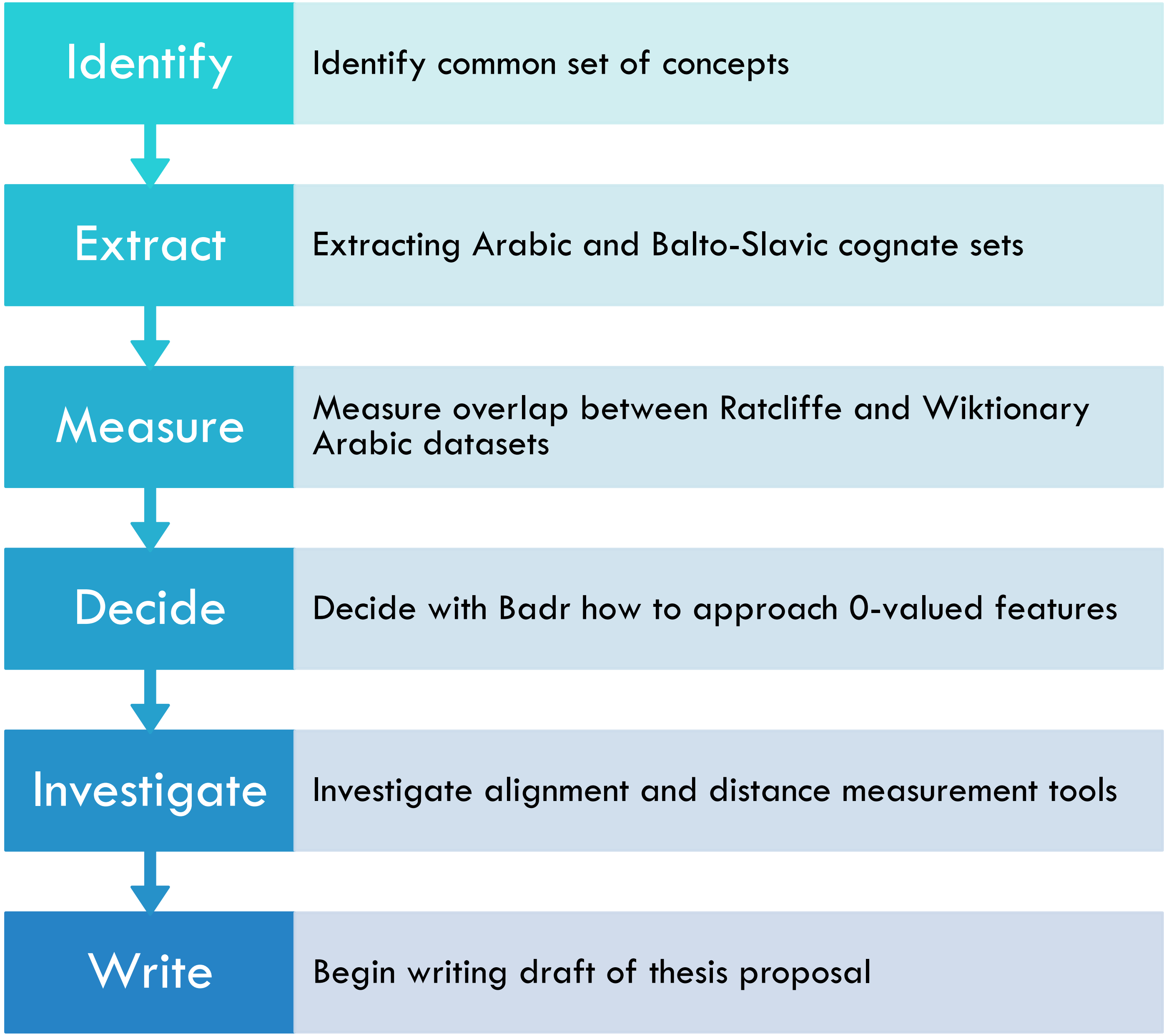
THESIS SEMINAR MEETING

Philip Georgis
July 12, 2021

TASKS FROM LAST TIME



Identify

Identify common set of concepts

Extract

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

Investigate

Investigate alignment and distance measurement tools

Write

Begin writing draft of thesis proposal

COMMON CONCEPT SET

- Different datasets use different labels for same concepts:

MEAT	Arabic, Balto-Slavic, Dravidian, Hokan, Turkic
FLESH	Sinitic
MEAT OR FLESH	Italic, Polynesian, Uralic

  - Other examples: “STONE” vs. “ROCK”, “WOMAN” vs. “FEMALE PERSON”, “WARM” vs. “HOT” vs. “WARM (OF WEATHER)”, etc.

- Matched each concept with “base concept”, e.g. “MEAT” for all label variations
- >1000 unique concepts so couldn’t inspect all, but I found the main “culprits”

