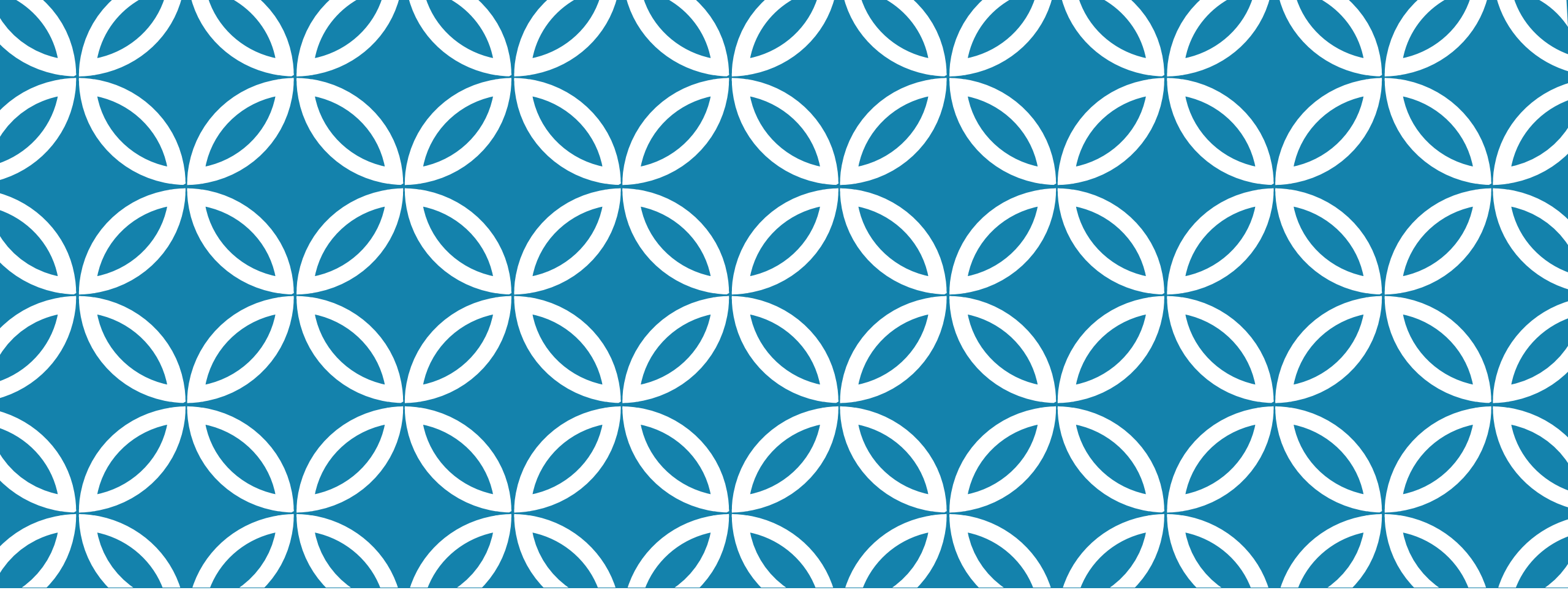
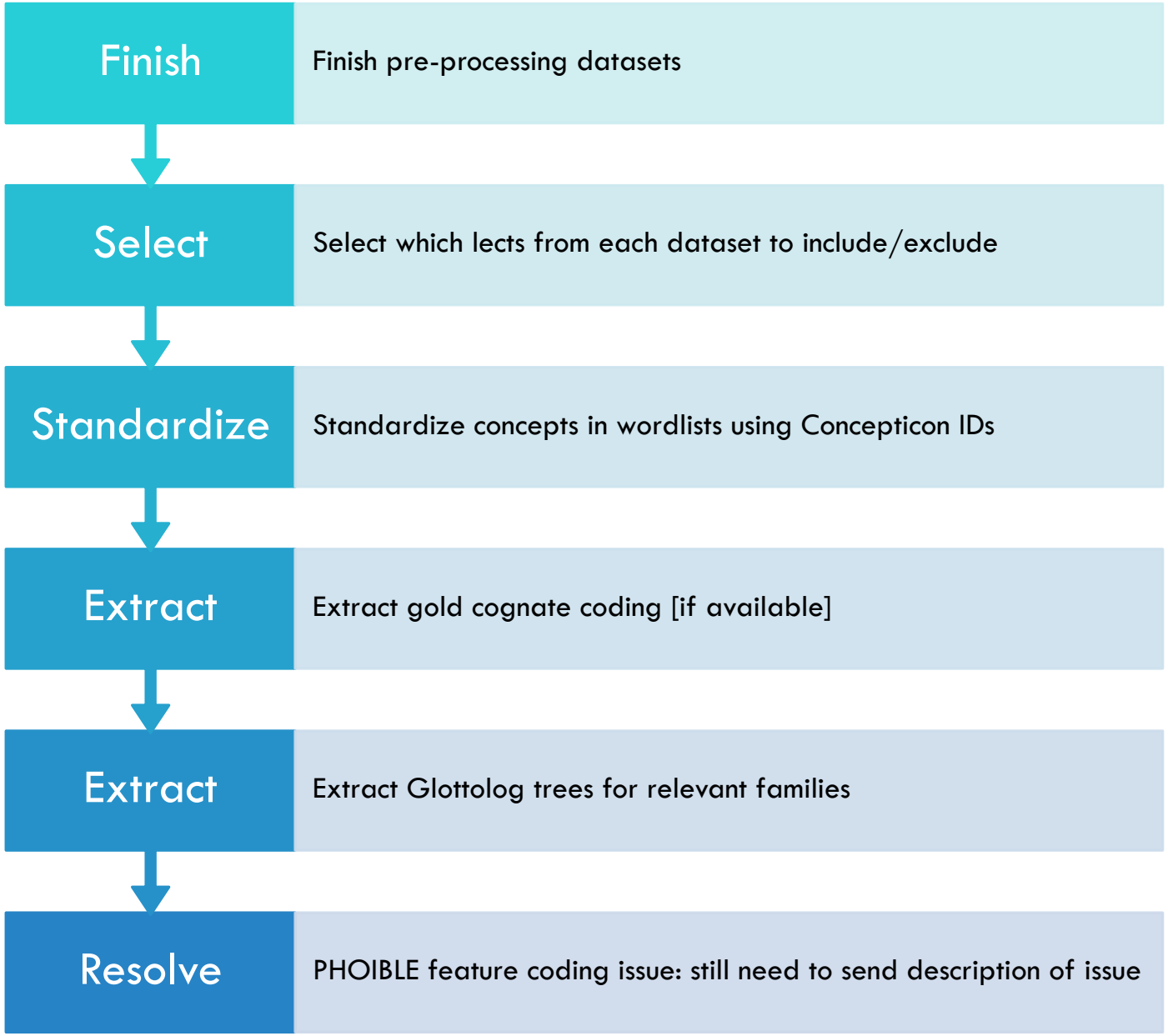


