DISTANCE METHODS FOR PHYLOGENETIC INFERENCE

Thesis Proposal

Philip Georgis

DISTANCE-BASED METHODS

Jäger (2018): Global-scale phylogenetic linguistic inference from lexical resources

“Aggregating over all families suggests that distance-based inference produces the best fit with the expert gold standard. […] Combining both types of characters in a partitioned model always leads to better results than the two character types individually.”

RESEARCH QUESTION

What types of distance-based methods are useful (or: improve) phylogenetic inference?

- Phonetic
  - Phonemes (IPA) or sound classes?
  - Which phonetic features to use?
    - e.g. PHOIBLE features, proposed features by Tresoldi (2016), other?
  - Weighting features or segment types? (e.g. weight consonants more than vowels, or weight certain phonetic features more than others)
  - Penalties for deletions?

RESEARCH QUESTION

What types of distance-based methods are useful (or: improve) phylogenetic inference?

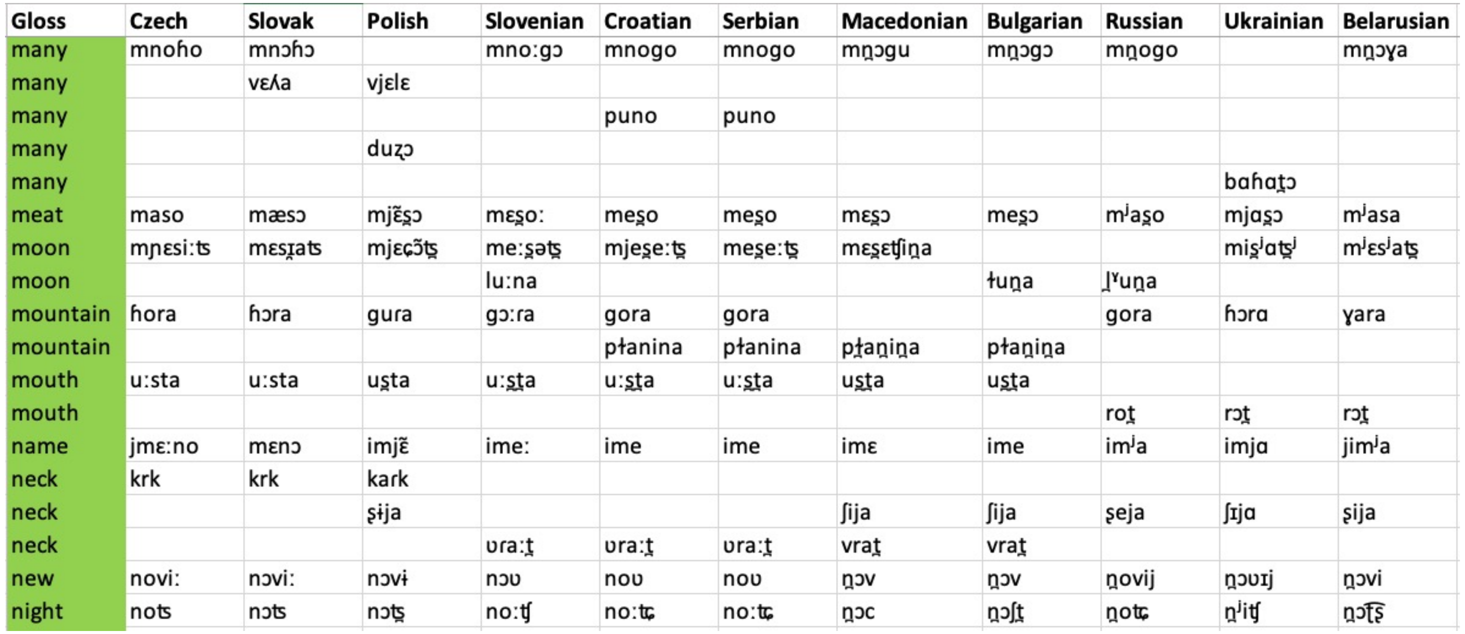
- Lexical (= cognate distance?)
  - Character-based binary lexical distances
  - PMI score *(method used by Jäger)*
  - Surprisal score → new method
    - Lower surprisal : total phoneme entropy ratio → closer cognates

