JASs *forum*
Journal of Anthropological Sciences
Vol. 88 (2010), pp. 231-233

The prospects for tracing deep language ancestry

Quentin D. Atkinson

Institute of Cognitive & Evolutionary Anthropology, University of Oxford, 64 Banbury Road, Oxford, OX2 6PN, U.K.
Department of Psychology, University of Auckland, Private Bag 92019, Auckland, New Zeland
e-mail: q.atkinson@auckland.ac.nz

> *"If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world; and if all extinct languages, and all intermediate and slowly changing dialects, had to be included, such an arrangement would, I think, be the only possible one."*
> *The Origin of Species,*
> *Charles Darwin (1859)*

Whilst there is now broad agreement that our genetic ancestry can be traced back to a late Pleistocene origin in Africa, there is no such consensus about the roots of the world's 6000 or so languages. Proposed language super-families – such as Amerind in the Americas and Nostratic and Eurasiatic in Eurasia – or global language classifications like those controversially linked to the human genetic tree (Cavalli-Sforza *et al.*, 1988), are viewed with scepticism by most linguists. Words are thought to evolve too rapidly to allow reliable identification of common ancestry beyond a limit of ~8ky BP (Ringe, 1998) and when apparent 'long-range' relationships are identified, proponents have been unable to provide statistical verification that any resemblances are beyond what would be expected by chance (Ringe, 1998). However, recent advances in the available data and methods (Dunn *et al.*, 2005; Pagel, 2000; Pagel *et al.*, 2007; Reesnik, Singer & Dunn, 2009) suggest the established ~8ky limit may need to be re-evaluated (Gray, 2005), potentially greatly extending the time depth over which language ancestry is informative about human prehistory.

Most claims for long-range language relationships rest on putative lexical homologues or 'cognates' identified on the basis of form and meaning correspondences across languages. One reason many have found this evidence hard to swallow is that the rate of replacement of cognates through time appears to be too rapid and too unpredictable to leave any reliable signal after just a few thousand years. For example, Morris Swadesh's (Swadesh, 1952) early attempts to derive a single lexical retention rate found that even among a set of 200 relatively stable basic vocabulary terms, on average roughly 20% of cognates are lost every 1000 years. As shown in Figure 1 (bold line), such a rate implies that a pair of languages that diverged just 4,500 years ago (separated by 9,000 years of change) is expected to share only five cognates from an initial 200 in the Swadesh list. After 7,000 years, this number drops below one. Under this scenario, proposals for language classifications stretching back to the early Neolithic and beyond seem completely untenable – the number of cognates at such time depths will be too few to allow genuine historical signal to be distinguished from chance resemblances.

However, not all words are created equal - some evolve more slowly than others. Pagel (2000) has shown that a model of lexical evolution that allows rates of change to differ across meanings fits the observed distribution of lexical divergence in Indo-European better than Swadesh's constant rate model. More recent work has revealed that the rate at which different

Swadesh list meanings evolve is correlated across language families (Pagel & Meade, 2006) and that the frequency with which a meaning is used in everyday speech, together with its part of speech, can explain almost 50% of the variation in rates of lexical replacement (Pagel *et al.*, 2007). Thus, commonly used pronouns (such as *I*, *you* and *we*) and numerals (*one*, *two*, *four* and *five*) evolve roughly 100 times slower than the rarer, more rapidly evolving Swadesh adjectives and verbs (such as *dirty*, or *to throw*) (Pagel *et al.*, 2007). This predictable variation in rates of lexical replacement dramatically increases the feasibility of reconstructing deep language ancestry.

Figure 1 (grey line) shows the expected number of surviving cognates shared between language pairs for a given separation time based on the empirically derived rate distribution from Pagel *et al.*, (2007). Whilst under a constant rate model it would take only 4,500 years to reduce the cognate pool from 200 to five, allowing for rate variation extends this threshold beyond 20,000 years. Even languages that separated 50kya, perhaps contemporaneous with the African exodus, are expected to share at least two cognates. Of course, even if cognates exist at such time depths, there remains the problem of identifying them and demonstrating that any similarities are beyond what would be expected by chance, but the predictability of rates across meanings may help here too. Based on information about word frequency, part of speech or rates of change within language families, one can predict not just how many cognates should be shared between a pair of languages given some time of separation, but which meanings are more likely to produce cognate forms. Finding cognate forms for two or three meanings from a possible 200 may not constitute convincing evidence for a relationship, but if those meanings are also *a priori* expected to be the most stable, then a case for common ancestry can be made.