THESIS SEMINAR MEETING

Philip Georgis

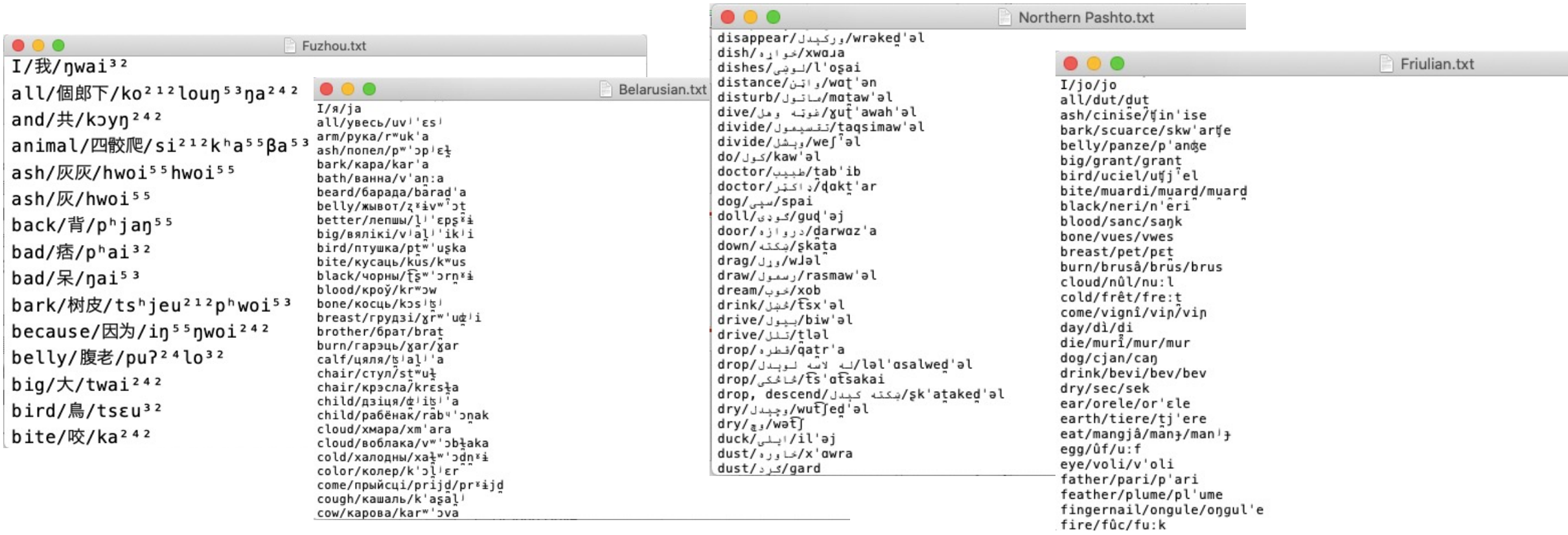
June 18, 2021

CURRENT TASKS



NEW STANDARD FORMAT

- Individual language documents → CLDF format dataset document



NEW STANDARD FORMAT

- Individual language documents → CLDF format dataset document

ID	Language_ID	Glottocode	ISO 639-3	Parameter_ID	Value	Form	Segments	Source_Form	Cognate_ID	Loan	Comment	Source
VaeakauTaumako_EIGHT_1	Vaeakau-Taumako	pile1238	piv	EIGHT	valu	valu	v a l u	valu	EIGHT_1	FALSE		Hovdhaugen-375-2009
Wallisian_EIGHT_1	Wallisian	wall1257	wls	EIGHT	valu	valu	v a l u	valu	EIGHT_1	FALSE	Pollex 06: Valu. :	POLLEX
Maori_EIGHT_1	Maori	maor1246	mri	EIGHT	waru	waru	w a r u	waru	EIGHT_1	FALSE		Biggs-85-2005
Kapingamarangi_EIGHT_1	Kapingamarangi	kapi1249	kpg	EIGHT	walu	walu	w a l u	walu	EIGHT_1	FALSE	Pollex 06: Walu.	POLLEX
Tahitian_EIGHT_1	Tahitian	tahi1242	tah	EIGHT	va'u	vaʔu	v a ʔ u	vaʔu	EIGHT_1	FALSE		Clark-173-2005
Emae_EIGHT_1	Emae	emae1237	mmw	EIGHT	βaru	βaru	β a r u	βaru	EIGHT_1	FALSE		52375
Rapanui_EIGHT_1	Rapanui	rapa1244	rap	EIGHT	va'u	vaʔu	v a ʔ u	vaʔu	EIGHT_1	FALSE		POLLEX
Mangareva_EIGHT_1	Mangareva	mang1401	mrv	EIGHT	varu	varu	v a r u	varu	EIGHT_1	FALSE		POLLEX
Luangiua_EIGHT_1	Luangiua	onto1237	ojv	EIGHT	valu	valu	v a l u	valu	EIGHT_1	FALSE	Pollex 06: Valu. :	POLLEX
TongatongaIslands_EIGHT_1	Tongan	tong1325	ton	EIGHT	valu	valu	v a l u	valu	EIGHT_1	FALSE	Pollex 06: Valu. :	117207
Tikopia_EIGHT_1	Tikopia	tiko1237	tkp	EIGHT	varu	varu	v a r u	varu	EIGHT_1	FALSE	Pollex 06: Varu. :	POLLEX
Sikaiana_EIGHT_1	Sikaiana	sika1261	sky	EIGHT	valu	valu	v a l u	valu	EIGHT_1	FALSE	Pollex 06: Valu. :	POLLEX
NorthMarquesan_EIGHT_1	North Marquesan	nort2845	mrq	EIGHT	va'u	vaʔu	v a ʔ u	vaʔu	EIGHT_1	FALSE	used on Nuku Hi	POLLEX
EastFutuna_EIGHT_1	East Futuna	east2447	fud	EIGHT	valu	valu	v a l u	valu	EIGHT_1	FALSE		POLLEX
Pukapuka_EIGHT_1	Pukapuka	puka1242	pkp	EIGHT	valu	valu	v a l u	valu	EIGHT_1	FALSE		Salisbury-152-2005
MeleFila_EIGHT_1	Mele-Fila	mele1250	mxe	EIGHT	eβaru	βaru	β a r u	βaru	EIGHT_1	FALSE		52375
AustralA_EIGHT_1	Austral A	aust1304	aut	EIGHT	vaGu	vaɢu	v a ɢ u	vaɢu	EIGHT_1	FALSE		Tamaititahio-1213-2015
Tuamotuan_EIGHT_1	Tuamotuan	tuam1242	pmt	EIGHT	varu	varu	v a r u	varu	EIGHT_1	FALSE	Pollex 06: Varu. :	POLLEX
Niuean_EIGHT_1	Niuean	niue1239	niu	EIGHT	valu	valu	v a l u	valu	EIGHT_1	FALSE	Pollex 06: Valu. :	POLLEX
AustralB_EIGHT_1	Austral B	aust1304	aut	EIGHT	vaʔu	vaʔu	v a ʔ u	vaʔu	EIGHT_1	FALSE		Meyer-128-2005
FutunaAniwa_EIGHT_1	Futuna-Aniwa	futu1245	fut	EIGHT	varu	varu	v a r u	varu	EIGHT_1	FALSE	Pollex 06: Varu. :	POLLEX
