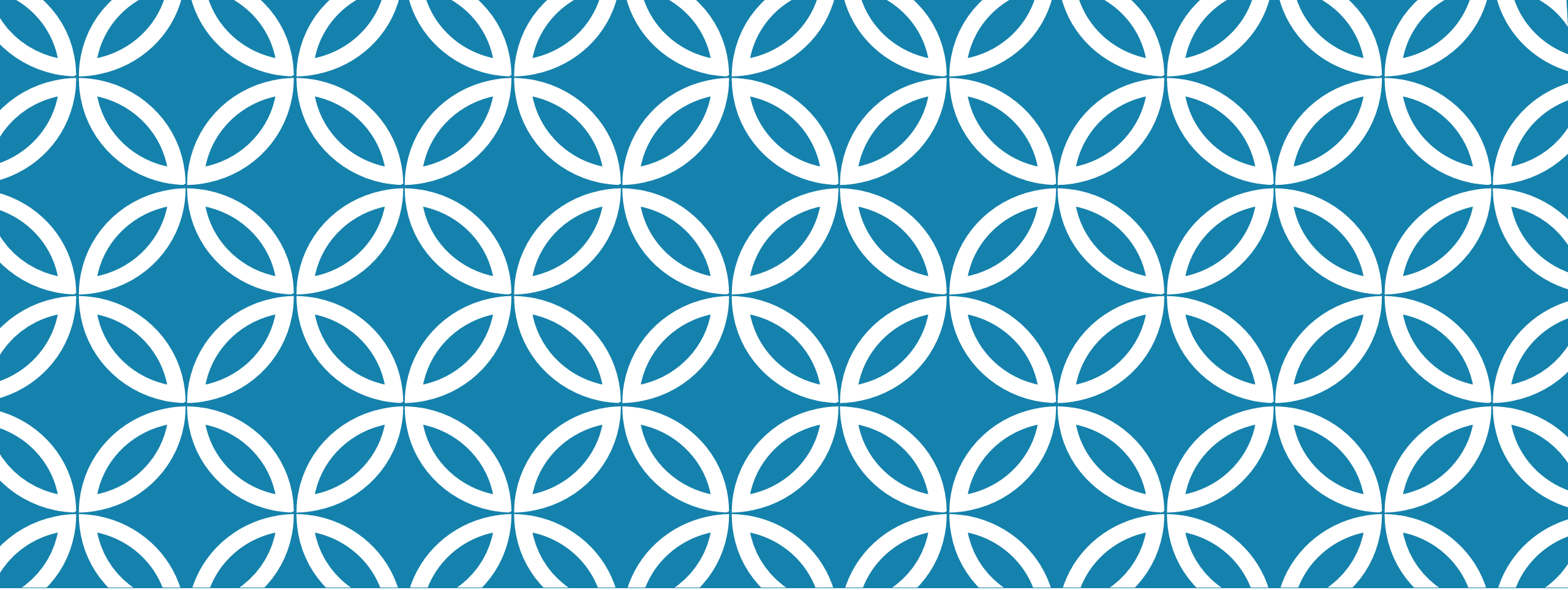
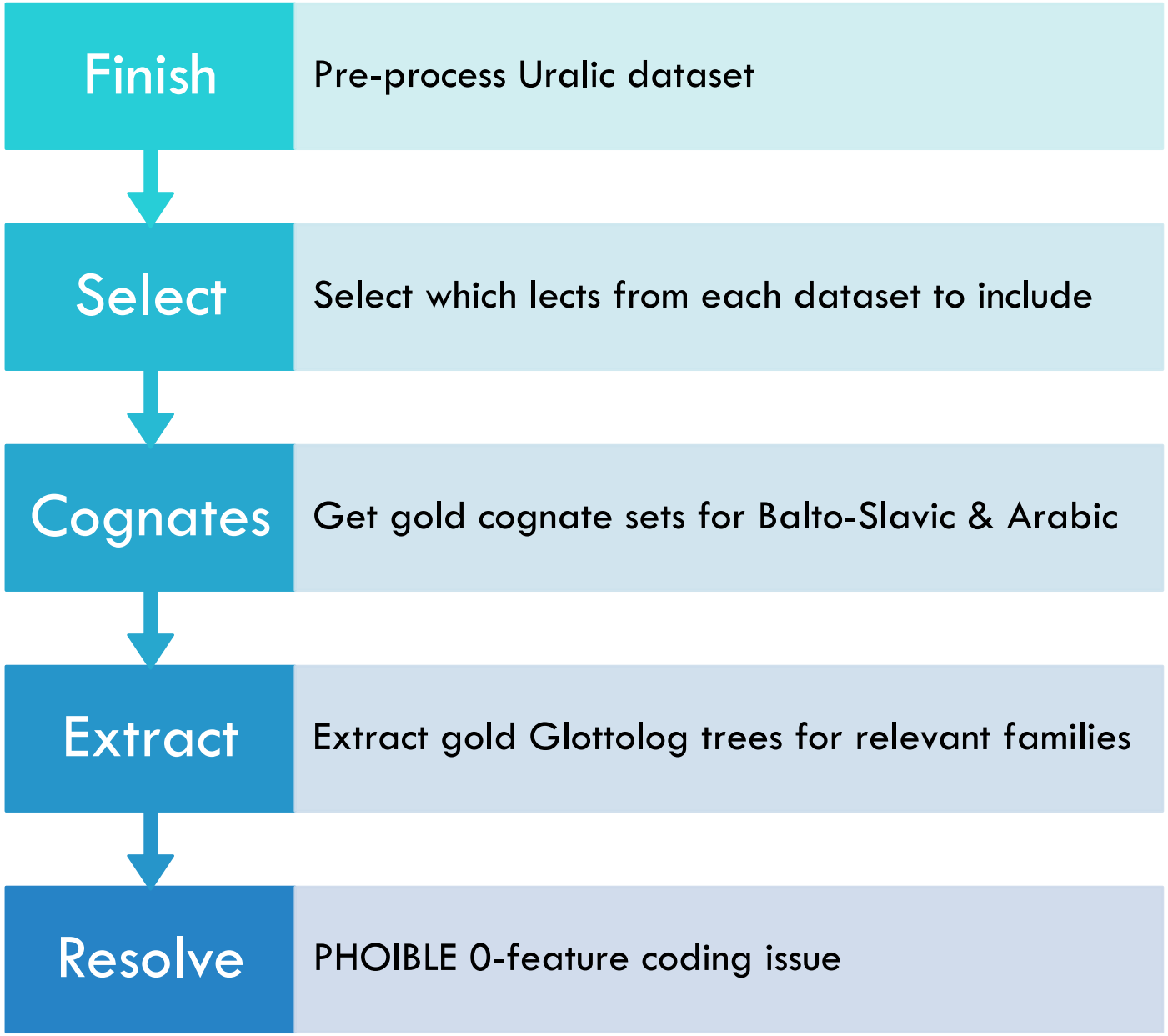


THESIS SEMINAR MEETING

Philip Georgis

June 28, 2021

TASKS FROM LAST TIME



URALIC (URALEX 2.0) DATASET

- Short story: fully preprocessed, standardized to CLDF, and gold cognate sets extracted ☺

1	Gloss	Erzya	Estonian	Finnish	Hungarian	Ingrian	Karelian	Khanty	Kildin Saami	Komi-Permyak	Komi-Zyrian	Livonian	Mansi	Meadow Mari	Nganasan	North Saami	Selkup	Skolt Saami	South Saami	Tundra Nenets	Udmurt	Veps	Votic
332	BLACK_1							pəɣtə															
333	BLACK_2								tʃɔ:hpe							tʃa:hp:es		tʃɔp:ed	tʃe:hpɛs				
334	BLACK_3																			pərʲidʲe:ɲa			
335	BLACK_4	rauʒo																					
336	BLACK_5																sæ:qə						
337	BLACK_6														tusajkaʔa								
338	BLACK_7		(must)	(musta)		(musta)	(mɔsta)					(musta:)										(mɔst)	(mus)
339	BLACK_8									ɕʏd	ɕʏd										ɕʏd		
340	BLACK_9												se:məl	ʃeme									
341	BLOOD_1																			we:ja			
342	BLOOD_2	verʲ	veri	ʊɐri	ve:r	veri	vɛri	wər	ve:r:	vir	vir	vɛr	wiɣər	βyr		vɑr:ɑ		vər:e	βir:ɛ		vir	vɛri	veri
343	BLOOD_3														kam		kʏm						
344	BLOOD_4												ke:lp										
345	BLOW (OF WIND)_1									pʏʌtnɯ	pʏʌalnɯ										peʌtɯnɯ		
346	BLOW (OF WIND)_3																				tʏlanɯ		
347	BLOW (OF WIND)_4	puhuta	puɦalta:		puhaɖa				pɔ:s:e														puhuma
348	BLOW (OF \	puvams			fu:j			pɵχ					puwi	puem	hyələsa		pu:						
349	BLUNT_1								mierʲesʲ														
350	BLUNT_10																			tɨnxa			
351	BLUNT_11				(tompɔ)																		
352	BLUNT_13															pɑst:ɔhɡæpme							
353	BLUNT_14			ʏylsæ			tylʧ:æ					tɨ:lza										tvlʲts	
354	BLUNT_15														tɯamə:								
355	BLUNT_16							pel:əɣ															
356	BLUNT_3																						tylpʲ:
357	BLUNT_4													(tokmak)									
358	BLUNT_6	noʃka	nyri			nyrhi				nɯʒ	nɯʒ			nyʃkø							nɯʒ		nyri
359	BLUNT_8																qamələʌ						
360	BLUNT_9												tɑʌəɣsup										
361	BONE_1	pakarʲ																					
362	BONE_2				ʧont																		
363	BONE_3																		mɔɑruʌ				
364	BONE_4									(koskɑ)													
365	BONE_5	lovaʒa	lu:	lu:		lu:	lu:	lɵχ		lɯ	lɯ	lu:	luw	lu	latə:		lɯ:			lɨ	lɯ	lɔ	lu:
366	BONE_6		kont																				
367	BONE_7								ta:xːʲt							ta:kti		taxʲtʲ:ɘ					

URALIC (URALEX 2.0) DATASET

- Longer story: lots of problems!
  - CLDF-formatted file was poorly constructed: inconsistent mixture of IPA, Uralic Phonetic Alphabet (UPA), and orthography for transcriptions, but all given in the IPA field
  - True IPA transcription sometimes listed in an unrelated field within CLDF file (e.g. “etym_notes”, “glossing_notes”), but also inconsistent in which field so not possible to extract automatically
  - Raw data file had proper IPA transcriptions in many (but not all) cases
  - Many entries lacked IPA transcriptions in both files

    → 11 languages had no transcriptions at all

URALIC (URALEX 2.0) DATASET

- Solutions
  - Created semi-automatic mapping between the raw and CLDF files to extract and/or fix the IPA transcriptions
  - Omitted word entries without IPA transcriptions in either file
    - 15 languages (of 27 total) still had >300 transcribed word forms
    - Võro language only had 1 transcribed word form → excluded

URALIC (URALEX 2.0) DATASET

- Solutions

  - Instead of excluding all 11 languages with 0 transcriptions in UraLex…
    - 7 are also included in NorthEuraLex dataset

      (Karelian, Livonian, Veps, North Saami, South Saami, Skolt Saami, Tundra Nenets)

      → combined UraLex and NorthEuraLex data

    - Other 4 languages not included in NorthEuraLex and thus needed to be excluded

      (Proto-Uralic, Ume Saami, Pite Saami, Inari Saami)

URALIC (URALEX 2.0 / NORTHEURALEX) DATASET

- Solutions
  - Combining UraLex and NorthEuraLex data
    - Word forms and IPA transcriptions extracted from NorthEuraLex
    - Cognate set and borrowing data taken from UraLex
    - Automatically combined only word entries whose Concepticon glosses and word forms matched *exactly* between the two databases
    - Generated table of word entries with matching Concepticon glosses but non-identical word forms to be manually matched
    - Manual matching sped up considerably by calculating Levenshtein distance between UraLex and NorthEuraLex word forms and sorting by this measure

      → Most genuine matches had length-normalized LD <0.4 (mean: 0.43)
    - Doesn't violate principle whereby transcriptions for a language should be taken from a single source, since this was only performed for languages with 0 UraLex transcriptions

URALIC (URALEX 2.0 / NORTHEURALEX) DATASET

- Solutions
  - Combining UraLex and NorthEuraLex data
    - Generated table of word entries with matching Concepticon glosses but non-identical word forms to be manually matched
    - Manual matching sped up considerably by calculating Levenshtein distance between UraLex and NorthEuraLex word forms and sorting by this measure

      → Most genuine matches had length-normalized LD <0.4 (mean: 0.43)

UraLex_Index	Language	UraLex_Form	UraLex_Value	NEL_Form	NEL_Value	NEL_Source_Form	LevenshteinDist	Accepted?
9309	Karelian	muurahaine	muurahaine	mu:rahaini	muurahaini	mu:rahaini	0.1	x
7868	Karelian	ukonkoari	ukonkoari	ʊkɔŋkʊari	ukonkuari	ʊkɔŋkʊari	0.111111111	x
10077	Karelian	kuvahaine	kuvahaine	kʊvahaini	kuvahaini	kʊvahaini	0.111111111	x
483	Karelian	henkitteä	henkitteä	hɛŋkit:yæ	henkittyä	hɛŋkit:yæ	0.111111111	x
5444	Karelian	hämehikki	hämehikki	hæmæhik:i	hämähikki	hæmæhik:i	0.111111111	x
7834	Karelian	ukonkoari	ukonkoari	ʊkɔŋkʊari	ukonkuari	ʊkɔŋkʊari	0.111111111	x
257	Karelian	šiivatta	šiivatta	si:vat:a	siivatta	si:vat:a	0.125	x
7623	Karelian	vihelteä	vihelteä	vihɛltyæ	viheltyä	vihɛltyæ	0.125	x
6377	Karelian	puistoa	puistoa	pʊistʊa	puistua	pʊistʊa	0.142857143	x
2973	Karelian	keärmis	keärmis	kiærmis	kiärmis	kiærmis	0.142857143	x
6100	Karelian	opastoa	opastoa	ɔpastʊa	opastua	ɔpastʊa	0.142857143	x
3540	Karelian	viskata	viskata	visata	visata	visata	0.142857143	x
3402	Karelian	tuolla	tuolla	tʊɔla	tuola	tʊɔla	0.166666667	x

URALIC (URALEX 2.0 / NORTHEURALEX) DATASET

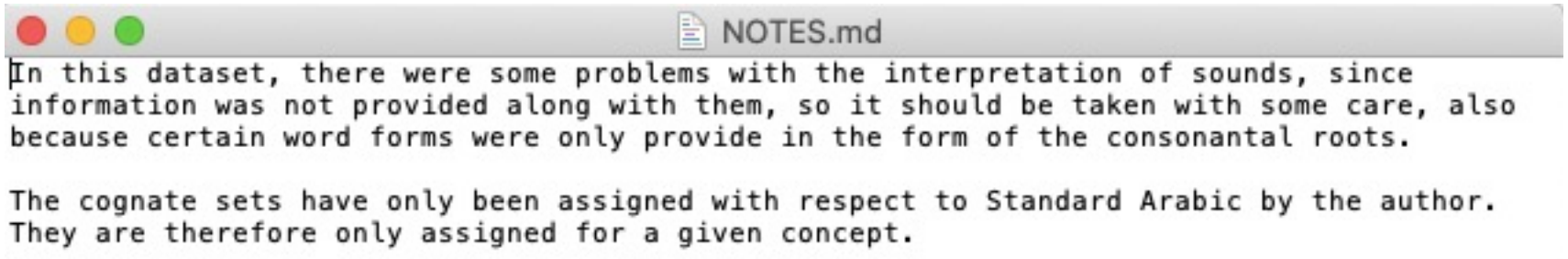
Language	Transcriptions in UraLex 2.0	Automatically Matched Transcriptions	Manually Confirmed Transcriptions	Total Extracted Transcriptions	Avg Levenshtein Dist. of Manually Matched Word Forms / Transcriptions
Karelian	0	146	56	202	0.30
Livonian	0	49	164	213	0.45
Veps	0	117	83	200	0.34
North Saami	0	178	32	210	0.33
South Saami	0	193	17	210	0.32
Skolt Saami	0	89	122	211	0.30
Tundra Nenets	0	15*	161	176	0.61*

* Fewer automatic matches and higher average LD for Tundra Nenets because orthographic word form in NEL was given in Cyrillic alphabet but in Latin alphabet in UraLex, so LD measured on word form/IPA instead

URALIC (URALEX 2.0 / NORTHEURALEX) DATASET

ARABIC DATASET(S)

