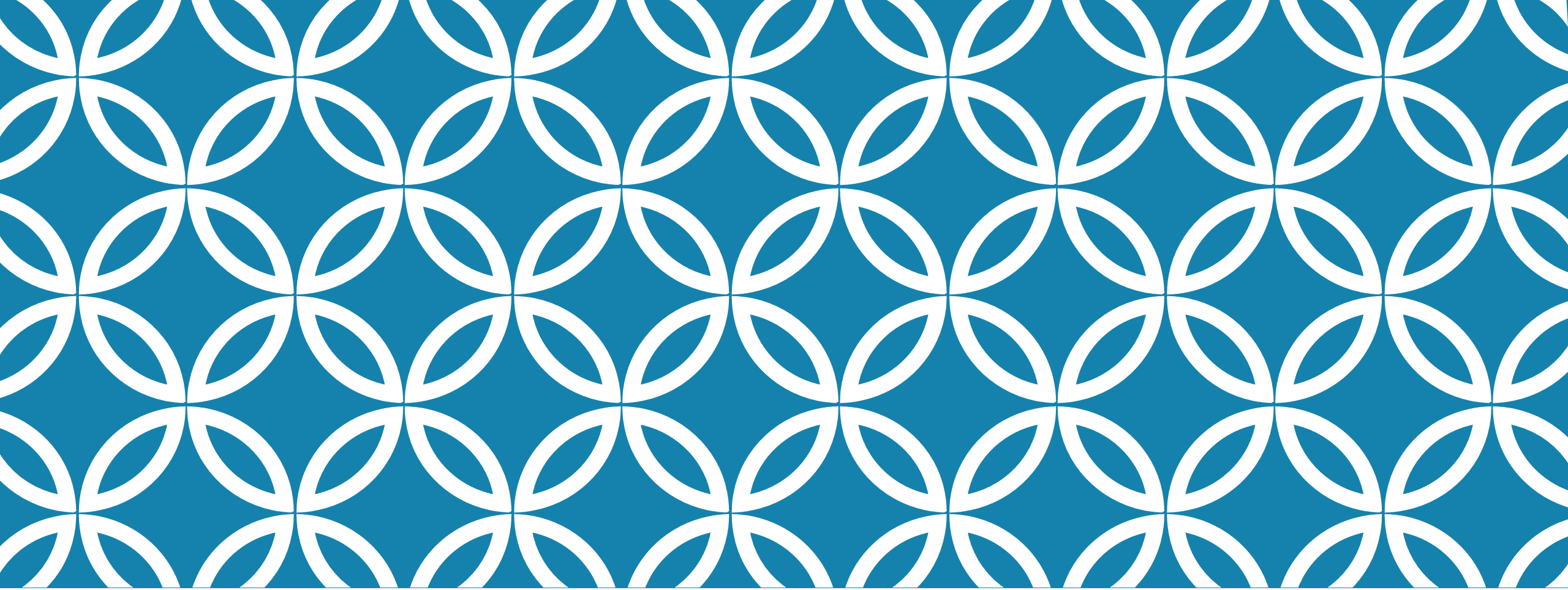


THESIS SEMINAR MEETING

Philip Georgis

May 26, 2021

OUTLINE OF THESIS

Research Questions:

1) Which types of distance methods are (most) useful for phylogenetic inference?

- Phonetic methods vs. information theory methods (PMI, surprisal) [+ mixed methods]
- Phonemes vs. sound classes

2) At what stage of phylogenetic inference are they useful?

(i.e. are they useful for cognate detection, or only for evaluating already-detected cognates?)

OUTLINE OF THESIS

(Sample) Abstract:

"Jäger (2018) reports that a hybrid method for phylogenetic inference, combining both binary character- and distance-based cognate detection methods, yields the best fit with expert-created gold standard trees. However, Jäger employs only one method for measuring the distance between cognate pairs, namely by means of pointwise mutual information on the basis of global sound class correspondences. **This thesis proposes several additional word pair distance metrics based on phonetic features and information content, in order to determine whether these might also prove useful, either for cognate detection in the first step or for subsequent generation of phylogenetic trees on the basis of cognate distance evaluation metrics.**"

OUTLINE OF THESIS

I) Introduction: Overview of language families, phylogenetic inference, cognates & cognate detection

II) Background / Literature Review

Survey of techniques in (computational) historical linguistics:

- comparative method, glottochronology, Bayesian phylogenetic methods, distance-based approaches

What else?

III) Methodology

Outline of cognate detection procedure
Phonetic and sound class distance metrics
Information-theory distance metrics
Clustering languages / generating phylogenetic trees

IV) Key Studies and Evaluation

Target families: Slavic, Romance, Turkic, Arabic, Sinitic
Cognate detection: evaluate against gold cognate sets
Distance measures: evaluate using tree topography comparison metrics, possibly also mutual intelligibility