Hawaiian_EIGHT_1	Hawaiian	hawa1245	haw	EIGHT	walu	walu	w a l u	walu	EIGHT_1	FALSE	Pollex 06: Walu.	71458
Rarotongan_EIGHT_1	Rarotongan	raro1241	rar	EIGHT	varu	varu	v a r u	varu	EIGHT_1	FALSE	Pollex 06: Varu. :	POLLEX
Penrhyn_EIGHT_1	Penrhyn	penr1237	pnh	EIGHT	varu	varu	v a r u	varu	EIGHT_1	FALSE	Pollex 06: Varu. :	POLLEX
Nukuria_EIGHT_1	Nukuria	nuku1259	nur	EIGHT	varu	varu	v a r u	varu	EIGHT_1	FALSE		Davletshin-1212-2015
Samoan_EIGHT_1	Samoan	samo1305	smo	EIGHT	valu	valu	v a l u	valu	EIGHT_1	FALSE	POllex 06: Valu. :	Blust-118-2005
RennellBellona_EIGHT_1	RennellBellona	renn1242	mnv	EIGHT	baŋgu	baŋgu	b a ᵑg u	baŋgu	EIGHT_1	FALSE	eight	POLLEX
Tuvalu_EIGHT_1	Tuvalu	tuva1244	tvl	EIGHT	valu	valu	v a l u	valu	EIGHT_1	FALSE		29903
Anuta_EIGHT_1	Anuta	anut1237	aud	EIGHT	varu	varu	v a r u	varu	EIGHT_1	FALSE	eight	POLLEX
VaeakauTaumako_FIFTY_3	Vaeakau-Taumako	pile1238	piv	FIFTY	gatoaelima	gatoaelima	g a t o a e l i m a	ga+toa+e+lima	FIFTY_3	FALSE		Hovdhaugen-375-2009
Maori_FIFTY_8	Maori	maor1246	mri	FIFTY	rima tekau	rimatekau	r i m a t e k a u	rima_te+kau	FIFTY_8	FALSE		Biggs-85-2005
Mangareva_FIFTY_1	Mangareva	mang1401	mrv	FIFTY	rima rongo'uru	rimaroŋoʔuru	r i m a r o ŋ o ʔ u r u	rima_roŋo+ʔuru	FIFTY_1	FALSE		POLLEX
TongatongaIslands_FIFTY_1	Tongan	tong1325	ton	FIFTY	nimangofulu	nimaŋofulu	n i m a ŋ o f u l u	nima_ŋo+fulu	FIFTY_1	FALSE		117207
EastFutuna_FIFTY_2	East Futuna	east2447	fud	FIFTY	kaulima	kaulima	k a u l i m a	kau+lima	FIFTY_2	FALSE		POLLEX
Pukapuka_FIFTY_4	Pukapuka	puka1242	pkp	FIFTY	tinolima	tinolima	t i n o l i m a	tino_lima	FIFTY_4	FALSE	50 people	Salisbury-152-2005
Pukapuka_FIFTY_2	Pukapuka	puka1242	pkp	FIFTY	laulima	laulima	l a u l i m a	lau+lima	FIFTY_2	FALSE	traditional	Salisbury-152-2005
Pukapuka_FIFTY_1	Pukapuka	puka1242	pkp	FIFTY	limangaulu	limaŋaulu	l i m a ŋ a u l u	lima_ŋa+ulu	FIFTY_1	TRUE	(Raro.)	Salisbury-152-2005
AustralA_FIFTY_9	Austral A	aust1304	aut	FIFTY	pae ʔahuGu	paeʔahuɢu	p a e ʔ a h u ɢ u	pae_ʔa+huɢu	FIFTY_9	TRUE	Likely a Tahitian l	Tamaititahio-1213-2015
Niuean_FIFTY_5	Niuean	niue1239	niu	FIFTY	lima fiha	limafiha	l i m a f i h a	lima_fiha	FIFTY_5	FALSE		POLLEX
AustralB_FIFTY_9	Austral B	aust1304	aut	FIFTY	pae ʔaʔuru	paeʔaʔuru	p a e ʔ a ʔ u r u	pae_ʔa+ʔuru	FIFTY_9	TRUE	Tahitian loan wit	Meyer-128-2005
Hawaiian_FIFTY_6	Hawaiian	hawa1245	haw	FIFTY	kanalima	kanalima	k a n a l i m a	kana+lima	FIFTY_6	FALSE		71458
Rarotongan_FIFTY_1	Rarotongan	raro1241	rar	FIFTY	rima ngauru	rimaŋauru	r i m a ŋ a u r u	rima_ŋa+uru	FIFTY_1	FALSE		POLLEX
Nukuria_FIFTY_7	Nukuria	nuku1259	nur	FIFTY	tipu rima	tipurima	t i p u r i m a	tipu_rima	FIFTY_7	FALSE		Davletshin-1212-2015

