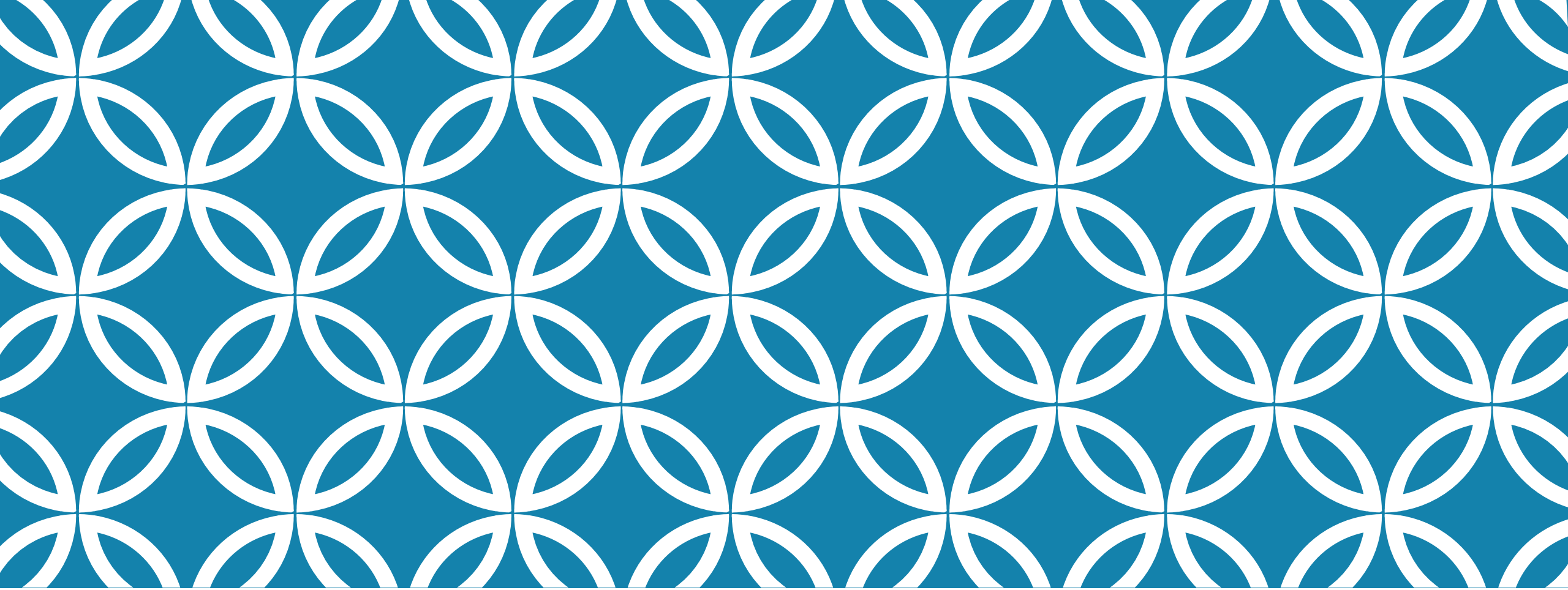
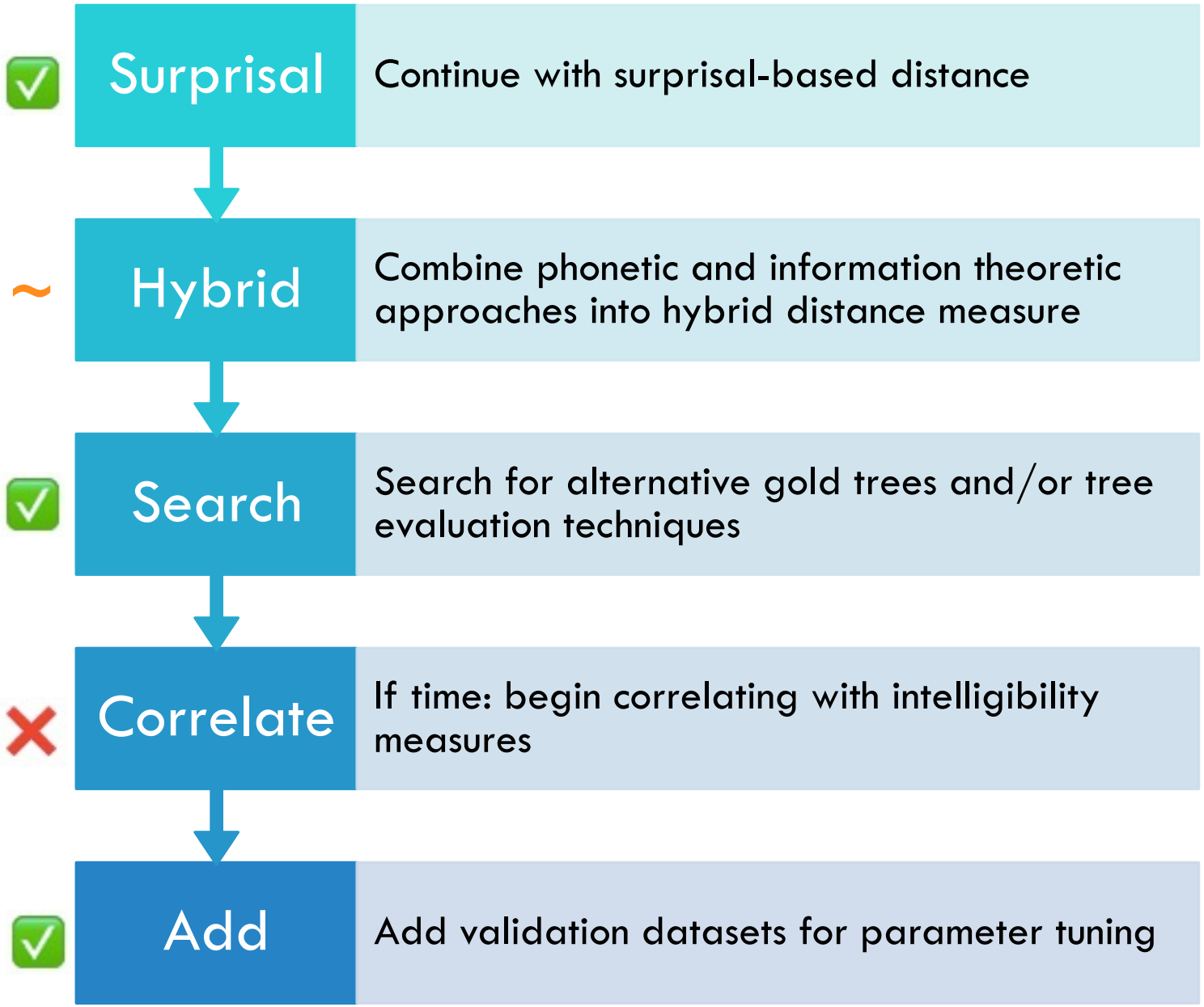


MASTER'S THESIS MEETING

Philip Georgis

October 13, 2021

CURRENT TASKS



VALIDATION DATASET

- Bantu
  - Sample of some of the largest among the >300 doculects from dataset
- Hellenic
  - 4 modern Hellenic doculects + 2 varieties of Ancient Greek
- Japonic
  - Japanese and Ryukyuan doculects + Old Japanese; omitted Middle Japanese
- Quechuan
  - Omitted doculects without glottocodes, and non-Quechuan languages from dataset
- Uto-Aztecan
  - Omitted non-Uto-Aztecan languages from dataset
- Vietic
  - Omitted non-Vietic languages, Proto-Vietic, and ambiguous doculects