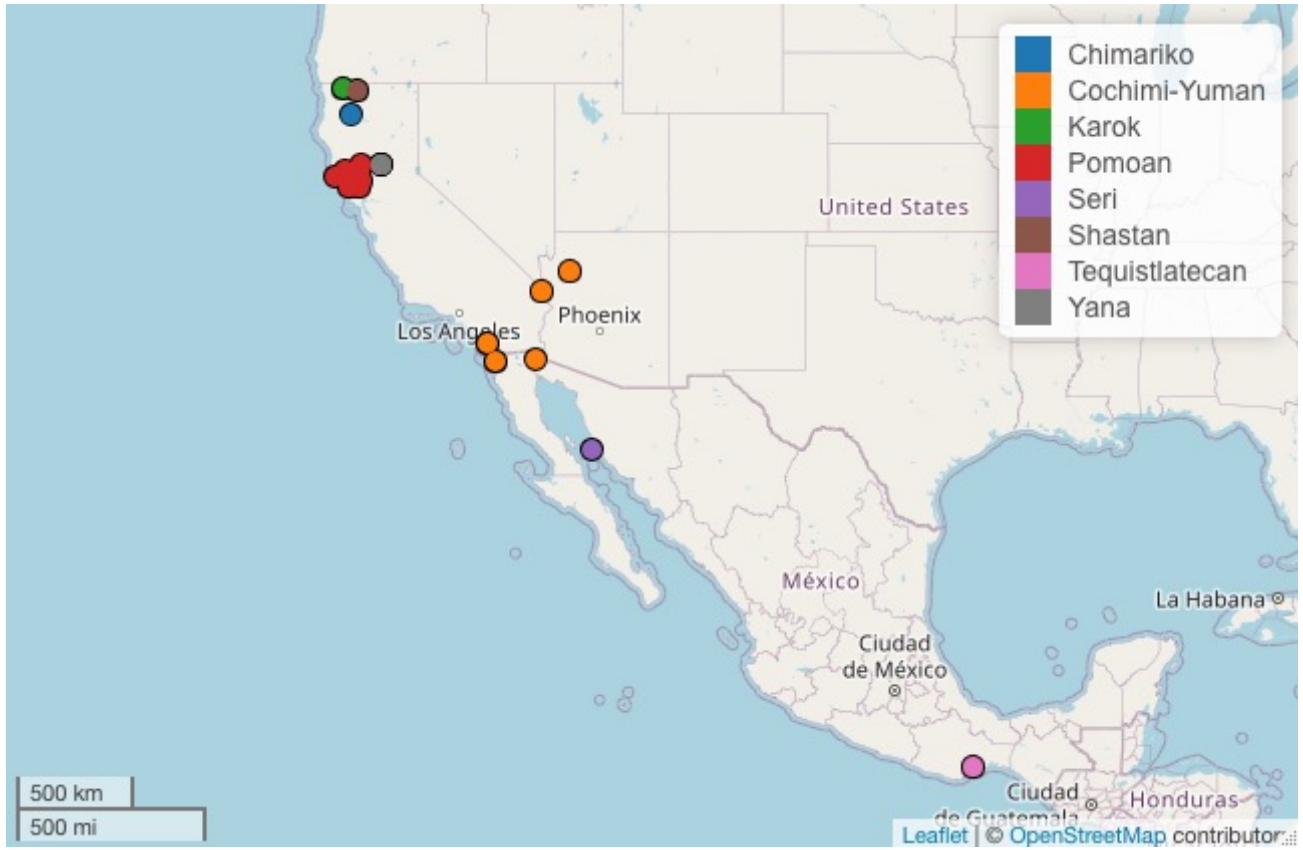
- Found CLDF version of Ratcliffe's dataset in GitHub/lexibank
  - Still no gold cognate coding, but could facilitate checking overlap with Wiktionary dataset better
  - Note by Johann-Mattis List:



- Wiktionary dataset
  - Seems to have been created/compiled largely by a single user (Qizilqurt)
  - But there doesn't seem to be any way to contact them to ask about sources

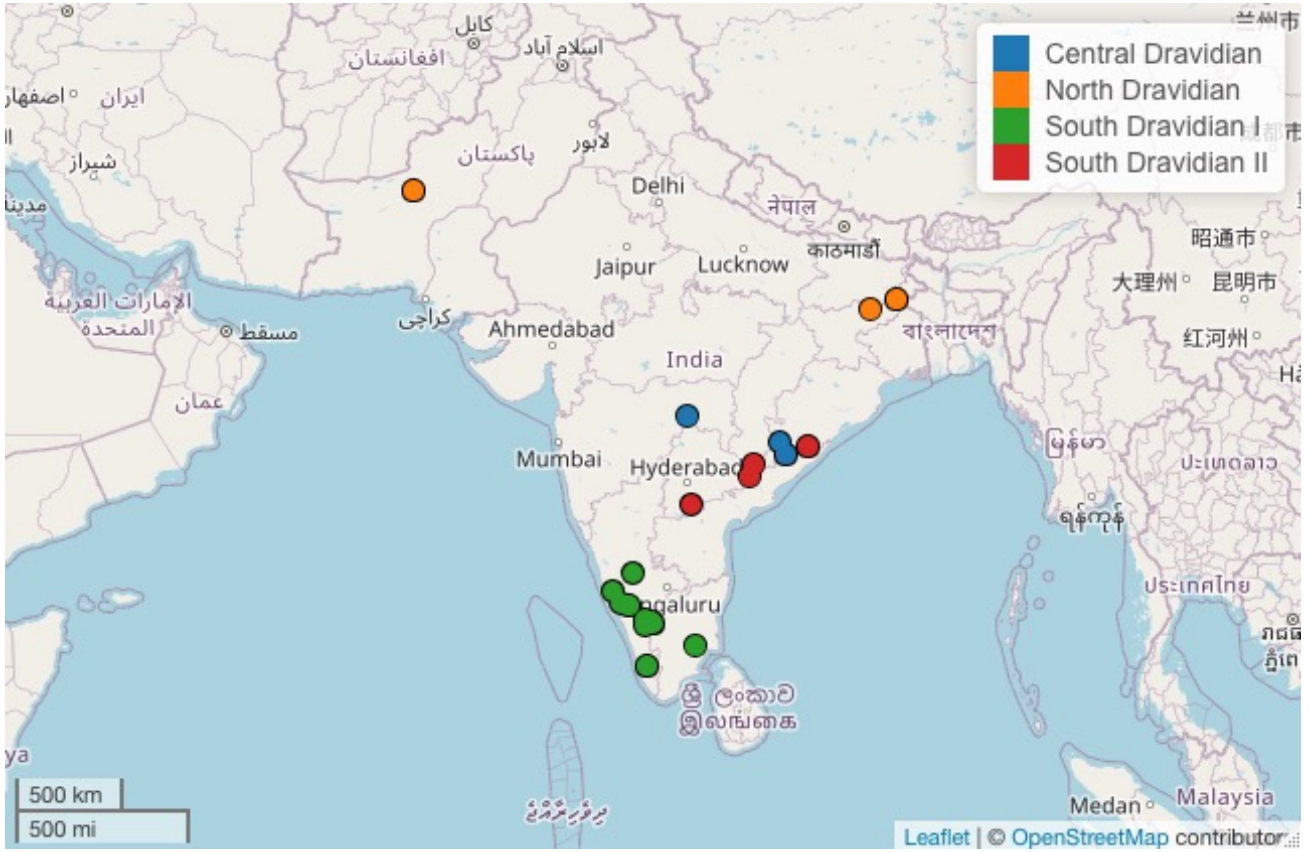
HOKAN DATASET (ZHIVLOV 2011-2015)

- Hokan: proposed language family comprising a handful of indigenous languages from California, Arizona, and Mexico
- Data taken from Global Lexicostatistical Database (same source as Italic data)
- Organized into 8 recognized (sub-)families/isolates, includes cognate set coding for each:
  - Chimariko
  - Cochimi-Yuman
  - Karuk
  - Pomo
  - Seri
  - Shastan
  - Tequistlatecan
  - Yana
- Idea: use as experimental case study for application to groups lacking consensus



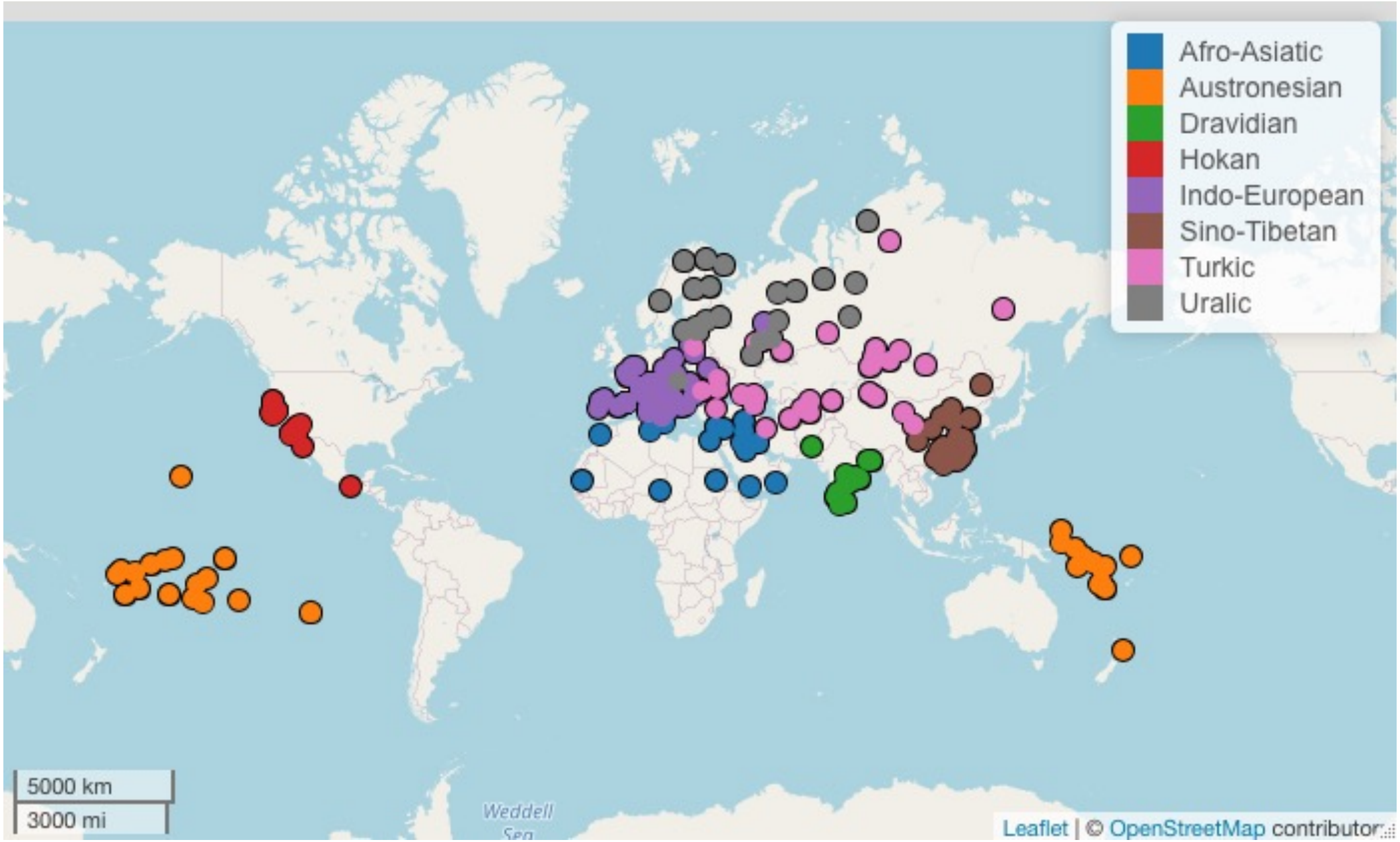
DRAVIDIAN DATASET (KOLIPAKAM ET AL., 2018)

- Final dataset: Dravidian
- Already in CLDF format, only a few minor/straightforward transcription changes needed
  - e.g. <ṭ> → <ʈ>, <ñ> → /ɲ/, <ā> → /aː/



OVERVIEW OF DATASETS

Family	Source Name	Reference	Number of Varieties
Arabic	Varieties of Arabic Swadesh lists	Wiktionary	16
Balto-Slavic	NorthEuraLex	Dellert et al. (2019)	11
Dravidian	DravLex: A Dravidian lexical database	Kolipakam et al. (2018)	20
Hokan	Global Lexicostatistic Database	Zhivlov (2011-2015)	20
Italic	Global Lexicostatistic Database	Saenko (2016)	58
Polynesian	Polynesian Segmented Data	Walworth (2018)	31
Sinitic	Collection of basic vocabulary words and characteristic dialect words in modern Chinese dialects	Líu et al. (2007)	19
Turkic	Basic vocabulary datasets for the Turkic languages	Savelyev & Robbeets (2020)	31
Uralic	UraLex 2.0: Uralic basic vocabulary with cognate and loanword information; NorthEuraLex	De Heer et al.; Syrjänen [submitted manuscript]; Dellert et al. (2019)	22


Afro-Asiatic
Austronesian
Dravidian
Hokan
Indo-European
Sino-Tibetan
Turkic
Uralic
5000 km
3000 mi
Weddell Sea
Leaflet | © OpenStreetMap contributors

STATUS OF DATASETS (PREVIOUSLY)

Dataset	Fully Preprocessed Transcriptions	ConceptIcon Cross-Reference	Standardized Format	Gold Cognate Sets Extracted	Extracted Glottolog Tree
Arabic	✓	✓	✓	X	X
Balto-Slavic	✓	✓	✓	X	X
Dravidian	X	X	X	X	X
Hokan	X	X	X	X	X
Italic	✓	✓	✓	✓	X
Polynesian	✓	✓	✓	✓	X
Sinitic	✓	✓	✓	✓	X
Turkic	✓	✓	✓	✓	X
Uralic	X	✓	✓	✓	X

STATUS OF DATASETS (NOW)

Dataset	Fully Preprocessed Transcriptions	ConceptIcon Cross-Reference	Standardized Format	Gold Cognate Sets Extracted	Extracted Glottolog Tree
Arabic	✓	✓	✓	partially	✓
Balto-Slavic	✓	✓	✓	partially	✓
Dravidian	✓	✓	✓	✓	✓
Hokan	✓	✓	✓	✓	✓
Italic	✓	✓	✓	✓	✓
Polynesian	✓	✓	✓	✓	✓
Sinitic	✓	✓	✓	✓	✓
Turkic	✓	✓	✓	✓	✓
Uralic	✓	✓	✓	✓	✓

COGNATE CODING: BALTO-SLAVIC AND ARABIC

- Ran LingPy LexStat cognate detection tool in order to get preliminary cognate sets
  - Still to do: correct manually (or semi-automatically)
  - References: (limited) cognate set coding from Ratcliffe's (2020) Arabic dataset, IELex for Balto-Slavic

- Matching cognate sets from IELex for Balto-Slavic
  - Original IELex website/database (https://ielex.mpi.nl/) no longer functional
  - Copy of data preserved on a third-party website by someone who had created cognate set maps from them (https://pappubahry.com/maps/ie_cognates/details.html)
  - Word forms are mix of IPA, orthography, block caps (sound classes?)

    → currently working on creating semi-automatic mapping similar to Uralic to extract gold cognate sets

CONCEPTS AND MUTUAL COVERAGE

Family	Number of Varieties	Min Number of Concepts	Average Number of Concepts	Mutual Coverage	Average Mutual Coverage
Balto-Slavic	11	1013 (474)*	1016 (476)*	1011	1.00
Uralic	22	172	265	103	0.74
Turkic	31	186	237	90	0.88
Arabic	16	179	203	162	0.96
Sinitic	19	201	202	201	1.00
Polynesian	31	178	200	109	0.91
Italic	58	103	110	98	0.99
Hokan	20	78	101	46	0.82
Dravidian	20	56	93	28	0.86

* Number of concepts found in at least one other dataset

CONCEPT SELECTION

- 228 total languages included in study

- 1186 unique concepts: all standardized to Concepticon glosses
  - 55% *only* appear in NorthEuraLex
  - → 546 total concepts, excluding the ones which appear only in NorthEuraLex
  - 64 concepts appear in all 9 datasets
  - Only 7 concepts appear in all 228 languages

- How to select common set of concepts?