COMMON CONCEPT SET

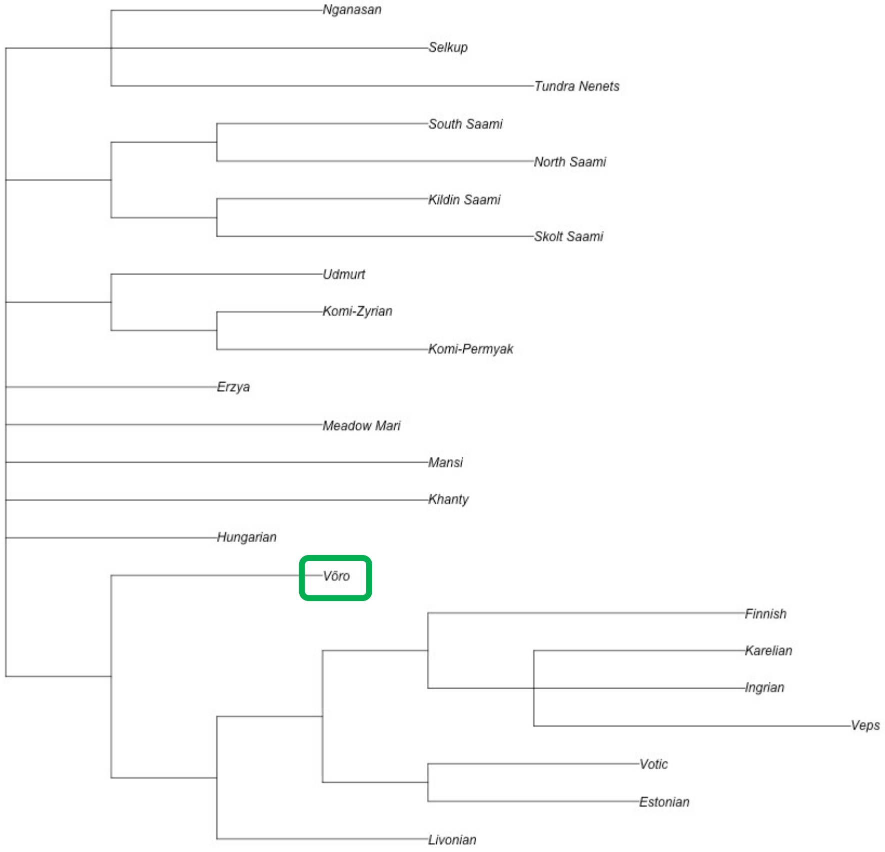
- Identified base concepts appearing in at least 7 out of 9 datasets (i.e. missing from maximum 2 datasets)
  - 110 concepts: includes all concepts from Swadesh 100 list

Dataset	Number of Concepts Included	Average Mutual Coverage
Arabic	110	0.96
Balto-Slavic	109	0.76
Dravidian	100	0.86
Hokan	110	0.82
Italic	110	1.00
Polynesian	97*	0.91
Sinitic	110	1.00
Turkic	108	0.88
Uralic	110	0.79

** Polynesian dataset oddly missing some very common concepts for some reason, incl. basic Swadesh list items, e.g. SUN, HEART, FINGERNAIL, TREE, etc.*