DATASETS OVERVIEW

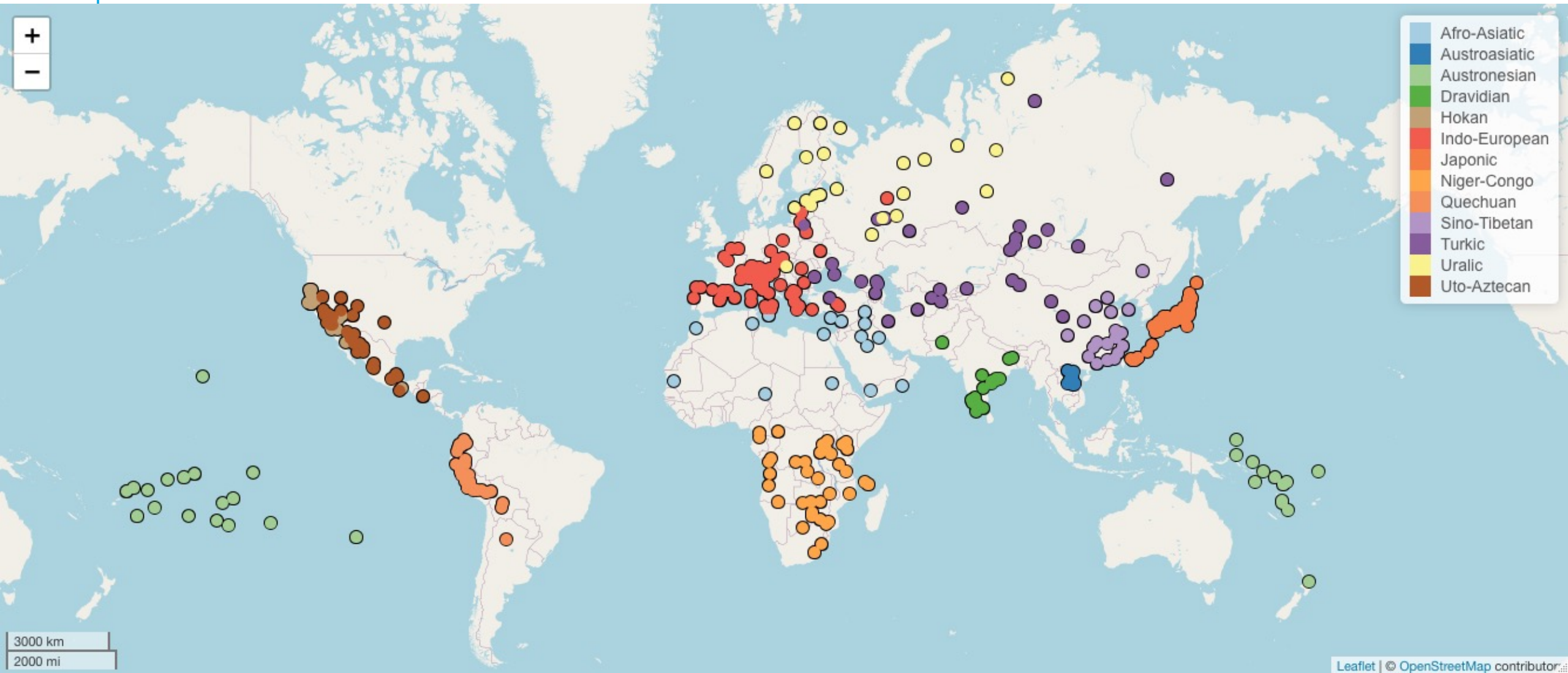
Validation Set

Family	Number of Languages	Time Depth	Macroarea
Bantu	37	5000	Africa
Hellenic	6	2400	Eurasia
Japonic	58	2400	Eurasia
Quechuan	33	1850	South America
Uto-Aztecan	34	7000	North America
Vietic	17	2750	Eurasia
Average	*31*	*3567*	

Test Set

Family	Number of Languages	Time Depth	Macroarea
Arabic	16	1600	Africa & Eurasia
Balto-Slavic	16	3000	Eurasia
Dravidian	20	6000	Eurasia
Hokan	20	5000	North America
Italic	58	2100	Eurasia
Polynesian	30	3000	Papunesia
Sinitic	19	2500	Eurasia
Turkic	32	2100	Eurasia
Uralic	23	8000	Eurasia
Average	*26*	*3700*	

DATASETS OVERVIEW



SURPRISAL-BASED SIMILARITY

- Word pairs aligned using typical alignment algorithm (costs based on phonetic similarity + PMI)
  - Ancient Attic Greek <αἷμα> /hâjma/ Modern Demotic Greek <αίμα> /ˈema/ ‘BLOOD’

AAG	h	a	j	m	a
MDG	-	-	e	m	a

SURPRISAL-BASED SIMILARITY

- Word pairs aligned using typical alignment algorithm (costs based on phonetic similarity + PMI)
  - Ancient Attic Greek <αἷμα> /hâjma/ Modern Demotic Greek <αίμα> /ˈema/ 'BLOOD'
- Adaptation surprisal calculated for each pair in alignment in both directions

AAG	h	a	j	m	a
MDG	-	-	e	m	a
Surprisal	1.41	2.61	3.17	1.32	0.61

TOTAL: 9.12

MDG	-	-	e	m	a
AAG	h	a	j	m	a
Surprisal	3.10	3.31	4.21	0.90	0.87

TOTAL: 12.39

SURPRISAL-BASED SIMILARITY

- Self-surprisal based on trigram probabilities calculated for each word

AAG	h	a	j	m	a
Self-Surprisal	3.29	2.00	1.91	0.92	2.98

TOTAL: 11.10

MDG	e	m	a
Self-Surprisal	3.35	2.14	2.10

TOTAL: *7.59*

SURPRISAL-BASED SIMILARITY

- Self-surprisal based on trigram probabilities calculated for each word

AAG	h	a	j	m	a
Self-Surprisal	3.29	2.00	1.91	0.92	2.98

TOTAL: 11.10

MDG	e	m	a
Self-Surprisal	3.35	2.14	2.10

TOTAL: *7.59*

- Adaptation surprisal normalized by self-surprisal
  - S(MDG|AAG) / S(MDG) = 9.12 / 7.59 = 1.20
  - S(AAG|MDG) / S(AAG) = 12.39 / 11.10 = 1.12
- Take mean of normalized adaptation surprisal in each direction
  - → 1.16