NEW STANDARD FORMAT

- Individual language documents → CLDF format dataset document

ID	Language_ID	Glottocode	ISO 639-3	Parameter_ID	Value	Form	Segments	Source_Form	Cognate_ID	Loan	Comment	Source
VaeakauTaumako_EIGHT_1	Vaeakau-Taumako	pile1238	piv	EIGHT	valu	valu	v a l u	valu	EIGHT_1	FALSE		Hovdhaugen-375-2009
Wallisian_EIGHT_1	Wallisian	wall1257	wls	EIGHT	valu	valu	v a l u	valu	EIGHT_1	FALSE	Pollex 06: Valu. :	POLLEX
Maori_EIGHT_1	Maori	maor1246	mri	EIGHT	waru	waru	w a r u	waru	EIGHT_1	FALSE		Biggs-85-2005
Kapingamarangi_EIGHT_1	Kapingamarangi	kapi1249	kpg	EIGHT	walu	walu	w a l u	walu	EIGHT_1	FALSE	Pollex 06: Walu.	POLLEX
Tahitian_EIGHT_1	Tahitian	tahi1242	tah	EIGHT	va'u	vaʔu	v a ʔ u	vaʔu	EIGHT_1	FALSE		Clark-173-2005
Emae_EIGHT_1	Emae	emae1237	mmw	EIGHT	βaru	βaru	β a r u	βaru	EIGHT_1	FALSE		52375
Rapanui_EIGHT_1	Rapanui	rapa1244	rap	EIGHT	va'u	vaʔu	v a ʔ u	vaʔu	EIGHT_1	FALSE		POLLEX
Mangareva_EIGHT_1	Mangareva	mang1401	mrv	EIGHT	varu	varu	v a r u	varu	EIGHT_1	FALSE		POLLEX

- Uses ConceptIcon glosses as Parameter IDs
- Cross-indexed with Glottocodes and ISO 639-3 codes
- Contains data about cognate sets, borrowings, sources, and comments (when available) all together to facilitate things later on
- Required revisiting all datasets and preprocessing again → revealed that the first cursory round of preprocessing was insufficient in many cases (esp. Italic & Polynesian)

VOCABULARY/COGNATE INDICES

- Using new format dataset files, vocabulary index files can be easily generated in order to visually examine/compare word forms within cognate sets
- Cognate word forms all in the same row, loanwords enclosed in parentheses

Gloss	Aromanian	Asturian	Catalan (Ca	Catalan (Ce	Catalan (M	Catalan (M	Catalan (Nc	Catalan (Va	Dalmatian	Emiliano (C	Emiliano (F	Emiliano (R	Franco-Pro	French	Friulian	Galician	Istro-Roma	Italian (Foli	Italian (Gro	Italian (Sta	Ladin (Fass	Ladin (Garc	Latin (Arch	Latin (Late	Ligurian (G	Ligurian
FULL_1	plˈinŭ~mpl	enʌˈeno	pɬe	pɬɛ	pɬe	pɬɛ	pɬe	pɬe	plajn	piːn	pin	piːn	pʌɛː	plɛ̃	plɛŋ	tʃˈeo	pʌir	pˈjenu~pˈir	pˈjɛno	pˈjɛno	pjeŋ	plaŋ	pɬˈeːnʊs	pɬˈeːnʊs	piŋ	piŋ
GIVE_1	daw	dar							dwor	dɛːr	dar	dɛːr			da:	dar	do	da	dˈare	dˈare	der	dɛ	dˈarɛ	dˈarɛ	da:	da
GIVE_2			donˈar	dunˈa	donˈar	dunˈa	donˈa	donˈar						dɔne												
GIVE_3													baʌˈi:													
GOOD_1	bˈunŭ	boŋ	bɔ~bɔn	bɔ~bɔn	bɔ~bɔn	bɔ~bɔn	bɔ~bɔn	bɔ~bɔn	buŋ	bɔwn	bon	bɔwn	bɔ̃	bɔ̃	bɔŋ	bo	bur	bˈonu	bˈɔno	bˈwɔno	boŋ	boŋ	bˈɔnʊs	bˈɔnʊs	buŋ	buŋ
GREEN_1	vˈgarde	bˈerde	vɛrt	bɛrt	bɛrt	vɛrt	bɛrt	bɛrt	(vjard)	verd	verd	vɛjrd	vɛ	vɛʁ	vɛrt	bˈerde		vˈerde	vˈerde	vˈerde	vert	vart	wˈɪrɪdɪs	vˈɪrɪdɪs	vˈerde	vˈerde
GREEN_2																	(zelˈen)									
HAIR_1			kaβˈeʌ	kəβˈɛʌ		kəβˈɛj	kaβˈeʌ	(kaβˈeʌ)	kapˈej	kavˈiː	kavˈi	kavˈiː	(ʃevˈøː)	ʃəvø	cavˈɛj	kaβˈelo		kˈapilːu	kapˈelːi	kapˈelːo	ɕavˈej	ɕavˈəj	kˈapɨtːʊs	kˈapɨtːʊs	kavˈelːi	kavˈelːi
HAIR_2	pˈerŭ	pˈelo														pˈelo	per									
HAIR_3					mˈoɲo																					
HAND_1	mˈinə	mˈano	ma	ma	ma	ma	ma	ma	mwoŋ	maŋ	man	maːn	mã	mɛ̃	maŋ	maŋ	mər	(mˈano)	mˈano	mˈano	maŋ	maŋ	mˈanʊs	mˈanʊs	maŋ	maŋ
HEAD_1	kˈapŭ	kabˈeθa	kap	kap	kap	kap	kap	kap	kup		kɔ				caːf	kaβˈeθa	kɒp	kapˈɔʧːa			ʧef	ʧa	kˈapʊt	kˈapʊt		
HEAD_2		tˈjesta							(tˈjasta)	tˈɛsta		tˈɛsta	tˈeta	tɛt					tˈɛsta	tˈɛsta					tˈesta	tˈesta
HEAD_3																										
HEART_1		koraθˈoŋ	kɔr	kɔr	kɔr	kɔ	kɔr	kɔr	kwor	koːr	kɔr	koːr	køː	kœʁ	kuːr	koraθˈoŋ		kˈɔre	kˈɔre	kˈwɔre	ker	kwer	kɔr	kɔr	kø	kø
HEART_2																										
HEART_3	ˈinimə																jˈirimæ									

