Sharper evaluation of deep-time hypotheses using both phonological and semantic divergence

Erich Round (University of Surrey; University of Queensland), James Elhindi (University of Sydney)

Keywords: long-distance relatedness; cognacy; semantic change; semantic divergence

Background: Over time, true cognates *a*) diverge phonologically; *b*) diverge semantically; and *c*) are subject to lexical replacement. Hypotheses of long-distance relatedness (LDR) can be contentious, as they push the interpretation of lexical evidence its limits [1, 2, 3]. Nevertheless, agreement has emerged [4, 5, 6] that any level-headed evaluation of LDR must compare levels of similarity in word pairs: *a*) phonologically, *b*) semantically and *c*) against chance. For task *c*) (comparison against chance) the agreed, best approach is a Monte Carlo test [4, 5], where word pairs of interest are compared against a large, baseline sample of randomly paired words. Since doing so entails thousands of pairwise comparisons, methods for tasks *a*) and *b*) (phonological and semantic comparison) must be scalable and automated. Computational methods for task *a*) (comparison of phonology) are now sufficient to meet these demands [7]. The remaining weak point is task *b*), semantic comparison. The current standard is to compare only word pairs that satisfy semantic *equivalence*, effectively adopting a hypothesis according to which deep-time cognates never diverge semantically. This *Stasis Hypothesis* is false, but has been widely adopted as a heuristic, 'least-bad' option, explicitly due to the unavailability of any viable, sufficiently non-subjective, scalable alternative [1].

Contribution: We present a scalable, non-subjective measure of semantic distance and demonstrate its application to LDR hypotheses among language families of Australia. Though approximate, our method appears to be more informative than reliance on the incorrect Stasis Hypothesis alone.

Semantic divergence: Semantic changes $*meaning_1 > meaning_2$ are often preceded by periods of polysemy [8], and thus paths of change can be inferred empirically from the typology of polysemy. Expertly hand-curated polysemy databases exist [9] but for only high-resource languages and only parts of the lexicon. Working with low-resource languages, we took English definitions from entire wordlists and dictionaries, applied parsing tools [10] to extract a bundle of keywords for each lexeme, and built a network of recurrent polysemies in 189 languages in Bowern's CHIRILA database [11] (35k edges, 9k nodes). A semantic *distance* between pairs of meanings can be read from their separation distance in the network (with weighted and unweighted methods producing similar results).

Evaluations of language relatedness: Starting with 63 Pama-Nyungan (PN) languages whose ages of relatedness are approximately known [12, 13], for a pair of languages we examined how similar their word pairs are phonologically when sampled at each of several semantic distances. In closely related languages these distances should correlate; and phonological matches at low semantic distances should exceed chance. We find these relationships as expected (Figure 1). Next, using the empirical distribution of phonological and semantic distances versus known ages of relationships in PN as a guide, we implemented a maximum-likelihood test to evaluate the highest-likelihood ages of relatedness for language pairs from five distinct non-PN families with various, speculated deep-time linkages [14, 15, 16]. In all cases, the most likely time depth (compared to PN-internal pairs) is 4–7,000y, but even more likely is a greater time depth, similar to PN/non-PN language pairs.

Conclusions: We demonstrate here that it is now possible, and useful, even for low-resource languages, to cease our reliance on the false Stasis Hypothesis when evaluating LDR hypotheses. Future directions include improved estimation of distances, more explicit modelling of data-generating processes, phylogenetic modelling for multi-lateral comparison [4]; and extension, to inference of semantic change and semantic reconstruction [17], and joint phonological and semantic inference.

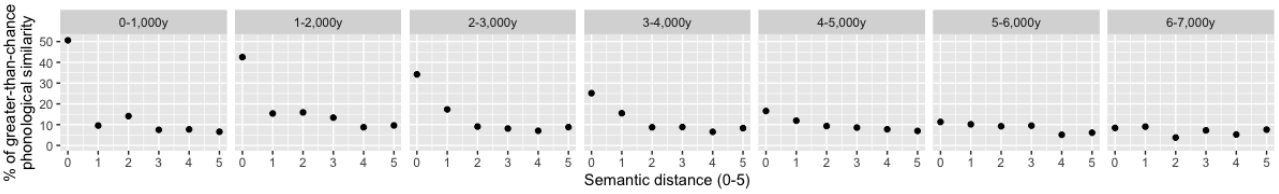

Figure 1: Phonological similarity (vertical) versus semantic distance (horizontal within panels) and time depth (between panels). Not only does the similarity of semantic equivalents (distance=0) drop with time, but the similarity of almost-equivalents (distance=1) first rises across millennia 0,1,2 and then falls again in millennia 3,4,5,6. At greater semantic distances (>1), no notable patterns emerge.