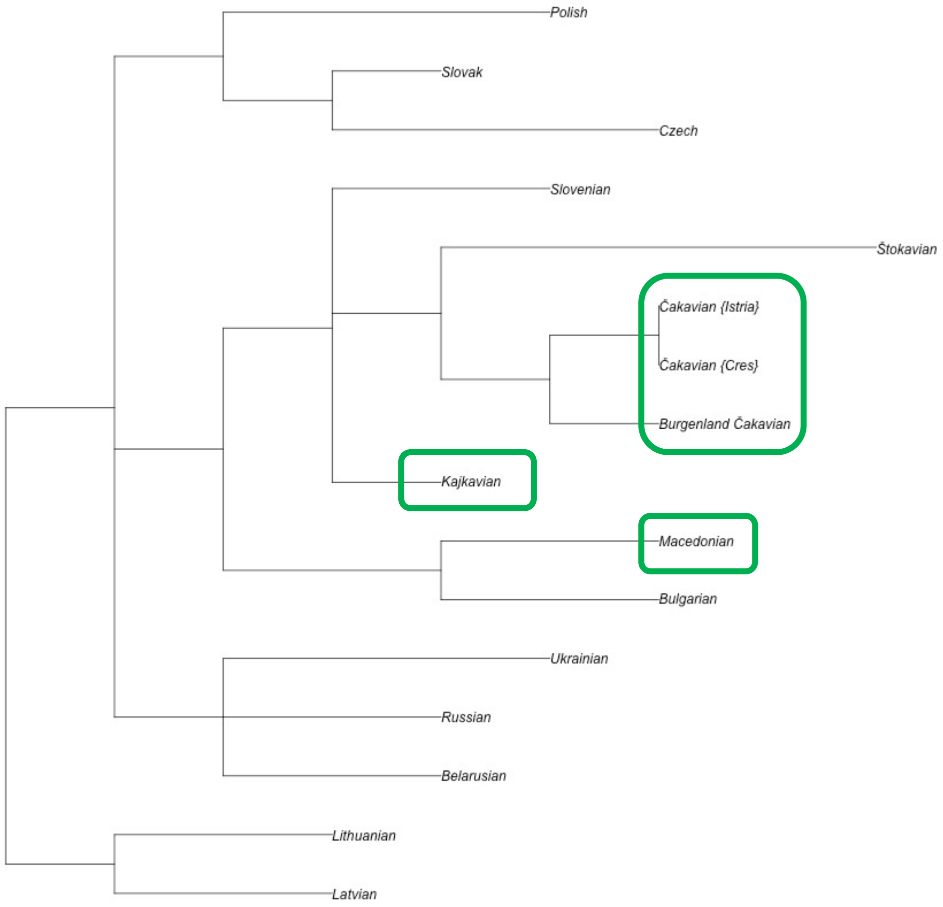
URALIC DATASET: VÕRO LANGUAGE

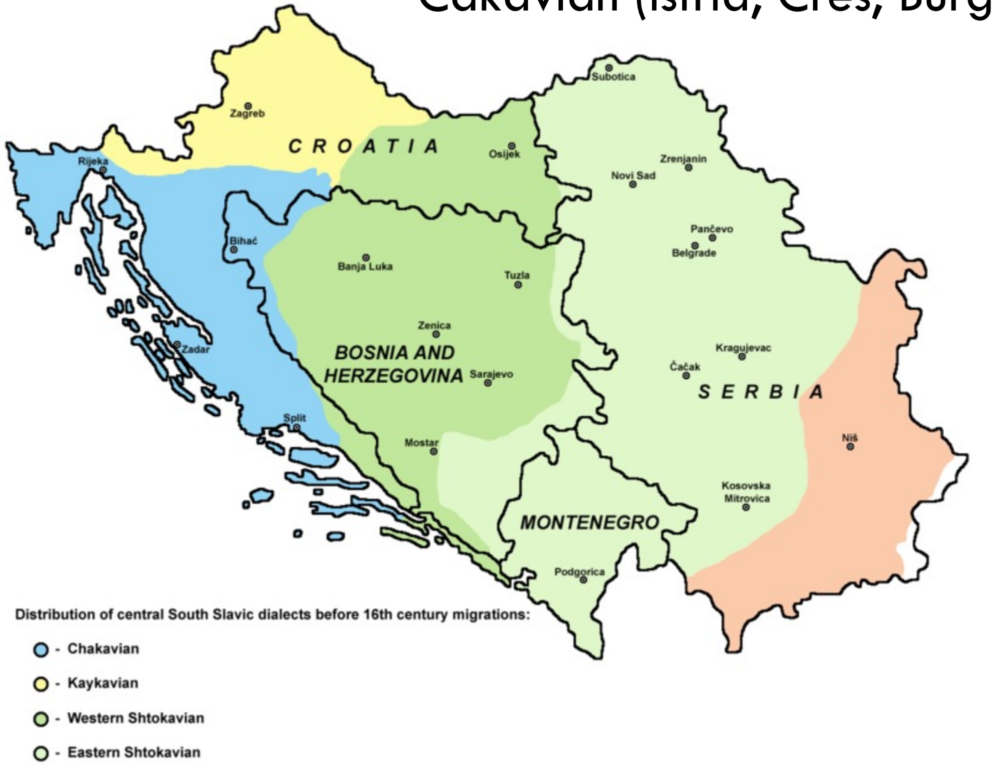
- Previously: excluded Võro because only 1 transcription was available in UraLex
- Looked into Võro data in more detail
- Orthographic data is available, but the Võro orthography has a nearly 1:1 correspondence with phonemes
- Instead wrote a G2P tool to automatically transcribe Võro lexical data into IPA
- Võro is the sole representative of a separate branch of Finnic languages, good to include it if possible
- Fun fact: Võro distinguishes 3 degrees of length in consonants and vowels, e.g. /k, kˑ, kː/



ADDITIONAL SLAVIC DATA

- Added data from Global Lexicostatistical Database for 5 new Slavic varieties:
  - Macedonian
  - Kajkavian
  - Čakavian (Istria, Cres, Burgenland)





Distribution of central South Slavic dialects before 16th century migrations:

BALTO-SLAVIC (NORTHEURALEX) DATASET

- Started manually correcting the LexStat-organized cognate sets, but quickly noticed that there was a problem with many NorthEuraLex transcriptions
- Many transcriptions had errors…

BALTO-SLAVIC (NORTHEURALEX) DATASET

- Started manually correcting the LexStat-organized cognate sets, but quickly noticed that there was a problem with many NorthEuraLex transcriptions
- Many transcriptions had errors…

e.g. Croatian <ije> represents /jeː/, the long version of <je>

<zvijezda>	/zʋjeːzda/	*/zʋ**i**jɛzda/	*phantom <i>!*
<pijesak>	/pjeːsak/	*/p**i**jɛːsak/	
<korijen>	/korjeːn/	*/kɔr**i**ɛn/	

BALTO-SLAVIC (NORTHEURALEX) DATASET

- Started manually correcting the LexStat-organized cognate sets, but quickly noticed that there was a problem with many NorthEuraLex transcriptions
- Many transcriptions had errors…

e.g. Russian

<имя>	/imʲə/	*/jɨˑmʲɪ/
<гладкий>	/gɫa**tkʲ**ɪ̯ɪ/	*/gɫaˑ**dk**ij/
<шесть>	/ʂ**ɛsʲ**tʲ/	*/ʃ**əs**tʲ/
<далёкий>	/dɐlʲɵkʲɪ̯ɪ/	*/dɐ**lokij**/
<считать>	/**ɕ**ːɪtatʲ/	*/**st͡ʃʲ**itaˑtʲ/

BALTO-SLAVIC (NORTHEURALEX) DATASET

- Started manually correcting the LexStat-organized cognate sets, but quickly noticed that there was a problem with many NorthEuraLex transcriptions
- Many transcriptions had errors…

e.g. Polish, Czech, Slovak, Ukrainian, Belarusian, Bulgarian…

PL	<śnieg>	/ɕ**ɲ**ɛk/	*/ɕ**ni**ɛk/
CZ	<stříbro>	/**str̝̊**iːbro/	*/**zdr̝**iːbro/
SK	<sedem>	/sɛ**ɟ**ɛm/	*/sɛ**d**ɛm/
UK	<ягода>	/jˈɑɦɔd**ɐ**/	*/jɑɦɔd**ɑ**/
BE	<блізкі>	/b**l**ʲi**s**ʲ**k**ʲi/	*/b**l**i**zk**i/
BG	<лежа>	/lɛʒˈ**ɤ**/	*/lɛʒ**a**/

BALTO-SLAVIC (NORTHEURALEX) DATASET

- Started manually correcting the LexStat-organized cognate sets, but quickly noticed that there was a problem with many NorthEuraLex transcriptions
- In some cases even the words were incorrect translations:

- e.g. ‘WHITE’ translated into Ukrainian as <сивий> (actually means ‘gray’)

  ‘HAND’ translated into Polish as <dłoń> (actually means ‘palm of the hand’)

  ‘DOG’ translated into Croatian as <pseto> (actually a derogatory term for a dog or a person)

  etc…

BALTO-SLAVIC (NORTHEURALEX) DATASET

- Read paper introducing NorthEuraLex dataset to investigate sources…
- Turns out it was compiled semi-automatically from dictionaries by non-experts/non-speakers of the languages
- Most transcriptions were done automatically using grapheme-to-phoneme conversion tools based on phonological descriptions of the language
- How to solve this?

