Transactions of the Philological Society Volume 101:1 (2003) 7–55

FINDING FAMILIES: QUANTITATIVE METHODS IN LANGUAGE CLASSIFICATION[1]

By April McMahon & Robert McMahon
Department of English Language and Linguistics, University of Sheffield
(Received 10 April 2002)

Abstract

Over the past two decades, many of the major controversies in historical linguistics have centred on language classification. Some of these controversies have been concentrated within linguistics, as in the methodological opposition of multilateral comparison to the traditional Comparative Method. Others have crossed discipline boundaries, with the question of whether correlations can be established between language families, archaeological cultures and genetic populations. At the same time, increasing emphasis on language contact has challenged the family tree as a model of linguistic relatedness. This paper argues that we must quantify language classification, to allow objective evaluation of alternative methods within linguistics, and of proposed cross-disciplinary correlations; and that a first step in this quantification is represented by the 'borrowing' of computational tools from biology.

[1] The work reported here was carried out as part of the research project 'Quantitative Methods in Language Classification'. We gratefully acknowledge the financial support of the AHRB (grant AN6720/APN12536). We also thank the other members of the project team, Paul Heggarty and Natalia Slaska, for extremely helpful discussion and advice. Parts of this work have been presented in talks at the University of Newcastle and the University of Surrey, Roehampton, and at the Oxford-Kobe Seminar on Language Change and Historical Linguistics; many thanks also to those audiences for their friendly reception and very useful comments, and to the editors of this journal and two anonymous referees, for suggestions which have considerably improved the paper.

1. Genetics and linguistics: shared problems, different methods

An early 21st-century word-association game, asking for a response to 'family' or 'relatedness', would almost certainly produce 'genetics'. Genetics is regularly in the news, with recent stories debating whether families should be allowed to request pre-implantation diagnosis of embryos, both to guarantee that the resulting child will not suffer from an inherited familial condition, and more controversially, to provide some hope of therapy or cure for existing affected children. Genetics is also becoming popularised, both in print and increasingly in the other media: recent programmes in the BBC's 'Gene Stories' series showed intrepid geneticists travelling fearlessly all the way to the Orkneys and taking mouthwash samples from a whole range of baffled but compliant Orcadian males, to see whether any could claim Viking ancestry. Genetics is also, increasingly, becoming big business. Within the academic community, there is serious concern over the practice of companies in patenting elements of the human genome. Some commercial exploitation of recent advances in genetics is altogether less dangerous, though it does build on a very common interest in genealogy. For instance, analysis of mitochondrial DNA, which is passed on only through the maternal line, from mothers to daughters, and of genes on the Y-chromosome, which conversely is passed on only through the paternal line, from fathers to sons, allows us to claim that particular individuals are related, since they share a DNA marker which must indicate a common ancestor in the remote and distant past. Through oxfordancestors.com, Professor Bryan Sykes, of the University of Oxford, offers a service whereby two or more men sharing a surname can have their Y-chromosomes analysed to discover if they share an ancestor. A recent report (*The Observer*, 14 January 2001) provided an illustration of this method, in which Bryan Sykes chose to compare chromosomes with Sir Richard Sykes, the chief executive of Glaxo Wellcome, discovering that they do share a common ancestor. An updated version of the Victorian practice of 'claiming kin' with rich and influential people (as in Hardy's *Tess of the d'Urbervilles*) now suggests itself; and any reader called

Getty, Branson or Gates may certainly wish to investigate what that Oxford website has to offer.

All these applications and popularisations are possible because geneticists can already tell us a good deal about relatedness between human populations, and even human individuals. Advances in human genetics allow us to correlate patterns of genetic variation with particular culturally or nationally defined groups, and indeed, increasingly, work on DNA itself allows us to pinpoint characteristics of human families, far beyond our conventional use of the term. Consequently, geneticists are already engaging with the issue of whether two individuals, or two populations, are related or not. They reach their conclusions using highly sophisticated computer technology, and objective quantitative methodologies, which remove the burden of decision, in the bulk of cases, from the shoulders of the individual scientist. Work is objective, statistically testable, and follows generally agreed methods. There is, of course, discussion in the academic community when a new or improved method is proposed; but there is a typical cycle of publication and replication of results to be followed before that method is accepted. Indeed, one might suggest that the existence of common criteria for the judgement of methodological success or progress, and the consequent general acceptance of a range of methods, gives us the luxury of debating ethical considerations rather than querying methods and results.

Although geneticists and historical linguists are both dealing with systems which change through time, their methods of comparison and classification could not be more different. There has been a strong tendency to see comparative linguistics as an art rather than a science, and as requiring sensitivity and depth of knowledge of one particular language group on the part of the individual scholar, rather than generalisable techniques which allow the processing of large quantities of data, regardless of the region or family from which they come. The result is that major controversy remains over the classification of languages into families, and even over whether such classification is possible for languages lacking a long written history. In this paper, we shall argue both that quantitative methodologies must be developed and accepted in comparative linguistics if the discipline is to progress, and, more specifically,

that adopting some of the perspectives and even computational tools of genetics and population biology may take us further towards finding (and refuting) language families.