References

[1] Lyle Campbell. “Distant genetic relationship and the Maya-Chipaya hypothesis”. In: *Anthropological Linguistics* 15.3 (1973), pp. 113–135.

[2] Sheldon P. Harrison. “On the limits of the comparative method”. In: *The handbook of historical linguistics*. Ed. by Brian D. Joseph and Richard Janda. Chicago: Blackwell, 2003, pp. 213–243.

[3] Lyle Campbell and William J Poser. *Language classification: History and method*. Cambridge: Cambridge University Press, 2008.

[4] Brett Kessler and Annukka Lehtonen. “Multilateral comparison and significance testing of the Indo-Uralic question”. In: *Phylogenetic methods and the prehistory of languages*. Ed. by Peter Forster and Colin Renfrew. Cambridge: McDonald Institute for Archaeological Research, 2006, pp. 33–42.

[5] Don Ringe and Joseph F Eska. *Historical linguistics: Toward a twenty-first century reintegration*. Cambridge: Cambridge University Press, 2013.

[6] Andrea Ceolin. “Significance testing of the Altaic family”. In: *Diachronica* 36.3 (2019), pp. 299–336.

[7] Johann-Mattis List, Simon Greenhill, and Robert Forkel. *LingPy. A Python library for quantitative tasks in historical linguistics*. Jena, 2017.

[8] Nicholas Evans and David Wilkins. “In the mind’s ear: the semantic extensions of perception verbs in Australian languages”. In: *Language* 76.3 (2000), pp. 546–592.

[9] Christoph Rzymski et al. “The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies”. In: *Scientific data* 7.1 (2020), pp. 1–12.

[10] Kenneth Benoit and Akitaka Matsuo. *spacyr: Wrapper to the ’spaCy’ ’NLP’ Library*. R package version 1.2.1. 2020. URL: `https://CRAN.R-project.org/package=spacyr`.

[11] Claire Bowern. “Chirila: Contemporary and Historical Resources for the Indigenous Languages of Australia”. In: *Language Documentation & Conservation* 10 (2016), pp. 1–44.

[12] Claire Bowern and Quentin Atkinson. “Computational phylogenetics and the internal structure of Pama-Nyungan”. In: *Language* 88.4 (2012), pp. 817–845.

[13] Remco R. Bouckaert, Claire Bowern, and Quentin D. Atkinson. “The origin and expansion of Pama–Nyungan languages across Australia”. In: *Nature ecology & evolution* 2 (2018), pp. 741–649.

[14] Nicholas Evans. “Comparative non-Pama-Nyungan and Australian historical linguistics”. In: *The Non-Pama-Nyungan languages of northern Australia: Comparative studies of the continents most linguistically complex region*. Ed. by Nicholas Evans. The Australian National University: Pacific Linguistics, 2003, pp. 3–28.

[15] Rebecca Green. “Proto Maningrida with proto Arnhem: evidence from verbal inflectional suffixes”. In: *The non-Pama-Nyungan languages of northern Australia : comparative studies of the continent’s most linguistically complex region*. Ed. by Nicholas Evans. Canberra: Pacific Linguistics, 2003, pp. 369–421.

[16] Mark Harvey and Robert Mailhammer. “Reconstructing remote relationships: Proto-Australian noun class prefixation”. In: *Diachronica* 34.4 (2017), pp. 470–515.

[17] Gerhard Jäger and Johann-Mattis List. “Using ancestral state reconstruction methods for onomasiological reconstruction in multilingual word lists”. In: *Language Dynamics and Change* 8.1 (2018), pp. 22–54.