BEYOND LEXICOSTATISTICS: HOW TO GET MORE OUT OF 'WORD LIST' COMPARISONS

Dr Paul Heggarty — McDonald Institute for Archaeological Research, University of Cambridge

ABSTRACT (500 WORDS)

If lexicostatistics could speak, it might justifiably assert: "reports of my death have been greatly exaggerated". For, glottochronology aside, various facets of Swadesh's basic lexicostatistical 'idea' seem alive and well, in a new breed of modern derivatives. This paper first reviews the dominant trends today, then presents alternative approaches to take quantitative lexical comparison in other new directions. Illustrative case-studies range from subfamilies of Indo-European to two major language families of the New World.

A persistent ambiguity has attended 'lexicostatistics', in that methods that go by this name have variously sought to answer two very different types of inquiry:

- An information-type question of *degree*: *how closely* related are languages A and B (within their *known* family)?
- A yes/no-type question: are languages A and B related or not?

These represent two opposing directions in which lexicostatistical methodology might be refined to extract more mileage out of it, extending its range at either the 'shallow' or the 'deep end' of its applicability. The thrust of recent work has been in the latter direction, seeking to isolate the most reliable signal diagnostic of deep-time relatedness, by excluding 'less stable' meanings (which also simplifies data collection). A raft of recent studies hone their lists down far beyond Swadesh's 200 and then 100 items, to a minimal 'most stable' core of just 55, 40, 35, or even 23 meanings.

In this paper I argue that we would do well *not* to discard the less stable meanings. Firstly, for 'shallow end' purposes the 'binary straightjacket' of lexicostatistics is already all too blunt a characterisation of *degree* of overlap in lexical semantics; *a fortiori* if we limit the list to the most stable, i.e. least variable, data. In phonetics, a new methodology offers a 'resolution per word' beyond the wildest dreams of lexicostatistics, discriminating even to the accent level. To extend quantification in lexical semantics likewise into dialectology, I propose a radically new method to extract, from each individual meaning, measurements considerably finer-grained than just a binary 'cognate, yes or no?' datum. Again, *less* stable meanings offer richer data. A lesson duly emerges for enthusiasts of phylogenetic analyses too: unrefined lexicostatistical 'encoding' inherently biases results towards more tree-like outputs than the real language data warrant, misrepresenting also the prehistory of speaker populations.

More unexpected is how useful the less stable meanings can prove even at the 'deep end'. Here it is the *contrast* with more stable meanings, and the detailed *gradient* between them, that provide a stark and highly informative perspective on whether given languages are distantly related. By again abandoning certain tenets of traditional lexicostatistics, the fraught case-by-case judgement of cognacy (automated or otherwise) can be sidestepped entirely — especially useful when wordforms are clearly correlate, but data and scholarship are inconclusive or insufficient to confirm whether contact or common origin is the explanation. The Andes provide an ideal test case: the method proposed adduces powerful evidence for the debates on Quechua-Aymara relatedness, and on how far 'borrowability' influences the stability of particular meaning slots.