2. Languages, cognates and family trees

The comments above do not question the general acceptance in historical linguistics of the hypothesis that languages form families. Rather, they raise the issues of how family membership is to be demonstrated, and how that demonstration can be tested, so family groupings can ideally either be established beyond reasonable doubt, or refuted as lacking sufficient evidence. Comparative linguists might argue that there already exists a generally accepted method of indicating and assessing family relationships, namely the Comparative Method (see Campbell 1998, Fox 1995, Durie and Ross 1996), and there is certainly a commonly adopted outline methodology. Normal practice involves documenting similarities in sound and meaning between two or more languages, like those shown in (1).

(1)

English	Latin	German	
mouse	mūs	Maus	'mouse'
father	pater	Vater	'father'
fish	piscis	Fisch	'fish'

What matters most here is that the similarities are not isolated events, like the chance resemblance of *aska* in the Jaqaru language of Peru, and English *ask*. Any method of comparison, in genetics or linguistics, must include mechanisms for ruling out statistical noise of this sort; but chance is a weak hypothesis, and while it is perfectly adequate for resemblances involving individual words, it is not sufficient in cases like (1), where we find instead correspondences of sounds across the different languages. That is, initial /m/ in English will very regularly correspond to initial /m/ in both Latin and German, in words with the same meanings; while initial /f/ in English *father*, *fish* and German *Vater*, *Fisch* frequently corresponds to initial /p/ in Latin. It follows that Jaqaru *aska* and English *ask* are chance resemblances, whereas the words in (1) are cognates, each set being derived from a common ancestral form. On the basis of these

regular and repeated correspondences (and self-evidently, using many more languages and much more data), we would first establish the existence of a Germanic subfamily including German and English, and ultimately the Indo-European family to which Latin and many others also belong. For many historical linguists, the application of the Comparative Method also crucially includes the reconstruction of most likely ancestral forms for the intermediate nodes in the tree, such as Proto-Germanic, and finally for the hypothesised common ancestor, here Proto-Indo-European. These reconstructions also rely on a series of hypotheses within the Comparative Method, notably the regularity of sound change, since random and sporadic developments would not allow regular reversal to recover original forms. The possibility of reconstruction, and the acceptability of the proposed ancestral forms, is taken in itself to constitute a demonstration of the correctness of the comparison and hence of the proposed family.

Textbooks of historical linguistics typically include arguments of this kind, along with more detailed demonstrations of reconstruction, and illustrative family trees, generally for the Indo-European group. However, there are three problems with existing work within this paradigm of comparison and reconstruction; and if these do not entirely undermine our knowledge, at least of firmly established families like Indo-European, they nonetheless make that knowledge less generalisable and ultimately less useful.

First, consider the quote in (2), from Patrick Leigh Fermor's (1986: 33) account of part of a journey he made on foot in 1933, at the age of 18, across Europe to Constantinople. In this passage, he has arrived in Hungary, and is trying to learn the language.

(2) I knew that Magyar belonged to the Ugro-Finnic group, part of the great Ural-Altaic family, 'Just', one of my new friends told me, 'as English belongs to the Indo-European.' He followed this up by saying that the language closest to Hungarian was Finnish.
'How close?'
'Oh, very!'
'What, like Italian and Spanish?'
'Well no, not quite as close as that . . .'

> 'How close then?'
> Finally, after a thoughtful pause, he said, 'About like English and Persian.'

Clearly, both linguists and non-linguists have intuitions about which languages are more similar to which others. Even if we know very little about Persian, we are likely to imagine it, almost instinctively, as so different from English as hardly to be related at all – much more different, indeed, than Italian and Spanish, or English and French. Unfortunately, however, comparative linguists at present are unable to formalise those intuitions beyond the broadest type of subclassification: although we may feel reasonably confident that we can recognise when languages are related, we cannot objectively measure closeness of relationship. Languages do appear in family trees collected into subgroups (such as Germanic, Romance, Indo-Iranian and so on, within Indo-European); and these judgements are made on the basis of features which only the subgrouped languages have in common. At present, however, there is very little scope for quantification of the degree of relatedness.

The second, and even more important problem, involves borrowing and contact between languages, or more accurately, between speakers of different languages. Lexical borrowing is serious enough (and more so, as we shall see below, in that many classifications are based, at least initially, on the use of word lists); but contact can affect the grammar much more generally. As pidginisation, creolisation, convergence and language mixing have been studied more intensively (see Thomason and Kaufman 1988; Singh 2000; Thomason 1997, 2001; Bakker 1997, 2000; Bakker and Mous 1994; Matras 2000), it has become apparent that contact can create situations which are strictly incompatible with the traditional family tree, and therefore with methods typically used to build it (see also Slaska 2002). For one thing, if there is enough borrowing, and especially if it comes consistently from one language, and/or we lack a long written history for the languages concerned, we might erroneously hypothesise a genealogical relationship, and hence build the wrong tree. Inter-borrowing in related languages is even harder to detect, and may also tend to involve more basic vocabulary, which is typically seen as relatively resistant to borrowing (see

4.2–4.3 below); situations like this, which may have held between English and Norse after the Viking settlements, for instance, might similarly lead to the hypothesis of inaccurate subtrees.

