The Comparative Method on a shoestring: Evaluating chance vs inheritance with a limited database

One of the less discussed limits of Comparative Method is the database requirements for establishing regularity of correspondences. We may consider a hypothetical example assessing relatedness among three languages – A, B, C. For Correspondence 1, there are 20 cognate sets where Langs A and B have /t/ and Lang C has /s/. For Correspondence 2, there are 2 sets where Lang A has /n/ and Lang B has /n̠/, 2 different sets where Langs A and C have /n/, and no other [anterior] nasal sets. Correspondence 1, A /t/, B /t/ C /s/, is more securely based than Correspondence 2, A /n/, B /n̠/, C /n/ and it is not clear that Correspondence 2 is sufficiently supported (Mailhammer 2015; Mailhammer & Harvey 2018).

The basic aim of the Comparative Method is to determine whether inheritance is better supported than chance or contact as explanations for similar forms (Harrison 2003; Weiss 2014). We propose that it is possible to statistically evaluate chance vs inheritance in cases where the database is limited, such as Correspondence 2. There are different mathematical approaches have been applied to assess correspondences in hypothesised distant genetic relationships (Ringe 1993; Oswalt 1993).

We propose a further development of Ringe (1992), assessing how likely it is that sound correspondences are accidental using a binominal distribution formula. The key information needed for this method is how frequent on average each relevant phoneme is in each relevant position across all the languages in the sample. Frequencies are multiplied for each phoneme considered and the formula determines whether a match in Y languages across a sample of X languages is accidental. We propose two innovations. The first is assessing the frequency of phonemes across entire lexicons to avoid false positives (Baxter 1993). The second is to compare only identical matches, i.e. identical phonological forms that share a common meaning.

We exemplify this method using a database from Australian languages, which provide a good testing ground, as there are proposals for extensive genetic relationships among Australian languages (Harvey & Mailhammer 2017; Koch 2014), but there are limited numbers of potential cognates and establishing correspondences is problematic (Miceli & Round 2022). We assembled phonologically identical forms expressing 25 lexical concepts across Australian languages, and we examined full lexicons from 35 Australian languages to determine average frequencies of phonemes in all phonotactic positions.

Results show that widespread identities are very unlikely to be due to chance. For example, there are 27 languages where the form /pu/ conveys the meaning 'hit'. It can be ruled out that this match is accidental on a 0.000 level, irrespective of whether average or the highest attested frequencies for /p/ and /u/ are used in the calculation. This even holds if the frequencies of /p/ in initial and /u/ in second position were 50%, which is of course unrealistic across the lexicon of any language. Given that chance is not supported, the remaining hypotheses on shared forms, such as /pu/ 'hit', are contact or inheritance. We do not consider the contact vs inheritance choice here, but note that there are well-established criteria bearing on the choice: e.g. continuity vs discontinuity of attestation (Harvey & Mailhammer 2017); variation in probability of borrowing by part-of-speech class and semantic domain (Tadmor, Haspelmath & Taylor 2010; Tadmor 2009).

References:

Baxter, William. 1993. Response to Oswalt and Ringe. In Joseph C. Salmons & Brian D. Joseph (eds.), *Nostratic: Sifting the evidence*, 217–236. Amsterdam & Philadelphia: John Benjamins.

Harrison, S. P. 2003. On the Limits of the Comparative Method. In Brian D. Joseph & Richard D. Janda (eds.), *The Handbook of Historical Linguistics*. Maldon, MA: Blackwell.

Harvey, Mark & Robert Mailhammer. 2017. Reconstructing remote relationships: Proto-Australian noun class prefixation. *Diachronica* 34(4). 470–515.

Koch, Harold. 2014. Historical relations among the Australian languages. In Harold Koch & Rachel Nordlinger (eds.), *The languages and linguistics of Australia: a comprehensive guide* (The World of Linguistics 3), 23–89. Berlin & Boston: Walter de Gruyter.

Mailhammer, Robert. 2015. Etymology. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge Handbook of Historical Linguistics*, 423–441. London: Routledge.

Mailhammer, Robert & Mark Harvey. 2018. A reconstruction of the Proto-Iwaidjan phoneme system. *Australian Journal of Linguistics* 38. 329–359.

Miceli, Luisa & Erich Round. 2022. Where have all the sound changes gone? Examining the scarcity of evidence for regular sound change in Australian languages. *Linguistics Vanguard* 8(s5). 509–518. https://doi.org/10.1515/lingvan-2021-0094.

Oswalt, Robert. 1993. A probabilistic evaluation of North Eurasiatic Nostratic. In Joseph C. Salmons & Brian D. Joseph (eds.), *Nostratic: Sifting the evidence*, 199–216. Amsterdam & Philadelphia: John Benjamins.

Ringe, Don. 1993. Probabilistic evidence for Indo-Uralic. In Joseph C. Salmons & Brian D. Joseph (eds.), *Nostratic: Sifting the evidence*, 153–197. Amsterdam & Philadelphia: John Benjamins.

Ringe, Donald. 1992. On calculating the factor of chance in language comparison. *Transactions of the American Philosophical Society* 82(1). 1–110.

Tadmor, Uri. 2009. Loanwords in the World's Languages: Findings and results. In Martin Haspelmath & Uri Tadmor (eds.), *Loanwords in the World's Languages*, 55–75. Berlin/New York: De Gruyter Mouton.

Tadmor, Uri, Martin Haspelmath & Bradley Taylor. 2010. Borrowability and the notion of basic vocabulary. *Diachronica* 27(2). 226–246.

Weiss, Michael. 2014. The Comparative Method. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge Handbook of Historical Linguistics*, 127–145. London: Routledge.