ARABIC DATASET: RATCLIFFE (2020)

- Known issues
  - Poor formatting of dataset (scraped from PDF)
  - Unclear transcription conventions
  - Obscure dialects of unclear classification
  - Cypriot Arabic forms often only as triconsonantal root

1.1 Data

		CA	Mor	Mlt	Cai	Dms	Irq	Skh	AqArb	Cyp	Glf	Ymn	Nig	Bux	Nub
1	ALL	kull	koll	Kollu	kull	kəll	kull	ʧill	kəll, sa:ɣi:n	kull	kil	kull	tʃat	kullu	kulu
2	ASH	rama:d	rˁmadˁ	rmi:d	ramaːdˁ*	rama:d	ruma:d	1	rˁama:d	ramatr–m–d	ramaad	1	1	ra'matt	ruman
3	BARK	qirfat	qeʃrˁa	ʔoʃra	ʔiʃr*	ʔəʃər	giʃra	kiʃr	səvi:ye	1	giʃra	giʃr	li:he, girfe	1	*kokobo lataka, (girifa)
4	BELLY	batˁn	kerʃ	zaʔʔ	batˁn*	batˁən	batˁin	batˁn, ʧarʃ	ʤoof	patnb–tˁ–n	batˁn	batˁn, karʃ	kirʃ	batin	batna
5	BIG	kabi:r	kbir	kibir	kibiir	kbi:r	ʧbi:r	ʧabi:r	gəbi:r	k–b–r	ʕood	Kabiir	kabi:r	ka'biir	kbir
6	BIRD	tˁa:ʔir	tˁir	ʕasfur	tˁeer*	tˁeer	tˁeer	tˁeer	ku:ʧka:ye	2	tˁeer	tˁajr	tˁe:ra	tayra	ter
7	BITE	ʕad	ʕed	gide	ʕadˁ	ʕadˁd	ʕaðˁð	ʕaðˁð	1	ʕaðð	ʕadˁd	lugus	adˁdˁa	yaʕazz	adi

ARABIC DATASET: RATCLIFFE (2020)

- New issue
  - Cognate coding only given wrt Classical Arabic
  - Only certain word forms cognate coded
  - Formatting means it would still largely need to be entered manually

1.2 Cognacy evaluations dialects and CA

Note: Ordinary black 0 represents non-cognation. Red 0 indicates non-attestation. In calculating word stability, dialects for which a particular item is unattested are factored out. When this type of pro-rating is used in calculating word stability, the result is also indicated in red.

ABC	ENGLISH	CA	CA-Mor	CA-Mlt	CA-Cai	CA-Dms	CA-Irq	CA-Skh	CA-AqArb	CA-Cyp	CA-Glf	CA-Ymn	CA-Nig	CA-Bux	CA-Nubi	word stability
1	ALL	kull	1	1	1	1	1	1	1	1	1	1	0	1	1	0.92
2	ASH	rama:d	1	1	1	1	1	0	1	1	1	0	0	1	1	1
3	BARK	qiʃr, qirfat	1	1	1	0	1	1	0	0	1	1	1	0	0	0.73
4	BELLY	batˤn	0	0	1	1	1	1	0	1	1	1	0	1	1	0.69
5	BIG	kabi:r	1	1	1	1	1	1	1	1	0	1	1	1	1	0.92
6	BIRD	tˤa:ʔir	1	0	1	1	1	1	0	0	1	1	1	1	1	0.83
7	BITE	ʕadˤdˤa	1	0	1	1	1	1	0	1	1	0	1	1	1	0.83
8	BLACK	ʔaswad	0	1	1	1	1	1	1	1	1	1	0	1	1	0.85
9	BLOOD	dam	1	1	1	1	1	1	1	1	1	1	1	1	1	1

ARABIC DATASET: RATCLIFFE (2020)

- New issue
  - Cognate coding only given wrt Classical Arabic
  - Only certain word forms cognate coded
  - Formatting means it would still largely need to be entered manually
- Looked into contacting Robert Ratcliffe about questions/issues…

  → He died in 2017, paper and dataset published posthumously in 2020

1.2 Cognacy evaluations dialects and CA

Note: Ordinary black 0 represents non-cognation. Red 0 indicates non-attestation. In calculating word stability, dialects for which a particular item is unattested are factored out. When this type of pro-rating is used in calculating word stability, the result is also indicated in red.

ABC	ENGLISH	CA	CA-Mor	CA-Mlt	CA-Cai	CA-Dms	CA-Irq	CA-Skh	CA-AqArb	CA-Cyp	CA-Glf	CA-Ymn	CA-Nig	CA-Bux	CA-Nubi	word stability
1	ALL	kull	1	1	1	1	1	1	1	1	1	1	0	1	1	0.92
2	ASH	rama:d	1	1	1	1	1	0	1	1	1	0	0	1	1	1
3	BARK	qiʃr, qirfat	1	1	1	0	1	1	0	0	1	1	1	0	0	0.73
4	BELLY	batˁn	0	0	1	1	1	1	0	1	1	1	0	1	1	0.69
5	BIG	kabi:r	1	1	1	1	1	1	1	1	0	1	1	1	1	0.92
6	BIRD	tˁa:ʔir	1	0	1	1	1	1	0	0	1	1	1	1	1	0.83
7	BITE	ʕadˁdˁa	1	0	1	1	1	1	0	1	1	0	1	1	1	0.83
8	BLACK	ʔaswad	0	1	1	1	1	1	1	1	1	1	0	1	1	0.85
9	BLOOD	dam	1	1	1	1	1	1	1	1	1	1	1	1	1	1