Family	Min Number of Concepts	Avg Number of Concepts	Avg Mutual Coverage
Balto-Slavic	1013 (474)*	**1016** (476)*	1.00
Uralic	172	**265**	0.74
Turkic	186	**237**	0.88
Arabic	179	**203**	0.96
Sinitic	201	**202**	1.00
Polynesian	178	**200**	0.91
Italic	103	**110**	0.99
Hokan	78	**101**	0.82
Dravidian	56	**93**	0.86

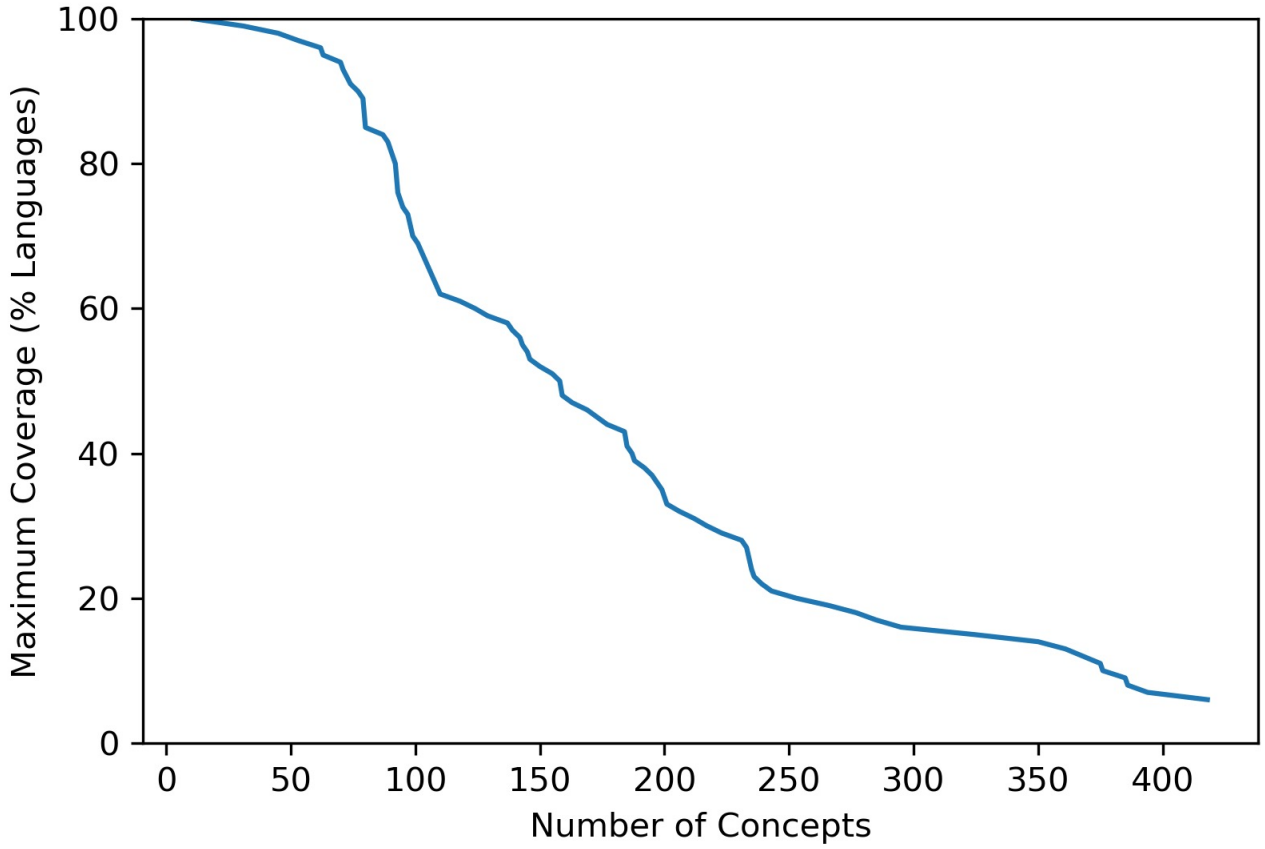
* Number of concepts found in at least one other dataset

CONCEPT SELECTION

- 228 total languages included in study

- 1186 unique concepts: all standardized to Concepticon glosses
  - 55% *only* appear in NorthEuraLex
  - → 546 total concepts, excluding the ones which appear only in NorthEuraLex
  - 64 concepts appear in all 9 datasets
  - Only 7 concepts appear in all 228 languages

- How to select common set of concepts?

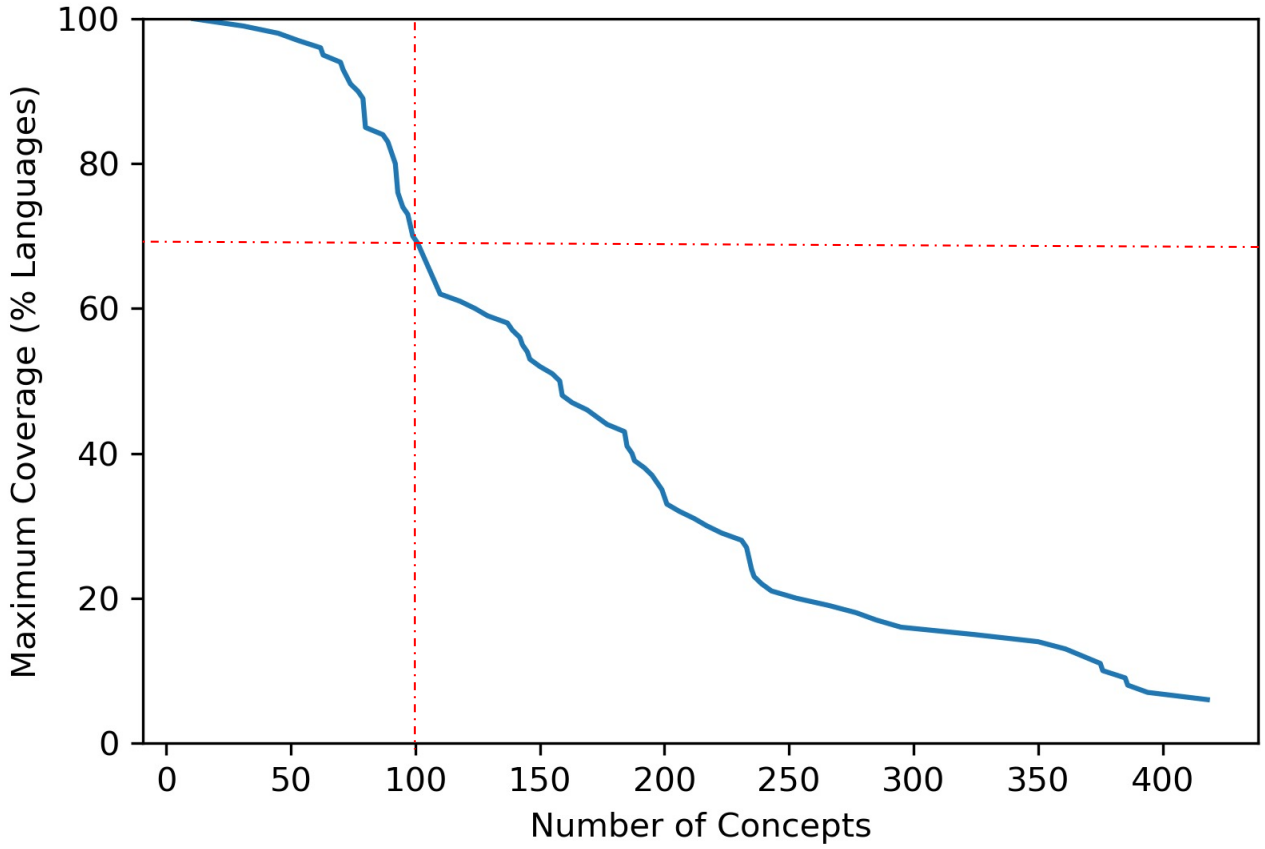


CONCEPT SELECTION

- 228 total languages included in study

- 1186 unique concepts: all standardized to Concepticon glosses
  - 55% *only* appear in NorthEuraLex
  - → 546 total concepts, excluding the ones which appear only in NorthEuraLex
  - 64 concepts appear in all 9 datasets
  - Only 7 concepts appear in all 228 languages

- How to select common set of concepts?



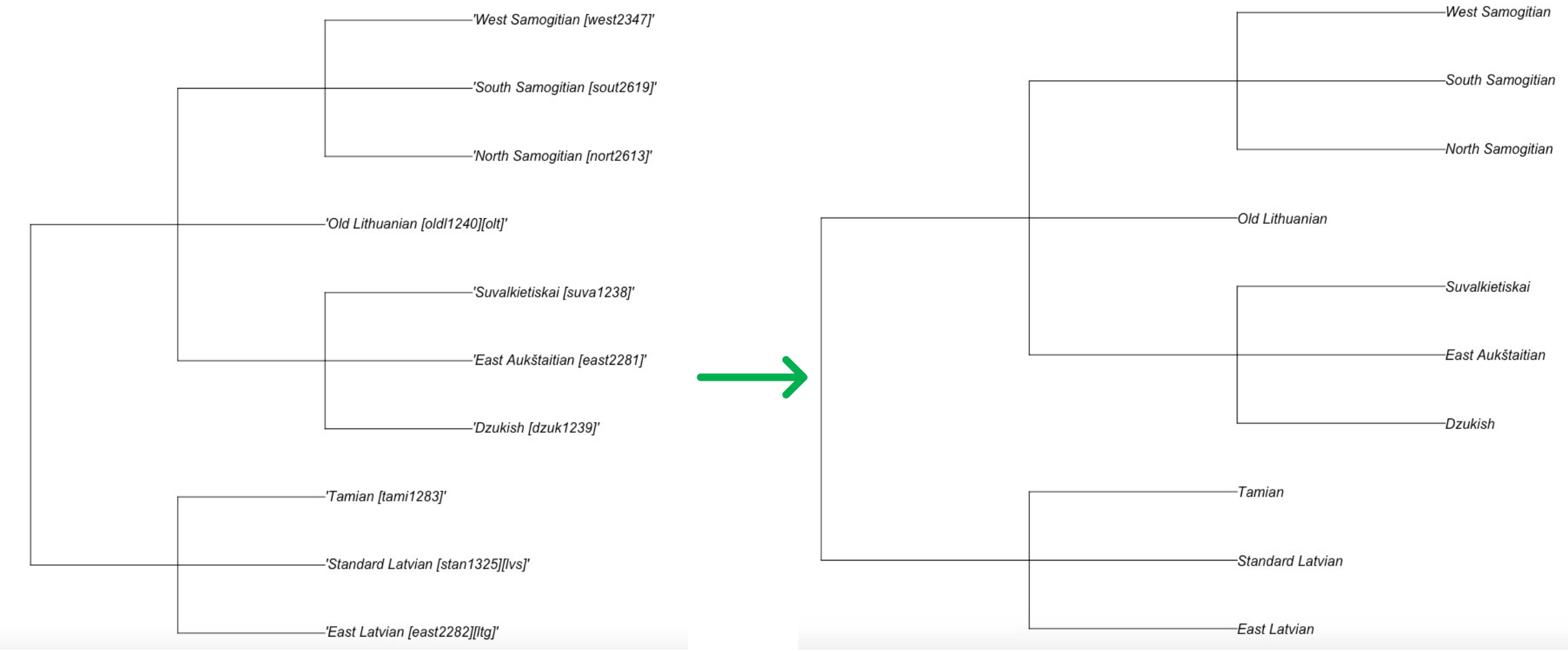
EXTRACTING GLOTTOLOG GOLD TREES

- Wrote Python script for extracting and preprocessing Glottolog trees into suitable format to manipulate with phylogenetic packages in R

- Python package pyglottolog
  - Given the Glottocode of a family's root, can extract full tree downwards from there in Newick format

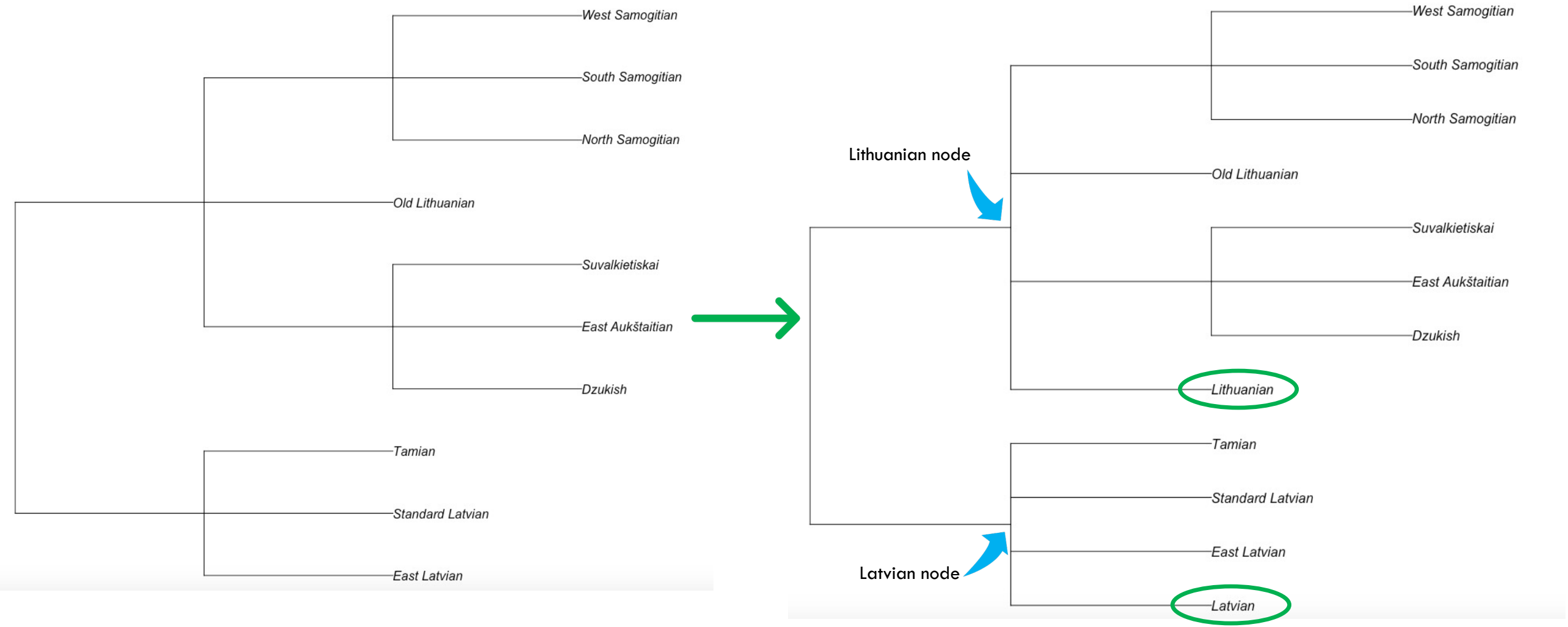
- Additionally cleaned the Newick tree by removing Glottocode and ISO code annotations, etc., leaving only the Glottolog name → converted to my designation so that all have unique names

- R script: using phytools package, can then remove unneeded varieties from tree and add tips when necessary