The third and, for the moment, final problem is that the Comparative Method, the main means of discerning linguistic relatedness, is limited. A family like Indo-European provides enough languages and data, and a sufficiently long written record for many subgroups, to arrive at a generally agreed consensus; but all our difficulties multiply when we are dealing with larger, or less well attested families. Even for Indo-European, however, application of the Comparative Method is typically based on a linguist's individual knowledge of a language group. While ideally one might discover the relationships between subgroups, and the changes responsible, by applying the Comparative Method, linguists usually begin the process with a fairly clear idea of the groupings involved, and the changes instantiated in particular cases, and understandably apply the method so as to reconstruct just these changes and groups. In other words, the Comparative Method rests on case law. It is taught in connection with its application in particular circumstances, rather than as a neutral and generalisable method; and students' perceptions of it are therefore inextricably linked with its results for certain families, notably Indo-European. It is true that Hoenigswald (1960) attempts to reduce the Comparative Method to a series of simple algorithms, which are intended to be generally applicable; but even here, there is inevitably a concentration on particular examples from Indo-European, and Hoenigswald often explicitly discusses interpretations rather than applications. Although Hoenigswald's book is certainly the basis of some courses in the Comparative Method, particularly in the United States, its impact on the field has been limited: those seeing comparative linguistics more as an art tend to distrust methods which claim to be scientific (a term prominent in the blurb for Hoenigswald 1960), while those who really want to develop quantitative and potentially programmable methods find that Hoenigswald's account, though undoubtedly helpful in presenting some degree of generalisation, still does not fall into this category. A recent exchange on the HISTLING discussion list, initiated by a request for 'the steps that would be followed in a full implementation of comparative

methodology' (L. V. Hayes, 30 April 2002), included one response that 'I'd very much like the answers to be amenable to extraction by an explicit step-by-step methodology, but I don't think they always are' (Bob Rankin, 2 May 2002). Rankin also notes that 'I think it is a mistake to assume that the CM is a set of airtight procedures which, if followed faithfully, will produce the desired answers – genetic relationships will automatically emerge . . . ', and argues that this follows from the nature of the Comparative Method, which is essentially a heuristic, and hence irreducibly knowledge- and experience-based. As it stands, therefore, there can be no statistical testing of the Comparative Method; and if the method were in some way altered or refined so as to become amenable to such testing, it would necessarily change, in a way which not all its proponents would approve of.

Stating the problem crudely, then, there is no rulebook for the Comparative Method, and no consensus on how much phonetic similarity, of which particular types, is or is not 'valid' (always assuming we had an agreed mechanism for measuring similarity in the first place). The best method we currently have is therefore inevitably subject to interference from individual linguists' opinions, either consciously or subconsciously, so that it is not absolutely clear that we could guarantee getting the same results from the same data considered by different linguists, especially for older, more isolated, or less abundantly attested language groups. Indeed, even for well established families, a range of different reconstructions will typically exist at any point; take the case of the glottalic hypothesis for Proto-Indo-European, for instance, which fundamentally affects the nature of the PIE stop system and the changes hypothesised for daughter groups, depending on whether one accepts or rejects it (Gamkrelidze and Ivanov 1995). To add to the difficulty, many linguists argue, admittedly without much hard evidence (see Renfrew, McMahon and Trask 2000), that the Comparative Method will not work beyond perhaps 8 or 10,000 years.

There are, then, inherent difficulties with the most generally accepted method within comparative linguistics, at least as it is currently practised. In the rest of this paper, we shall discuss two rather general reasons why comparative linguistics must urgently adopt a quantitative approach: either to test and confirm existing

methods, rule out less robust ones or develop new approaches within linguistics; or to contribute to the formulation and assessment of correlations between language families and potentially parallel constructs in archaeology and genetics. We shall go on to report ongoing joint work, which shows in a very preliminary way what the outcome of such quantitative work might be, and to suggest several ways in which this pilot research might be extended.

3. Two arguments for quantitative methods in comparative linguistics

3.1. Testing and developing methods within comparative linguistics

A common reaction to the use of quantification in linguistics, or to arguments for such quantification, is encapsulated in Ross (1950: 59) (quoted in Embleton 1986: 25), who considers that:

> In comparison to old and established techniques, numerical methods must surely always be either inefficient or supererogatory. That is to say, on the one hand, if no solution to a problem of this kind can be reached with the old methods then I would not trust a numerical solution, and on the other hand, if a solution can be reached with the old methods, then a numerical solution is unnecessary.

On the other hand, an opposing view is provided by Kroeber and Chrétien (1937: 85), who believe that statistical analysis *can* contribute, in that it may '. . . validate and correct insight, or, where insight judgements are in conflict, help to decide between them. In short, it increases objectivity, sharpens findings, and sometimes forces new problems.' Our view, and our procedure as reported below, is much more in line with Kroeber and Chrétien's: the first requirement of quantitative work is to allow objective testing of existing results and methods, which is not always currently feasible in an independent way. As we have already seen, it is not viable simply to repeat the Comparative Method for the same data: even if the Method were not such a gradual, cumulative process, part of the problem, and the reason why linguists may feel the need for testing, is that there is a suspicion of the results being determined by the

linguist concerned. Embleton (1986: 22) notes that, 'Intentionally or unintentionally, IE historians may discuss only features which tend to reinforce their prior conclusions'; and while this need not mean that we mistrust the particular colleagues who may have worked seriously and cumulatively to produce results in a particular domain, it does entail that it is simply not possible to guarantee that the results are objective and repeatable, and indeed extendable to other language groups, without some kind of truly independent testing. We are also unable to demonstrate that the family tree produced for a particular putative group is significant among the mass of possible trees not considered. For precisely parallel reasons, these uncertainties would not be removed by simply asking another linguist to repeat the procedure.