BALTO-SLAVIC (NORTHEURALEX) DATASET

- Solutions

1) Incorrect translations: replace or remove obviously incorrect translations

PL	‘HAND’	*dłoń	→	ręka
UK	‘WHITE’	*сивий	→	білий
HR	‘DOG’	*pseto	→	Ø

2) Incorrect transcriptions: fix transcriptions

BALTO-SLAVIC (NORTHEURALEX) DATASET

- Solutions

1) Incorrect translations: replace or remove obviously incorrect translations

2) Incorrect transcriptions: fix transcriptions

- How to fix transcriptions?
  - Modify existing transcriptions: e.g. Croatian <ije> issue is easy to fix automatically
  - Issues in other languages are more complex…

e.g. impossible to fix Czech <přijít> */br̝ɪjiːt/ → /pr̝̊ɪjiːt/ without reference to orthography

(if the word had been spelled <břijít> this transcription would have been correct)

BALTO-SLAVIC (NORTHEURALEX) DATASET

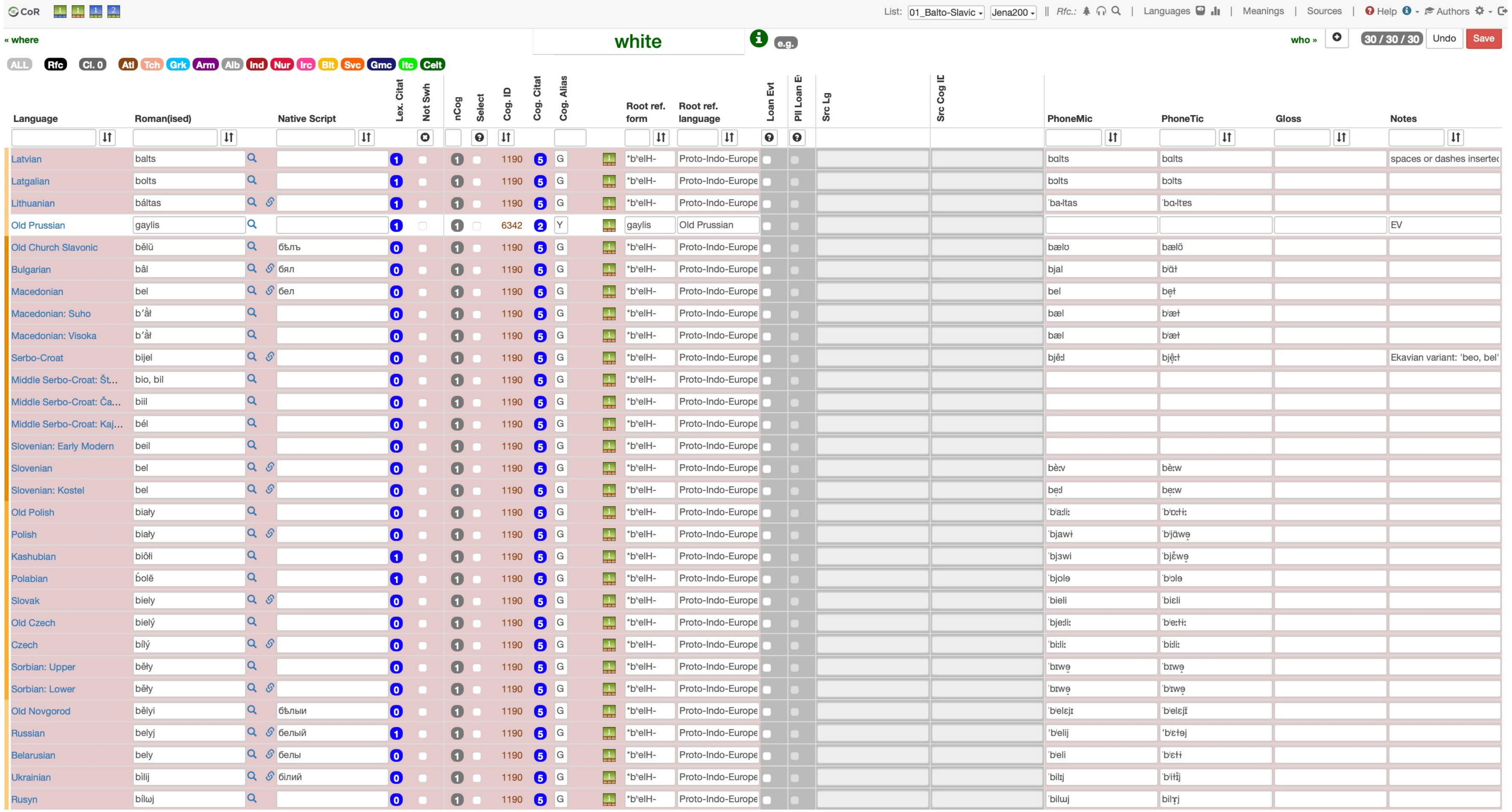
- Solutions

1) Incorrect translations:	replace or remove obviously incorrect translations
2) Incorrect transcriptions:	fix transcriptions

- How to fix transcriptions?
  - Wrote improved G2P conversion tools for languages where the transcriptions issues were too complex, addressed the shortcomings of the original transcriptions
  - Russian: orthography is not fully phonetic, instead extracted transcriptions from Wiktionary entries
  - Issues and fixes documented in Appendix of thesis proposal

BALTO-SLAVIC DATASET

- Cognate codes from IE-CoR (thank you Cormac! ☺)
- Haven’t had a chance yet to use them to correct the cognate sets
- IE-CoR data provide an additional reference for mistaken transcriptions or translations



