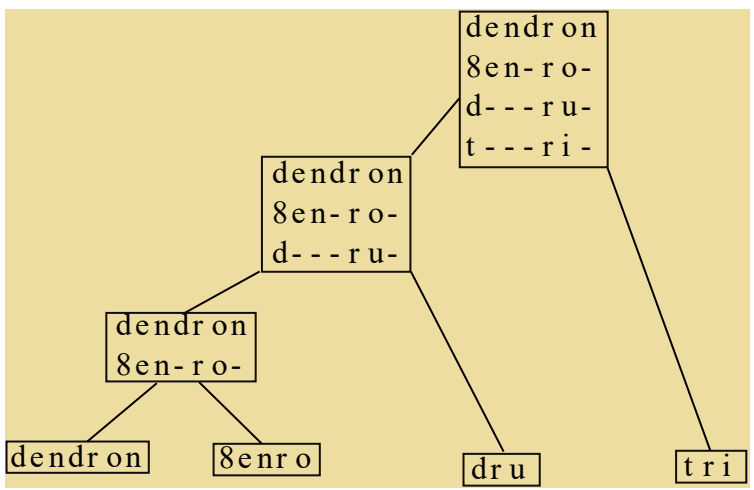


T-Coffee

1. Pairwise alignment for all word pairs using Viterbi/pHMM
2. Ternary alignments via relation composition
3. Indirect alignment scores between sound **occurrences**
4. Progressive alignment using those scores







Pilot study

- Lexibank: 35 Lexibank datasets
- Grambank: 35 5097 binary entries from 2 467 languages and 195 typological features
- subset selection:
  - only glottocodes present in both datasets
  - entries for Concepticon_Gloss and Cognateset_ID
  - Concepticon_Gloss from 110 concepts with largest coverage
- 113,671 entries from 928 languages remaining

Phylogenetic inference

- three types of character matrices:
    1. binarized Cognate_IDs
    2. automatic cognate clustering + soundclass-concept characters (Jäger 2018)
    3. binarized MSA as described above
- maximum-likelihood phylogenetic inference via *raxml-ng* (Kozlov et al. 2019)

Evaluation

- three types of evaluation:
    1. Compare the inferred phylogenetic tree with the Glottolog expert classification.
    2. Compare the inferred phylogenetic tree with the Grambank typological features.
    3. Assess the phylogenetic difficults of the data ($\approx$ inversely related to strength of phylogenetic signal) via **pypythia** (Haag et al. 2022).

What does difficult mean?

Difficulty = ruggedness of the tree space

- Few highly similar tree topologies
- Single likelihood peak

- Highly distinct topologies, statistically indistinguishable
- Multiple likelihood peaks

Source: https://cme.h-its.org/exelixis//pubs/pythia_erga_nov_2023.pdf

Results: full dataset

Method	GQD (Glottolog)	AIC (Grambank)	Difficulty
Cognate classes	0.188	105,340	0.59
PMI	0.062	104,903	0.63
MSA	**0.042**	**104,752**	**0.45**