However, once that testing and possible confirmation of existing methods is achieved, we may also, as Kroeber and Chrétien suggest, continue to propose other methods. We have no wish to argue against the use of rigorous Comparative-Method-based work where that is feasible; but we may have to accept that there are certain cases, where for instance there is no significant written history for a group of languages, and genealogical affiliations are particularly unclear, where the standard represented by the Comparative Method simply cannot be attained for essentially environmental reasons. Of course, a written history of any length is not essential to the application of the Comparative Method. However, it does impose limits on how far we can extend our hypotheses into prehistory; and working with present-day data only will be seriously problematic unless there is plenty of it. Where a number of the languages in a putative group have already been lost or where the data are limited by endangerment, for instance, the absence of a written record could be a serious matter. In such cases, we must either accept that we cannot be sure what the affiliations of those languages are or work towards a method which seems to give the right results for Indo-European and which can be generalised on the basis of purely present-day data to languages where we do not have the luxury of the kind of witnesses that exist to earlier stages within Indo-European. This might never give us the nuanced understanding of precise details of language-family history which we have in the case of Indo-European; but it might well allow us to approximate

that history, at least in terms of the outline of an appropriate family tree, better than other approaches. It is also likely to provide a framework for further, intensive research.

In turn, this means that we have to modify Ross's (1950) statement on both counts. If there is already an available solution using existing methods (in this case, the Comparative Method), then we can ensure that our computational methods approximate it; and where they do not, we can perhaps interrogate those results and hence improve the method. On the other hand, where there is no possible solution in terms of the old methods, this may not be a methodological concern, but an 'environmental', or external historical one. That is, there is no reason in principle why the Comparative Method should not work on groups of Amazonian languages, for example, except that they lack the written history and hence early evidence to feed into the model. The problem is not directly a challenge to the method, but an obstacle to that particular application. It is in precisely cases of this sort that a computational method using less, but relatively robust, data might allow us at least to make some headway in classification.

In other words, we are arguing here that comparative linguists need a more secure method for the sake of the discipline itself. We might want to see the traditional Comparative Method, with its painstaking analysis and reconstruction of individual words and sounds, as the Gold Standard for comparative and classificatory work, especially if it turns out that we can test the results of this method statistically; but we are never going to meet that standard in cases where languages are poorly attested (and recall the significant factor of endangerment for a considerable and increasing proportion of the world's languages), or where there is no written history enabling us to take surer steps into the past. Perhaps even more importantly, some exceptionally controversial historical linguistic methods have been proposed as alternatives to the Comparative Method. The most obvious case for discussion in this context is Greenberg's method of mass, or multilateral comparison, especially as used in his controversial tripartite classification of native American languages (Greenberg 1987). Although this is not the place for a detailed consideration of the controversy surrounding mass comparison, the main difficulty is the fact that the criteria for

allowing a phonetic or semantic match between languages are not made explicit (Campbell 1988; McMahon and McMahon 1995), so that it is even harder to see how this work could be repeated by other scholars than in the case of the Comparative Method. In consequence, mass comparison could produce a whole range of results, depending on the linguist's personal judgement. As Ringe (1999) has shown, a statistical evaluation of the method is therefore strictly impossible, since the number of possible matches is undeterminable; furthermore, adding more languages and more data, far from refining the results as appears to be true of the Comparative Method, simply increases the chances of a false match and hence of counterhistorical results. A method which claims success partly on the grounds of the amount of material included must be sure to step up the rigour of the analysis as more data are considered, and the complete absence of internal testing from mass comparison, beyond the control exerted by the individual scholar's knowledge and subjective judgement, shows that this is emphatically not the case here. The problem is that, at present, arguing for the Comparative Method over mass comparison comes down to the forceful articulation of preferences, and the resulting fight has as a consequence been rather a dirty one. If we are to assess the relative merits of the two methods, or indeed of any potential third method, we must have a neutral means of evaluation; and our argument is that this must necessarily involve a quantitative approach.

3.2. Linguistics, genetics and 'the new synthesis'

The second reason for encouraging quantitative approaches to comparative linguistics is that linguistics is not an isolated discipline; its results are relevant to colleagues working in other fields, and conversely, their results are relevant to linguists. Hopes for the so-called 'new synthesis' of disciplines are high: Cavalli-Sforza (2000: vii), for instance, introduces his recent book as follows:

> This book surveys the research on human evolution from the many different fields of study that contribute to our knowledge. It is a history of the last hundred thousand years, relying on archaeology, genetics, and linguistics. Happily, these three

> disciplines are now generating many new data and insights. All of them can be expected to converge toward a common story, and behind them must lie a single history. Singly, each approach has many lacunae, but hopefully their synthesis can help to fill the gaps.

The new synthesis, however, is not unproblematic – for a careful and critical review, see Sims-Williams (1998). Similarly, Renfrew (1999) is rather restrained in his assessment, expressing cautious hope for the future in the matter of disciplinary interinfluence when he notes that, 'We may be on the brink of seeing some convergence in our understanding of issues of genetic diversity, cultural diversity and linguistic diversity. It may be possible, then, to work toward a unified reconstruction of the history of human populations. It is much needed, because certainly we do not have such a unified history at the moment' (1999: 1–2).

