

Making the Most of Limited Data: Network methods and language documentation.

April McMahan
University of Edinburgh

In some cases of language documentation, linguists (and likewise colleagues from other disciplines) may encounter limitations in the scope, type and quantity of language data they are able to collect. Difficulties of access to data are of many kinds, and in this paper I shall focus on methodological innovations which allow us to make maximal use of limited data to answer linguistic and cross-disciplinary questions.

I shall focus on the use of network-based quantitative and computational methods in recent and ongoing research, and specifically on their relevance to two sets of issues. First, I shall show that such methods, applied to Swadesh-type basic vocabulary lists, may help answer previously intractable questions about language family affiliations, and will illustrate these points using the case of the Andean languages Quechua and Aymara. A second live issue concerns the degree of similarity between varieties of a single language (dialect classification rather than language classification, if you like); and here I shall argue that phonetic data represent the best way forward, as a considerable quantity of comparative phonetic data is contained even in a small number of words. Data here come from an ongoing project on accents of English, but I shall also discuss possible applications to cases of language documentation.

I shall begin from the assumption that quantitative methods of this kind are relatively unfamiliar; but colleagues wishing to explore issues or methods in advance of the workshop are invited to read the papers in special issue 103.2 of *Transactions of the Philological Society*, which includes contributions by a range of groups currently working in this area; or April McMahan and Robert McMahan (2005) *Language Classification by Numbers* (OUP), a general introduction to quantitative approaches and their applications.