SURPRISAL-BASED SIMILARITY

- Self-surprisal based on trigram probabilities calculated for each word

AAG	h	a	j	m	a
Self-Surprisal	3.29	2.00	1.91	0.92	2.98

TOTAL: 11.10

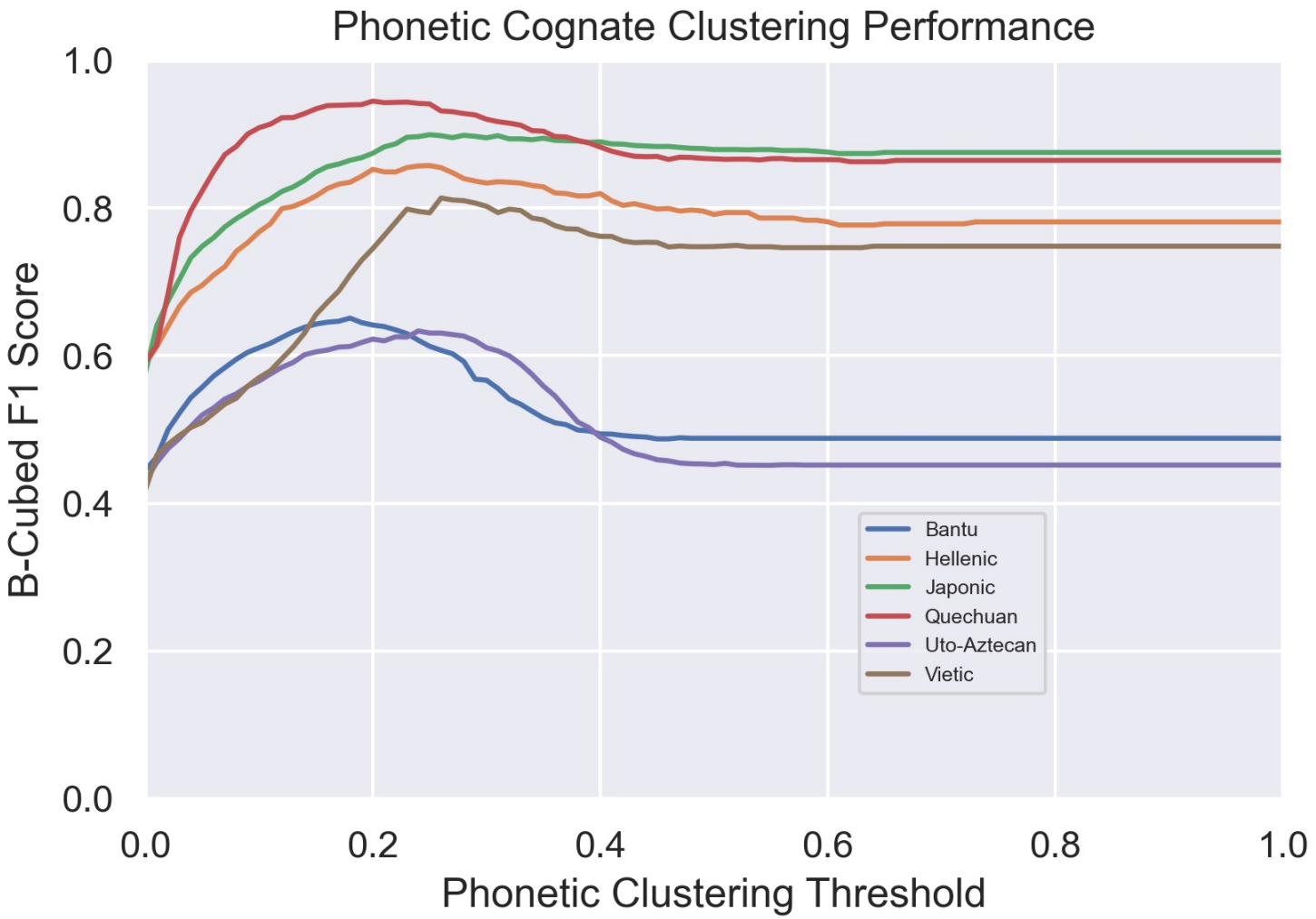
MDG	e	m	a
Self-Surprisal	3.35	2.14	2.10

TOTAL: *7.59*

- Adaptation surprisal normalized by self-surprisal
  - S(MDG|AAG) / S(MDG) = 9.12 / 7.59 = 1.20
  - S(AAG|MDG) / S(AAG) = 12.39 / 11.10 = 1.12
- Take mean of normalized adaptation surprisal in each direction
  - → 1.16 → turn into similarity: e^^-1.16 = 0.31