Of course, arguing for such a synthesis does not suggest that the possession of particular genetic material makes us susceptible to learning and using a particular language (without prejudice to the proposal of a genetic component behind the predisposition of our species to learn and use language in general). On the other hand, since we are talking here about the histories of populations, which consist of people who both carry genes and use languages, it might be more surprising if there were no correlations between genetic and linguistic configurations. The observation of this correlation, like so many others, goes back to Darwin (1996 [1859]: 342), who suggested that, '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'. The norm today is to accept a slight tempering of this hypothesis, such that, 'The correlation between genes and languages cannot be perfect . . .', because both languages and genes can be replaced independently, but, 'Nevertheless . . . remains positive and statistically significant' (Cavalli-Sforza 2000: 167). This correlation is supported by a range of recent studies. For instance, Barbujani (1997: 1011) reports that, 'In Europe, for example, . . . several inheritable diseases differ, in their incidence, between geographically close but linguistically distant populations', while Poloni et al.

(1997) show that a group of individuals fell into four non-overlapping classes on the basis of their genetic characteristics and whether they spoke an Indo-European, Khoisan, Niger-Congo or Afro-Asiatic language. In other words, there is a general and telling statistical correlation between genetic and linguistic features, which reflects interesting and investigable parallelism rather than determinism. Genetic and linguistic commonality now therefore suggests ancestral identity at an earlier stage: as Barbujani (1997: 1014) observes, 'Population admixture and linguistic assimilation should have weakened the correspondence between patterns of genetic and linguistic diversity. The fact that such patterns are, on the contrary, well correlated at the allele-frequency level . . . suggests that parallel linguistic and allele-frequency change were not the exception, but the rule.'

If there is to be a new synthesis, with genetics and archaeology together telling unified population histories, then it is surely a matter of some concern to linguists that comparative linguistics should also be involved, and ideally as an equal partner. At present, however, one of the main obstacles to the development of the new synthesis lies precisely with language classification, which currently does not typically produce data which are interpretable and usable by neighbouring disciplines. Both archaeology and, to an even greater extent, genetics are quantitative in their approaches and methods, and in their evaluations of results; and if their practitioners are to understand and use historical linguistic data, linguists must therefore deal in probabilities and degrees of relatedness. At the moment, in not doing this, mainstream comparative linguistics is at a double disadvantage.

First, the most ambitious hypotheses of relatedness typically now come from the multilateral comparison camp (Greenberg 1987; Ruhlen 1991), and these are perhaps naturally the hypotheses that attract archaeologists and geneticists. If we wish to alert such colleagues to the unreliable nature of Greenberg's method, and to prevent them from gravitating towards such problematic results, we must be able to argue against those methods numerically. In failing to do this, linguists risk looking simply old-fashioned and unwilling to accept big ideas; and this is not a theoretical risk, but an actual one. Cavalli-Sforza (2000: 137–8), for instance, contends that, 'By

disallowing reliable measurements, and by limiting the relationship between two languages only to "related or not related", the American linguists opposing Greenberg have ruled out the possibility of hierarchical classifications, an essential prerequisite to taxonomy.' Archaeologists and geneticists will inevitably favour what they see as bold hypotheses, which offer the chance of reconstructing linguistic and population histories at greater time-depths, unless linguists can convince them that the bold hypotheses are not sound hypotheses, and do so in quantitative terms.

Second, we must face the prospect that if linguists do not attempt to provide quantification in our own terms, archaeologists and geneticists will increasingly be forced to supply their own. If these do not correspond to linguists' intuitions about degrees of relatedness among languages, we are condemned to fighting a perpetual rearguard action against these externally imposed figures; and if we still do not supply linguistically coherent alternatives, we are not likely to be taken seriously by scholars in other disciplines. Taking one recent example, Poloni and her colleagues (1997) assigned grades of relatedness of 0, 1, 2 and 3 to pairs of languages, depending on their proximity in terms of nesting within the family tree, with a score of 8 for pairs generally thought to be unrelated. Linguists might feel that these are extremely crude approximations to the real linguistic situation; but in that case, the onus is very firmly on us to provide a system we are content to work with.

4. Developing quantitative methods

4.1. An outline research strategy

Embleton (1986: 3) suggests that the research strategy which should be employed in developing and testing quantitative methodologies in linguistics is much the same as that generally found in applied mathematics; and the crucial point is that this is a three-stage process, as shown in (3).

(3) Embleton (1986: 3): steps in quantitative analysis:
   (i) to devise a procedure, based on theoretical grounds, on a particular model, or on past experience . . . ;

(ii) to verify the procedure by applying it to some data where there already exists a large body of linguistic opinion for comparison, often Indo-European data . . . This may lead to revision of the procedure of stage (i), or at the extreme to its total abandonment;
(iii) to apply the procedure to data where linguistic opinions have not yet been produced, have not yet been firmly established, or perhaps are even in conflict. In practice, this usually means application to non-Indo-European data . . .

In our approach to introducing quantitative methodology into linguistic classification, we accept this strategy. In the spirit of Embleton's three-way division, the first stages should therefore address the testing and automatisation of existing approaches to language comparison and classification. It follows that one important goal of the work reported here is to explore ways of confirming what we feel we already know, but by different, repeatable and statistically testable means. Consequently, most of the discussion below will focus on new approaches to two well worn tools of language classification, namely, word lists and family trees. In the sections below, we shall report on, first, the length of word list used, then the construction and testing of trees produced from these lists, and, finally, some issues involving the composition of the lists, and the possibility that different meanings are making different contributions to the eventual trees. All this work to date is limited to Embleton's step 2, since it involves reanalysis of relatively clear cases; as she notes, this typically means Indo-European. We shall, however, comment towards the end of the paper on our plans to take these analyses forward into step 3.