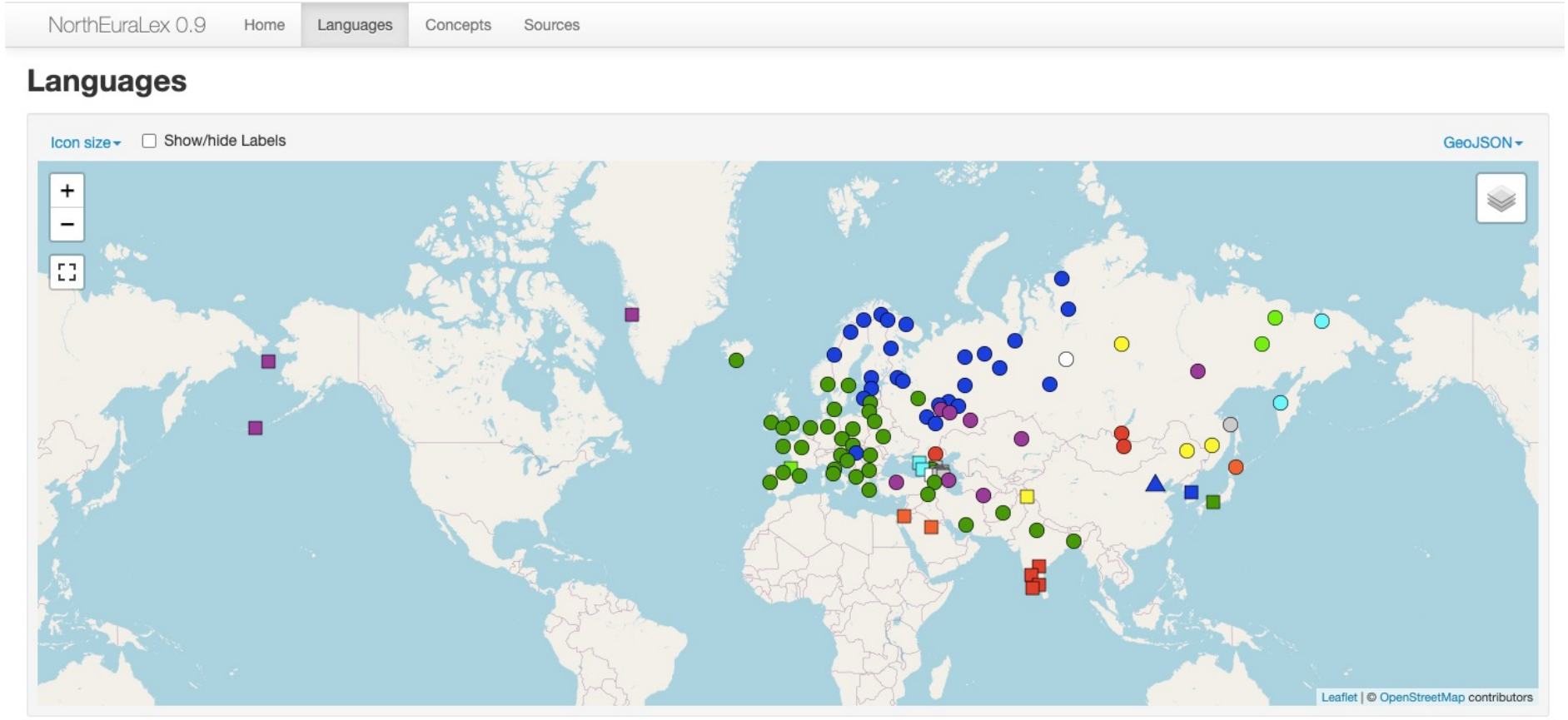
THESIS PROCEDURE

- Decide on input datasets (wordlists) and corresponding gold trees/cognate sets for evaluation
- Develop and implement distance metrics
- Develop/extend/adapt data structures for processing inputs, implementing cognate detection and evaluation steps, and comparing results with gold standard
- Perform cognate detection using each method, evaluate results against gold cognate sets
- Take results from best-performing cognate detection method, perform cognate evaluation
- Compare calculated linguistics distances against expert trees and intelligibility data

DATASETS

NorthEuraLex: 107 languages (Indo-European, Uralic, Turkic, Dravidian, Eskimo-Aleut, etc.)

→ use for (Balto-)Slavic, Uralic, *any others?*



DATASETS

NorthEuraLex: 107 languages (Indo-European, Uralic, Turkic, Dravidian, Eskimo-Aleut, etc.)

→ use for (Balto-)Slavic, Uralic, *any others?*

Sinitic (Líu et al., 2007): 19 Sinitic varieties

Turkic (Savelyev & Robeets, 2020): 31 Turkic languages

Arabic (Ratcliffe, 2020): 14 dialects of Arabic

Global Lexicostatistic Database: Romance (58 varieties), and many others…

SINITIC, ROMANCE, NORTHEURALEX DATASETS

- Downloaded and preprocessed all of the data to yield input files in the format needed for my implementation