NEW ARABIC DATASET

"Varieties of Arabic Swadesh lists" [Wiktionary]

- https://en.wiktionary.org/w/index.php?title=Appendix:Varieties_of_Arabic_Swadesh_lists&oldid=62250595

English	Arabic (MSA)	Najdi (Riyadh)	Gulf (Emirati)	North Levantine (Central/Beirut)	Basra (Southern Iraq)	Baghdadi (Central Iraq)	Moslawi (Northern Iraq)	Sanaani (Sana'a)	Dhofari (Dhofar)	Cypriot (Kormakitis)	Egyptian (Delta)	Sudanese (Khartoum)	Chadian (Western Sudanic)	Tunisian (Tunis)	Moroccan (Casablanca)	Moroccan (Northern pre-Hilalian)	Hassaniya (Mauritanian)	Maltes
I	ʔanā	ana	āna	ʔanɑ, ʔana	ʔānah	ʔānī	anā	ʔane(h)	ānā	ana	ana, dana (statements)	ana	ana	ēne, āna	anɑ, anayɑ	āna	āna	jien
you (singular)	ʔanta (m), ʔant(i)	ənta (m), ənti (f)	entɑ (m), enti (f)	entɑ (m), ente (f)	atah (m), atī (f)	ente (m), enti (f)	ăntɑ (m), ăntē (f)	ant (m), anti (f)	intɑ (m), inti (f)	int (m), inti (f)	enta (m), enti (f)	intɑ (m), inti (f)	inta, inte (m), inti (f)	inti	ntɑ, ntayɑ (m), nti, ntiyɑ (f)	ntina	(ă)nta (m), (ă)ntiyya (f)	int, inti
he	huwa	huww(a)	(ə)hu	huwwe	hūh	huwwɑ	ōwe, -we	hū	hō	uo	howwɑ	hū	hū	huwwā	həwá	huwwa	huwwɑ, hūwɑ	huwa
we	naḥnu, ʔinnā	əḥna, (ḥ)ənnā	neḥen	naḥnā	aḥnh	eḥne	niḥnā	əḥnə(h)	naḥana	naxni	eḥna	(n)eḥna	ʔanīna	(a)ḥənā	ḥnɑ, ḥnayɑ	ḥna	(n)ăḥna (m), (n)ăḥnāti (f)	aḥna
you (plural)	ʔantum (m), ʔantunna (f)	əntum (m), ənten (f)	entu (m), enten (f)	ʔento	antm (m), antn (f)	entu (m), enten (f)	ăntim	anto	intū (m), intēn (f)	intu	entu	intu	intu(m)	ntūma	ntumɑ	ntūma	(ă)ntūmɑ (m), ntūmāti (f)	intom
they	hum (m), hunna (f)	humm (m), hənn (f)	(ə)hum (m), (ə)hen (f)	hennē	hamh (m), hanh (f)	huṃṃɑ (m), heṇṇ (f)	hăyyəm, imme	hom (m), hen (f)	hum (m), hēn (f)	innen	homma	hum (m), hun (f, rare)	human (m/f), hinna (f, rare)	hūma	humɑ	hūma	hūṃɑ (m), hūmāti (f)	huma
this	hāðā (m), hāðihī (f)	(hā)ða (m), (hā)ði (f)	hāða (m), hāði (f)	haydɑ (m), hayde (f)	hʾāḏh (m), hʾāīh (f)	hāðɑ (m), hīātā (f)	hāðā (m), hāðī (f)	ðayyāh (m), tayyih (f)	hāðī	aða (m), aðī (f)	da(h) (m), di(h) (f)	da (m), de(h) (f)	dā (m), dī (f)	hāða (m), hāði (f)	hada (m), hadi (f)	hāda (m), hādi (f)	(hā)ða (m), (hā)ði (f)	dan (m), (f)
that	ðālika (m), tilka (f), tīka (f)	(ha)ðāk (m), (ha)ðīk (f)	haðāk (m), hāðīč (f)	haydīk (m), hīdēk (f)	hḏʾākh (m), hḏīčh (f)	ðākɑ (m), ðíeče (f)	hāðēk (m), hāðīk (f)	ðayyik (m), tayyik (f)	hāðāk (m), hāðīk (f)	áðak (m), áðik (f)	dawwat (m), dayyat (f), (variants) dawwan, dawwak, dawwa, etc.	dāk (m), dēk (f)	dāk (m), dīk (f)	hāðāka (m), hāðīka (f)	hadak, dak (m), hadik, dik (f)	hadāk (m), hadīk (f)	ðāk, ðowk-huwwa (m), ðīk, ðēk-hiyyɑ (f)	dak (m), (f)
here	hunā	fiðā, həna	əhni	hown ~ hōn	hnʾāh	əhnāna, əhnā	hūnī, hāwn	hāna	hinnī	ɑwnɑ, annaxula	hena; aho, henaho(wwat) (m), ahe,	hina	hine	hūnī, hnē	hnɑ, hnayɑ	hna	hūn, hūnāti(yya)	hawn

NEW ARABIC DATASET

"Varieties of Arabic Swadesh lists" [Wiktionary]

- https://en.wiktionary.org/w/index.php?title=Appendix:Varieties_of_Arabic_Swadesh_lists&oldid=62250595
- Contains 16 dialects with sufficient coverage, mostly overlap with the 14 in Ratcliffe's dataset
- Full word forms for Cypriot Arabic vs. only triconsonantal roots
- Explicit transcription conventions in most cases

- Data (more) easily extractable automatically
  - Scraped table from Wiktionary
  - Preprocessed into proper IPA and removed all annotations
  - Maltese: automatically converted orthography → IPA using transcriptions given at Maltese Wiktionary entries (with Wiktionary parser)

TURKIC DATASET (SAVELYEV & ROBEETS, 2020)

- Extremely messy dataset format, inconsistent annotations and transcriptions
- Some mistakes in transcriptions