OPTIMAL COGNATE CLUSTERING THRESHOLDS

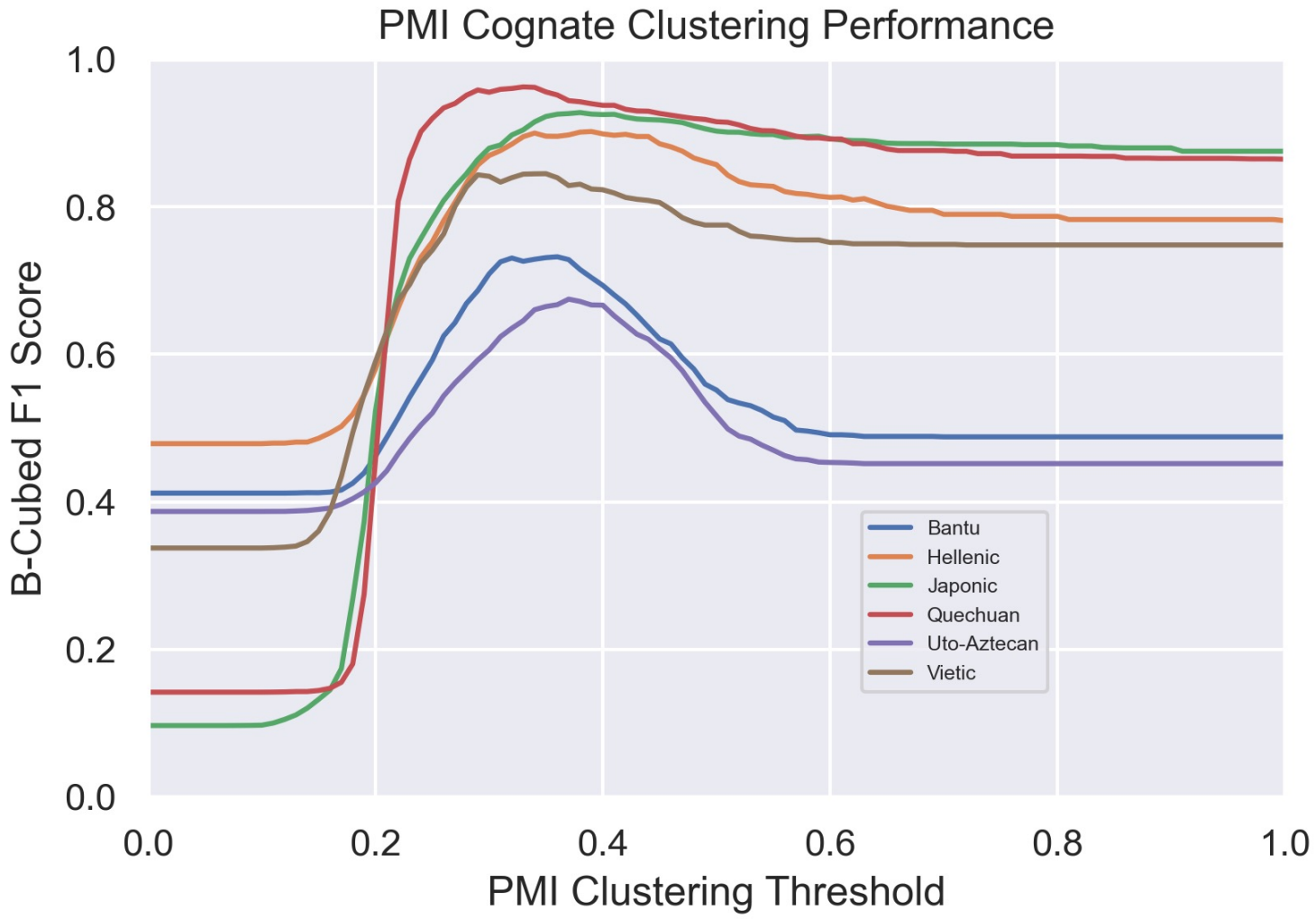
BASED ON VALIDATION DATASETS



Optimal threshold = 0.23

OPTIMAL COGNATE CLUSTERING THRESHOLDS

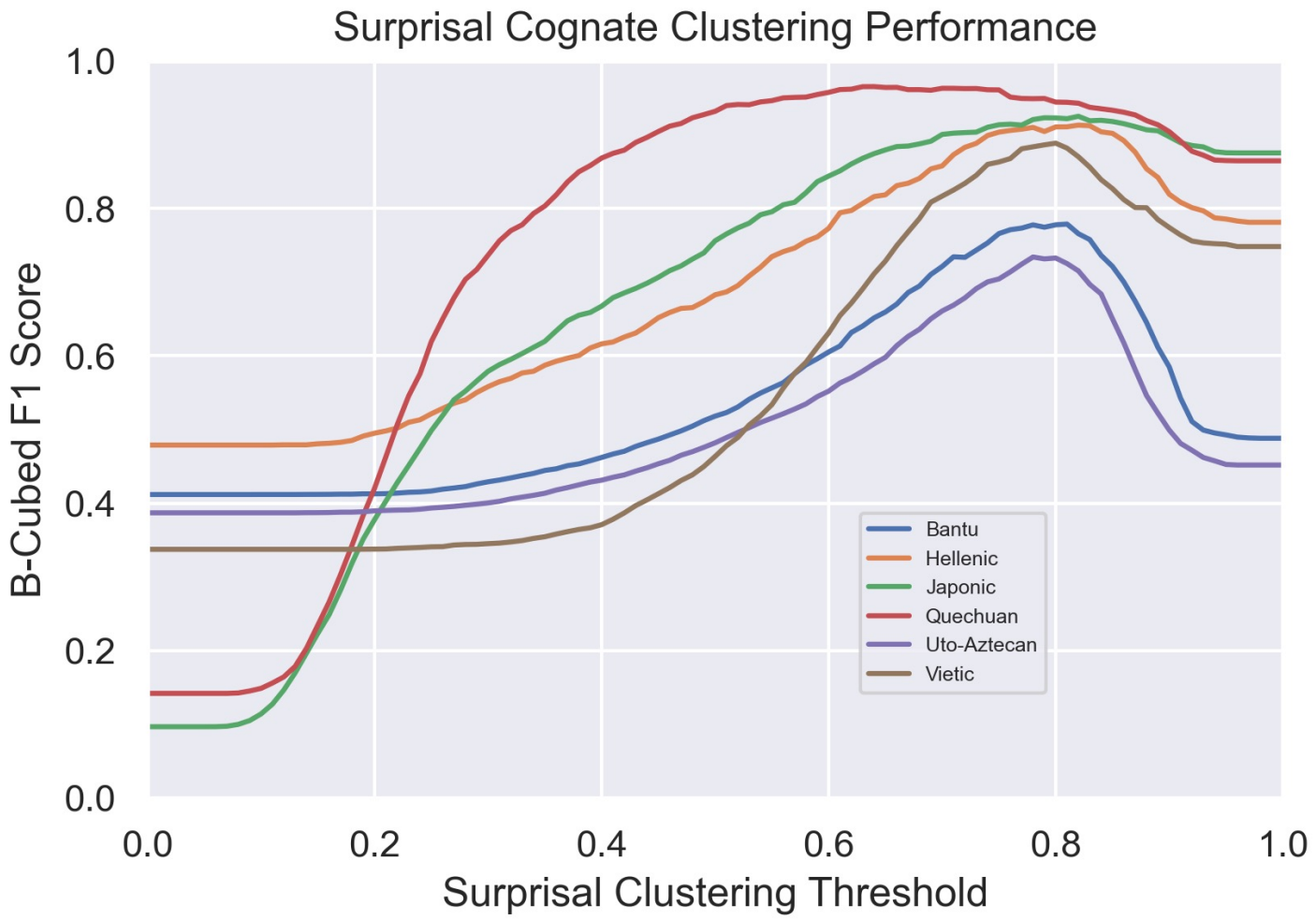
BASED ON VALIDATION DATASETS



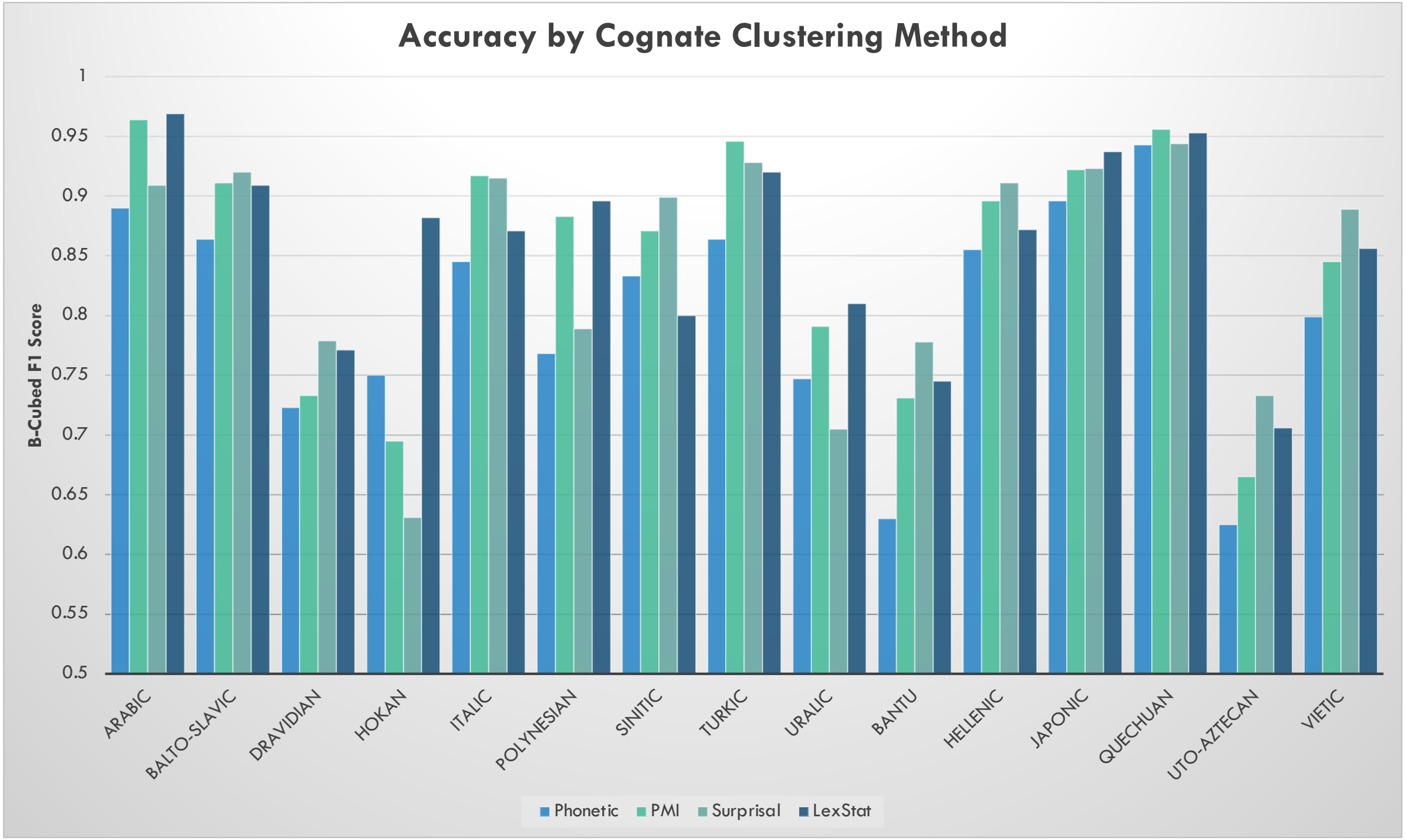
Optimal threshold = 0.35

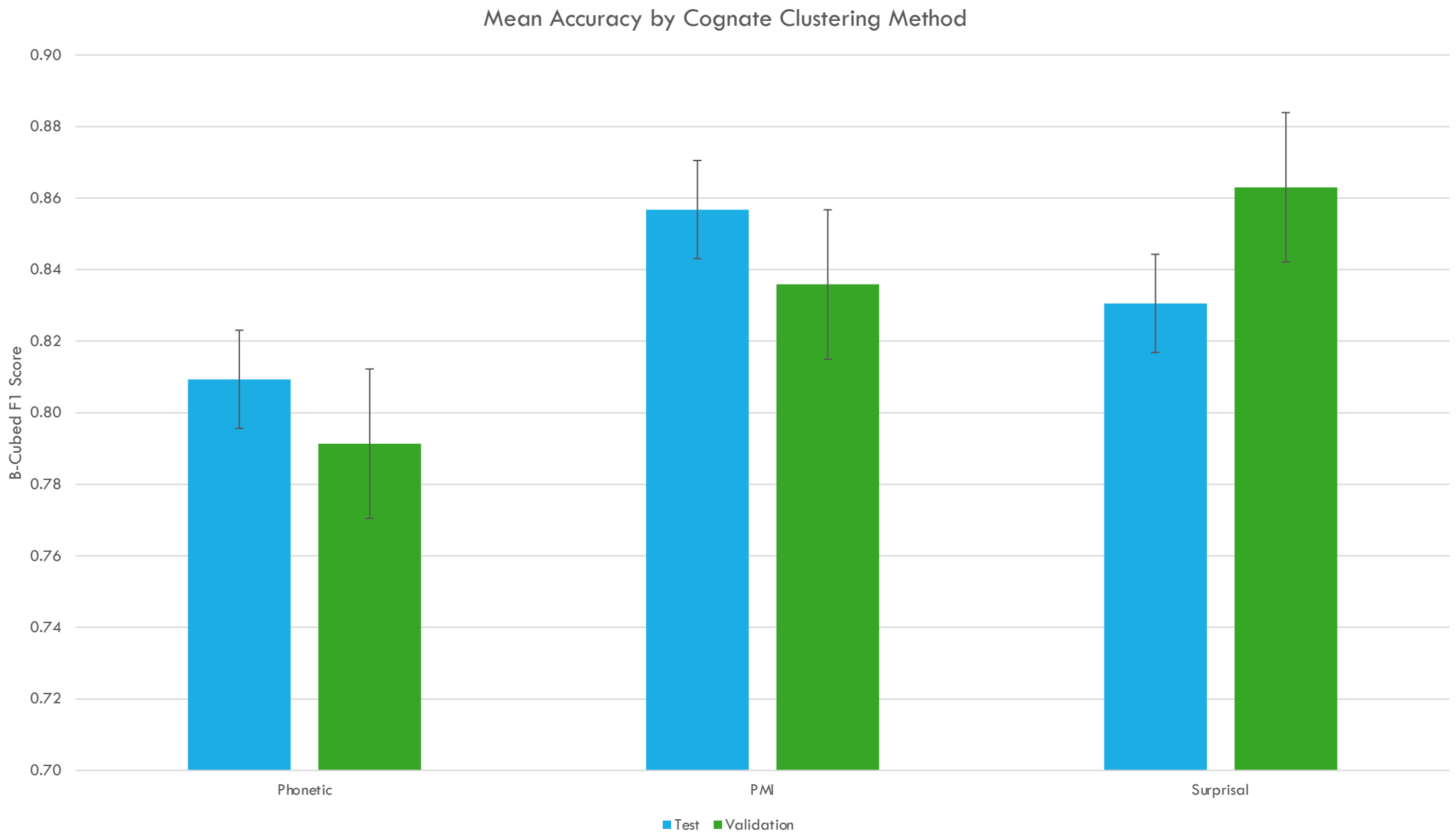
OPTIMAL COGNATE CLUSTERING THRESHOLDS

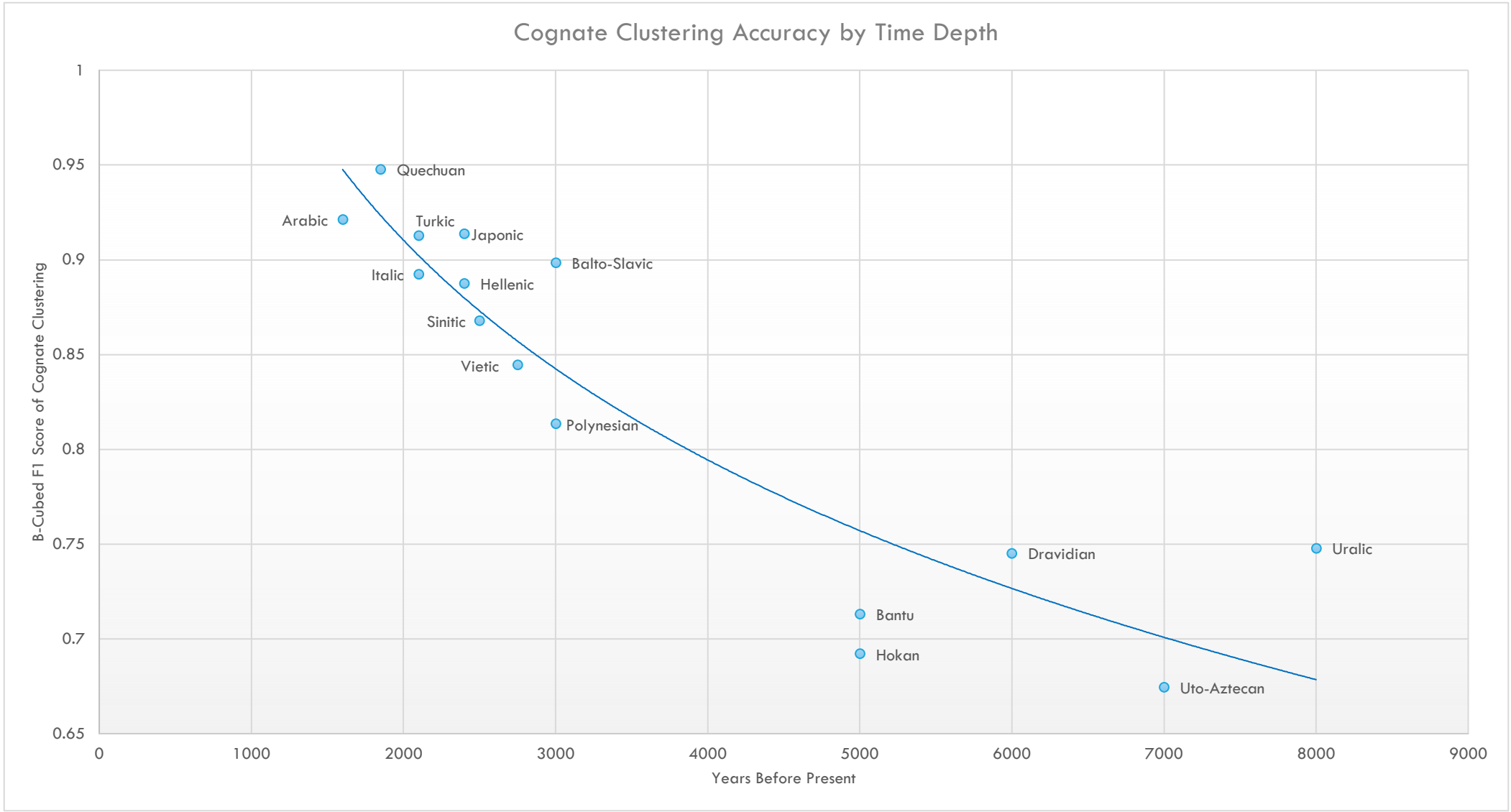
BASED ON VALIDATION DATASETS



Optimal threshold = 0.80

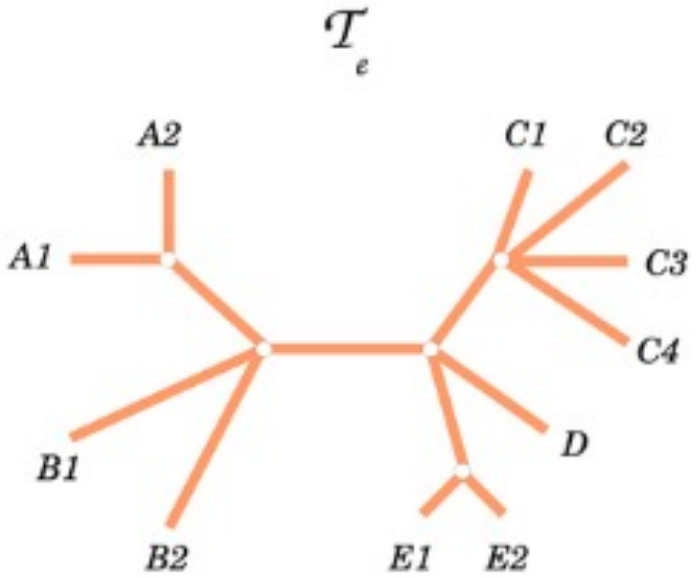

Accuracy by Cognate Clustering Method
B-Cubed F1 Score
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
ARABIC