liusinitic.tsv

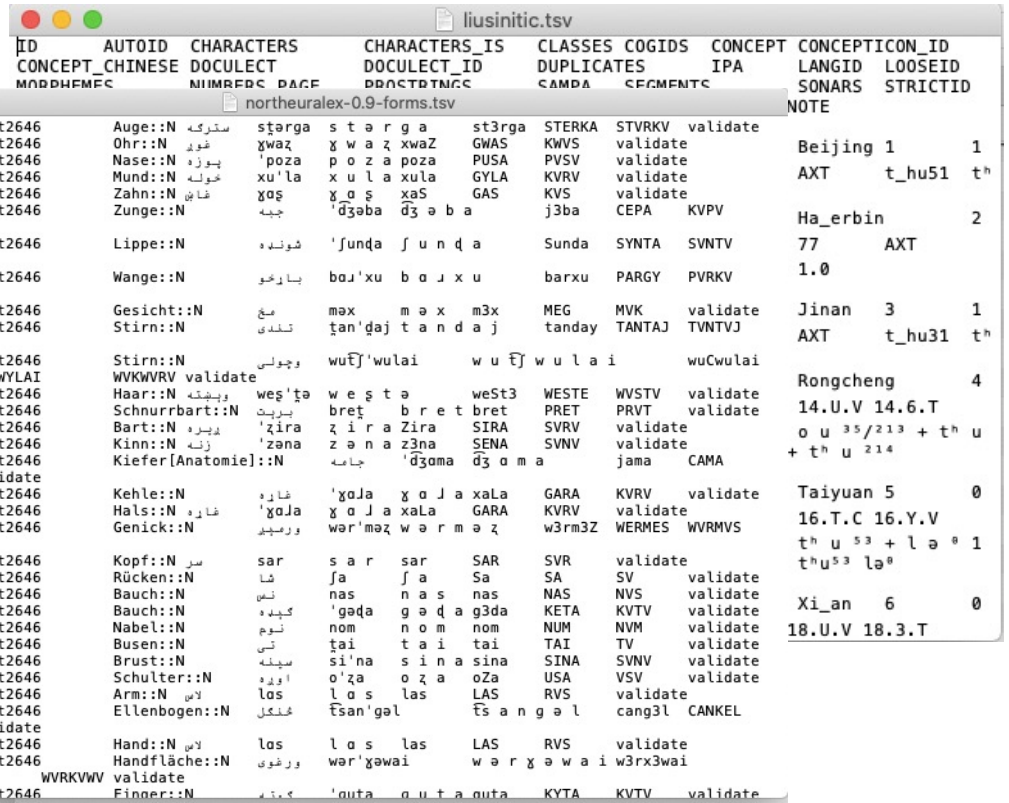
ID	AUTOID	CHARACTERS	CHARACTERS_IS	CLASSES	COGIDS	CONCEPT	CONCEPTICON_ID
CONCEPT_CHINESE	DOCULECT		DOCULECT_ID	DUPLICATES		IPA	LANGID	LOOSEID
MORPHEMES	NUMBERS	PAGE	PROSTRINGS	SAMPA	SEGMENTS		SONARS	STRICTID
	STRUCTURE		STRUCTURES	TOKENS	VALUE	WEIGHTS	LEXEME_NOTE	
CHARACTER_NOTE	NOTE							
1	999	吐	吐	TY3	1	vomit	1278	嘔吐	Beijing 1	1
	tʰu⁵¹	1	487	spit/吐	1.T.C 1.Y.V 1.3.T		77	AXT	t_hu51	tʰ
u ⁵¹	1 7 8	717	i n t	i n t	tʰ u ⁵¹ tʰu⁵¹		1.6 3.0 1.0			
2	999	吐	吐	TY3	1	vomit	1278	嘔吐	Ha_erbin	2
	1	tʰu⁵³	7	487	spit/吐	7.T.C 7.Y.V 7.3.T		77	AXT	
t_hu53	tʰ u ⁵³	1 7 8	717	i n t	i n t	tʰ u ⁵³ tʰu⁵³		1.6 3.0 1.0		
3	999	吐	吐	TY3	1	vomit	1278	嘔吐	Jinan	3	1
	tʰu³¹	8	487	spit/吐	8.T.C 8.Y.V 8.3.T		77	AXT	t_hu31	tʰ
u ³¹	1 7 8	717	i n t	i n t	tʰ u ³¹ tʰu³¹		1.6 3.0 1.0			
4	999	嘔 吐	嘔吐	U6_TY6	2 1	vomit	1278	嘔吐	Rongcheng	4
	0	ou²¹³+tʰu²¹⁴	14	487	nausea/嘔 spit/吐		14.U.V 14.6.T			
14._._ 14.T.C 14.Y.V 14.6.T			77	XT_AXT	ou213-35 t_hu214		o u ³⁵/²¹³ + tʰ u			
²¹⁴	7 8 9 1 7 8		868	n t + i n t	n t + i n t		ou ²¹³ + tʰ u ²¹⁴			
ou²¹³⁻³⁵ tʰu²¹⁴	3.0 1.0 0.0 1.6 3.0 1.0									
5	1000	嘔 嗷	嘔嗷	TY3_LE6	2 5	vomit	1278	嘔吐	Taiyuan 5	0
	tʰu⁵³+lə⁰		16	487	nausea/嘔 _:PERFECTIVE/了		16.T.C 16.Y.V			
16.3.T 16._._ 16.L.C 16.E.V 16.6.T			77	AXT_AXT	t_hu53 l@0		tʰ u ⁵³ + l ə ⁰ 1			
7 8 9 5 7 8		869	i n t + i n t	i n t + i n t	tʰ u ⁵³ + l ə ⁰ tʰu⁵³ lə⁰					
1.6 3.0 1.0 0.0 1.6 3.0 1.0										
6	999	嘔 吐	嘔吐	NU3_TY3	2 1	vomit	1278	嘔吐	Xi_an	6	0
	ŋou²¹+tʰu⁵³		18	487	nausea/嘔 spit/吐		18.N.C 18.U.V 18.3.T			

Fuzhou.txt

```
I/我/ŋwai³²
all/個郎下/ko²¹²louŋ⁵³ŋa²⁴²
and/共/kɔyŋ²⁴²
animal/四骰爬/si²¹²kʰa⁵⁵βa⁵³
ash/灰灰/hwoi⁵⁵hwoi⁵⁵
ash/灰/hwoi⁵⁵
back/背/pʰjaŋ⁵⁵
bad/痞/pʰai³²
bad/呆/ŋai⁵³
bark/树皮/tsʰjeu²¹²pʰwoi⁵³
because/因为/iŋ⁵⁵ŋwoi²⁴²
belly/腹老/puʔ²⁴lo³²
big/大/twai²⁴²
bird/鳥/tsɛu³²
bite/咬/ka²⁴²
```

SINITIC, ROMANCE, NORTHEURALEX DATASETS

- Downloaded and preprocessed all of the data to yield input files in the format needed for my implementation



ARABIC DIALECT DATASET (RATCLIFFE, 2020)

- Can't access paper: https://brill.com/view/journals/ldc/11/1/article-p1_1.xml
- Extracted data table from PDF of paper's appendix
- Working on manually cleaning/reformatting extracted data