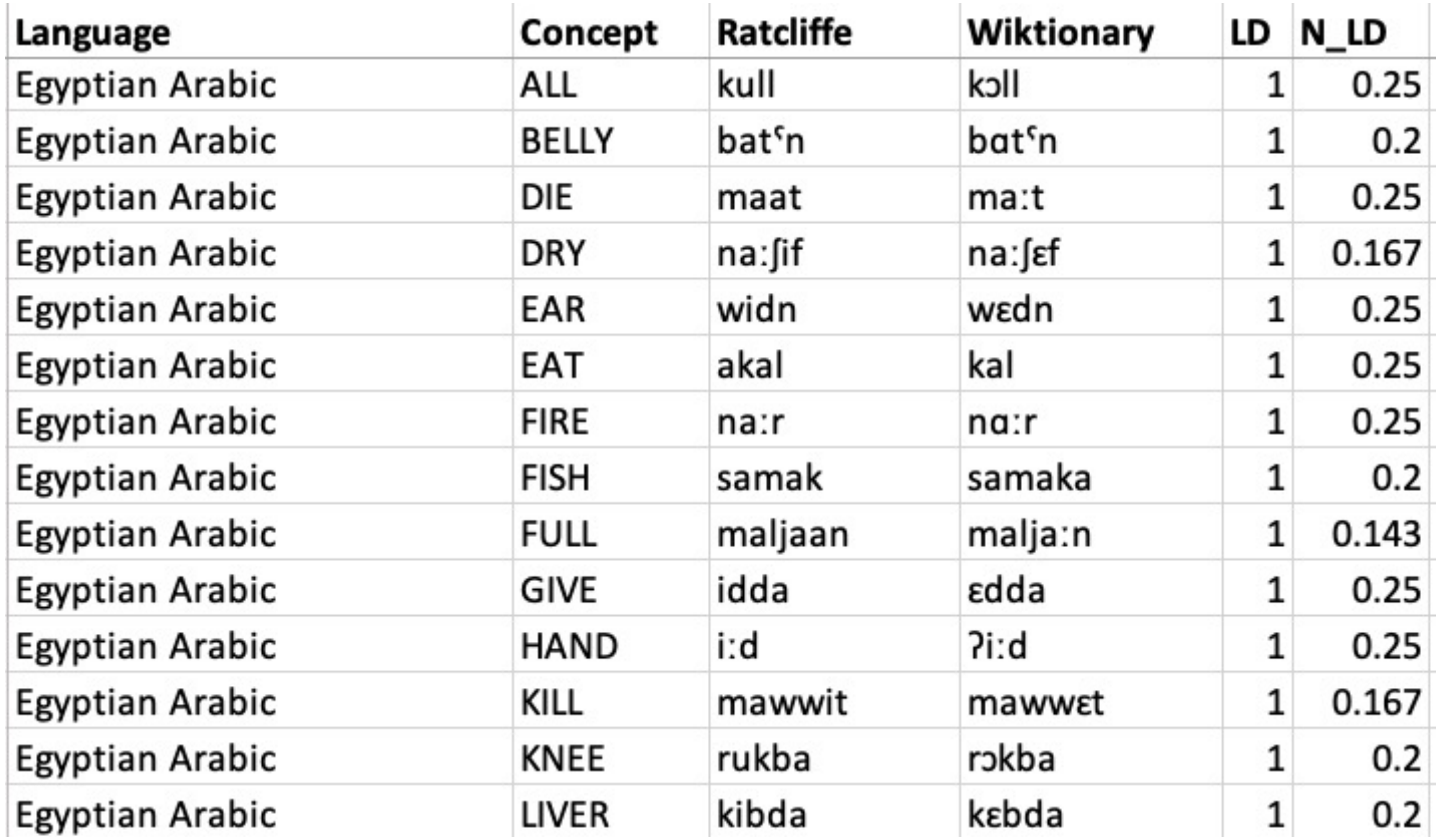
List: 01_Balto-Slavic Jena200

white

Language	Roman(ised)	Native Script	Lex. Citat	Not Swh	nCog	Select	Cog. ID	Cog. Citat	Cog. Alias	Root ref. form	Root ref. language	Loan Evt	Pll Loan E	Src Lg	Src Cog ID	PhoneMic	PhoneTic	Gloss	Notes
Latvian	balts		1	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			bɑlts	bɑlts		spaces or dashes insertec
Latgalian	bolts		1	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			bɔlts	bɔlts		
Lithuanian	báltas		1	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			ˈbaˑltas	ˈbɑˑltɐs		
Old Prussian	gaylis		1	☐	1	☐	6342	2	Y	gaylis	Old Prussian	☐	☐						EV
Old Church Slavonic	bělŭ	бѣлъ	0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			bælʊ	bælʊ̆		
Bulgarian	bâl	бял	0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			bjal	bʲɑ̈ɫ		
Macedonian	bel	бел	0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			bel	be̞ɫ		
Macedonian: Suho	b’ằł		0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			bæl	bʲæɫ		
Macedonian: Visoka	b’ằł		0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			bæl	bʲæɫ		
Serbo-Croat	bijel		0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			bjêːl	bjê̞ːɫ		Ekavian variant: 'beo, bel'
Middle Serbo-Croat: Št…	bio, bil		0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐						
Middle Serbo-Croat: Ča…	biil		0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐						
Middle Serbo-Croat: Kaj…	bél		0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐						
Slovenian: Early Modern	beil		0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐						
Slovenian	bel		0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			bèːv	bèːw		
Slovenian: Kostel	bel		0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			be̝ːl	be̝ːw		
Old Polish	biały		0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			ˈbʲaːliː	ˈbʲɒːɫɨː		
Polish	biały		0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			ˈbjawɨ	ˈbʲjäwɘ̟		
Kashubian	biôłi		1	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			ˈbjɜwi	ˈbjɛ̽wɘ̟		
Polabian	ḃolě		1	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			ˈbjolə	ˈbʲɔlə		
Slovak	biely		0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			ˈbieli	ˈbiɛli		
Old Czech	bielý		0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			ˈbjeːliː	ˈbʲeːɫɨː		
Czech	bílý		0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			ˈbiːliː	ˈbiːliː		
Sorbian: Upper	běły		0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			ˈbɪwɘ̟	ˈbɪwɘ̟		
Sorbian: Lower	běły		0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			ˈbɪwɘ̟	ˈbʲɪwɘ̟		
Old Novgorod	bělyi	бѣлыи	0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			ˈbʲelɛjɪ	ˈbʲelɛjɪ̆		
Russian	belyj	белый	1	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			ˈbʲelij	ˈbʲɛɫəj		
Belarusian	bely	белы	0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			ˈbʲeli	ˈbʲɛɫɨ		
Ukrainian	bìlij	білий	0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			ˈbilɪj	ˈbʲiɫɪ̽j		
Rusyn	bílɯj		0	☐	1	☐	1190	5	G	*bʰelH-	Proto-Indo-Europe	☐	☐			ˈbilɯj	bilɤ̞j		