- Morphosyntactic
  - How to represent features? (binary or continuous?)
  - Where to take features from? (e.g. WALS, grammars, UD corpora)
  - Which features to include?
    - → need to research which MST features are most associated with genetic relatedness

EVALUATION

Validation with gold cognate sets to measure performance in cognate detection step

- F1 score or MCC for classification accuracy
- Purity and completeness measures for cluster identification

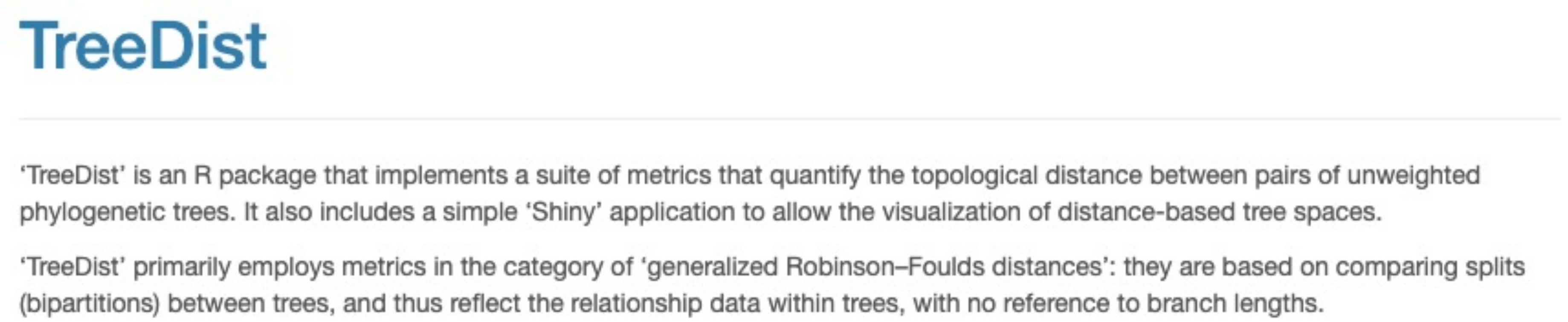


Gloss	Czech	Slovak	Polish	Slovenian	Croatian	Serbian	Macedonian	Bulgarian	Russian	Ukrainian	Belarusian
many	mnoɦo	mnɔɦɔ		mno:gɔ	mnogo	mnogo	mn̪ɔgu	mn̪ɔgɔ	mn̪ogo		mn̪ɔɣa
many		vɛʎa	vjɛlɛ								
many					puno	puno					
many			duʐɔ								
many										bɑɦɑt̪ɔ	
meat	maso	mæsɔ	mjɛ̃s̪ɔ	mɛs̪o:	mes̪o	mes̪o	mɛs̪ɔ	mes̪ɔ	mʲas̪o	mjɑs̪ɔ	mʲasa
moon	mɲɛsi:ʦ	mɛsɪ̯aʦ	mjɛɕɔ̃ts̪	me:s̪əts̪	mjes̪e:ts̪	mes̪e:ts̪	mɛs̪ɛʧin̪a			mis̪ʲɑts̪ʲ	mʲɛsʲats̪
moon				lu:na				ɫun̪a	l̠ˠun̪a		
mountain	ɦora	ɦɔra	gura	gɔ:ra	gora	gora			gora	ɦɔrɑ	ɣara
mountain					pɫanina	pɫanina	pɫ̪an̪in̪a	pɫan̪in̪a			
mouth	u:sta	u:sta	us̪ta	u:s̪t̪a	u:s̪t̪a	u:s̪t̪a	us̪t̪a	us̪t̪a			
mouth									rot̪	rɔt̪	rɔt̪
name	jmɛ:no	mɛnɔ	imjɛ̃	ime:	ime	ime	imɛ	ime	imʲa	imjɑ	jimʲa
neck	krk	krk	kark								
neck			ʂɨja				ʃija	ʃija	ʂeja	ʃɪjɑ	ʂija
neck				ʋra:t̪	ʋra:t̪	ʋra:t̪	vrat̪	vrat̪			
new	novi:	nɔvi:	nɔvɨ	nɔʋ	noʋ	noʋ	n̪ɔv	n̪ɔv	n̪ovij	n̪ɔʋɪj	n̪ɔvi
night	noʦ	nɔʦ	nɔts̪	no:ʧ	no:ʨ	no:ʨ	n̪ɔc	n̪ɔʃt̪	n̪oʨ	n̪ʲiʧ	n̪ɔt͡ʂ