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	ʧat	kullu	kulu
2	ASH	rama:d	rˁmadˁ	rmi:d	rama:dˁ*	rama:d	ruma:d	ɪ	rˁama:d	ramat r–m–d	ramaad	ɪ	ɪ	ra'matt	ruman
3	BARK	qirfat	qeʃrˁa	ʔoʃra	ʔiʃr*	ʔəʃər	giʃra	kiʃr	səvi:ye	ɪ	giʃra	giʃr	li:he, girfe	ɪ	*kokobo lataka, (girifa)
4	BELLY	batˁn	kerʃ	zaʔʔ	batˁn*	batˁən	batˁin	batˁn, ʧarʃ	ʤoof	patn b–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	Sood	Kabiir	kabi:r	ka'biir	kbir
6	BIRD	tˁa:ʔir	tˁir	Sasfur	tˁeer*	tˁeer	tˁeer	tˁeer	ku:ʧ ka:ye	2	tˁeer	tˁajr	tˁe:ra	tayra	ter
7	BITE	Sad	Sed	gide	Sadˁ	Sadˁd	Saðˁð	Saðˁð	ɪ	Saðð	Sadˁd	lugus	adˁdˁa	yaSazz	adi

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Index	Gloss	CA	Mor	Mlt	Cai	Dms	Irq	Skh	AqArb	Cyp	Glf	Ymn	Nig	Bux	Nub
2	1	ALL	kull	koll	Kollu	kull	kəll	kull	ʧill	kəll	kull	kil	kull	ʧat	kullu	kulu
3										sa:ɣi:n						
4	2	ASH	rama:d	rˁmadˁ	rmi:d	rama:dˁ	rama:d	ruma:d		rˁama:d	ramat	ramaad			ra'matt	ruman
5	3	BARK	qirfat	qeʃrˁa	ʔoʃra	ʔiʃr*	ʔəʃər	giʃra	kiʃr			giʃra	giʃr	girfe		girifa
6																*kokobo
7										səvi:ye						
8														li:he		
9																lataka
10	4	BELLY	batˁn			batˁn*	batˁən	batˁin	batˁn		patn	batSn	batˁn		batin	batna
11				kerʃ					ʧarʃ				karʃ	kirʃ		
12					zaʔʔ											
13										ʤoof						
14	5	BIG	kabi:r	kbir	kibir	kibiir	kbi:r	ʧbi:r	ʧabi:r	gəbi:r	k-b-r		Kabiir	kabi:r	ka'biir	kbir
15												Sood				
16	6	BIRD	tˁa:ʔir	tˁir	Sasfur	tˁeer*	tˁeer	tˁeer	tˁeer			tˁeer	tˁajr	tˁe:ra	tayra	ter
17										ku:ʧ						
18										ka:ye						
19	7	BITE	Sadˁdˁa	Sedˁdˁ	gidem	Sadˁdˁ	Sadˁdˁ	Saðˁðˁ	Saðˁðˁ		Saðð	Sadˁdˁ		adˁdˁa	yaSazz	adi
20													lugus			
21	8	BLACK	ʔaswad	kħel	iswed	iswid	ʔaswad	ʔaswad	ʔaswad	əswad	isfet	ʔaswad	ʔaswad	axadˁar	ʔaswad	*asue
22	9	BLOOD	dam	demm	demm	damm	damm	damm	dam	Dam	d-m-y	damm	Dam	damm	dam	dum
23																
24	10	BONE	Saðˁm	Sdˁem	Sadma	Sadma	Sadˁme	Saðˁma	Saðˁm	Sadˁəm		Saðʕm	Saðˁma	adˁum	Sazəm	ladim
25																
26	11	BREAST	sˁadr	bezzula	Sider	sidir*	sˁadˁər	nahid	θidi(f.)sˁidr		pzaz9pl)b-(	deed(f)sSa	sˁadr	de:d	sadər;bu3u	*sudur
27																
28	12	BURN	ħaraqa	ħreq	ħaraʔ	ħaraʔ	ħaraʔ	ħirag	ħarak	ħ-r-q	ħ-r-q	ħarag	ħarag	hirig,taʃʃa (	Salag	haragu
29																
30	13	CLAW	ðˁufr	dˁferˁ	Difer	dˁifrˁ*	dˁəfər	ðˁifir	iðˁfir	dˁəfər	ðˁ-f-r	mixlab	θˁufrin**	xumfur		*dufuru
31												ðʕifr				
32	14	CLOUD	yama:mat	ymama	shaba	saħaab	yeeme	yeema	yeemi		ɣ-y-r	sihaab,yeem		dˁa(ha)wij(	yeem,qeen	*soobura
33																
34	15	COLD	ba:rid	bared	bi:red	bard	barəd	barda:n	bi:rid	beerəd	*peret(brd	baarid	baarid	barda:n	bard	bari
35																
36	16	COME	ʤa:ʔa	ʒa	ʤi:	ga	ʔəʤa	ʤa:	atsa	ʤa:	ʤ-y-y	ʤa	ʤaaʔ	ʤa	d3akki	ʤa
37																
38																
39	17	DIE	ma:ta	mat	mi:t	maat	ma:t	ma:t	ma:t	ma:t	m-w-t	maat	maat	ma:t	maat	mutu
40																
41	18	DOG	kalb	kelb	kelb	kalb	kalb	ʧalib	ʧalb	kalb	k-l-b	ʧalb	kalb	kalb	kalb	keli

FEATURAL REPRESENTATION OF SOUNDS

- Goal: encode any IPA sequence as a vector or sequence of vectors representing the distinctive phonological features of the sound(s) in question