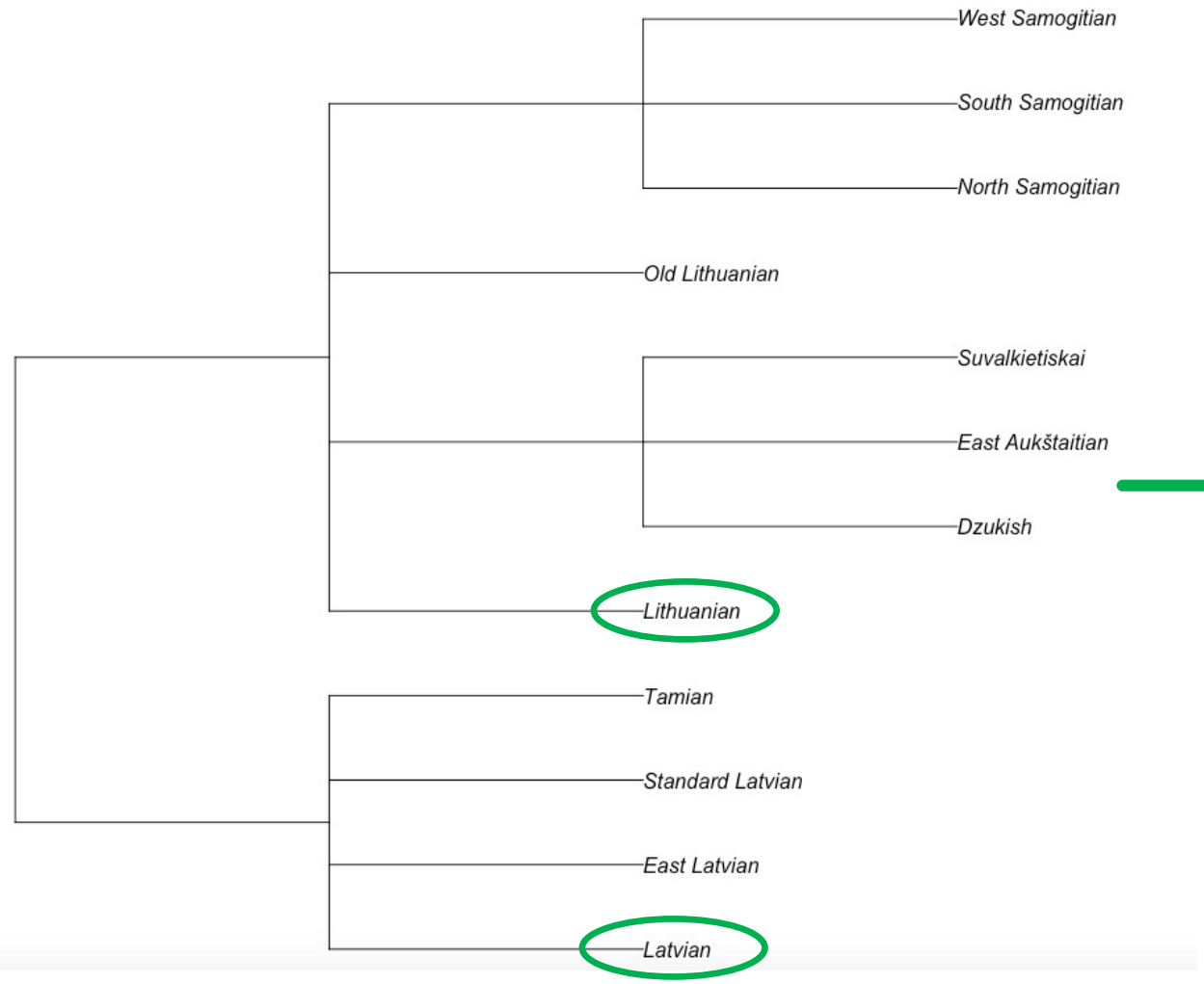
Example: East Baltic branch of Balto-Slavonic tree

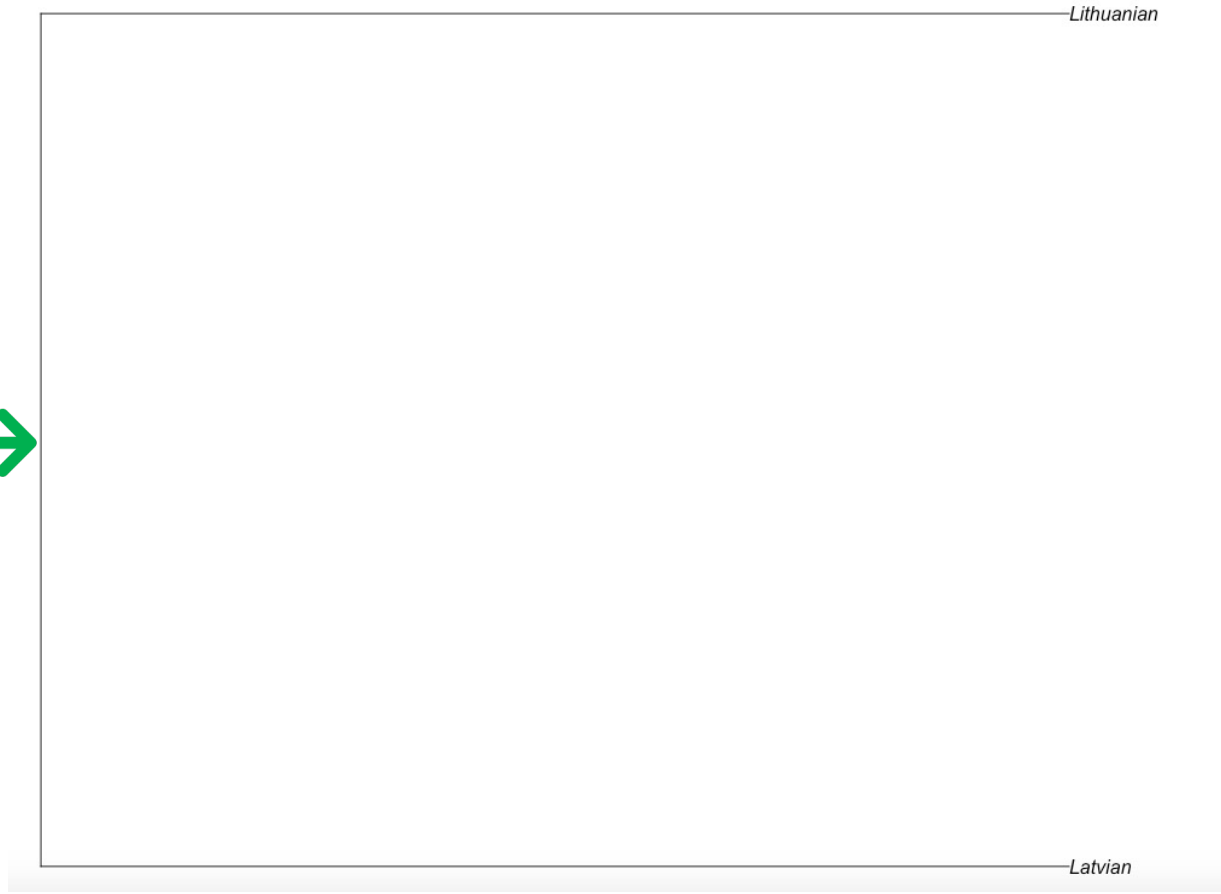
Newick preprocessing to remove Glottocodes and other annotations



Example: East Baltic branch of Balto-Slavonic tree

Adding missing languages (nodes) as tips under themselves

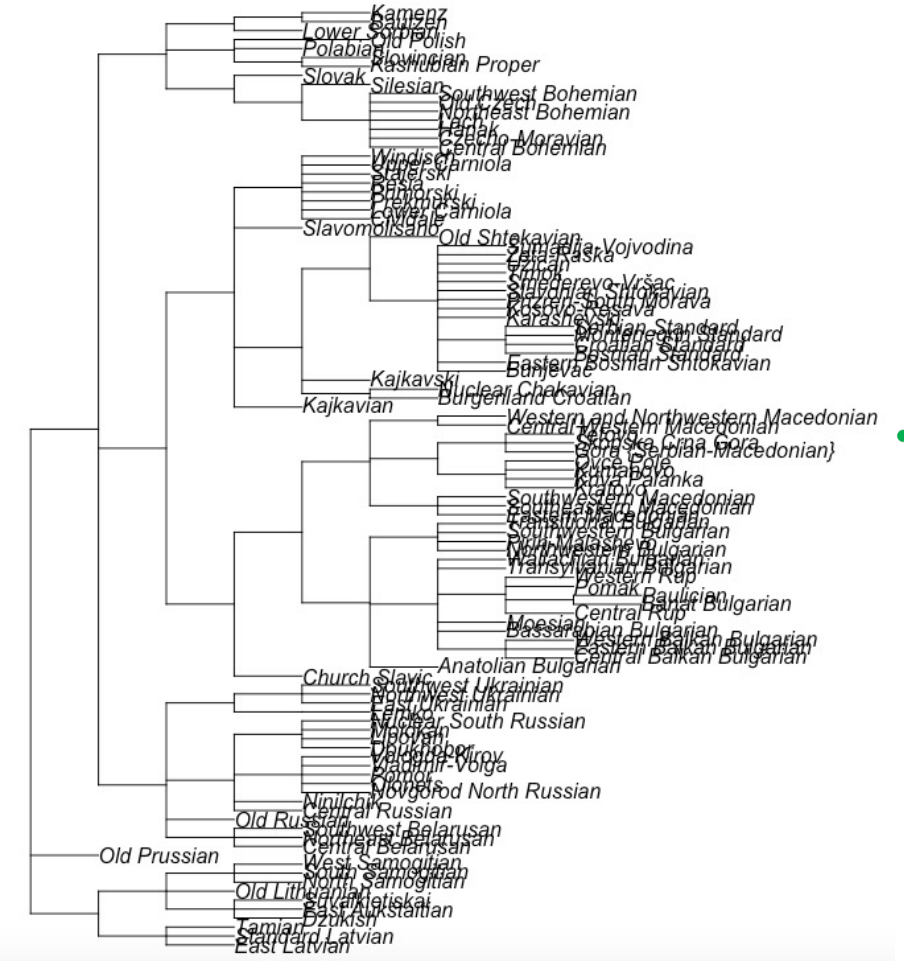


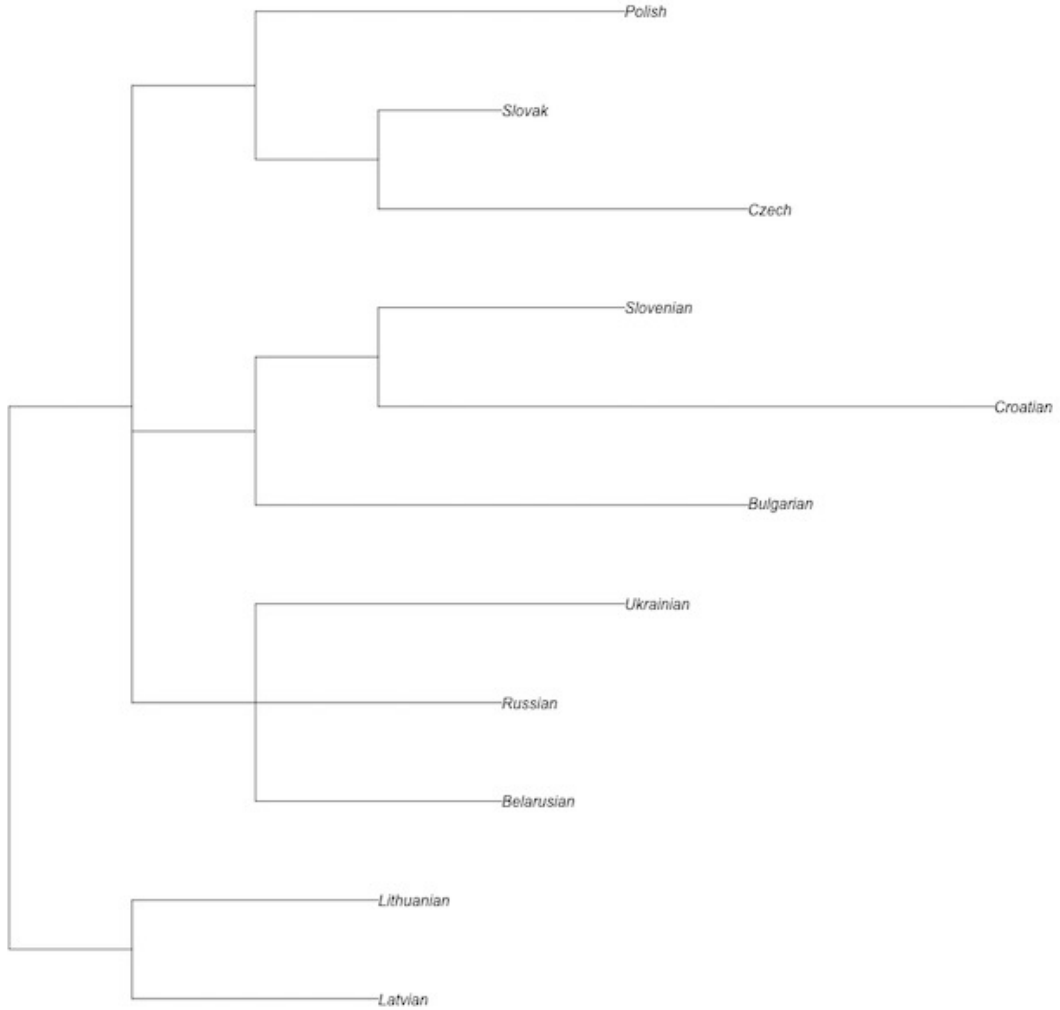


Example: East Baltic branch of Balto-Slavonic tree

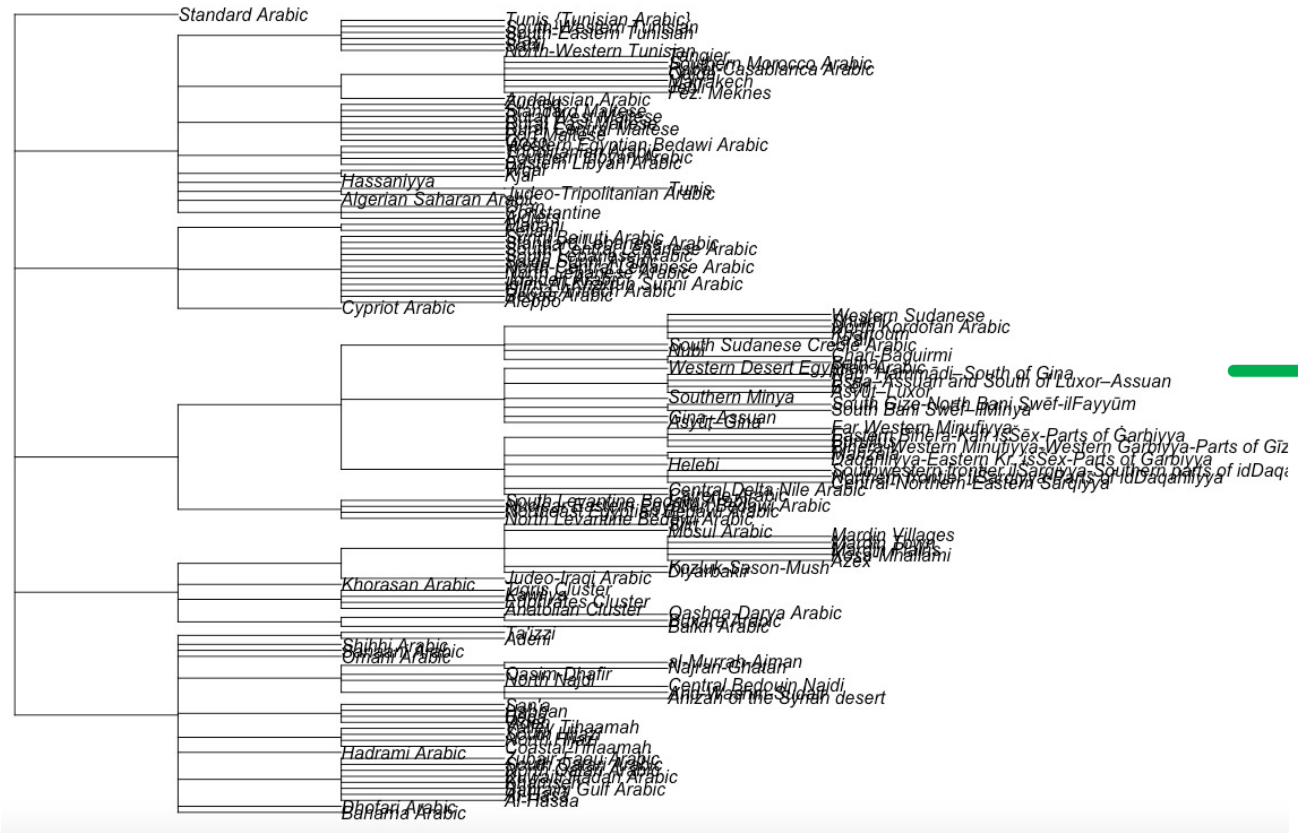
Prune all varieties not included in the specified list