	A	B	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	
1	**Meaning**	**Root**	**Gagauz**		**KaraKalpak**		**KarachayBalka**		**Karaim**		**Kazakh**		**Khakas**	
2			**Standard roman**	**IPA**	**Standard roman**	**IPA**	**Standard roman**	**IPA**	**Standard roman**	**IPA**	**Standard roman**	**IPA**	**Standard roman**	**IPA**
3	**fire (n.)**	*o:t			ot	ot	ot	ot	ot	ot	ot	ot	ot	ot
4		*kaŋ-												
5	**nose (n.)**	*burUn	burnu	burnu	murïn	muruun	burun	burun	burun	burun	murïn	muruun	purun	purun
6		*tumčuk												
7		*bar-			? bar	? bar	bar	bar	bar	bar	bar	bar	par	par
8	**go (v.)**	*jorï- / *jör(i)-												
9		*kejt-	git	git	? ket	? ket			ket	ket	ket	ket		
10	**water (n.)**	*sıb	su	su	suw	suw	suw	suw	suv	suv	su	su	suɣ	suɣ
11		*agıŕ	a:z	a:z	awïz	awuuz	awuz	awuz	avuz	avuz	awïz	awuuz	aas	aas
12	**mouth (n.)**	*ańak												
13		*jü:ŕ												
14	**tongue (n.)**	*tIl	dil	dil	til	til	til	til	til	til	tïl	təl	təl	təl
15	**blood (n.)**	*kia:n	kan	kan	ḳan	qan	ḳan	qan	kan, (k) ḳan	kan, (k) qan	ḳan	qan	χan	χan
16	**bone (n.)**	*siŋök			süyek	syjek	süyek	syjek	sivek, (k) süyek	sivek, (k) syjek	süyek	syjek	söök	søøk
17		*kemük	kemik	kemik										
18	**2SG pronoun**	*sę												
19		*sen	sän	sɐn	sen	sen	sen	sen	sen	sen	sen	sen	sin	sin
20		*tąmor			tamïr	tamuur	tamïr	tamuur	tamur	tamur	tamïr	tamuur		
21		*jïldïŕ												
22		*taŕıl = *tasıl												
23	**root (n.)**	*kök	kök	køk										
24		*tüp												
25		*ö:ŕ												
26	**come (v.)**	*gęl	gel	gel	kel	kel	kel	kel	kel	kel	kel	kel	kil	kil
27	**breast (n.) // (chest) (n.)**	*köküR(ek)	gü:s	gy:s	kökürek	køkyrek	kökürek	køkyrek	kökräk, kekrek, k	køkrɐk, kekrek, k	kökrek	køkrek	kögɔs, köksɔ	køgɔs,
28		*tiö:Ĺ			? tös	? tøs	töš	tøʃ	töš, tes	tøʃ, tes	tös	tøs		
29		*kebde			gewde	gewde					kewde	kewde		
30		*jag-	ya:mur	ja:mur	žawïn, (žamɣïr)	ʒawuun, (ʒamɣuur	jaŋur, (Balk.) jaw	dzaŋur, (Balk.) dz	yamɣur	jamɣur	žaŋbïr, žawïn	ʒaŋbuur, ʒawuun	naŋmïr	naŋmu

TURKIC DATASET (SAVELYEV & ROBEETS, 2020)

- Extremely messy dataset format, inconsistent annotations and transcriptions
- Some mistakes in transcriptions
- Fully cleaned and standardized now
- Solutions to some issues:
  - **Turkish:**
    - /ɣ, q/ → /ɯ, k/
    - Corrected any individual vocabulary entries which I noticed were wrong (yʃtynde → ystynde)
  - **Turkmen:**
    - /s, z/ → /θ, ð/
  - Checked with NorthEuraLex's Turkic and with Turkic dataset at http://turkic.elegantlexicon.com/ when in doubt for other languages

URALIC DATASETS

- Originally planned on using NorthEuraLex's Uralic dataset (26 languages)
- Problem: doesn't include gold cognate set coding, I'm not familiar enough with Uralic languages to attempt it myself
- Found new Uralic dataset: De Heer et al., Syrjänen [submitted manuscript]
  - https://zenodo.org/record/4777568#.YMXcH5MzbjA
  - Includes cognate coding
  - CLDF format
  - Language coverage is more or less the same as NorthEuraLex's Uralic
  - Still need to preprocess transcriptions

COGNATE CODING: ITALIC

- Words already coded into cognate sets, but all words marked as borrowings lumped together into one set, regardless of source (code = -1)
- Manually re-coded the words marked as borrowings (dataset notes provided the source of borrowing in most cases)
- New cognate ID is the negative cognate set ID which it was borrowed from
- Important in case loanwords are not to be excluded from analysis

Language	Word_ID	Gloss	New CognateID	IPA	Notes	Source Notes
Istro-Romanian	IstroRomanian_BURN (SOMETHING)_-1	BURN (SOMETHING)	-5	pal'i		Borrowed from Croatian {paliti} 'to burn'. Distinct from 'ɒrde {ârde} 't
Occitan	ProvençalOccitan_BURN (SOMETHING)_-1	BURN (SOMETHING)	-1	bryl'a		There are two expressions for 'to burn': brül-'a {brula} and krem-'a {cre
Istro-Romanian	IstroRomanian_CLOUD_-1	CLOUD	-2	obl'ɒk		There is no word for 'cloud' either in Kovačec's dictionary or in the tex
Istro-Romanian	IstroRomanian_DOG_-1	DOG	-4	brek		Probably borrowed from Italian {bracco} 'hound' of Germanic origin.
Occitan	ProvençalOccitan_DOG_-1	DOG	-1	tʃiŋ		Borrowed from French {chien} 'dog'.
Catalan (Castelló de la Plana)	CastellódelaPlanaCatalan_EAT_-1	EAT	-2	meɲdʒ'ar		Borrowed from Old French {mangier} 'to eat'.
Catalan (Central)	CentralCatalan_EAT_-1	EAT	-2	məɲʒ'a		Borrowed from Old French {mangier} 'to eat'.
Catalan (Manises)	ManisesCatalan_EAT_-1	EAT	-2	meɲtʃ'ar		Borrowed from Old French {mangier} 'to eat'.
Catalan (Minorcan)	MinorcanCatalan_EAT_-1	EAT	-2	məɲʒ'a		Borrowed from Old French {mangier} 'to eat'.

