Short, frequent words are more likely to appear genetically related by chance

An important question in historical linguistics is whether deep genetic relationships exist across language families. Although specific families can be reconstructed back to around 6,000 y ago, Pagel et al. (1) claim that seven Eurasian families arose from a common ancestor 15,000 y ago. Pagel et al. develop a phylogenetic model, starting with a subset of the Swadesh basic word list for seven language families in the Languages of the World Etymological Database, which lists reconstructed proto-words and cognates. Because these reconstructions are potentially unreliable, Pagel et al. treat each reconstructed cognate pair as a binary random variable. They find a robust correlation between the size of the cognate class and the word replacement rate (i.e., how fast the word is likely to be replaced in the vocabulary), which is closely related to frequency. As predicted, words with a slower replacement rate show deeper relationships across language families, which they take as evidence that there are deep relationships among the seven families.

Pagel et al.'s model critically requires that judgments of cognates not be confounded with frequency. Because there are known correspondences between frequency and wordform, this assumption is suspect. Pagel et al. underestimate the possibility that the relationship between cognate class size and frequency is due to chance by not accounting for a word length bias in how cognate pairs are assigned. There is a robust inverse correlation between word frequency and word length (2, 3), so words like "I" or "me" that are frequent across languages are also likely to be short. Even slightly shorter words are much more likely to be phonologically similar simply by chance: the likelihood of finding a minimal phonological pair for randomly generated words over an alphabet increases exponentially with the number of letters in the pair: two 3-letter words randomly sampled from a uniform distribution over 26 letters are 20 times more likely to be one edit apart than two 4-letter words sampled from the same distribution. Because, for articulatory reasons, short words are cross-linguistically likely to consist of simple sequences like CV or CVC, the space of possible variation across languages is small.

Pagel et al. address this concern and suggest that their result holds even when closed classes of words (such as pronouns, which tend to be short) are excluded, but they do not include word length information in their model. For the 188 unique words in Pagel et al.'s table S1 (excluding infinitive verb markers), there is a −0.24 correlation between English phonological word length from CELEX (4) and cognate family size. The ultraconserved words have mean phonological length 3.4 compared with 3.0 for all other words in the table. This bias could artificially inflate the link between frequency and cognate class size by encouraging false positives: cognitive science research shows that humans consistently find structure in random data.

Although Pagel et al.'s research raises many interesting questions, the data presented are not sufficient to conclude that the seven language families have a common ancestor.

Kyle Mahowald[1] and Edward Gibson
Department of Brain and Cognitive Sciences,
Massachusetts Institute of Technology,
Cambridge, MA 02139

1 Pagel M, Atkinson QD, Calude AS, Meade A (2013) Ultraconserved words point to deep language ancestry across Eurasia. *Proc Natl Acad Sci USA* 110(21):8471–8476.

2 Zipf GK (1949) *Human behavior and the principle of least effort* (Addison-Wesley, New York).

3 Piantadosi ST, Tily H, Gibson E (2011) Word lengths are optimized for efficient communication. *Proc Natl Acad Sci USA* 108(9):3526–3529.

4 Baayen RH, Piepenbrock R, Gulikers L (1995) *The CELEX Lexical Database* (release 2; CD-ROM) (Linguistic Data Consortium, Univ of Pennsylvania, Philadelphia).

Author contributions: K.M. and E.G. designed research, and K.M. wrote the paper.

The authors declare no conflict of interest.

[1] To whom correspondence should be addressed. E-mail: kylemaho@mit.edu.

www.pnas.org/cgi/doi/10.1073/pnas.1308822110