ARABIC DATASET

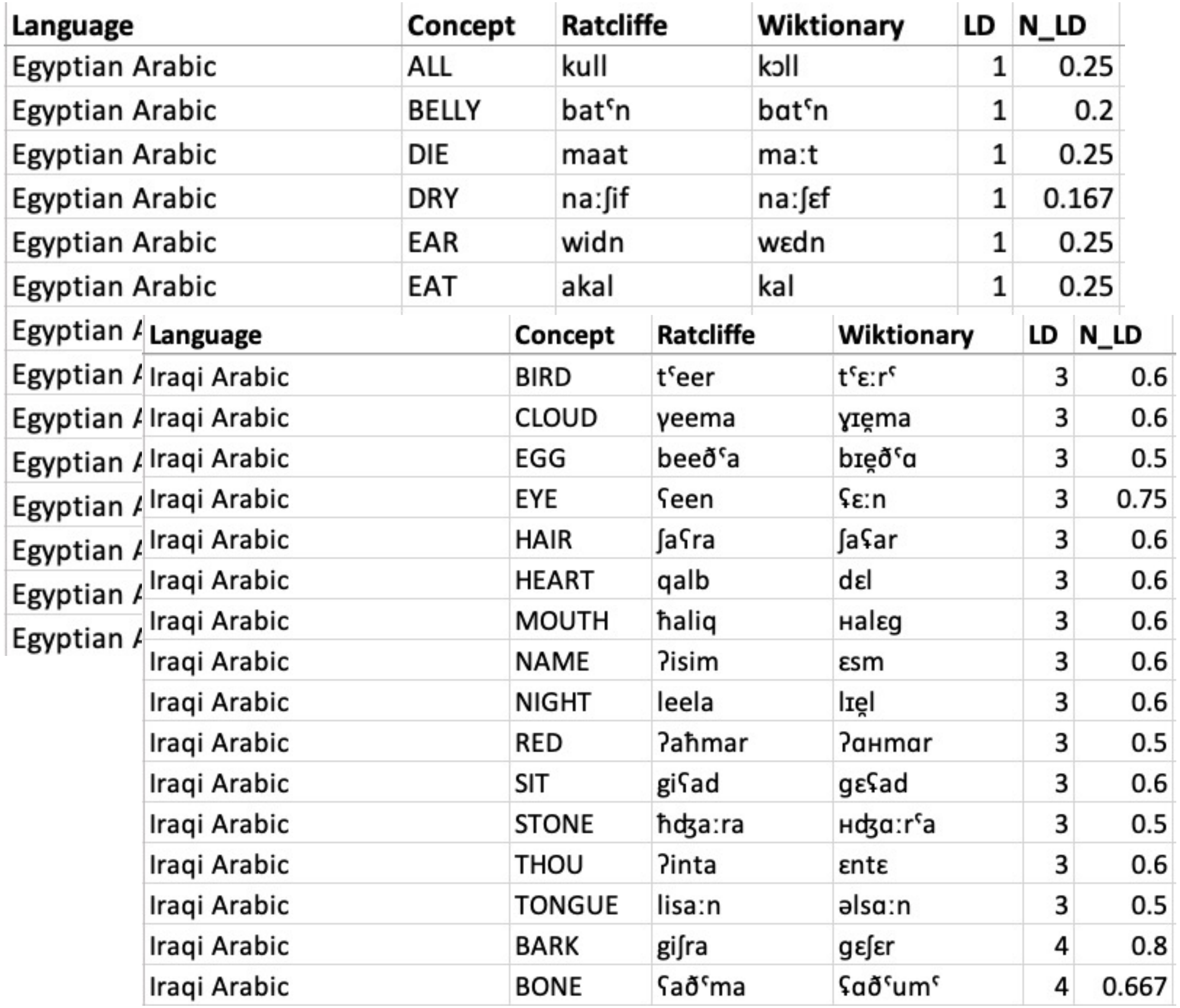
- Checked overlap of Ratcliffe’s (2020) dataset and Wiktionary Swadesh lists
- Almost all of Ratcliffe’s forms are represented in the Wiktionary dataset, minor differences in transcription
  - Considerations: Arabic dialects are not standardized, may not refer to exact same variety
  - Arabic expert would be needed to verify the details of the transcriptions, but seem close enough
- Wiktionary lists include more words than Ratcliffe’s dataset → no good way to check those



Language	Concept	Ratcliffe	Wiktionary	LD	N_LD
Egyptian Arabic	ALL	kull	kɔll	1	0.25
Egyptian Arabic	BELLY	batˤn	bɑtˤn	1	0.2
Egyptian Arabic	DIE	maat	ma:t	1	0.25
Egyptian Arabic	DRY	na:ʃif	na:ʃɛf	1	0.167
Egyptian Arabic	EAR	widn	wɛdn	1	0.25
Egyptian Arabic	EAT	akal	kal	1	0.25
Egyptian Arabic	FIRE	na:r	nɑ:r	1	0.25
Egyptian Arabic	FISH	samak	samaka	1	0.2
Egyptian Arabic	FULL	maljaan	malja:n	1	0.143
Egyptian Arabic	GIVE	idda	ɛdda	1	0.25
Egyptian Arabic	HAND	i:d	ʔi:d	1	0.25
Egyptian Arabic	KILL	mawwit	mawwɛt	1	0.167
Egyptian Arabic	KNEE	rukba	rɔkba	1	0.2
Egyptian Arabic	LIVER	kibda	kɛbda	1	0.2