COGNATE CODING: SLAVIC AND ARABIC DATASETS

- No gold cognate set coding included for NorthEuraLex Slavic or Wiktionary Arabic datasets

- Idea: semi-automatic with manual correction
  - Try using automatic cognate detection prototype, manually correct results
  - Familiar enough with Slavic that I could do this (have done it previously)
  - Could attempt same for Arabic, generally clear when words are related across Arabic dialects; also can use Ratcliffe's cognate coding as partial reference
  - BUT would need to limit the number of concepts if manual, max 100-200 (NorthEuraLex has >900 concepts)

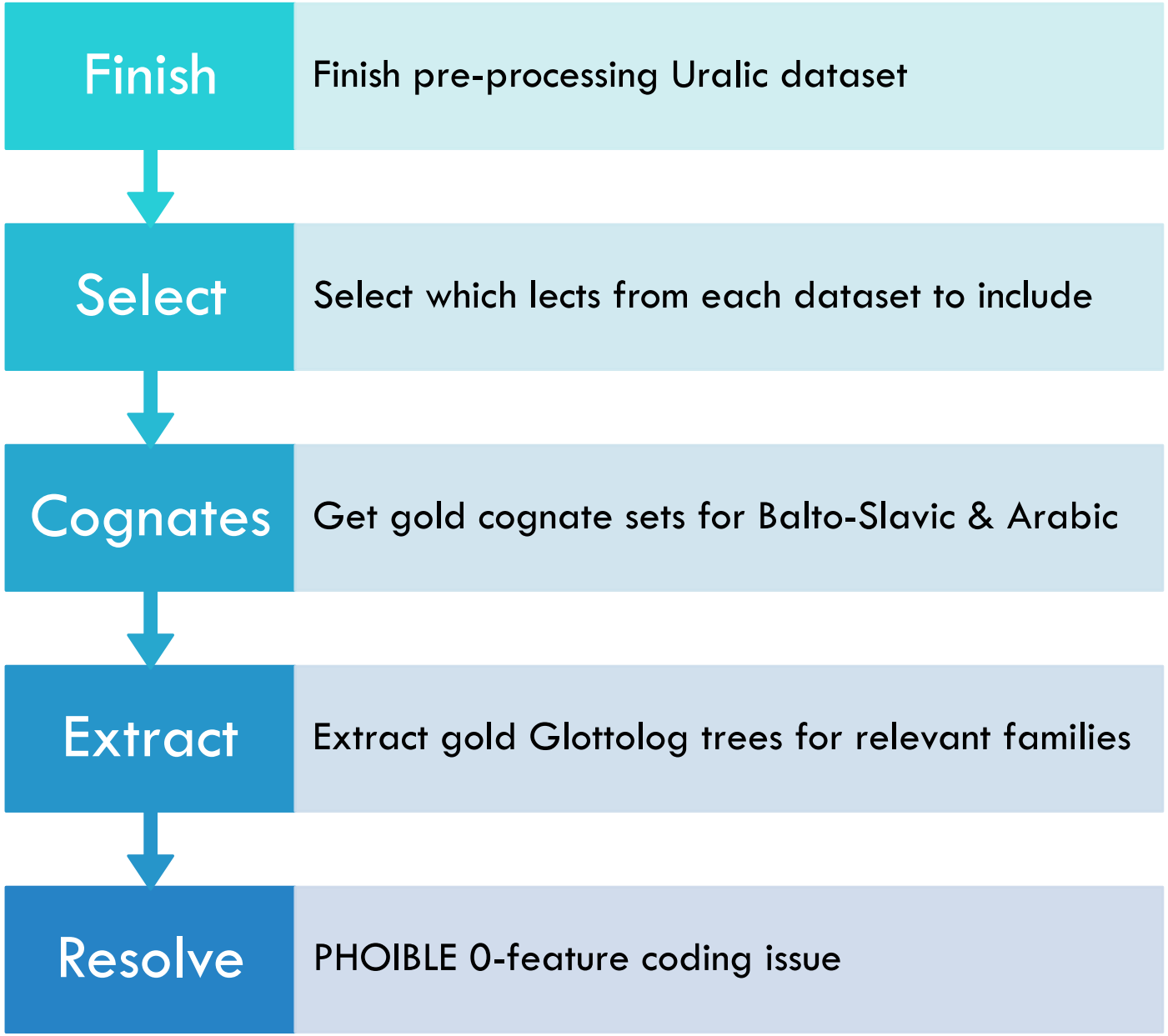
OVERVIEW OF DATASETS

Family	Source Name	Reference	Number of Varieties
Arabic	Varieties of Arabic Swadesh lists	Wiktionary	16
Italic	Global Lexicostatistic Database	Saenko (2016)	58
(Balto-)Slavic	NorthEuraLex	Dellert et al. (2019)	9 (+2 Baltic)
Uralic	Uralic basic vocabulary with cognate and loanword information	De Heer et al.; Syrjänen	27
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

STATUS OF DATASETS

Dataset	Fully Preprocessed Transcriptions	Concepticon Cross-Reference	Standardized Format	Gold Cognate Sets Included	Extracted Glottolog Tree
Arabic (Ratcliffe)	✓	✓	X	partially	X
Arabic (Wiktionary)	✓	✓	✓	X	X
Italic	✓	✓	✓	✓	X
(Balto-)Slavic	✓	✓	✓	X	X
Polynesian	✓	✓	✓	✓	X
Sinitic	✓	✓	✓	✓	X
Turkic	✓	✓	✓	✓	X
Uralic (NorthEuraLex)	✓	✓	✓	X	X
Uralic (De Heer, et al.; Syrjänen)	X	✓	✓	✓	X

STILL TO DO...


Finish
Finish pre-processing Uralic dataset
Select
Select which lects from each dataset to include
Cognates
Get gold cognate sets for Balto-Slavic & Arabic
Extract
Extract gold Glottolog trees for relevant families
Resolve
PHOIBLE 0-feature coding issue