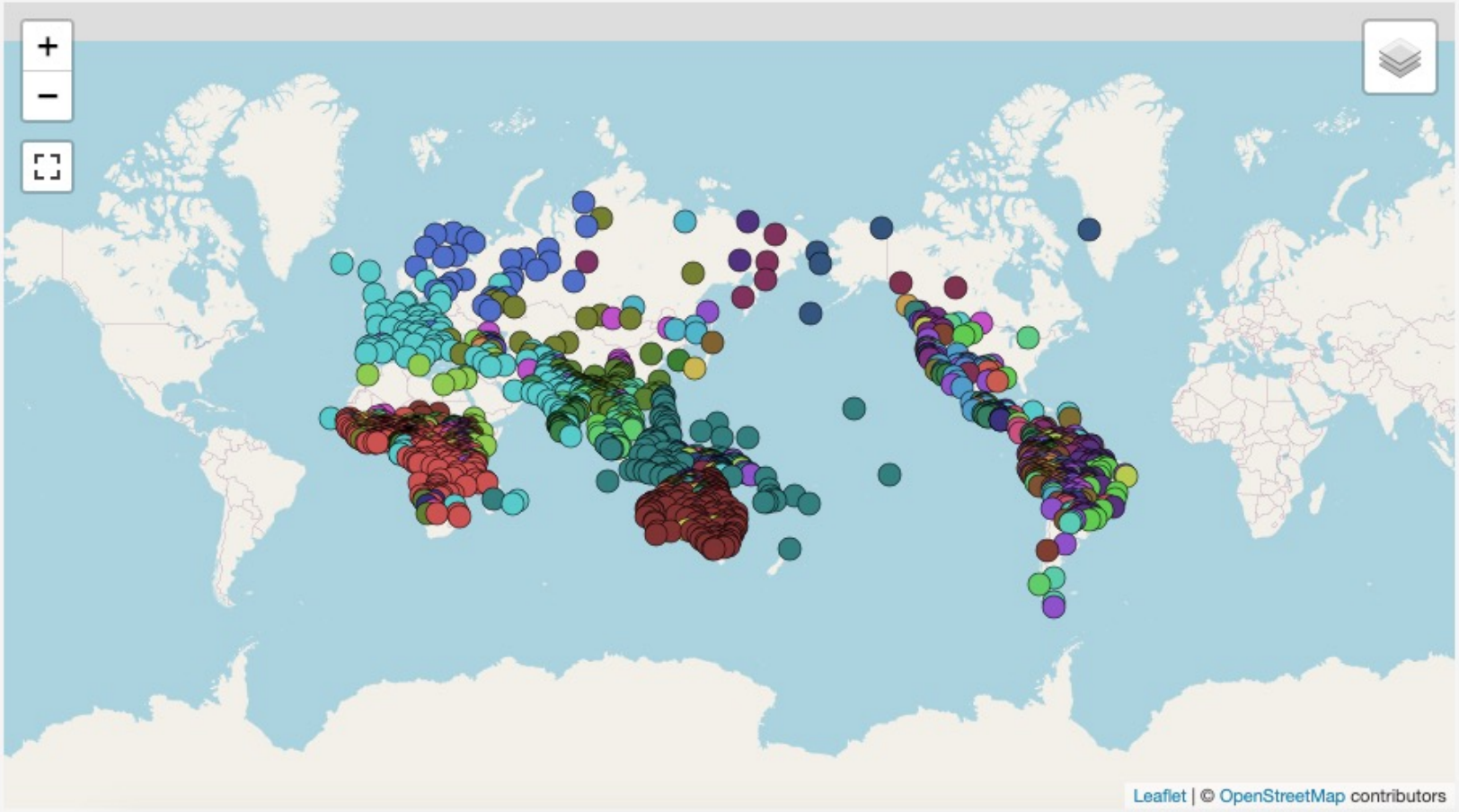
- Challenges/Considerations
  - Should be able to handle all possible IPA characters and any relevant combinations
  - Should be able to deal with different equivalent transcriptions, e.g. [p^{jh}] = [p^{hj}], [$\widehat{tʃ}$] = [$\widehat{ʧ}$]
  - Ideally: no redundancy in features, i.e. use as few features as possible for all distinct sounds to have distinct representations

- Source
  - PHOIBLE feature set: https://phoible.org/parameters

Consonant p W

LATIN SMALL LETTER P

Icon size ▾ Show/hide Labels GeoJSON ▾



Features

advanced tongue root	0
anterior	0
approximant	-
back	0
click	-
consonantal	+
constricted glottis	-
continuant	-
coronal	-
delayed release	-
distributed	0
dorsal	-
epilaryngeal source	-
fortis	-
front	0
high	0
labial	+
labiodental	-
lateral	-
long	-
low	0

EXPLORING SEGMENTS/FEATURES

- Examined phonetic symbols used in NorthEuraLex, Sinitic, and Romance datasets, plus set of PHOIBLE features
- All standard IPA characters used in all 4 datasets already covered by my earlier implementation of phonetic distance
  - Only a few diacritics and characters specific to Sinitic transcription conventions missing

    → easy to add or replace with equivalent characters, e.g. /ɻ/ = /z̩/

- 20 of 38 PHOIBLE features have direct equivalents to features in my implementation
  - Some of the remaining PHOIBLE features may be redundant, e.g. “labial” and “labiodental” features are not needed as they can equivalently be expressed by combinations of other features