BALTO-SLAVIC TREE

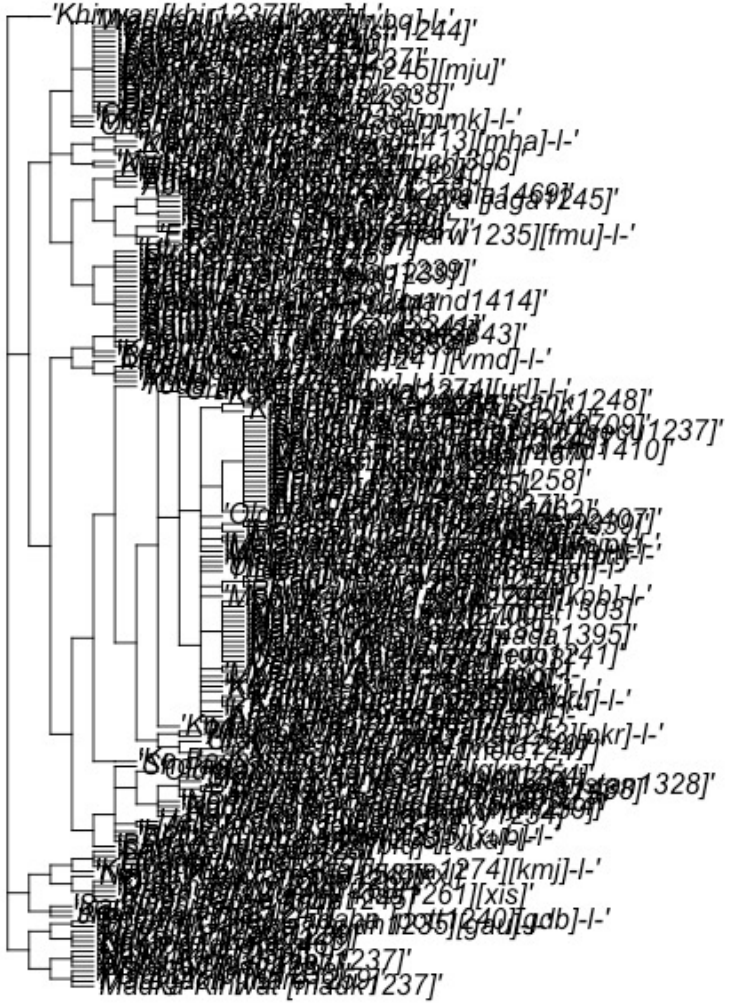


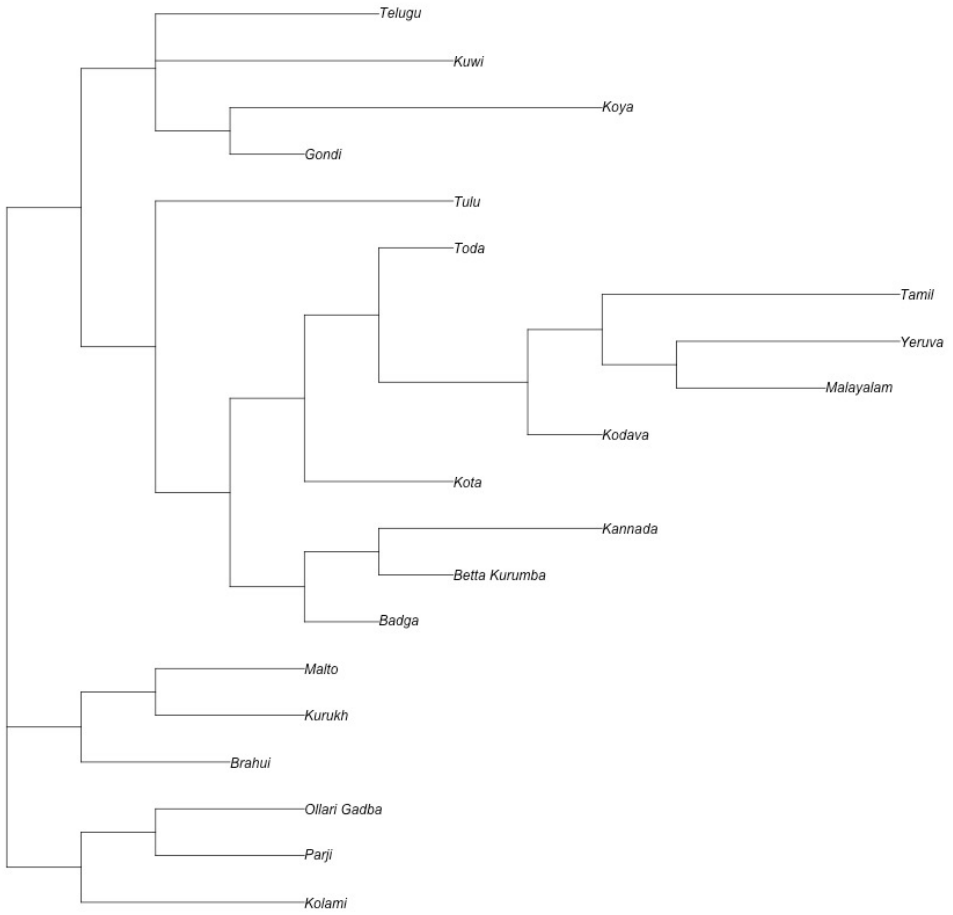


ARABIC TREE

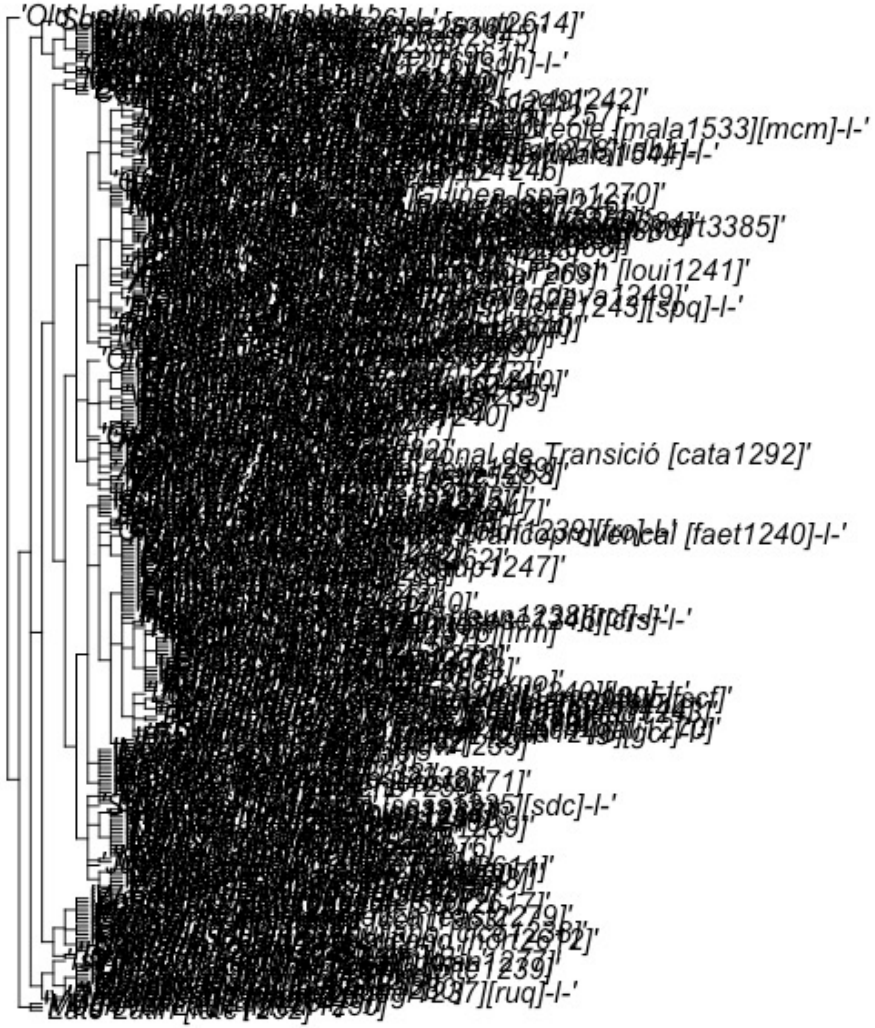

Standard Arabic
Cypriot Arabic
Khorasan Arabic
Hadrami Arabic
Dhofari Arabic
Hassaniyya
Algerian Saharan Arabic
Judeo-Iraqi Arabic
Judeo-Tripolitanian Arabic
Andalusian Arabic
Western Sudanese
Western Desert Egyptian Arabic
Southern Minya
Asyūṭ–Sīnā
Helebi
Central Delta Nile Arabic
Mosul Arabic
Diyarbakır
Oashqa-Darya Arabic
Baikh Arabic
Aden
Najran-Ghalan
Central Bedouin Najdi
Al-Hasaa
Tunisian Arabic
Moroccan Arabic
Maltese
Hassaniya Arabic
Cypriot Arabic
Levantine Arabic
Sudanese Arabic
Chadian Arabic
Egyptian Arabic
North Mesopotamian Arabic
Iraqi Arabic
Sanaani Arabic
Najdi Arabic
Gulf Arabic
Dhofari Arabic