Family trees are particularly relevant here because of their very close association with the Comparative Method, though the realism of the tree model has been increasingly criticised as the attention paid to contact-induced linguistic change has grown. Furthermore, family trees are also extensively used in biology, at the species, population and individual levels, to represent common inheritance of entities and groups. However, as we have already seen, there is a major difference between tree-drawing methods in biology and

linguistics: while biologists regard classification as a quantitative discipline, and use computer programs to identify the best, or most parsimonious tree from the whole range of possible trees, linguists tend to work more intuitively, so that classifications are acceptable, broadly speaking, if our colleagues agree that they are acceptable. As we have already seen, this intuitive approach is inherently problematic, and we shall consider below the feasibility of adopting some tree-drawing and tree-selection programs from biology.

Word lists do not directly form a part of the Comparative Method, which operates primarily on the basis of regular correspondences of sound and meaning across the lexicon; but they are very commonly used as a precursor to this more detailed comparison, and are also the mainstay of lexicostatistics and glottochronology. However (and perhaps, for some, because of their association with glottochronology in particular), word lists seem to be regarded as something of a necessary evil even by those who use them. This is not the only case where linguists have sought to distance themselves from the use of a technique which is readily grasped by non-linguists: one parallel would involve the phoneme, which was effectively expelled from phonological theory following Chomsky and Halle (1968), but was and remains highly useful in, for instance, many discussions of L2 acquisition. Despite its rather marginal theoretical status, introductory textbooks and courses almost without exception include discussion of the phoneme, and it is still very typically used, with a token apology, by most phonological theorists, too.

We do not propose to enter into a detailed discussion of word lists (though see Kessler 2001), except to note that they have the very considerable advantage of being collectable even when available data is seriously incomplete, for whatever reason. This means that, whatever our attitude to comparison based on word lists, we should be willing to consider at least testing the approach. If we can provide a sufficiently rigorous and independent test, we can perhaps make a decision on the basis of that evidence as to whether word lists should be retained or rejected as a basis for language classification and subclassification. We can also use quantitative techniques to reveal and evaluate some assumptions which are inherent in many uses of these lists, though they are hardly ever specified and tested as actual

hypotheses. However, we will make one terminological change to common practice, in preferring *meaning lists* to *word lists*, at least in the discussion of our own research: what stays constant across the languages compared is the meaning for a particular slot in the list, and this is more transparently reflected in the revised label. None of this addresses the entirely reasonable objection that comparing word lists is not the same as comparing languages, and that a classification based on word lists may not, for any number of reasons, yield the same tree as a wider-ranging and more extensive comparison including data from other levels of the grammar. We are currently pursuing additional methods which involve phonetic and morphosyntactic comparison for precisely this reason (Heggarty 2000a, b; Heggarty and McMahon 2002); and there has been interest in such methods outside of vocabulary for some time (Allen 1953; Grimes and Agard 1959; Nichols 1992). For a quantitative approach to a mixed data set, involving a character-based methodology and based on the concept of the 'perfect phylogeny', see Ringe, Warnow and Taylor (2002). However, there are equally valid reasons for not writing off lexical comparisons altogether, especially before they have been adequately tested. Ultimately, a comparison of languages which excluded lexical information might be as likely to provide an erroneous classification as one including only word lists: an integration of all factors is the ultimate goal, and this necessarily means testing and exploring each separate tool individually.

4.2. Length of meaning lists

The first important issue involves the optimal length for a meaning list. Embleton (1986: 89–93) compares results for simulated data using 100-, 200- and 500-meaning lists. Embleton modelled the sequential bifurcation of an original 'language' into ten daughters, allowing borrowing between neighbours, only, with continual turnover of items in the list in all the languages, at a constant rate of change. At the end point of each simulation, she attempted to reconstruct the appropriate tree on the basis of the three different list lengths. Embleton reports that the general pattern of results is the same for all three lists, but also that accuracy improves with list

length. However, the magnitude of the difference in resolution is far greater for a 200-word over a 100-word list than it is for the comparison of 500- with 200-word lists. That is, 'The improvement is sufficiently small . . . that, when faced with the practical problems of real language data, it is doubtful whether the researcher will find it worth the time and trouble to work with a 500-word list in preference to a 200-word list' (Embleton 1986: 92). On the other hand, in comparison with a 200-word list, 'accuracy . . . is considerably decreased by using a 100-word list. Hence the conscientious researcher must prefer a 200-word list over a 100-word list, even for the construction of provisional family trees. This is an important result, as both 200- and 100-word lists are presently in common use' (Embleton 1986: 92–3).

This conclusion, of course, is predicated on the unstated (and strictly invalid, as we shall see below) premises that lexical replacement rate is the same for each word, but also that the history of each word is independent of all other words, so that a change in one individual meaning will not entail a change in any other particular meaning. In other words, these results will only hold if each individual meaning in each list contributes equally to the results obtained. If some meanings turn out to be less stable than others, then increasing the length of the list could increase or decrease the number of meanings in that category, and therefore significantly skew the results. However, we can guard against this possibility only if we accept that such differences in rate of change exist, and establish which meanings are more or less retentive. We return to this problem of variability within the meaning list in 4.5, but should for the moment assume that the 200-meaning list is preferable as the best compromise between practicality and accuracy.