ARABIC DATASET

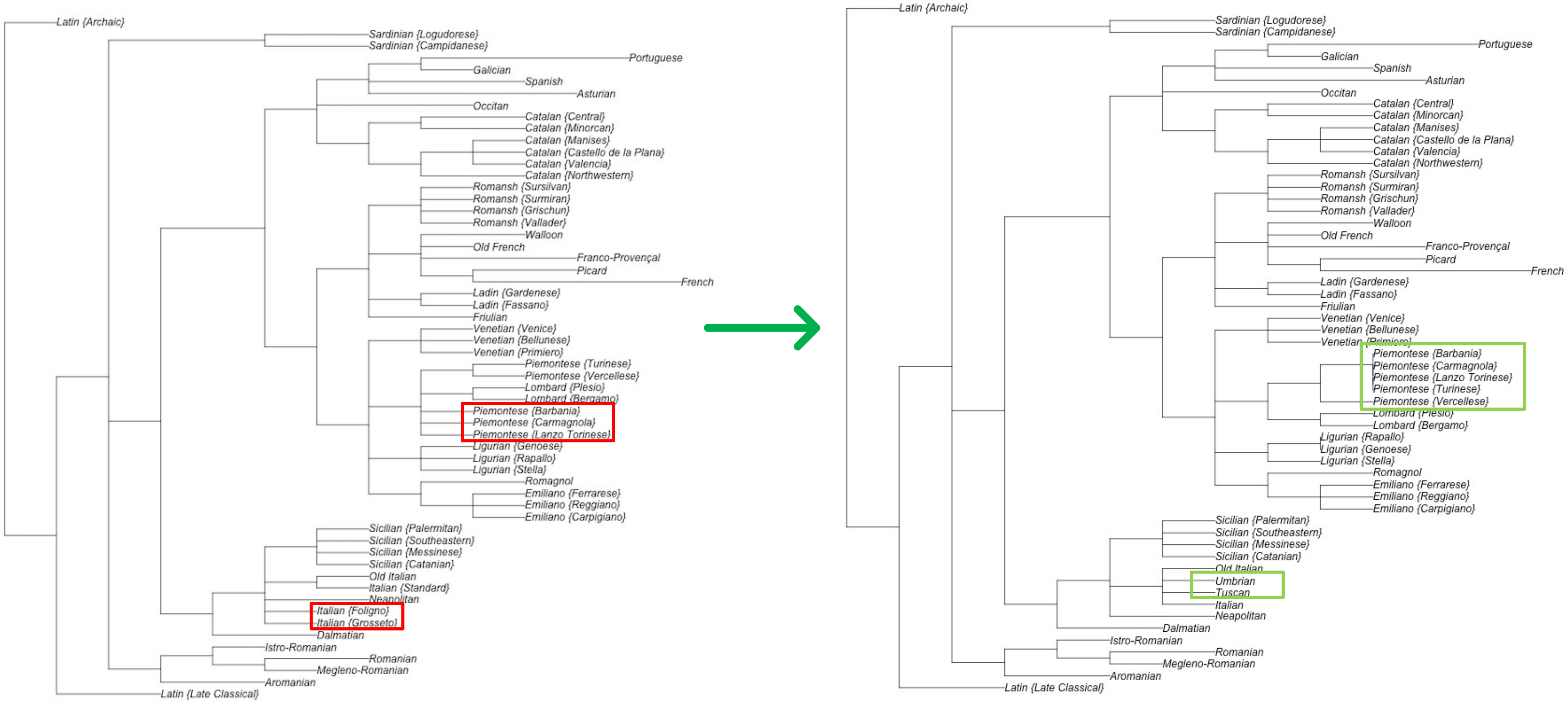
- Checked overlap of Ratcliffe’s (2020) dataset and Wiktionary Swadesh lists
- Almost all of Ratcliffe’s forms are represented in the Wiktionary dataset, minor differences in transcription
  - Considerations: Arabic dialects are not standardized, may not refer to exact same variety
  - Arabic expert would be needed to verify the details of the transcriptions, but seem close enough
- Wiktionary lists include more words than Ratcliffe’s dataset → no good way to check those



Language	Concept	Ratcliffe	Wiktionary	LD	N_LD
Egyptian Arabic	ALL	kull	kɔll	1	0.25
Egyptian Arabic	BELLY	batˁn	batˁn	1	0.2
Egyptian Arabic	DIE	maat	ma:t	1	0.25
Egyptian Arabic	DRY	na:ʃif	na:ʃɛf	1	0.167
Egyptian Arabic	EAR	widn	wɛdn	1	0.25
Egyptian Arabic	EAT	akal	kal	1	0.25
Egyptian A					
Egyptian A					
Egyptian A					
Egyptian A					
Egyptian A					
Egyptian A					
Egyptian A					
Egyptian A					

Language	Concept	Ratcliffe	Wiktionary	LD	N_LD
Iraqi Arabic	BIRD	tˁeer	tˁɛ:rˁ	3	0.6
Iraqi Arabic	CLOUD	ɣeema	ɣɪe̯ma	3	0.6
Iraqi Arabic	EGG	beeðˁa	bɪe̯ðˁɑ	3	0.5
Iraqi Arabic	EYE	ʕeen	ʢɛ:n	3	0.75
Iraqi Arabic	HAIR	ʃaʕra	ʃaʢar	3	0.6
Iraqi Arabic	HEART	qalb	dɛl	3	0.6
Iraqi Arabic	MOUTH	ħaliq	ʜalɛg	3	0.6
Iraqi Arabic	NAME	ʔisim	ɛsm	3	0.6
Iraqi Arabic	NIGHT	leela	lɪe̯l	3	0.6
Iraqi Arabic	RED	ʔaħmar	ʔɑʜmɑr	3	0.5
Iraqi Arabic	SIT	giʕad	gɛʢad	3	0.6
Iraqi Arabic	STONE	ħdʒa:ra	ʜdʒɑ:rˁa	3	0.5
Iraqi Arabic	THOU	ʔinta	ɛntɛ	3	0.6
Iraqi Arabic	TONGUE	lisa:n	əlsɑ:n	3	0.5
Iraqi Arabic	BARK	giʃra	gɛʃɛr	4	0.8
Iraqi Arabic	BONE	ʕaðˁma	ʢɑðˁumˁ	4	0.667

GLOTTOLOG TREES: BUG FIXED



PHONETIC DISTANCE

- Some literature review: see thesis proposal for more details
  - Séguy (1973): dialectometry using lexical, phonetic, and morphological features as binary characters
  - Kessler (1995): clustering Irish Gaelic dialects using phonetic transcriptions
  - Covington (1996): phonetic distance using hand-crafted alignment costs
  - Kondrak's ALINE algorithm (2000-2002): aligning phonetic sequences using feature-based phonetic similarity
    - Features weighted by salience
    - Not distinctive phonological features: multi-valued, ordinal rather than binary features
    - COGIT: cognate-detection using composite phonetic/semantic similarity score