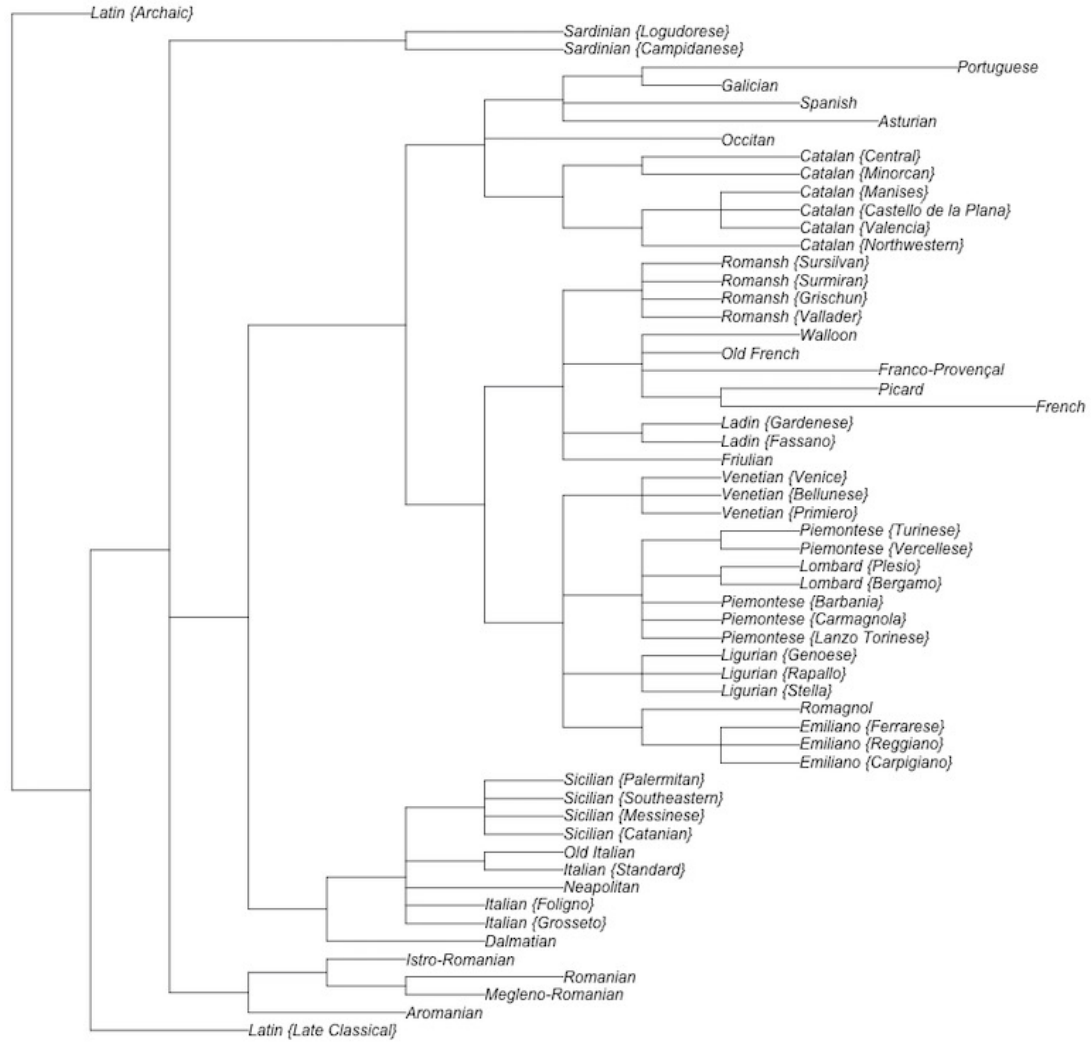
DRAVIDIAN TREE



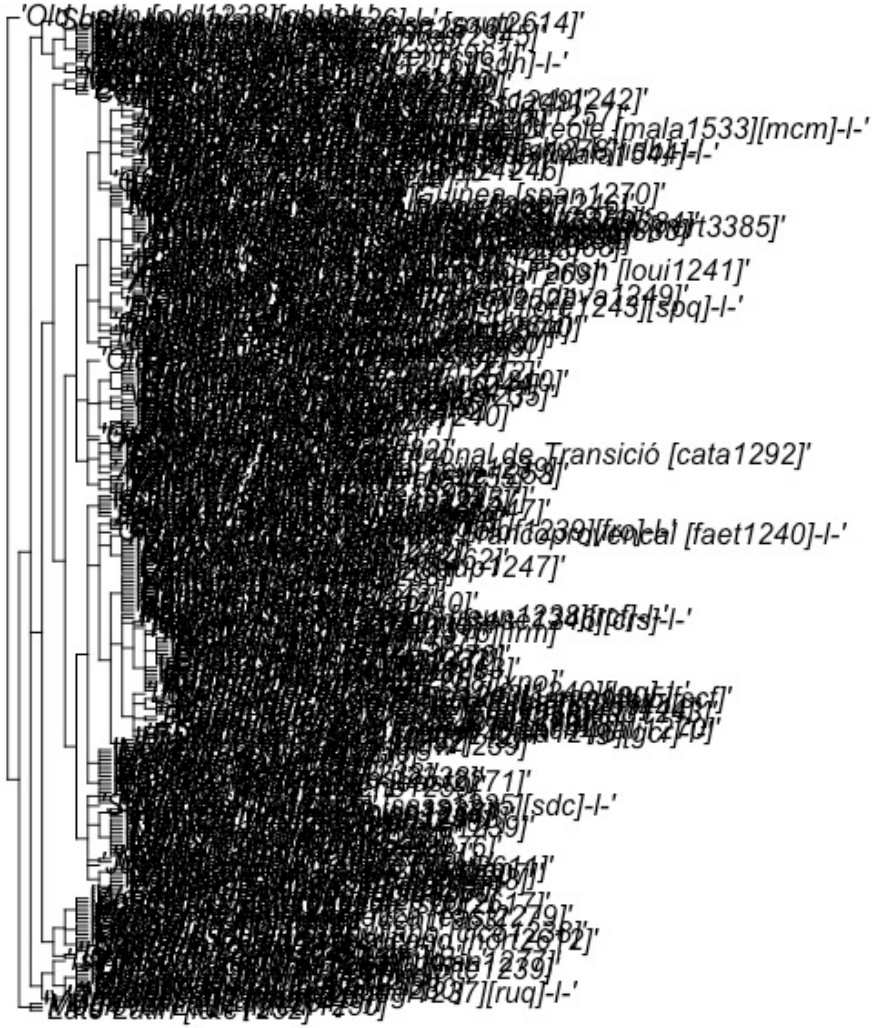


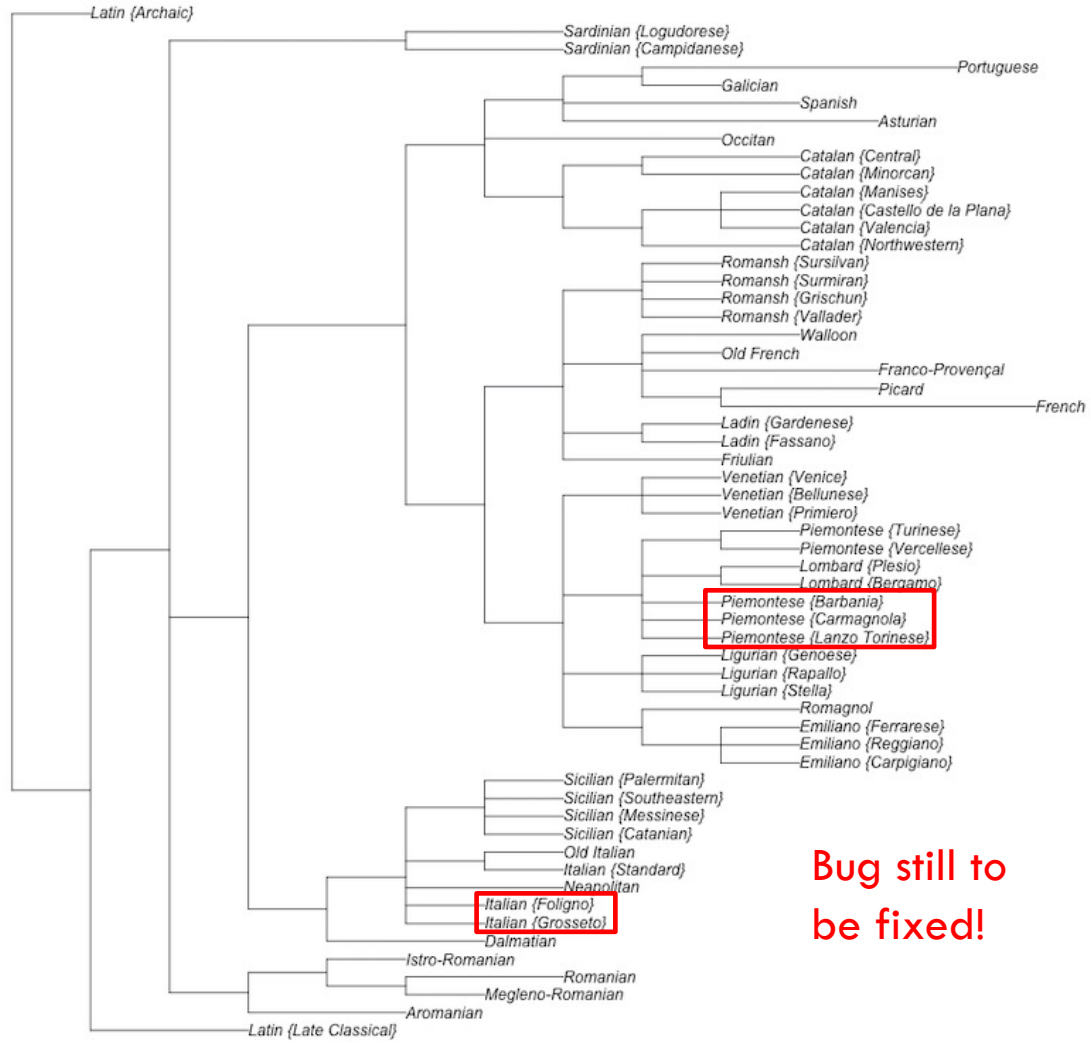
ITALIC TREE




Latin {Archaic}
Sardinian {Logudorese}
Sardinian {Campidanese}
Portuguese
Galician
Spanish
Asturian
Occitan
Catalan {Central}
Catalan {Minorcan}
Catalan {Manises}
Catalan {Castello de la Plana}
Catalan {Valencia}
Catalan {Northwestern}
Romansh {Sursilvan}
Romansh {Surmiran}
Romansh {Grischun}
Romansh {Vallader}
Walloon
Old French
Franco-Provençal
Picard
French
Ladin {Gardenese}
Ladin {Fassano}
Friulian
Venetian {Venice}
Venetian {Bellunese}
Venetian {Primiero}
Piemontese {Turinese}
Piemontese {Vercellese}
Lombard {Plesio}
Lombard {Bergamo}
Piemontese {Barbania}
Piemontese {Carmagnola}
Piemontese {Lanzo Torinese}
Ligurian {Genoese}
Ligurian {Rapallo}
Ligurian {Stella}
Romagnol
Emiliano {Ferrarese}
Emiliano {Reggiano}
Emiliano {Carpigiano}
Sicilian {Palermitan}
Sicilian {Southeastern}
Sicilian {Messinese}
Sicilian {Catanian}
Old Italian
Italian {Standard}
Neapolitan
Italian {Foligno}
Italian {Grosseto}
Dalmatian
Istro-Romanian
Romanian
Megleno-Romanian
Aromanian
Latin {Late Classical}

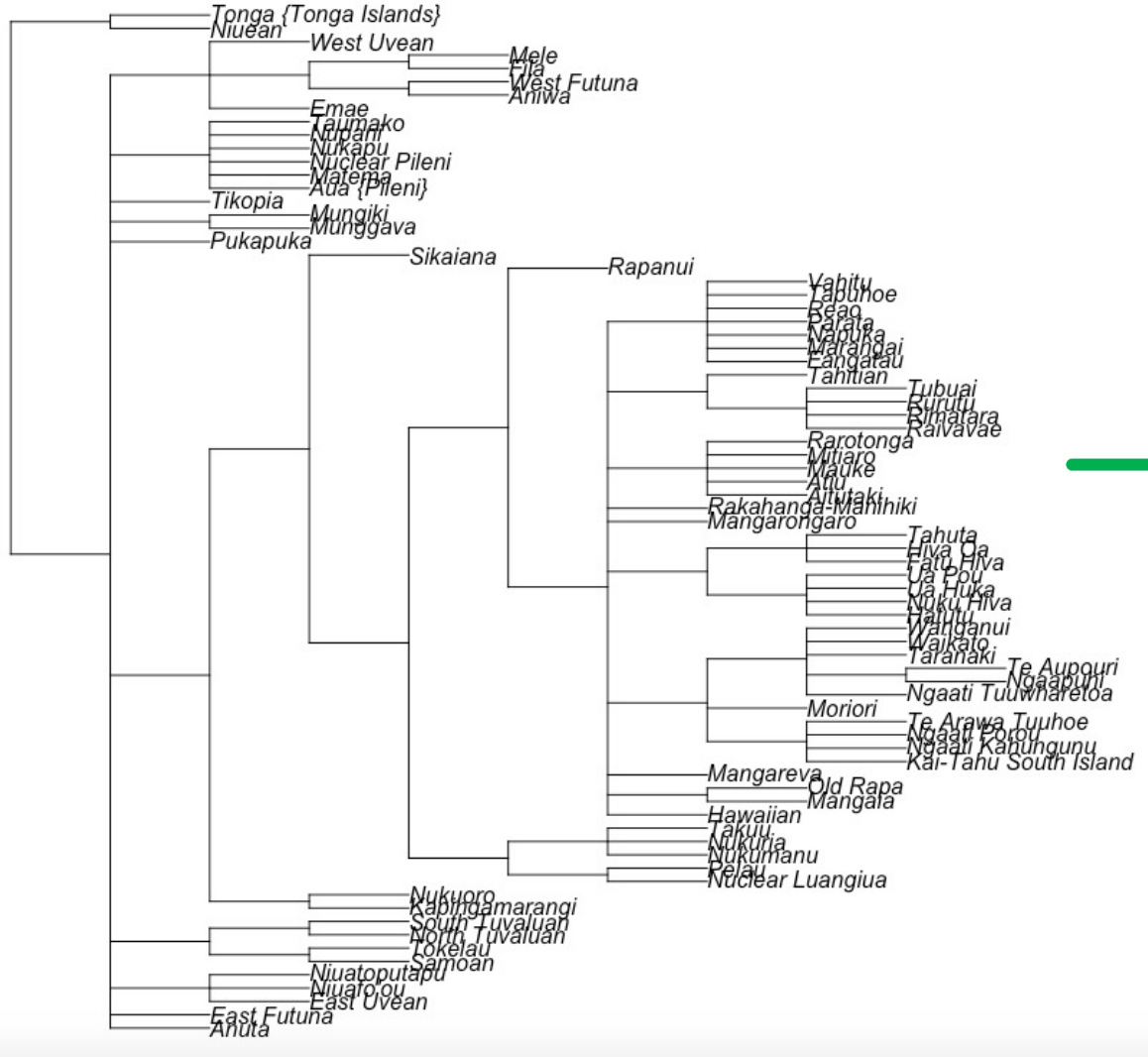
ITALIC TREE

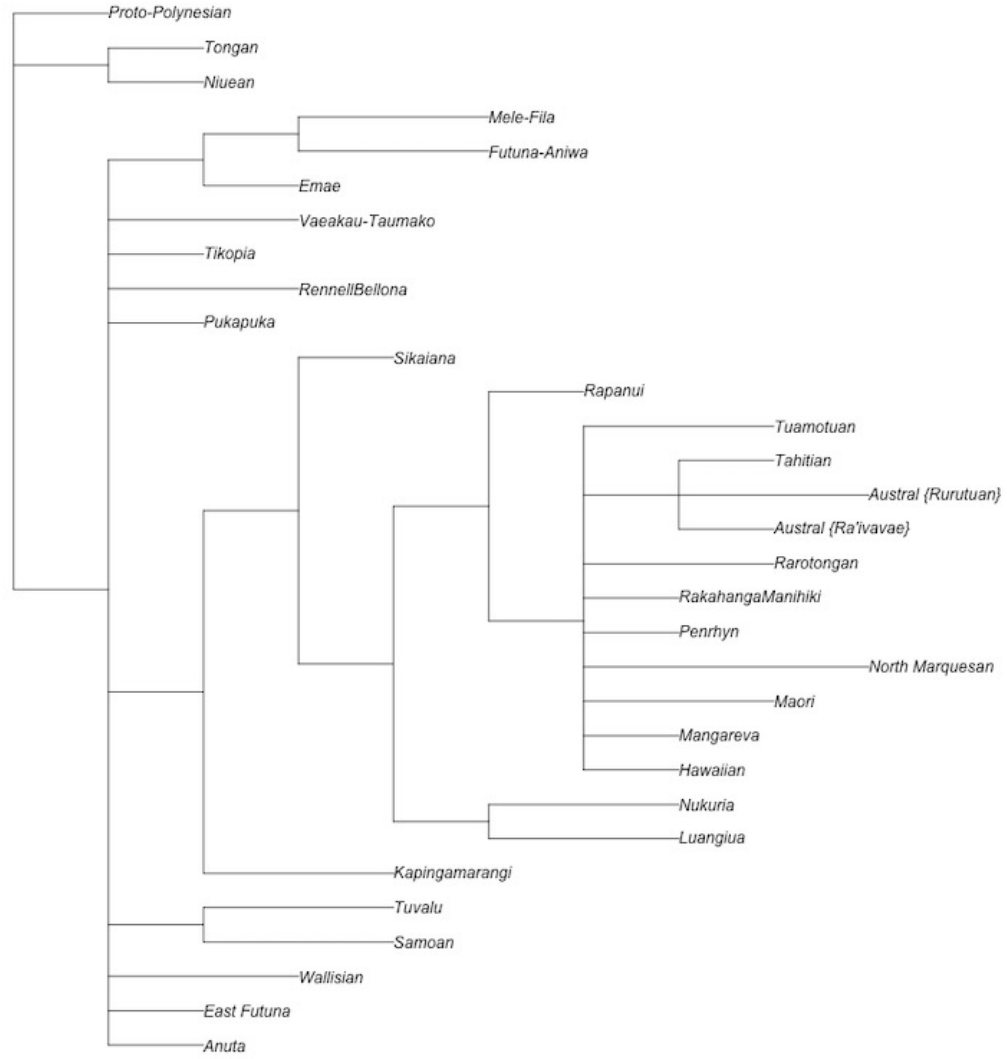




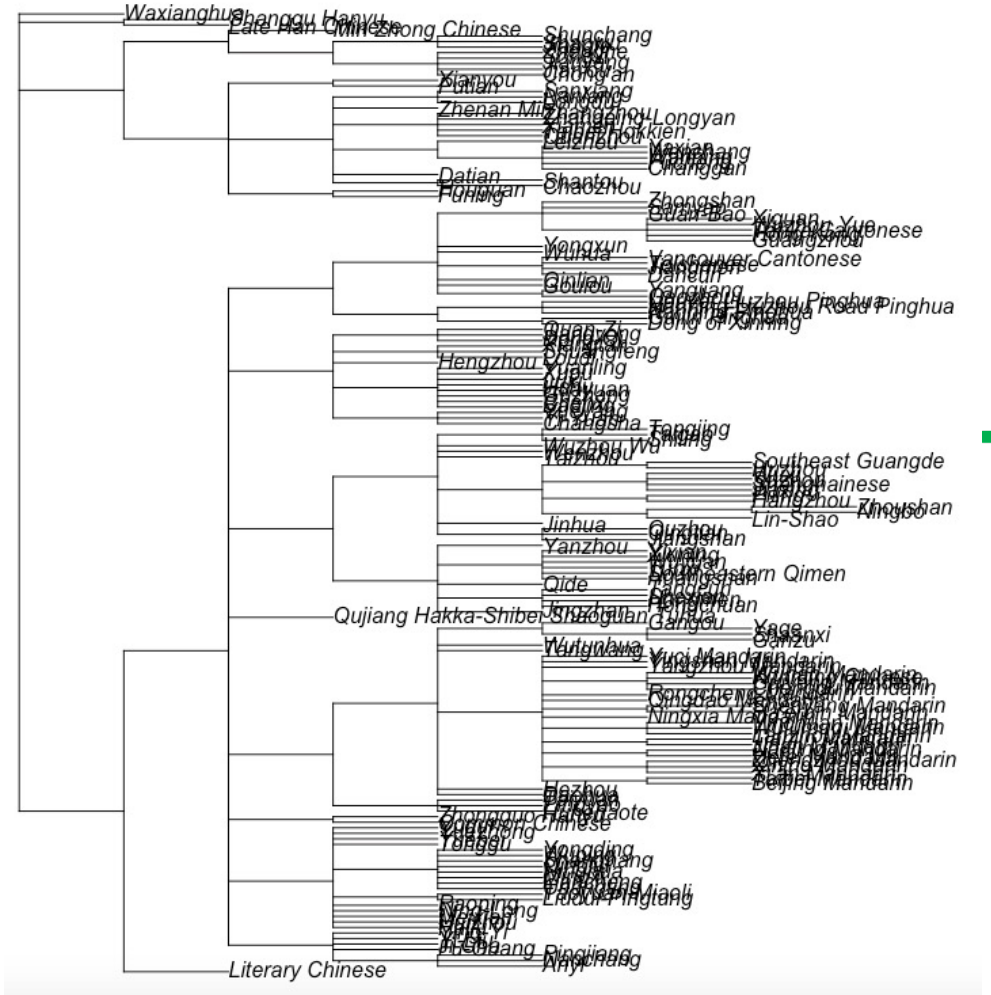
Bug still to be fixed!