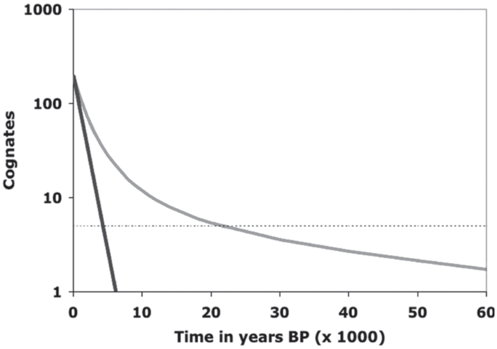

Fig. 1 – the expected number of Swadesh 200 meaning list cognates surviving to the present plotted against language divergence time for Swadesh's (1952, 1955) constant rate model (bold line) and for a model incorporating the empirical distribution of rates derived from Indo-European (grey line; Pagel, Atkinson and Meade, 2007). For comparison, a dashed line is drawn at five surviving cognates. The colour version of this figure is available at the JASs website.

As well as words, structural features of language, such as the set of phonemes a language uses, its gender system or favoured word order, can also provide information about language ancestry. Although we currently lack rate estimates for structural data of the kind mentioned above, some structural features are claimed to be highly stable (Nichols, 1992) and so may prove decisive in identifying long-range language relationships. Indeed, some of the most promising recent research testing deep ancestry hypotheses makes use of structural language features. Dunn *et al.,* (2005), for example, were able to use structural data together with phylogenetic inference techniques from evolutionary biology to identify historical signal in the Papuan languages likely to date back over 10,000 years. More recently, Reesnik et al. (2009), have used structural data to classify the languages of the ancient super-continent Sahul into recognized major groups, some of which are likely to be just as old or perhaps much older. These findings are among the first to demonstrate language relationships beyond the traditionally held ~8ky limit. As in the case of the lexical data, if a set of highly stable structural features can be identified, it should be possible to push this time horizon back substantially further.

From our origins in Africa, the story of human evolution is largely one of cultural change. Language genealogies track cultures in a way that genes cannot (Friedlaender *et al.,* 2009) and so are crucial to our understanding of human prehistory. The findings discussed here suggest that we should in principle be able to trace language ancestry back beyond the Neolithic, perhaps even as far as our expansion from Africa. Comparative analysis and hypothesis testing on a global scale will require high-quality and easily accessible lexical and structural language databases covering a large fraction of the world's languages. Some important steps are now being taken in this direction (e.g., the World Atlas of Language Stuctures (Haspelmath *et al.,* 2005)) but more work is needed along these lines if we are to fully capitalise on the linguistic legacy of our cultural past.

References

Cavalli-Sforza L.L., Piazza A., Menozzi P. & Mountain J. 1988. Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. *Proc. Natl. Acad. Sci. U.S.A.*, 85:6002-6006.

Dunn M., Terrill A., Reesink G., Foley R.A. & Levinson S.C. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science,* 309:2072-2075.

Friedlaender J., Hunley K., Dunn M., Terrill A., Lindstrom E., Reesink G. & Friedlaender F. 2009 Linguistics More Robust Than Genetics. *Science,* 324:464-465.

Gray R. D. 2005. Pushing the Time Barrier in the Quest for Language Roots. *Science,* 309:2007-2008.

Haspelmath M., Dryer M. S., Gil D. & Comrie B. 2005. *The World Atlas of Language Structures*. Oxford University Press, Oxford.

Nichols J. 1992. *Linguistic Diversity in Space and Time*. University of Chicago Press, Chicago.

Pagel M. 2000. Maximum-likelihood models for glottochronology and for reconstructing linguistic phylogenies. In C. Renfrew, A. McMahon & L. Trask (eds): *Time Depth in Historical Linguistics*, pp. 189-207. MacDonald Institute of Archaeological Research, Cambridge.

Pagel M., Atkinson Q. D. & Meade A. 2007 Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature,* 449:717-720.

Pagel M. & Meade A. 2006. Estimating Rates of Lexical Replacement on Phylogenetic Trees of Languages. In J. Clackson, P. Forster & C. Renfrew (eds): *Phylogenetic methods and the prehistory of languages*, pp. 173-182. MacDonald Institute for Archaeological Research, Cambridge

Reesink G., Singer R. & Dunn M. 2009. Explaining the Linguistic Diversity of Sahul Using Population Models. *Plos Biol.*, 7, e1000241.

Ringe D. 1998. A probabilistic Evaluation of Indo-Uralic. In J. C. Salmons & B. D. Joseph (eds): *Nostratic: Sifting the Evidence*, pp. 143-198. John Benjamins Publishing, Amsterdam.

Swadesh M. 1952 Lexico-statistic dating of prehistoric ethnic contacts. *Proc. Am. Philos. Soc.*, 96:453-463.