4.3. Composition of meaning lists

Recall that Embleton's three stages (see (3) above) always involve progressing from the known to the unknown, and from cases where the results are already fairly well established, to cases which are less clear. As we have already noted, it was therefore inevitable that we should have begun work on Indo-European languages. We have used the largest single set of 200-meaning lists for Indo-European

languages we could identify, namely, that of Dyen, Kruskal and Black (1992), who provide 200-meaning lists for 84 Indo-European languages and dialects, with a rather larger complement of 95 lists available at http://www.ldc.upenn.edu. Dyen, Kruskal and Black use a slightly modified version of Swadesh's (1952) 200-meaning list, which itself was intended as a culture-neutral collection of basic meanings; these in turn are hypothesised to be relatively resistant to borrowing and to loss. Note also that all the varieties included are modern ones: Dyen, Kruskal and Black (1992) themselves note that the inclusion of earlier stages or ancestral languages like Latin might have resolved some ambiguities, but we have chosen not to include these, since the use of present-day languages, only, more accurately reflects the situation we would encounter, perforce, in extending these methods beyond Indo-European. Dyen, Kruskal and Black's (1992) modified list appears in (4).

(4) Dyen, Kruskal and Black's (1992) modified 200-word list

all	and	animal	ashes
at	back	bad	bark
because	belly	big	bird
black	blood	bone	child
cloud	cold	day	dirty
dog	dry	dull	dust
ear	earth	egg	eye
far	fat	father	feather
few	fire	fish	five
flower	fog	foot	four
fruit	good	grass	green
guts	hair	hand	he
head	heart	heavy	here
hold	how	husband	I
ice	if	in	know
lake	leaf	left (hand)	leg
liver	long	louse	man
many	meat	mother	mountain
mouth	name	narrow	near
neck	new	night	nose
not	old	one	other

person	red	right	right (hand)
river	road	root	rope
rotten	rub	salt	sand
scratch	sea	seed	sharp
short	skin	sky	small
smoke	smooth	snake	snow
some	star	stick	stone
straight	sun	tail	that
there	they	thick	thin
this	thou	three	to bite
to blow	to breathe	to burn	to come
to count	to cut	to die	to dig
to drink	to eat	to fall	to fear
to fight	to float	to flow	to fly
to freeze	to give	to hear	to hit
to hunt	to kill	to laugh	to lie
to live	to play	to pull	to push
to rain	to say	to see	to sew
to sing	to sit	to sleep	to smell
to spit	to split	to squeeze	to stab
to stand	to suck	to swell	to swim
to think	to throw	to tie	to turn
to vomit	to walk	to wash	tongue
tooth	tree	two	warm
water	we	wet	what
when	where	white	who
wide	wife	wind	wing
wipe	with	woman	woods
worm	ye	year	yellow

Meaning-list comparisons operate on the basis of cognacy judgements: the meaning list is translated into the two languages to be compared, and a decision is reached as to whether the resulting forms, for each slot in the list, are cognate or not. As we shall see below, this mechanism is intended to filter out borrowings and chance resemblances, and to identify retentions from the common ancestor. The disadvantage of using an existing, coded set of lists like that in Dyen, Kruskal and Black (1992) is that this involves

dependence on pre-existing cognacy judgements; and these are vital, since we are working in such comparisons with a very restricted set of possibilities, namely, that two forms either are cognate or not. However, the advantages of using this set were that the material is in the public domain, so that independent evaluation of the judgements is possible – the raw data are available on the web at http://www.ldc.upenn.edu. Although we might have succeeded in producing 200-meaning lists for a larger number of languages by combining material from a range of sources, this would have raised the additional difficulty of controlling across a range of linguists for cognacy judgements, and also for the actual composition of the lists; although lists of the same length in the literature are broadly similar, they are not always absolutely identical (and we have already seen that Dyen, Kruskal and Black's list is a modified version of Swadesh's). In any case, the 95 lists provided by Dyen, Kruskal and Black are, we would argue, quite extensive enough to allow a reasoned evaluation of the meaning-list approach. Later in the project, we propose to collect a range of our own lists. These lists will be somewhat modified, for reasons given below; but more generally, we would not be following the strategy in (3) if we embarked on this kind of data collection without first establishing whether working with meaning lists was worthwhile.

4.4. Drawing and selecting trees

Although Dyen, Kruskal and Black (1992) provide a great deal of useful raw material in the form of these lists, they do not generate trees, but a highly complex series of nested box diagrams. Percentages of cognate material are given for each language pair; and boxes corresponding to conventional Indo-European subfamilies are drawn round those sets with the highest cognacy percentages for the same meanings. Dyen, Kruskal and Black also provide a distance matrix on their website: this again is a pairwise comparison, giving a percentage cognacy score for each language pair, expressed as a measure of difference between them. We have used this distance matrix as the basis for our further analysis.