Results: 14 largest families

μ: sample mean σ: sample standard deviation

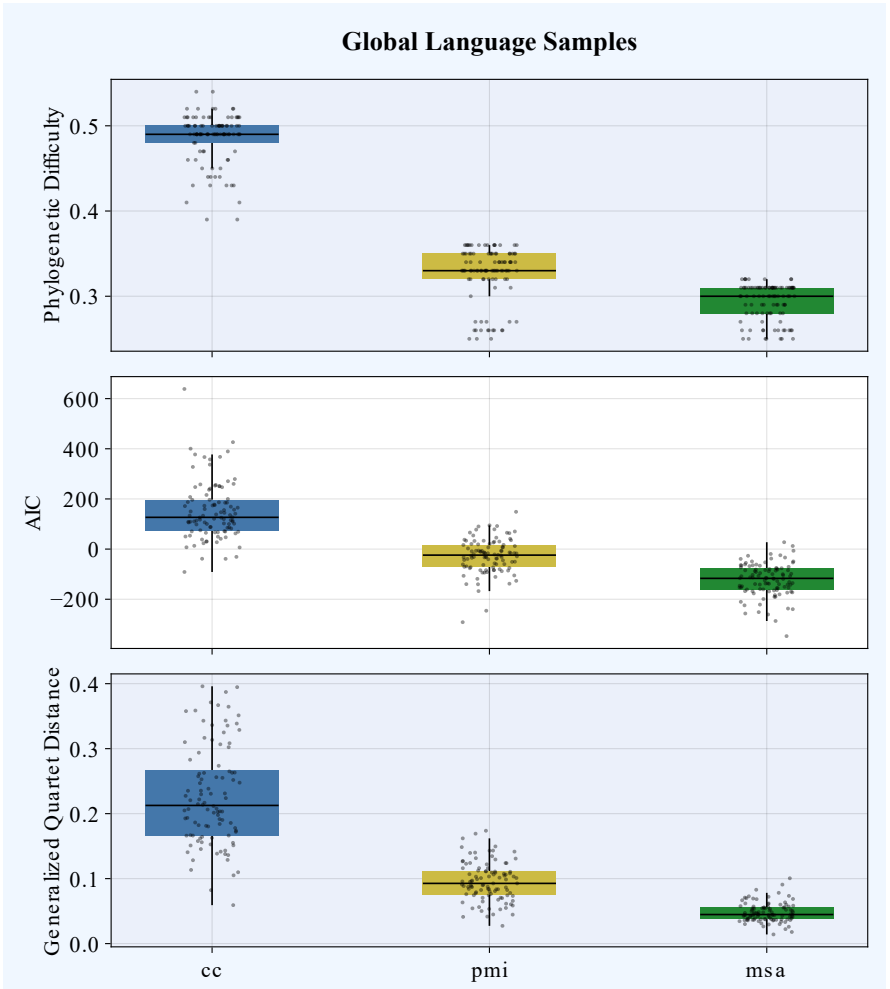
Method	μ GQD	σ GQD	μ AIC	σ AIC	μ Difficulty	σ Difficulty
Cognate classes	0.223	0.130	**−1.73**	17.01	0.401	0.164
PMI	0.221	0.109	3.42	20.43	0.280	0.187
MSA	**0.218**	0.109	−1.69	14.30	**0.203**	0.159

Results: 100 random samples, each containing 100 languages

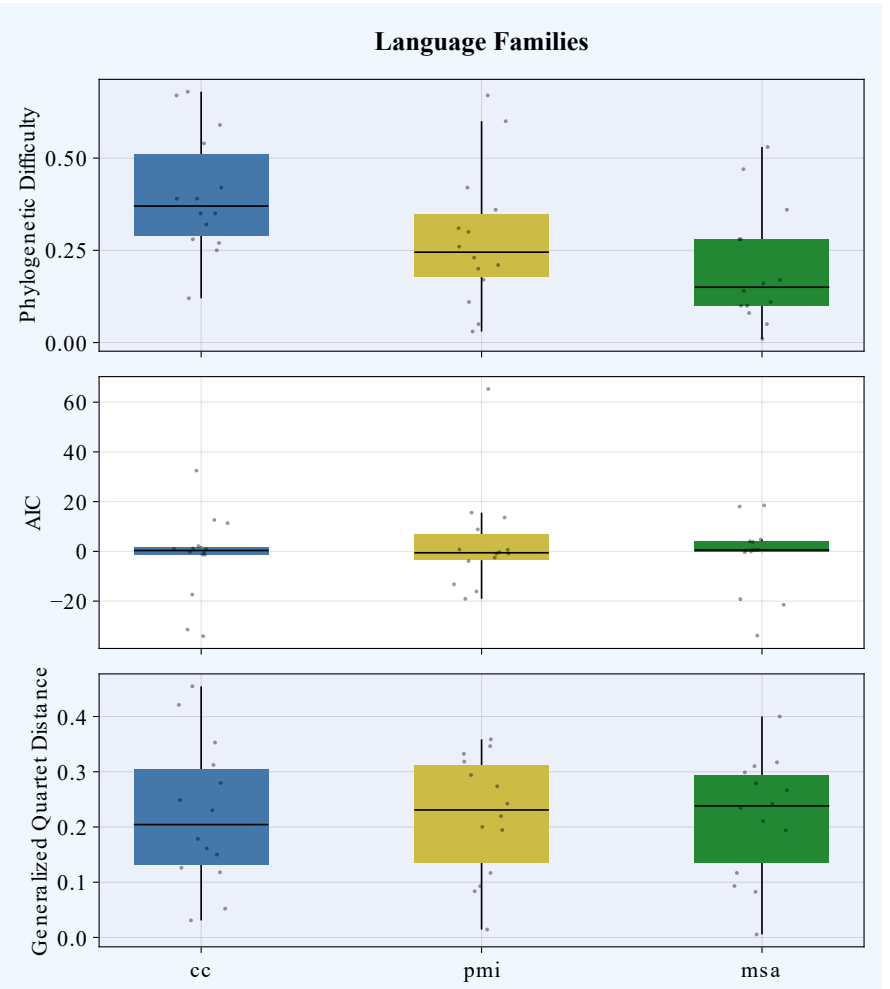
μ: sample mean σ: sample standard deviation