EVALUATION

Comparison of generated phylogenetic trees with gold-standard trees from experts

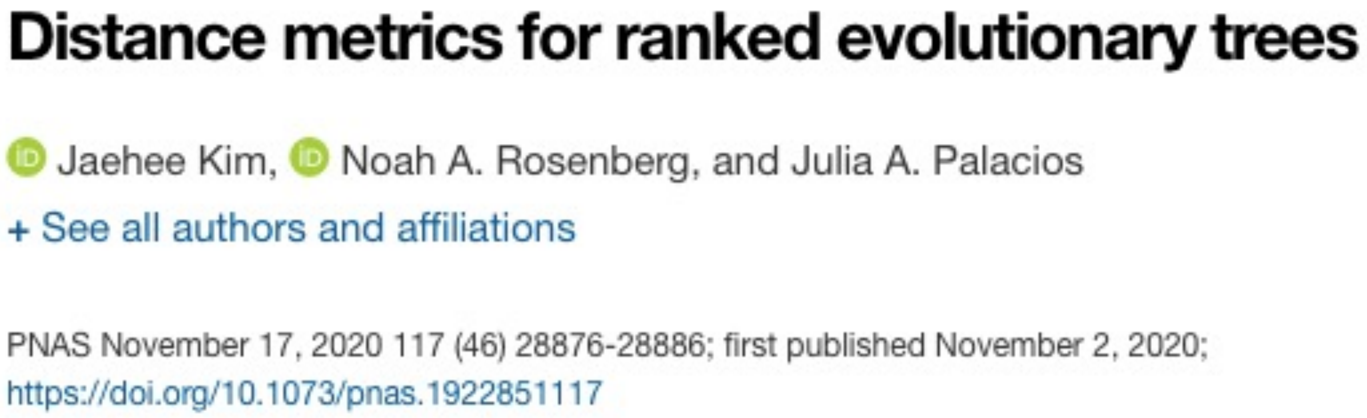
- e.g. Glottolog trees



TreeDist

'TreeDist' is an R package that implements a suite of metrics that quantify the topological distance between pairs of unweighted phylogenetic trees. It also includes a simple 'Shiny' application to allow the visualization of distance-based tree spaces.

'TreeDist' primarily employs metrics in the category of 'generalized Robinson–Foulds distances': they are based on comparing splits (bipartitions) between trees, and thus reflect the relationship data within trees, with no reference to branch lengths.



Distance metrics for ranked evolutionary trees

Jaehee Kim, Noah A. Rosenberg, and Julia A. Palacios

+ See all authors and affiliations

PNAS November 17, 2020 117 (46) 28876-28886; first published November 2, 2020; https://doi.org/10.1073/pnas.1922851117

EVALUATION

How to evaluate the **distance measures?**

- Correlation with mutual intelligibility measures
    - Available to some extent for:
        - Slavic (SFB C4 experimental results; Lindsay; Golubović & Gooskens, 2015)
        - Germanic (Gooskens, 2017)
        - Turkic (Lindsay)
        - …possibly others?

- Other ideas?

FORMAL ASPECTS

First step: Get clearer idea research question and contact supervisor

Master seminar

- Meet regularly with supervisor to refine topic
- Write thesis proposal
- Give ~30 minute presentation of proposal
- Supervisor then signs thesis registration form

Thesis

- Continue meeting with supervisor
- Write up findings
- Submission: minimum 3 months, maximum 6 months from date of registration

FORMAL ASPECTS

Second advisor

- **Bernd Möbius**
  - → if the focus would be more on phonetic aspects generally

- **Tania Avgustinova**
  - → if the focus would be more on correlation with inter-comprehension, morphosyntax, and/or Slavic as a key example family