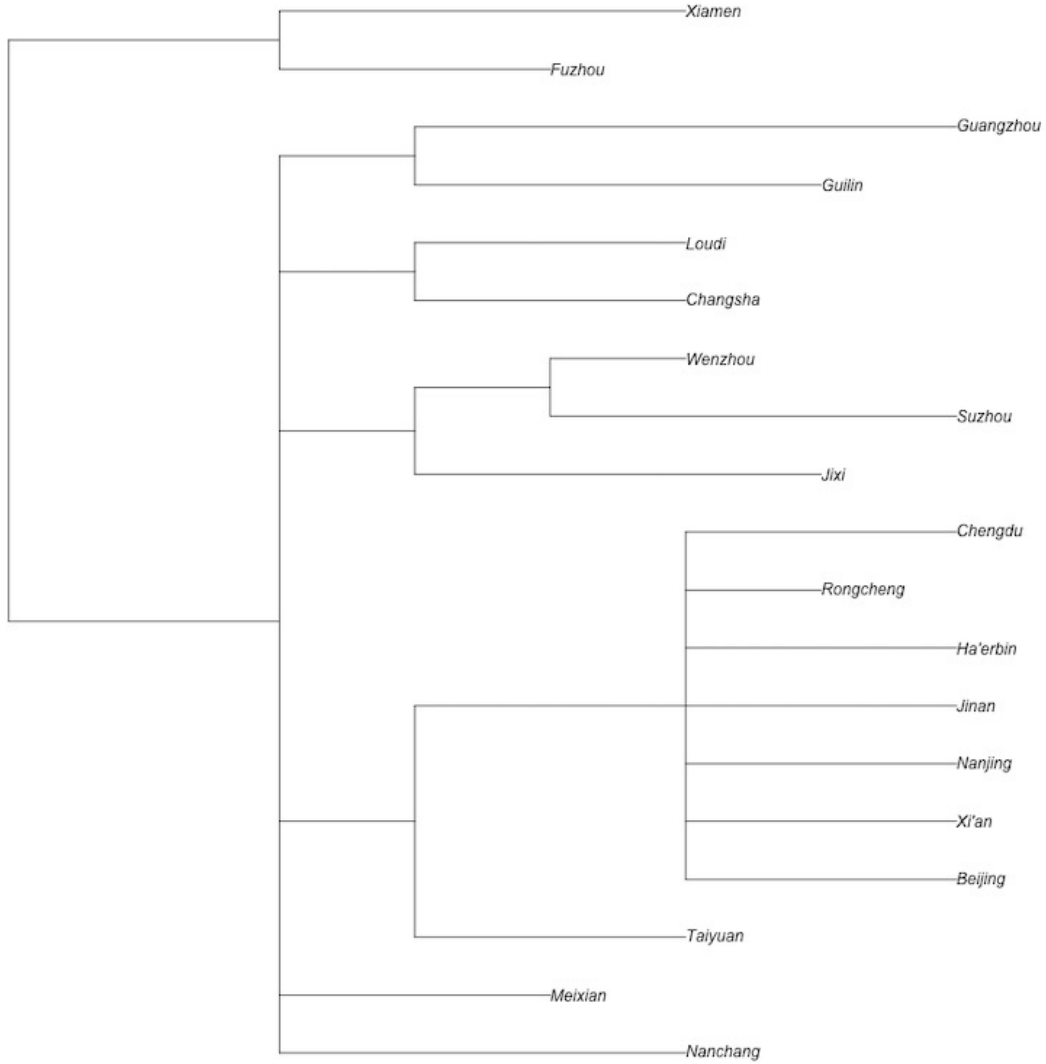
POLYNESIAN TREE



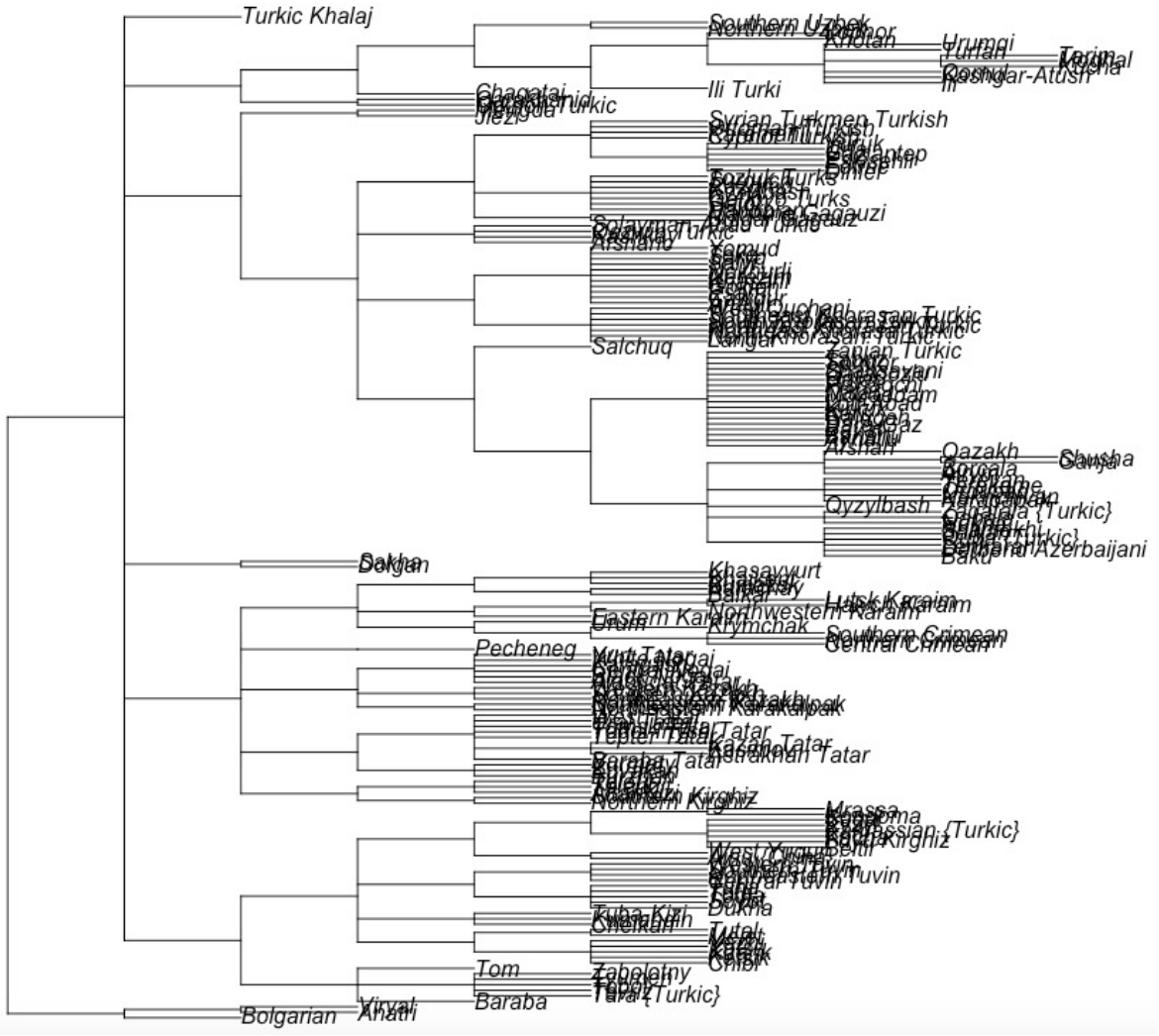


SINITIC TREE

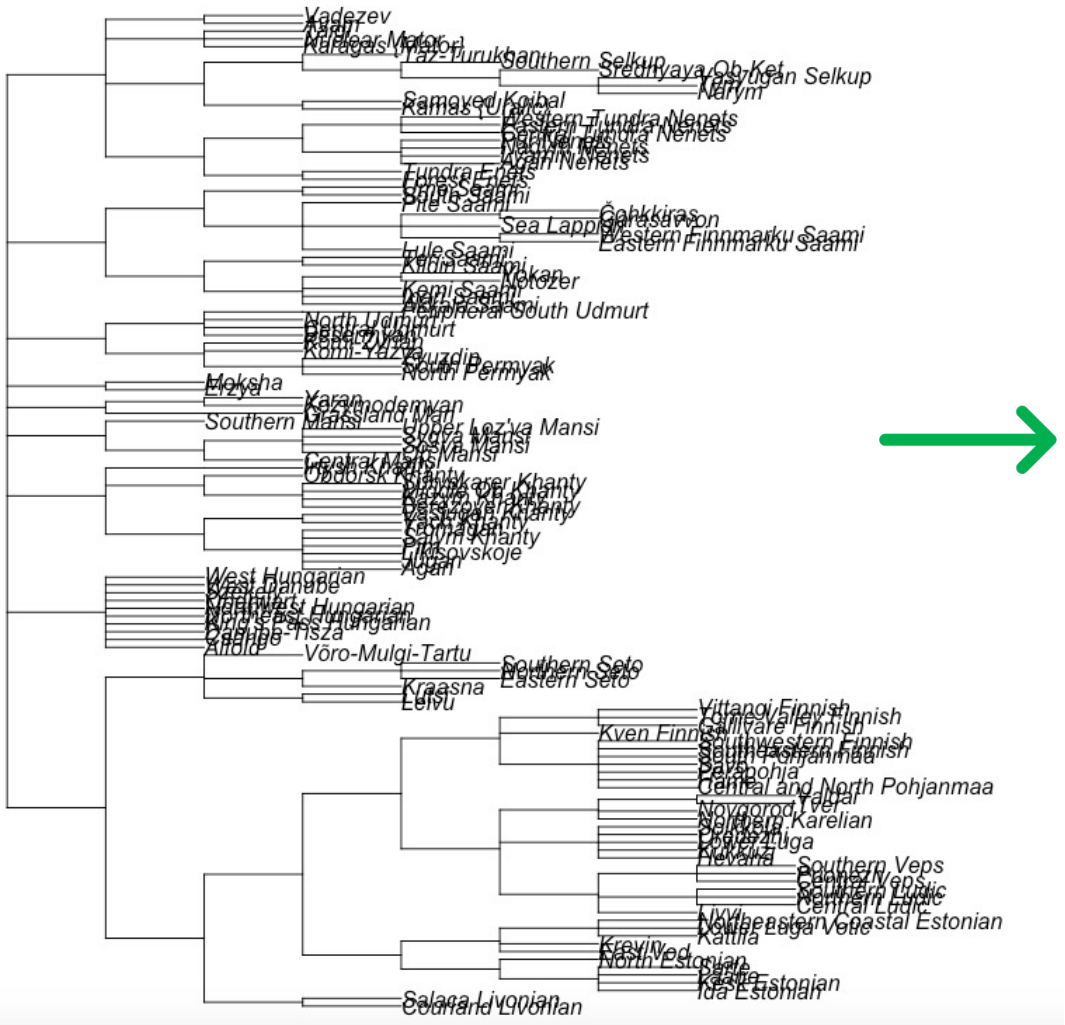


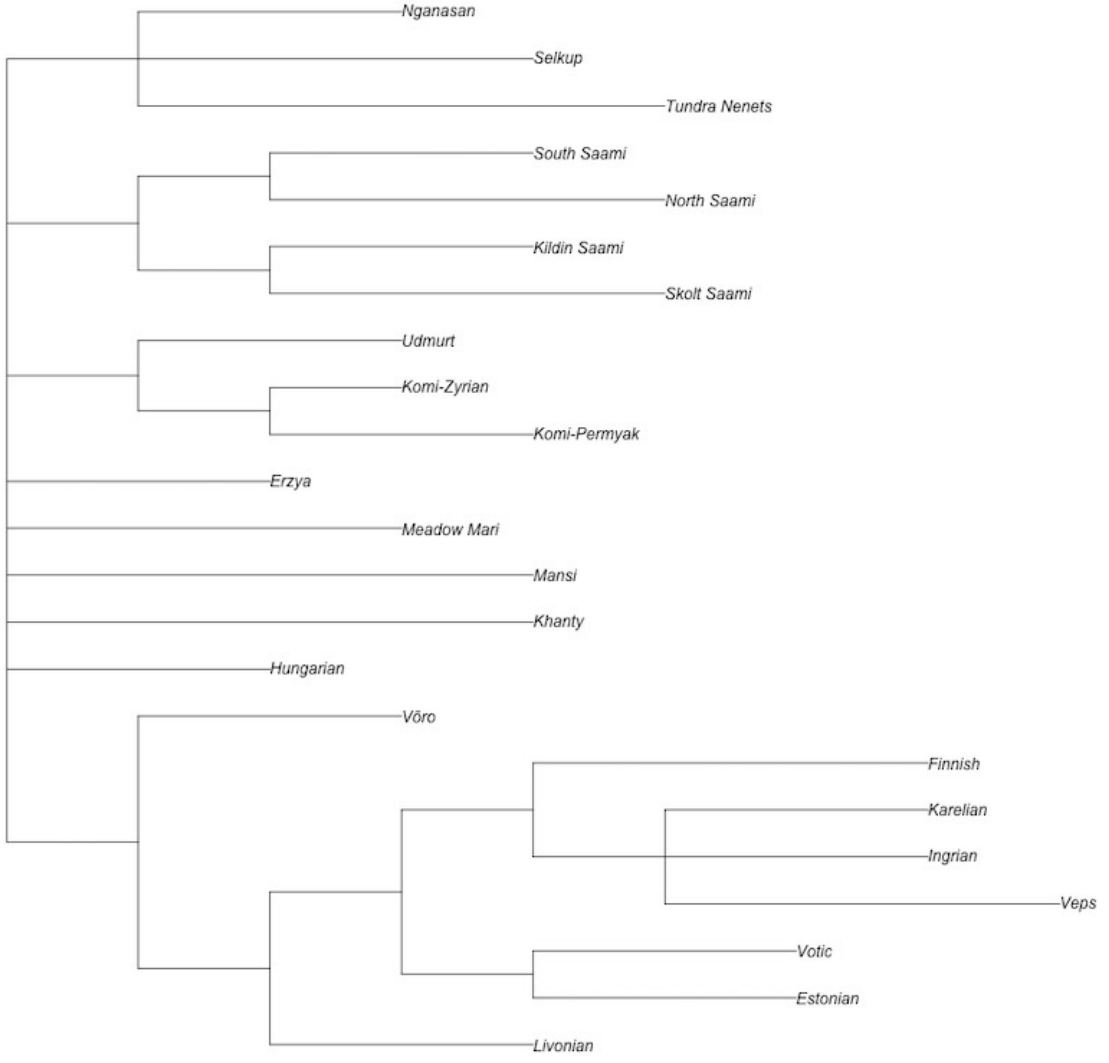


TURKIC TREE

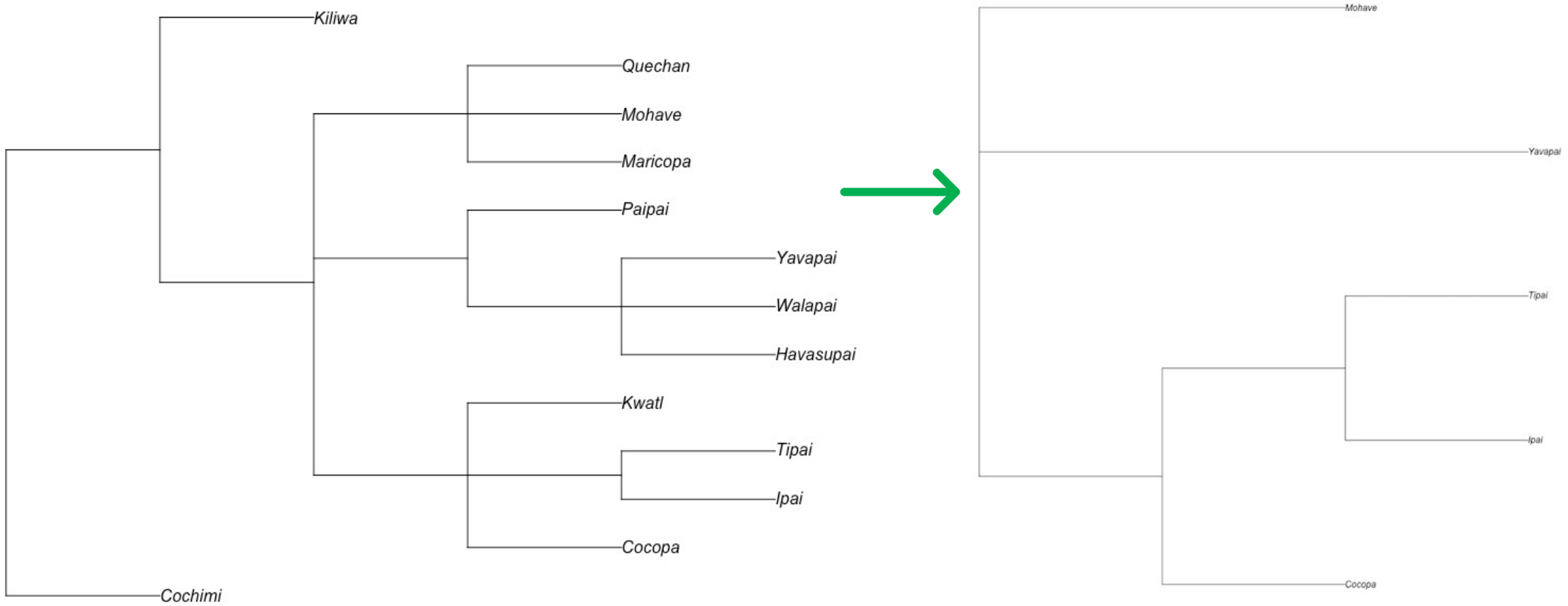

Turkic Khalaj
Chagatai
Uzbek
Uyghur
Ili Turki
Salchuq
Qazakh
Qyzylbash
Baku
Dolgan
Pecheneg
Khasavyurt
Eastern Karaim
Krymchak
Northern Kyrgyz
Tuvin
Tom
Baraba
Bolgarian
Khalaj
Uzbek
Uyghur
Old Turkic
Salar
Turkish
Gagauz
Turkmen
Azerbaijani
Yakut
Dolgan
Kumyk
Karachay-Balkar
Karaim
Crimean Tatar
Cuman
Nogai
Kazakh
Kara-Kalpak
Tatar
Baraba Tatar
Bashkir
Southern Altai
Kyrgyz
Shor
Khakas
West Yugur
Tuvan
Tofa
Northern Altai
Chulym
Baraba Tatar
Chuvash