BALTO-SLAVIC
DRAVIDIAN
HOKAN
ITALIC
POLYNESIAN
SINITIC
TURKIC
URALIC
BANTU
HELLENIC
JAPONIC
QUECHUAN
UTO-AZTECAN
VIETIC
Phonetic
PMI
Surprisal
LexStat

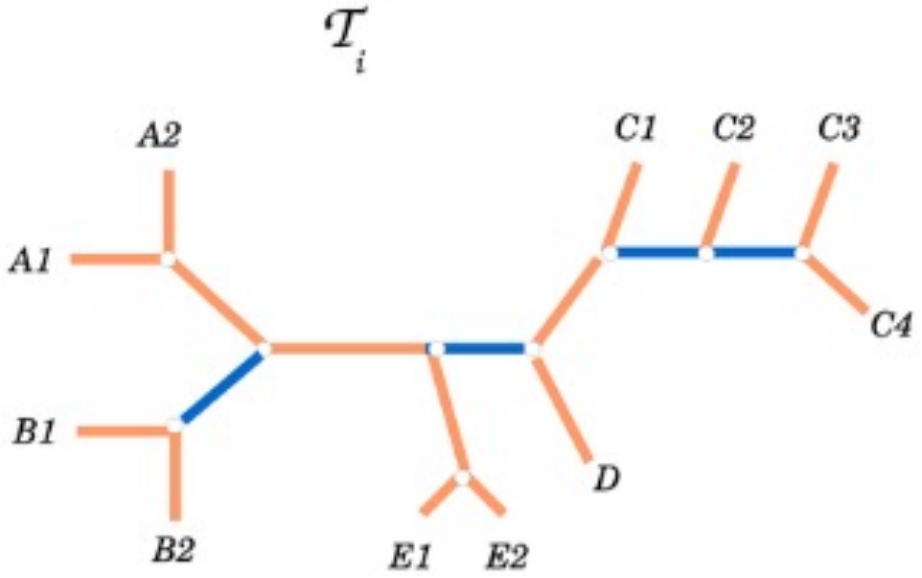

Mean Accuracy by Cognate Clustering Method
0.90
0.88
0.86
0.84
0.82
0.80
0.78
0.76
0.74
0.72
0.70
B-Cubed F1 Score
Phonetic
PMI
Surprisal
Test
Validation


Cognate Clustering Accuracy by Time Depth
B-Cubed F1 Score of Cognate Clustering
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
Years Before Present
Quechuan
Arabic
Turkic
Japonic
Balto-Slavic
Italic
Hellenic
Sinitic
Vietic
Polynesian
Dravidian
Uralic
Bantu
Hokan
Uto-Aztecan

IMPROVED TREE COMPARISON

- Found paper which discusses impact of comparing resolved, binary trees with unresolved, non-binary trees on Quartet Distance and Robinson-Foulds distance
  - Pompei, S., Loreto, V., & Tria, F. (2011). On the Accuracy of Language Trees. *PLoS ONE*, *6*(6). https://doi.org/doi:10.1371/journal.pone.0020109
  - Modification necessary to distinguish inherent contradictions in tree structure from differences due to resolution of trees → ***Generalized* Quartet Distance**
  - GQD seems to be the measure which is used in other phylogenetic inference papers
- Tree Dist presumably has the same issue but found no modification → GQD as primary evaluation now