Method	μ GQD	σ GQD	μ AIC	σ AIC	μ Difficulty	σ Difficulty
Cognate classes	0.227	0.077	151	115	0.486	0.030
PMI	0.095	0.030	–28	69	0.326	0.032
MSA	**0.048**	0.015	**–123**	66	**0.294**	0.021

Evaluation Summary



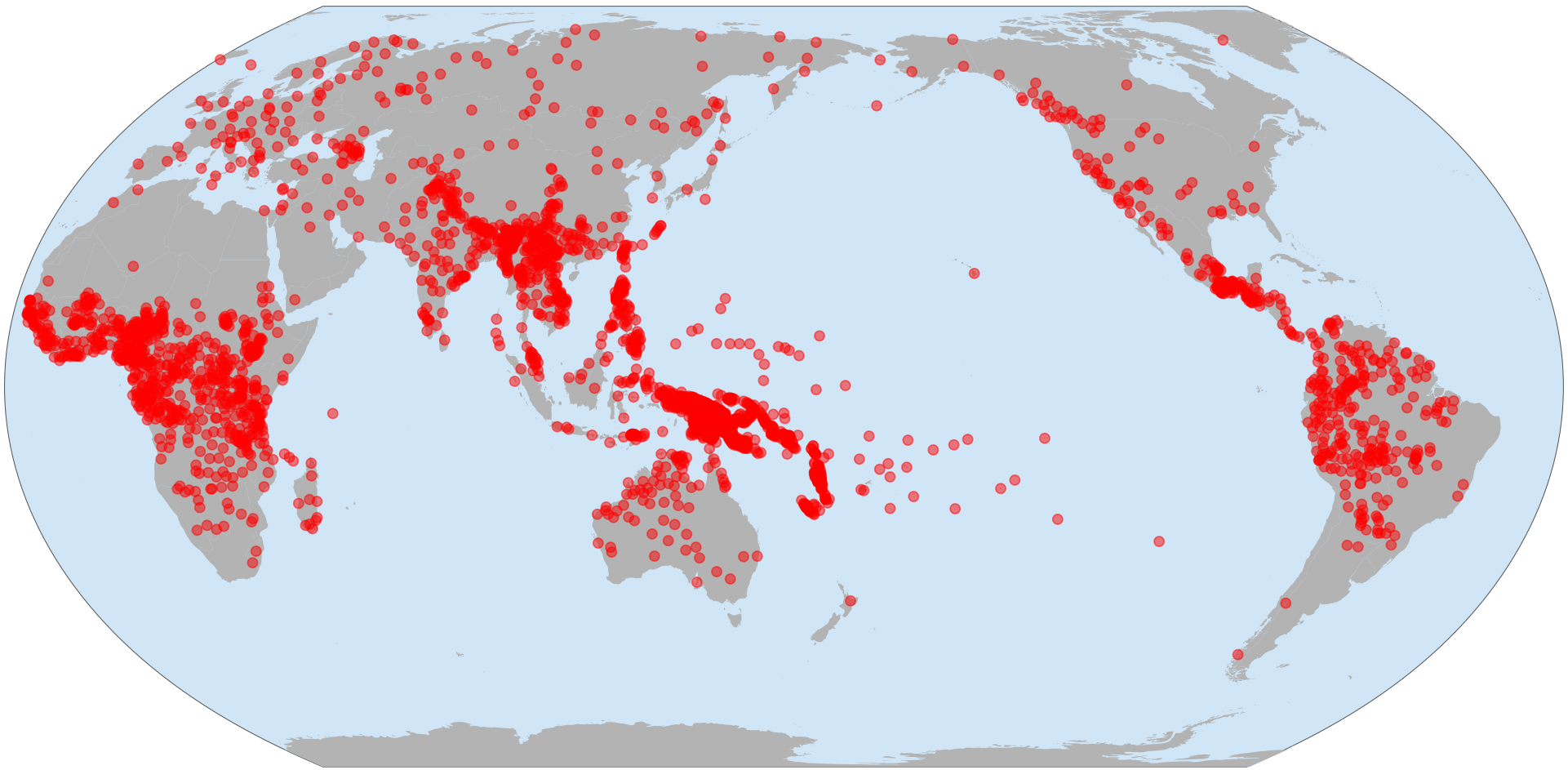
Left panel: Comparison of methods across three evaluation metrics for the 100 random samples. The boxplots show distribution per method, while the overlaid points represent individual samples.