EXPLORING SEGMENTS/FEATURES

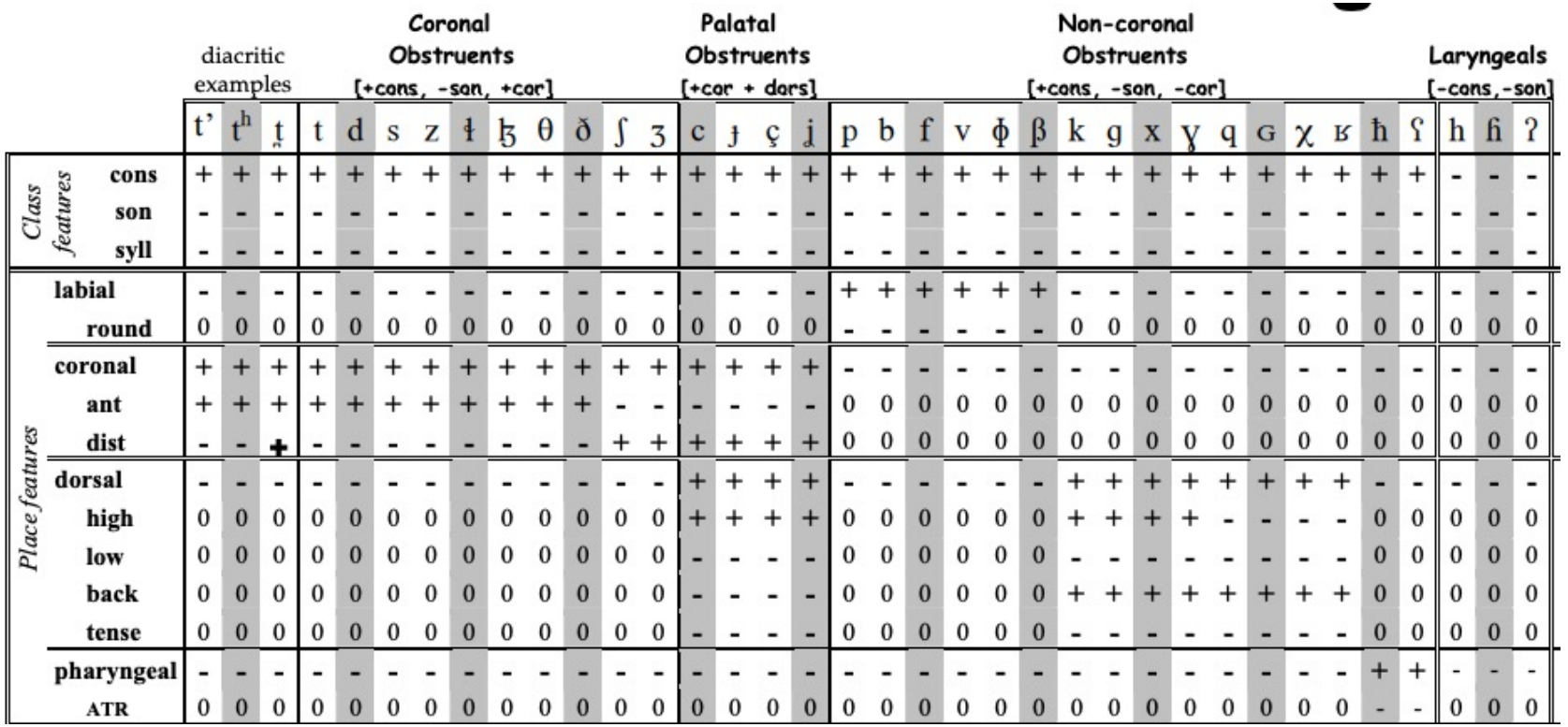
- Strategy for dealing with phonetic diacritics (e.g. ◌̃, ◌ʰ, ◌ʲ, ◌ʷ, ◌ː, etc.)
  - PHOIBLE: each character with diacritics has a separate entry in feature matrix
    - e.g. /p/ has a separate entry from /pʰ/ and another for /pʰː/, etc.
    - 2162 segment entries
  - My method: each base character has a unique featural entry; diacritics associated with specific features and values, modify the base feature vector
    - e.g. /p/ has unique representation; adding /ʰ/ adds the +SPREAD GLOTTIS feature; /ː/ adds +LONG feature
    - Only 116 segment entries, plus diacritics

FEATURAL REPRESENTATION OF SOUNDS

- **Plan: adapt PHOIBLE segment/feature dataset**
  - Use PHOIBLE feature representation for basic IPA characters (i.e. segments without diacritics)
  - Remove any redundant features
  - Extend current method for allowing diacritics to modify base characters' feature vectors
    - Eliminates need for an entry in feature matrix for every combination of IPA characters and diacritics
    - Allows for more flexibility in transcription conventions