PHONETIC DISTANCE

- Badr’s feedback on encoding distinctive features
  - No clear “correct” approach – would depend on the needs of the application
  - Ideally would develop a perceptual similarity study to compare against, but this is impractical and would be necessarily biased towards the languages of the participants
  - Sufficient to find a solution which performs as well or better than existing methods
  - Try all of the approaches with some toy data and see if the results make sense

PHONETIC SEQUENCE ALIGNMENT

- Until now, I’ve used my own implementation of phonetic similarity with Needleman-Wunsch alignment algorithm
  - Find optimal sequence alignment according to similarity of phones (distinctive features + sonority)
  - Seems similar to Kondrak’s ALINE tool
  - Also incorporates aspect of information content (next)

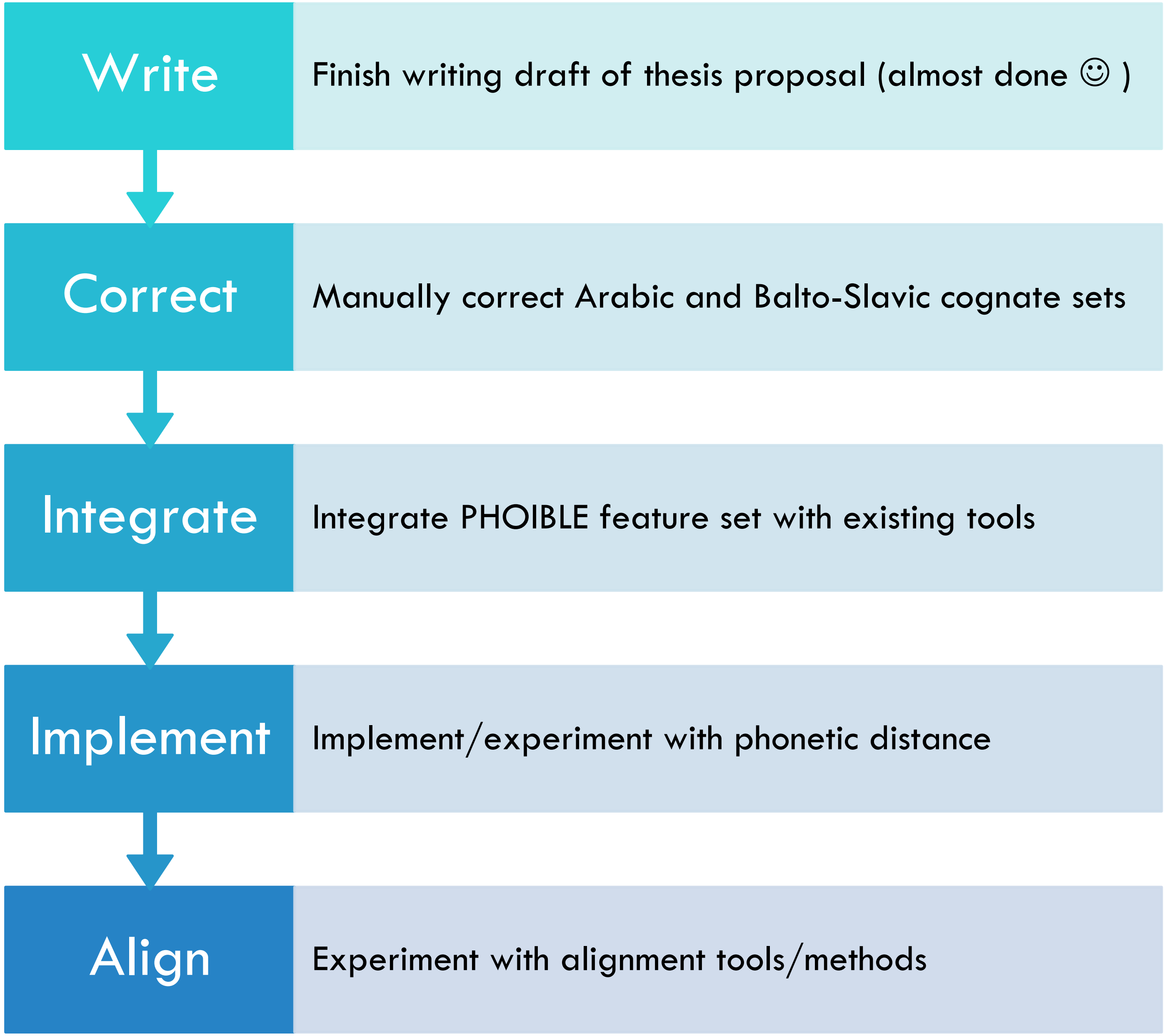
PHONETIC SEQUENCE ALIGNMENT

- So far: I’ve used my previous implementation of phonetic similarity with Needleman-Wunsch alignment algorithm
  - Find optimal sequence alignment according to similarity of phones (distinctive features + sonority)
  - Seems similar to Kondrak’s ALINE tool
  - Also incorporates aspect of information content (below)
- Dellert (2018) information content alignment
  - Use trigram frequency within language’s transcriptions to determine relative informativity of trigrams
  - Add difference in informativity to whatever other cost (e.g. phonetic distance)
  - Idea: penalize alignments of information-heavy with information-light segments to discourage, e.g. alignment of verb suffix with verb root

PHONETIC SEQUENCE ALIGNMENT

- So far: I've used my previous implementation of phonetic similarity with Needleman-Wunsch alignment algorithm
  - Find optimal sequence alignment according to similarity of phones (distinctive features + sonority)
  - Seems similar to Kondrak's ALINE tool
  - Also incorporates aspect of information content (below)
- Another possibility: alignment informed by historical sound changes
  - Extract all valid alignments for true cognates, get costs for alignment of phonemes or sound classes from correspondence probabilities
  - Might yield more accurate alignments than ones based on phonetic similarity alone
    - e.g. /s/ → /h/ is a common sound change, but not phonetically similar
  - Similar to method used by Jäger (2018)

NEXT STEPS



Write — Finish writing draft of thesis proposal (almost done ☺)

Correct — Manually correct Arabic and Balto-Slavic cognate sets

Integrate — Integrate PHOIBLE feature set with existing tools

Implement — Implement/experiment with phonetic distance

Align — Experiment with alignment tools/methods