URALIC TREE

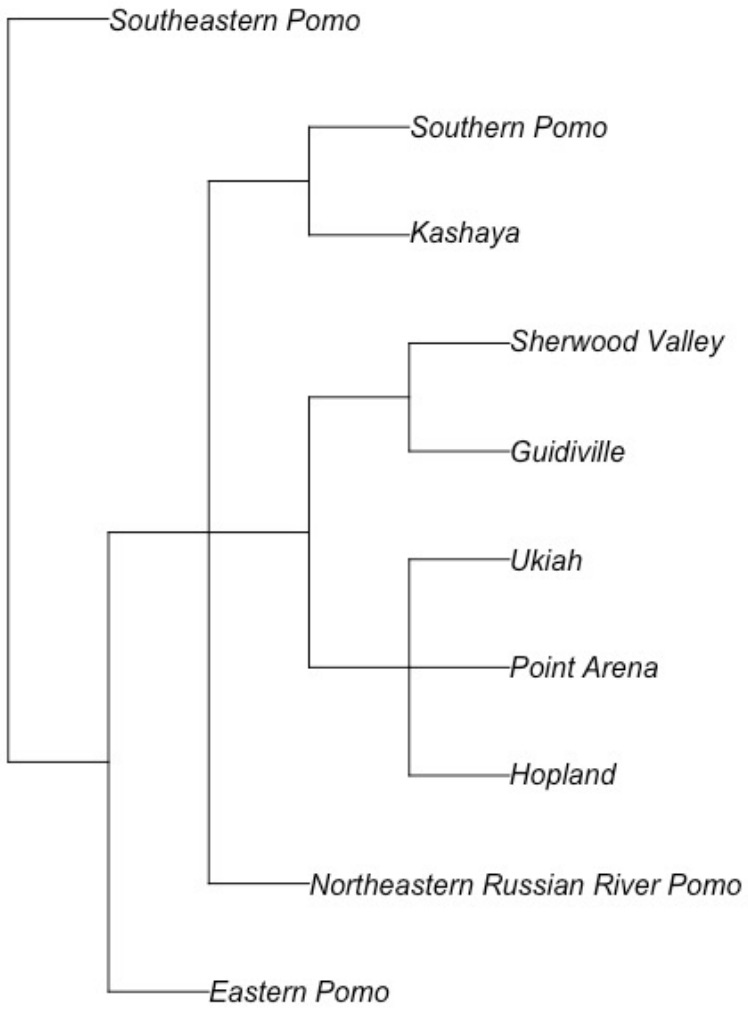


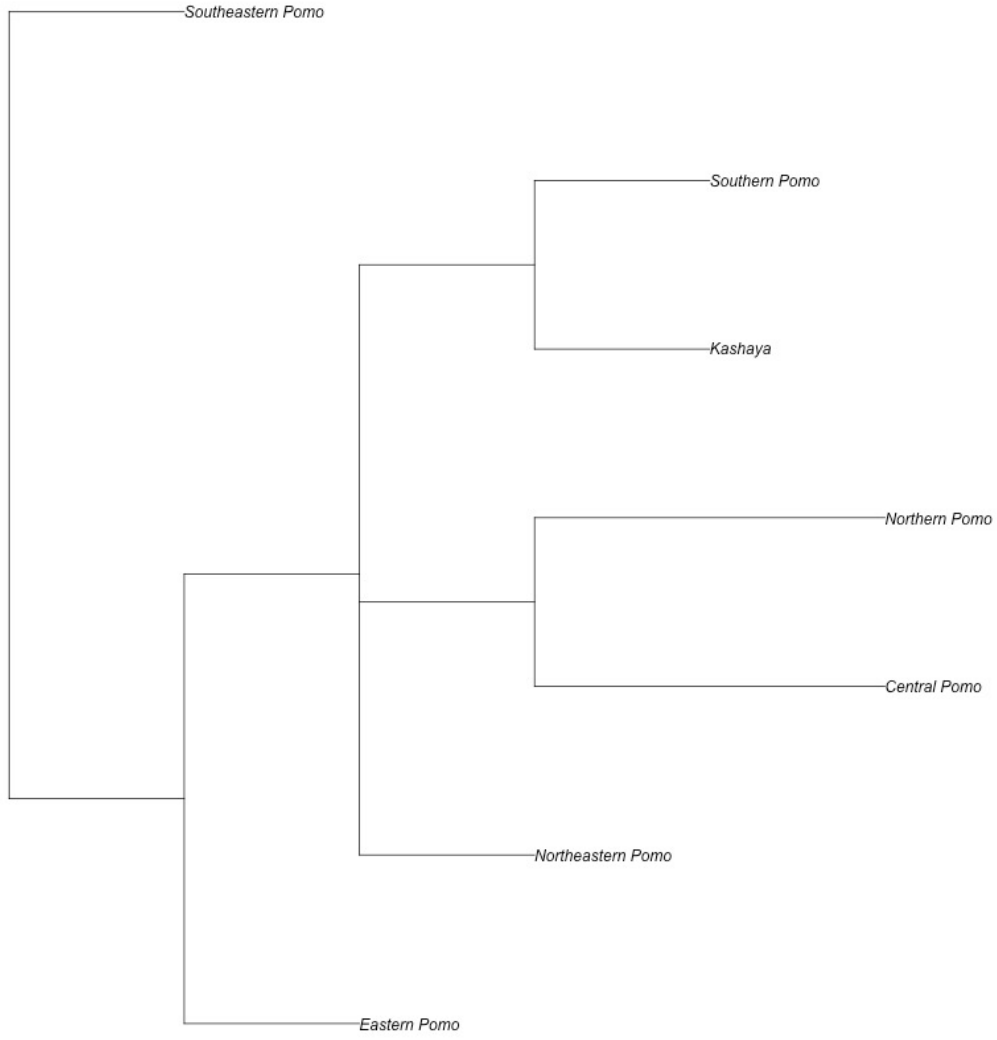

Nganasan
Selkup
Tundra Nenets
South Saami
North Saami
Kildin Saami
Skolt Saami
Udmurt
Komi-Zyrian
Komi-Permyak
Erzya
Meadow Mari
Mansi
Khanty
Hungarian
Võro
Finnish
Karelian
Ingrian
Veps
Votic
Estonian
Livonian

HOKAN: COCHIMI-YUMAN TREE

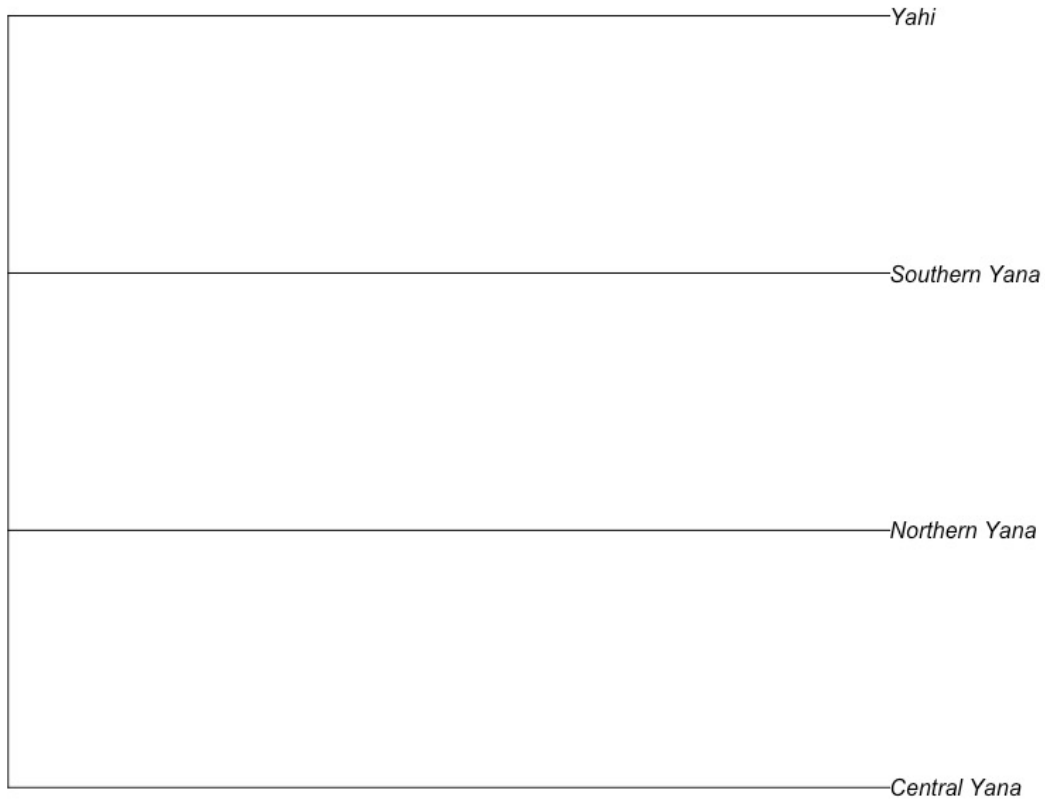


HOKAN: POMOAN TREE



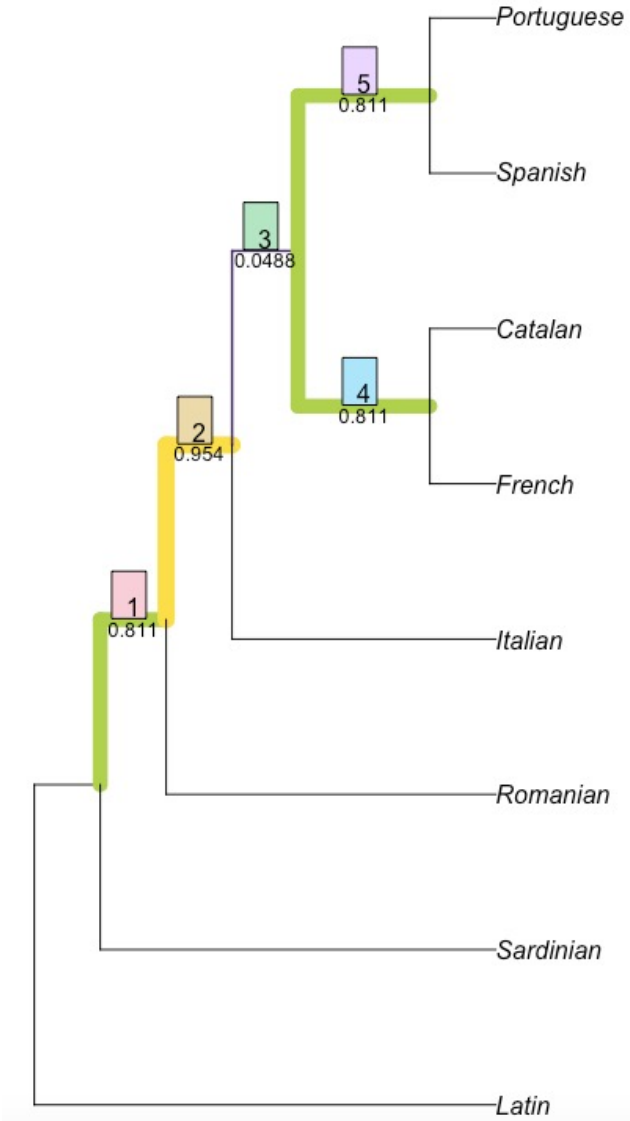


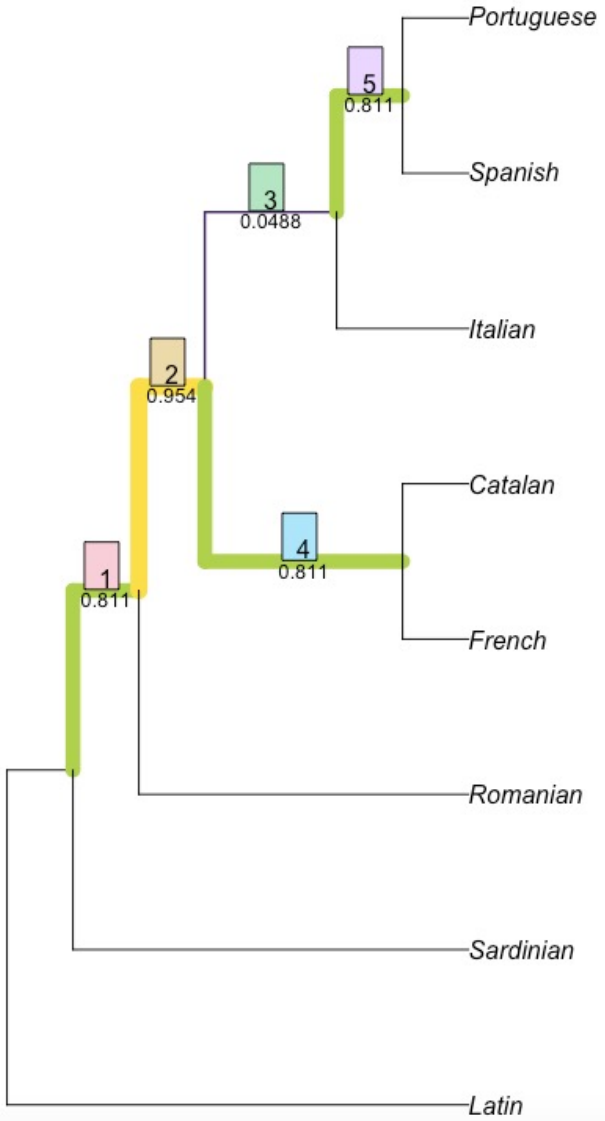
HOKAN: YANA TREE



COMPARING TREES IN R

- TreeDist package in R allows comparison between trees
- Measures topological distance between trees, various metrics
- Identical trees have distance of 0
- Based around matching and scoring splits





CONTACTS

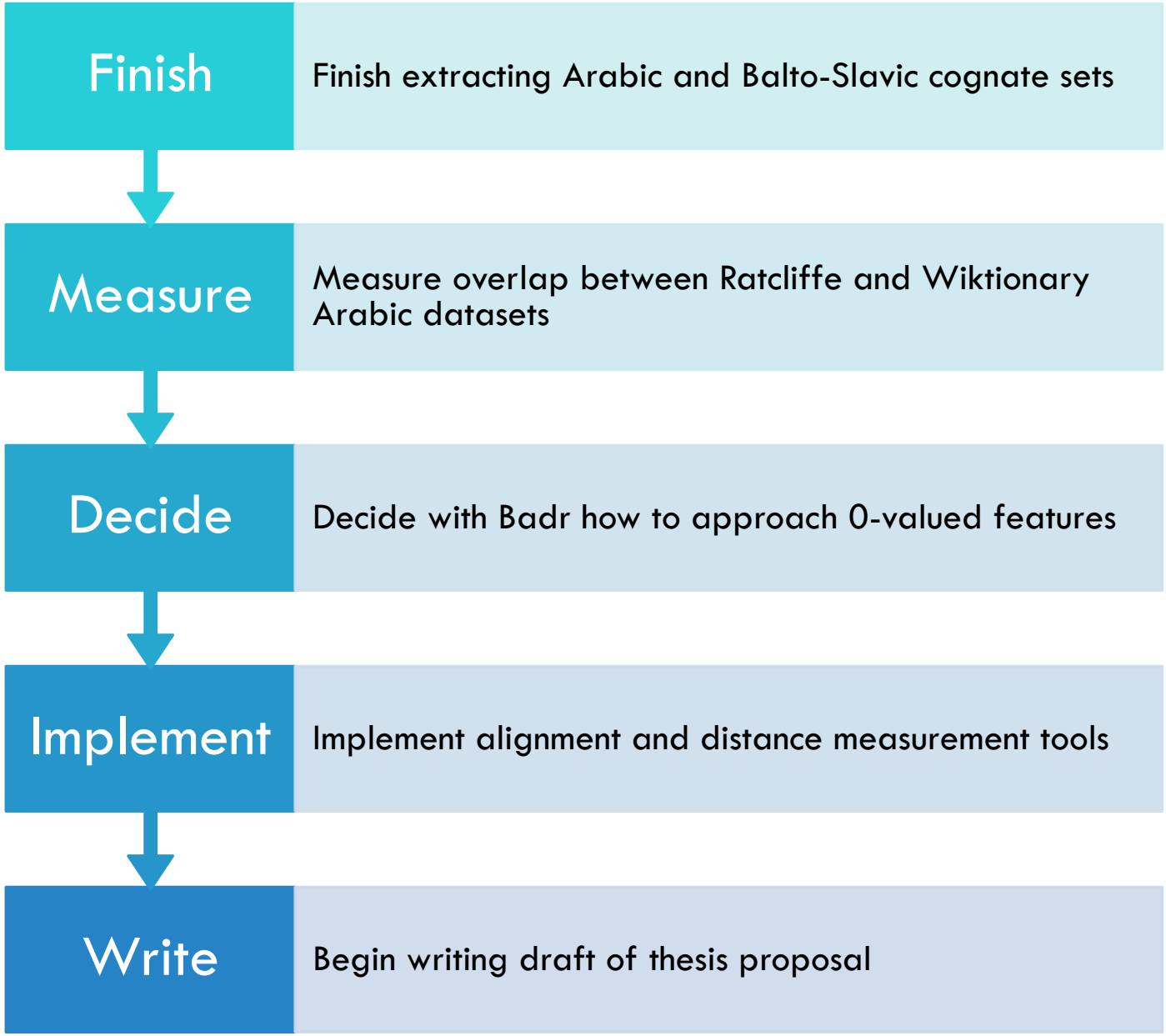
- Wrote to Badr detailed description of PHOIBLE feature coding issue
  - He will get back to me this week
  - Organized thoughts about the issue are ready to forward to Steven Moran in case there is no clear solution

- Wrote to Prof. Möbius about serving as my second thesis advisor
  - He has accepted ☺
  - Suggested that the 3 of us meet once I have prepared a draft of the thesis proposal

OFFICIAL THESIS PROPOSAL

- Want to write it within next 1-2 weeks
- Length: ~10 pages
- What exactly should it include? How detailed?
- How far in advance does the associated talk need to be scheduled?

NEXT STEPS


Finish
Finish extracting Arabic and Balto-Slavic cognate sets
Measure
Measure overlap between Ratcliffe and Wiktionary Arabic datasets
Decide
Decide with Badr how to approach 0-valued features
Implement
Implement alignment and distance measurement tools
Write
Begin writing draft of thesis proposal