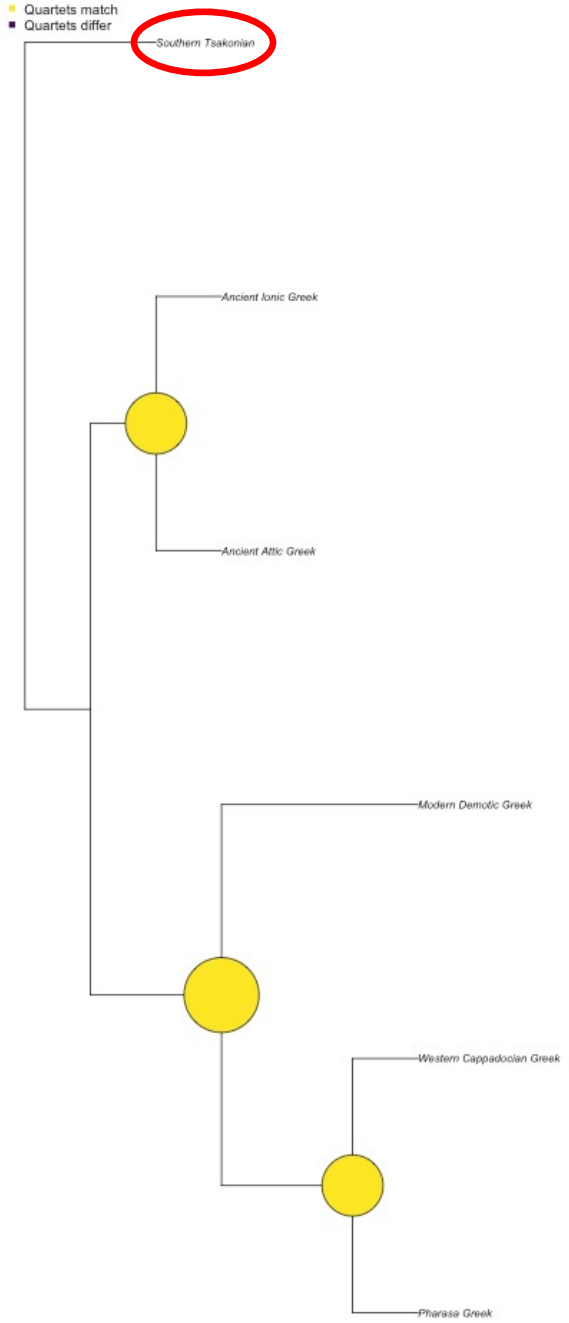


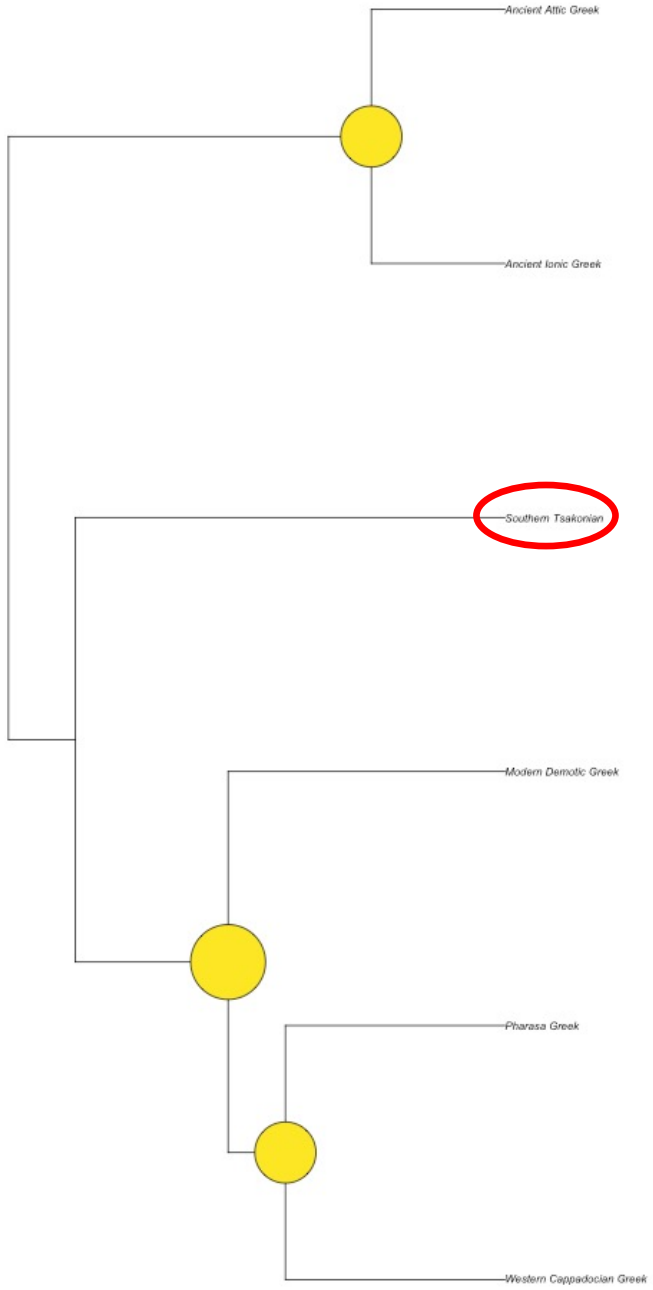


WEIRD BUG?

- Quartet Distance and Tree Distance seem not to penalize an outgroup item being grouped elsewhere in the tree
- 0 distance for both measures between these two trees, despite Southern Tsakonian appearing in different positions

Why? How to fix?

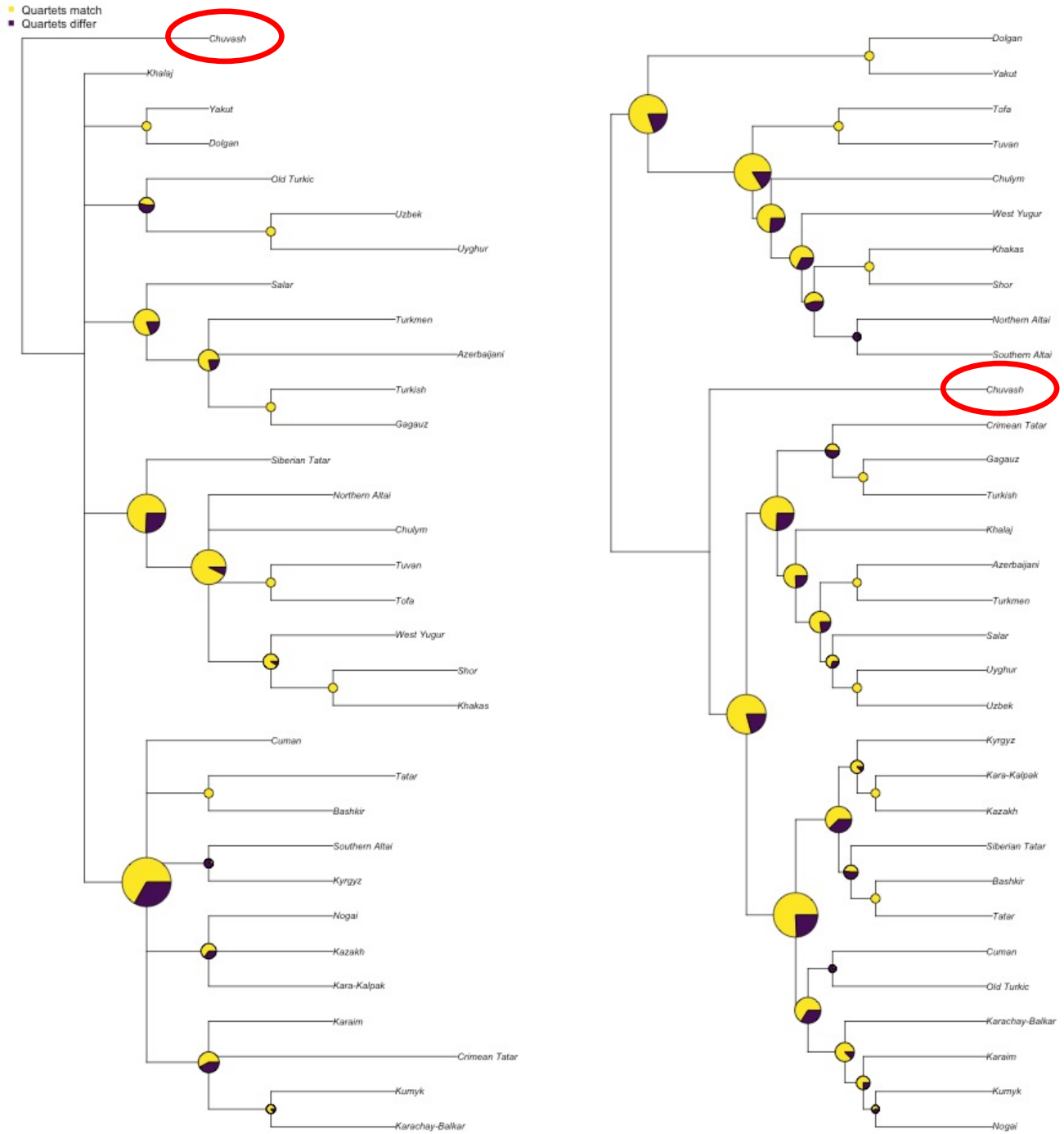




WEIRD BUG?