Right panel: Comparison of methods across three evaluation metrics for the 14 largest language families. The boxplots show distribution per method, while the overlaid points represent individual samples.

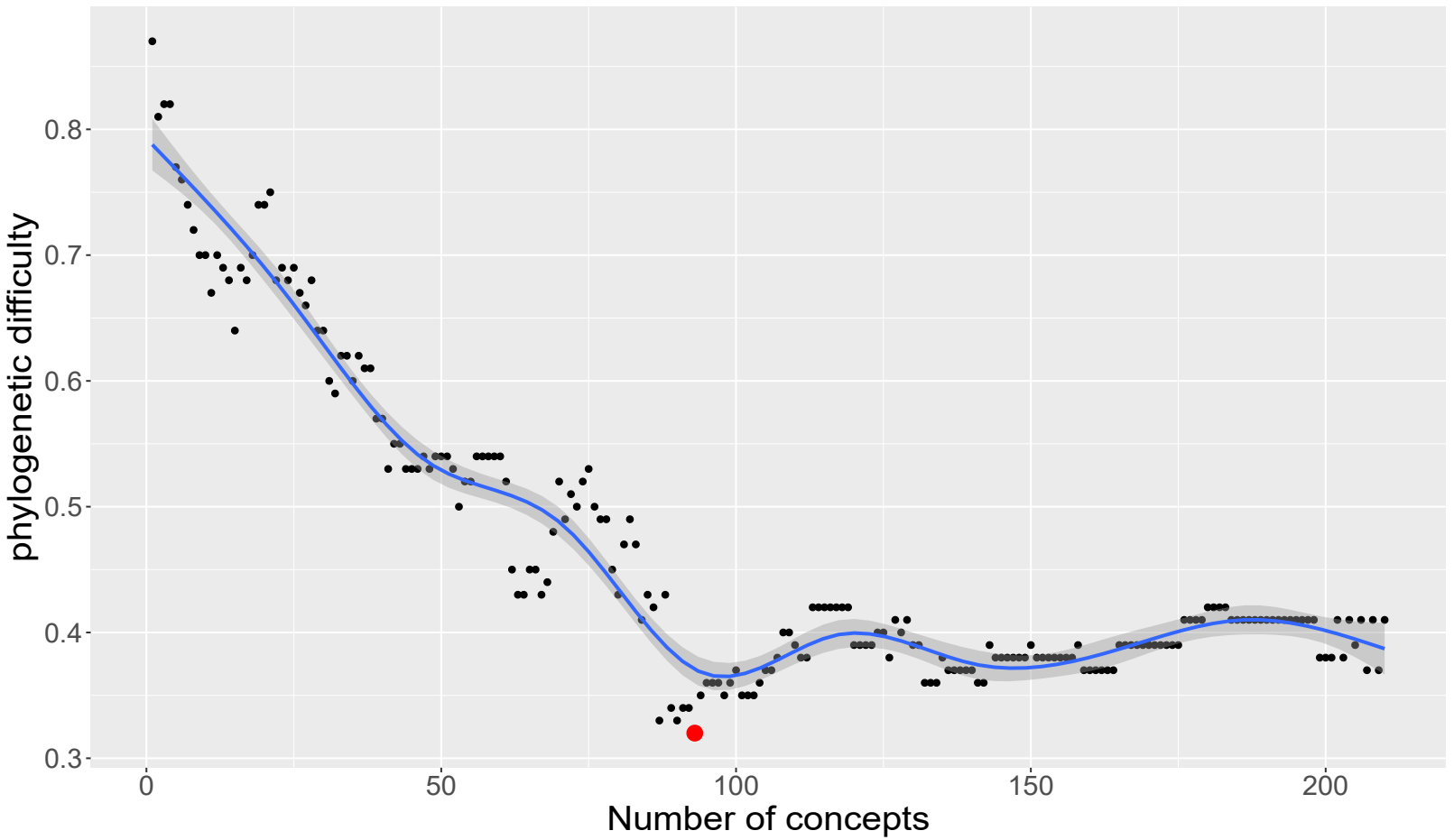
Building the Lexibank World Tree

- Lexibank data as of March 25, 2025
- doculects identified via Glottocode
- only doculects that are extant spoken L1 languages according to Glottolog
- only entries with a Concepticon_Gloss within the 210 most stable concepts according to Dellert & Buch (2018)
- results in
  - 76,4531 entries
  - from 3,397 Glottocodes
  - from 123 Lexibank datasets
  - from 153 families + 142 isolates



How many concepts do we really need?