MEASURING PHONETIC DISTANCE

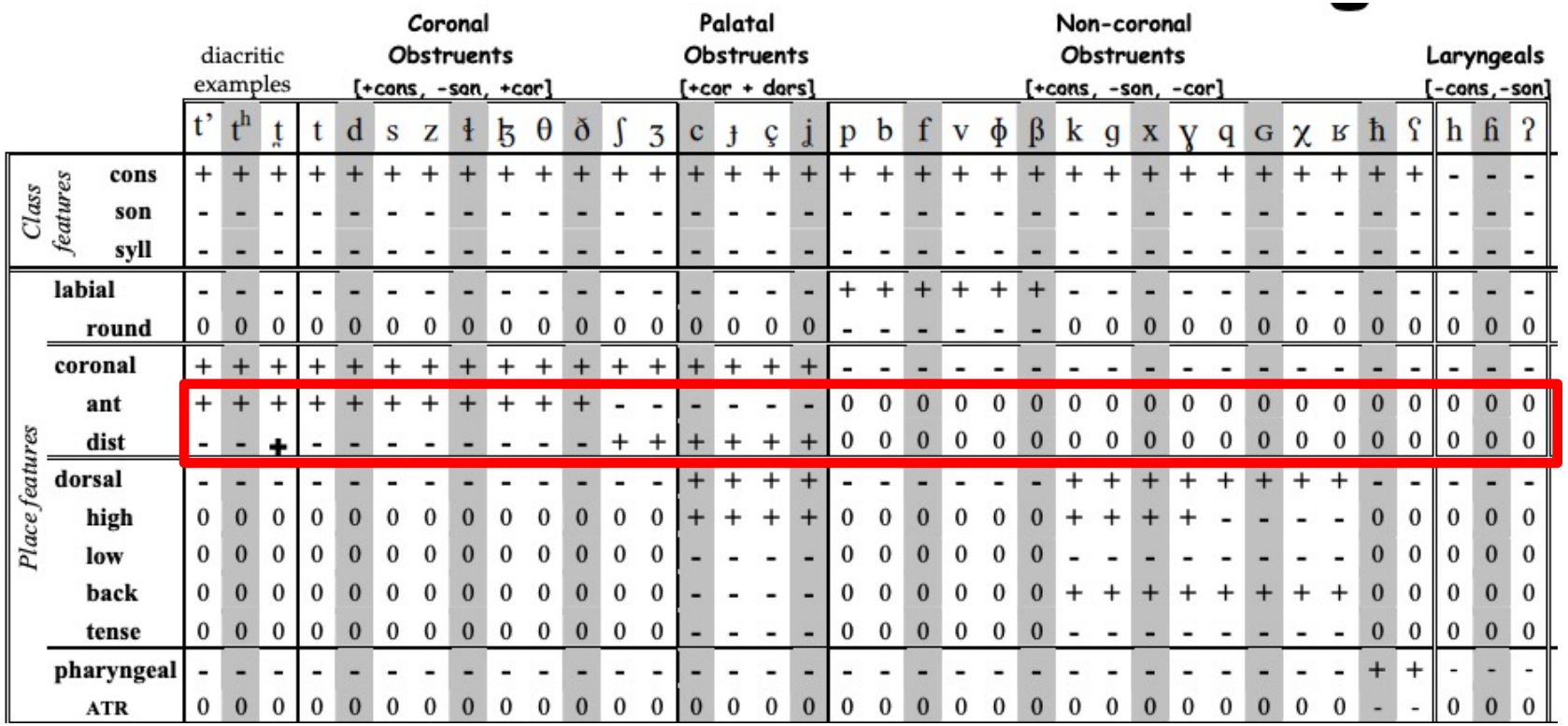
- Values of distinctive features can be:
  - \+ the sound has this feature
  - \- the sound does not have this feature
  - 0 this feature is not relevant for this sound



		diacritic examples			Coronal Obstruents [+cons, -son, +cor]										Palatal Obstruents [+cor + dors]				Non-coronal Obstruents [+cons, -son, -cor]															Laryngeals [-cons,-son]		
		t’	t^h	t̪	t	d	s	z	ɬ	ɮ	θ	ð	ʃ	ʒ	c	ɟ	ç	ʝ	p	b	f	v	ɸ	β	k	g	x	ɣ	q	ɢ	χ	ʁ	ħ	ʕ	h	ɦ	ʔ
Class features	cons	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-
	son	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	syll	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Place features	labial	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-
	round	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0
	coronal	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	ant	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	dist	-	-	+	-	-	-	-	-	-	-	-	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	dorsal	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	-	-	-	-	-	+	+	+	+	+	+	+	+	-	-	-	-	-
	high	0	0	0	0	0	0	0	0	0	0	0	0	0	+	+	+	+	0	0	0	0	0	0	+	+	+	+	-	-	-	-	0	0	0	0	0
	low	0	0	0	0	0	0	0	0	0	0	0	0	0	-	-	-	-	0	0	0	0	0	0	-	-	-	-	-	-	-	-	0	0	0	0	0
	back	0	0	0	0	0	0	0	0	0	0	0	0	0	-	-	-	-	0	0	0	0	0	0	+	+	+	+	+	+	+	+	0	0	0	0	0
	tense	0	0	0	0	0	0	0	0	0	0	0	0	0	-	-	-	-	0	0	0	0	0	0	-	-	-	-	-	-	-	-	0	0	0	0	0
	pharyngeal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-
	ATR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	-	0	0	0	0

MEASURING PHONETIC DISTANCE

- Values of distinctive features can be:
  - \+ the sound has this feature
  - \- the sound does not have this feature
  - 0 this feature is not relevant for this sound



		diacritic examples	Coronal Obstruents [+cons, -son, +cor]											Palatal Obstruents [+cor + dors]				Non-coronal Obstruents [+cons, -son, -cor]																Laryngeals [-cons,-son]		
		t’	t^h	t̪	t	d	s	z	ɬ	ɮ	θ	ð	ʃ	ʒ	c	ɟ	ç	ʝ	p	b	f	v	ɸ	β	k	g	x	ɣ	q	ɢ	χ	ʁ	ħ	ʕ	h	ɦ	ʔ
Class features	cons	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-
	son	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	syll	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Place features	labial	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-
	round	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0
	coronal	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	ant	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	dist	-	-	+	-	-	-	-	-	-	-	-	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	dorsal	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	-	-	-	-	-	+	+	+	+	+	+	+	+	-	-	-	-	-
	high	0	0	0	0	0	0	0	0	0	0	0	0	0	+	+	+	+	0	0	0	0	0	0	+	+	+	+	-	-	-	-	0	0	0	0	0
	low	0	0	0	0	0	0	0	0	0	0	0	0	0	-	-	-	-	0	0	0	0	0	0	-	-	-	-	-	-	-	-	0	0	0	0	0
	back	0	0	0	0	0	0	0	0	0	0	0	0	0	-	-	-	-	0	0	0	0	0	0	+	+	+	+	+	+	+	+	0	0	0	0	0
	tense	0	0	0	0	0	0	0	0	0	0	0	0	0	-	-	-	-	0	0	0	0	0	0	-	-	-	-	-	-	-	-	0	0	0	0	0
	pharyngeal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-
	ATR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	-	0	0	0

MEASURING PHONETIC DISTANCE

- Values of distinctive features can be:
  - \+ the sound has this feature
  - – the sound does not have this feature
  - 0 this feature is not relevant for this sound

- Problem: how to code the '0' (unspecified) value when quantifying differences?
  - **Strategy 1:** Omit 0-valued features entirely when comparing a pair of sounds