- Quartet Distance and Tree Distance seem not to penalize an outgroup item being grouped elsewhere in the tree
- Similar issue in Turkic tree with placement of Chuvash

Why? How to fix?



BEST AUTOMATIC TREES (NOT USING GOLD COGNATE SETS)

Dataset	TreeDist	Generalized Quartet Distance	Cognate Clustering Method	Word Form Evaluation Method	Linkage Method
Arabic	0.67	0.22	PMI	surprisal	Ward
Balto-Slavic	0.16	0.02	none	surprisal	complete
Dravidian	0.41	0.22	none	phonetic	average
Hokan	0.40	0.04	surprisal	phonetic	average
Italic	0.39	0.16	none	surprisal	Ward
Polynesian	0.58	0.12	PMI	phonetic	Ward
Sinitic	0.61	0.16	none	phonetic	average
Turkic	0.56	0.31	surprisal	PMI	Ward
Uralic	0.32	0.04	none	phonetic	complete

NOTES:

- Significantly lower tree distances using Generalized Quartet Distance
- No separation of cognate/non-cognates before evaluation often produces best trees
- Phonetic (5/9) and surprisal (3/9) are the most effective evaluation methods

BEST GOLD TREES (USING GOLD COGNATE SETS)

Dataset	TreeDist	Generalized Quartet Distance	Word Form Evaluation Method	Linkage Method
Arabic	0.64	0.30	PMI	Ward
Balto-Slavic	0.16	0.02	surprisal	complete
Dravidian	0.44	0.26	surprisal	complete
Hokan	0.36	0.02	phonetic	average
Italic	0.41	0.21	surprisal	Ward
Polynesian	0.57	0.12	PMI	Ward
Sinitic	0.63	0.27	surprisal	Ward
Turkic	0.57	0.32	surprisal	weighted
Uralic	0.40	0.05	phonetic	average

NOTES:

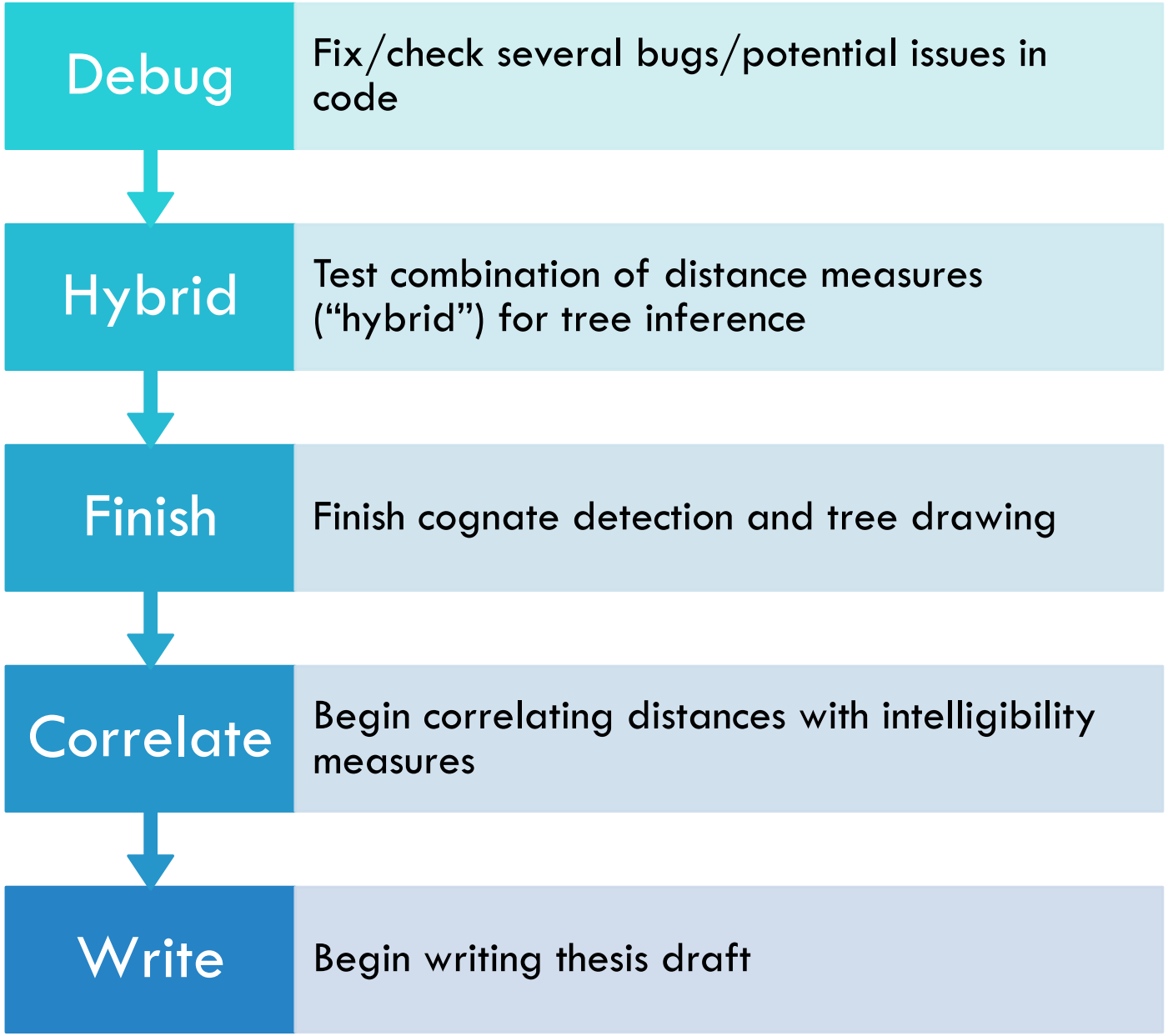
- **Most trees based on gold cognate sets are less accurate** than those using either automatic cognate sets or no separation of cognates/non-cognates before evaluation
- Only Hokan tree is better using gold cognate sets
- **Surprisal** seems to be the most effective word evaluation technique (5/9)

TIME PLAN…

- Finish coding distance measures: OCTOBER 19
- Finish testing on cognate detection: OCTOBER 22
- Finish testing on trees: OCTOBER 29
- Finish all data collection/experiment: NOVEMBER 1
- First draft complete: NOVEMBER 15 *
- Second draft complete: DECEMBER 1

* Will have begun writing sooner

NEXT TASKS


Debug
Fix/check several bugs/potential issues in code
Hybrid
Test combination of distance measures ("hybrid") for tree inference
Finish
Finish cognate detection and tree drawing
Correlate
Begin correlating distances with intelligibility measures
Write
Begin writing thesis draft