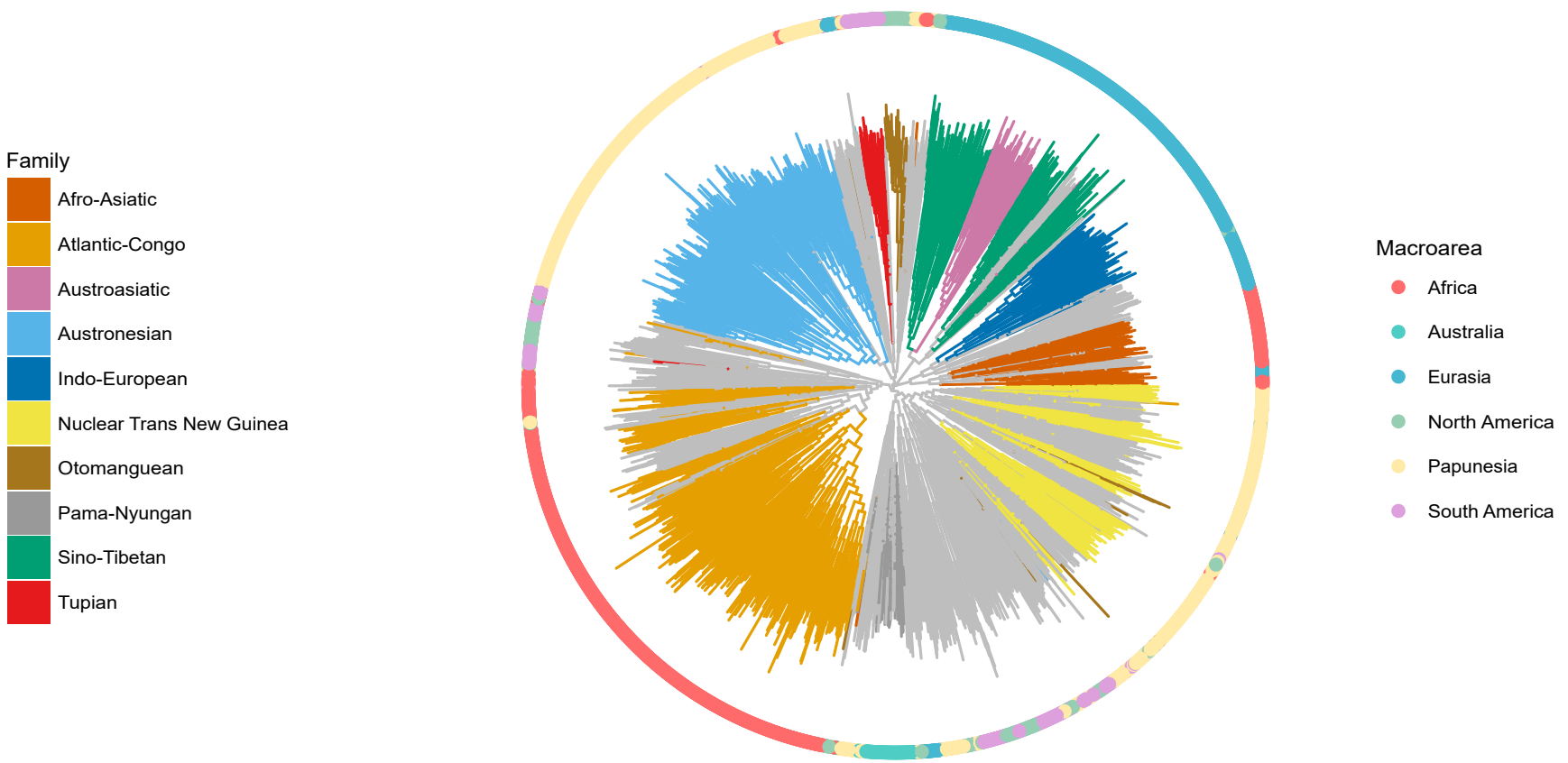
- ordering concepts according to Dellert stability