    Rationale: doesn't make sense to compare using a feature that is unspecified

    Problem: different pairs of sounds would be compared using vectors of different lengths
  - **Strategy 2:** Encode 0-valued features the same as – [sound doesn't have this feature]

    Rationale: 0-valued features are actually sub-features of other features; if the main feature has value –, then the sub-features would by default have the same – value

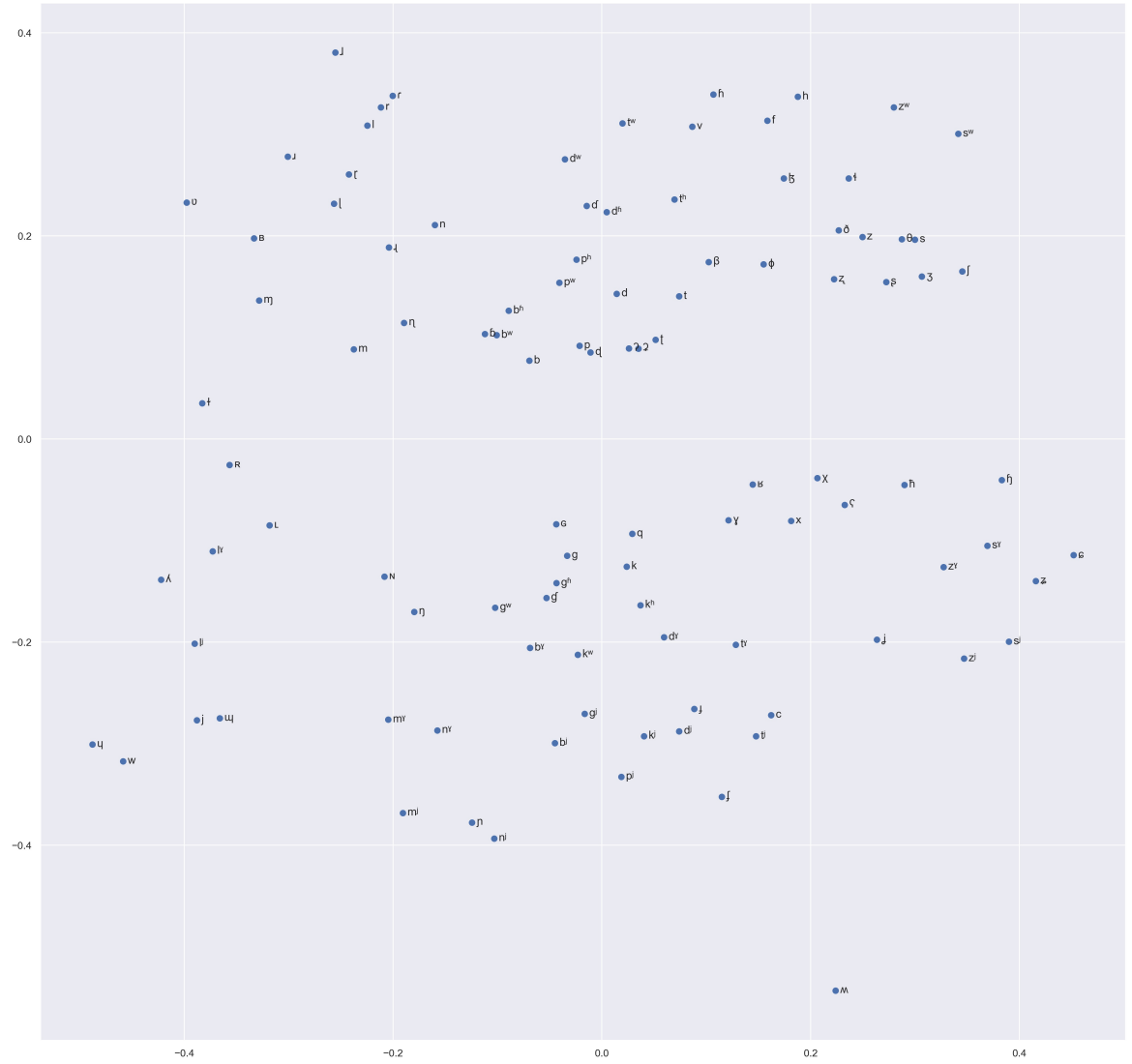
    Problem: what is actually the difference between value 0 and value –?
  - **Strategy 3:** Create 3-way distinction in feature values and consider 0 different from both + and –

    Rationale: avoids problem of either ignoring 0 features or needing to re-encode them

    Problem: is it a valid comparison between 0 and +/-?

MEASURING PHONETIC DISTANCE

- How to pick the appropriate method for dealing with 0 values?
- Compare MDS projections of segments based on features (once adapted)



WORKS CITED

Dellert, J., Daneyko, T., Münch, A. et al. Lang Resources & Evaluation (2019). https://doi.org/10.1007/s10579-019-09480-6 (version 0.9).

Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data*, *5*(180189), 1–16. https://doi.org/10.1038/sdata.2018.189

Líu, L.; Wáng, H.; Bǎi, Y. (2007): Xiàndài Hànyǔ fāngyán héxīncí, tèzhēng cíjí 现代汉语方言核心词·特征词集 [Collection of basic vocabulary words and characteristic dialect words in modern Chinese dialects]. Nánjīng: Fènghuáng.

Moran, Steven & McCloy, Daniel (eds.) 2019. PHOIBLE 2.0. Jena: Max Planck Institute for the Science of Human History. http://phoible.org.

Ratcliffe, R. R. (2020). The glottometrics of Arabic: Quantifying linguistic diversity and correlating it with diachronic change. *Language Dynamics and Change*, *11*(1), 1–29.

Savelyev, A., & Robbeets, M. (2020). Bayesian phylolinguistics infers the internal structure and the time-depth of the Turkic language family. *Journal of Language Evolution*, *5*(1), 39–53.