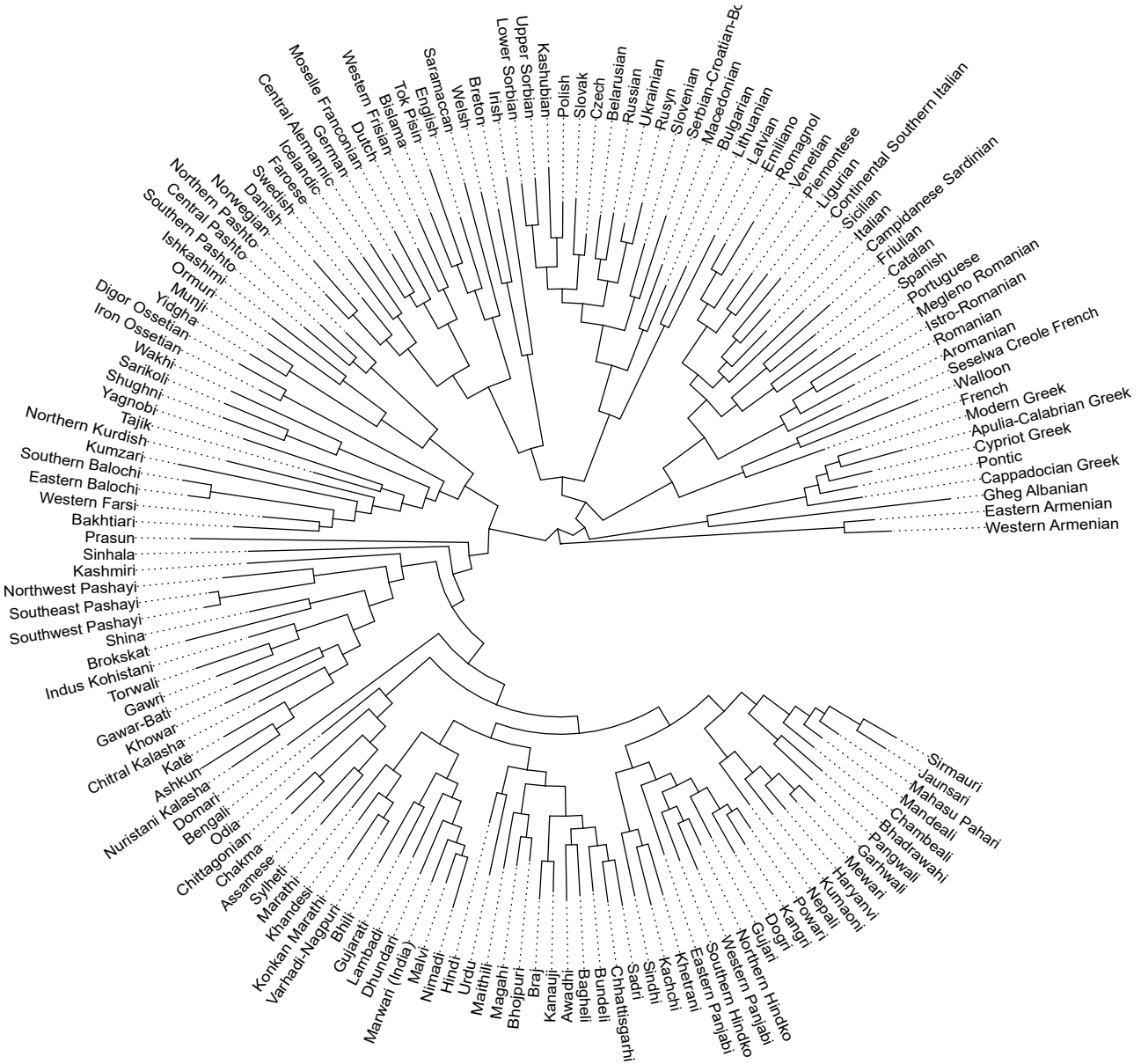
- Lowest difficulty is achieved with 93 concepts
- 76,909 remaining characters in total – most of them very sparse

The Lexibank World Tree

Generalized Quartet Distance (GQD) to Glottolog expert classification: 0.038



Indoeuropean Subtree



Conclusion

- MSA-based phylogenetic inference outperforms cognate-based methods
- MSA-based phylogenetic inference is feasible for large datasets
- Lexibank World Tree is a first step towards a large-scale phylogenetic tree of the world’s languages

Downsides

- branch lengths are not meaningful

Todo

- assess branch support
- filter out non-informative characters