The next step was actually to draw trees using these data. Although historical linguists have been drawing family trees since

at least the late nineteenth century, and may in fact have predated the biological use of trees (Koerner 1983), there is now a fundamental difference between the way the tree is approached in comparative linguistics and in comparative biology. Linguists tend to use the data to draw *the* tree which fits the pattern, whether the data are derived from cognacy scores alone, or from more extensive comparison and reconstruction; there is no agreed mechanism, or series of steps in tree drawing, and any argument in the discipline centres on the tree itself, not the method by which it was constructed. On the other hand, tree drawing in biology is seen as a quantitative discipline: there are books devoted to different types of tree and different ways of arriving at them (see for instance Page and Holmes 1998); and in addition, and more relevantly from the present point of view, there are computer programs which draw and select the most parsimonious tree. In consequence, there is an accepted means of testing which tree is the best, and of evaluating the differences between trees.

As noted above, we took as our starting point the Dyen, Kruskal and Black (1992) distance matrix, which is based on the percentages of non-cognate forms between each pair of languages. This was converted into a tree, using three different programs from the PHYLIP package (Felsenstein 2001), a suite of programs developed for the reconstruction of phylogenies, or evolutionary histories, in biology. It is, of course, entirely irrelevant to these programs whether they are dealing with frequencies of alleles for particular genes or percentage cognacy scores in word lists; and it matters even less, from a methodological perspective, whether the final trees are labelled with species or language names. We have therefore simply treated our data as if they were analogous to genetic information. The great advantage of this approach is that the PHYLIP programs are not simply drawing trees: we are not replacing one linguist and a pencil with a computational equivalent. These programs are selecting from the population of possible trees for a given data set, evaluating either all of these or a selection of those trees which prove most consistent with the data, depending on the complexity of the operation and the amount of data involved. That is, the PHYLIP programs generate all or many of the possible trees,

and then select from this set the tree where branch lengths and order of branching are most consistent with the distances in the data matrix. It is not possible using this method to privilege particular data points which are 'known' to be especially salient in subgrouping within a family, or to accord extra weight to an ostensive similarity across family lines: the entire process is automatised, and the result can be evaluated statistically, as well as on the basis of the final diagram achieved.

The three PHYLIP programs used (namely, Neighbour, Fitch and Kitch) approach the problem of selecting trees in three rather different ways. Given a population of 95 languages, the matrix of percentage distances is fairly significant in size, and there are consequently very many trees which could fit the data, and which must therefore be evaluated by the programs. The Neighbour-Joining approach taken in the Neighbour program most accurately reflects the route which would be taken by a linguist drawing a single tree, since each step involves clustering the closest two languages (in terms of percentage similarity), then adding the next closest, and so on up the tree. This is a fairly crude method computationally, and may not give absolutely the best tree, but has the advantage of using rather simple computational operations, so that the procedure takes less than ten minutes to run for the Dyen, Kruskal and Black data on a 700 MHz PC. On the other hand, the Maximum-Likelihood approach of the Fitch and Kitch programs attempts to minimise the differences between the branch lengths in the tree and the distances in the matrix; it also allows the entire tree to be globally rearranged after each addition, rather than regarding the previously drawn branchings as sacrosanct, as Neighbour Joining does. This means the population of possible trees considered is substantially larger under the Maximum-Likelihood approach; and although the latter is therefore likely to give more accurate results (because in considering a larger population of trees, it is less likely to miss the true tree), these come at the cost of greater complexity in computing, with a run taking between three and six hours. For completeness, note that within the Maximum-Likelihood approach, Kitch differs from Fitch in assuming a constant rate of change throughout the tree, while Fitch allows for different rates down each individual branch, meaning that the relative lengths of branches in Fitch trees

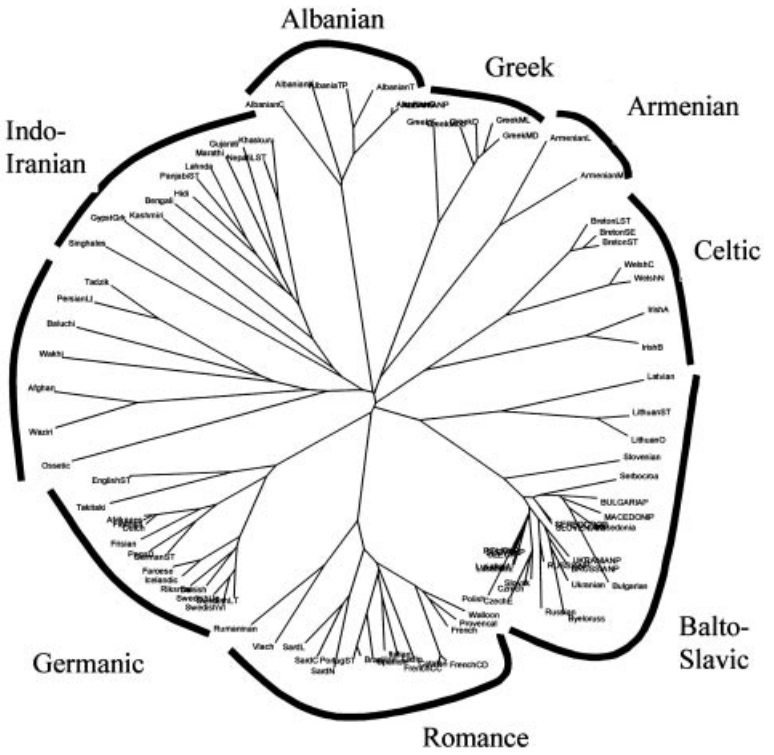

Figure 1. Radial unrooted tree based on the output from the Neighbour program. The tree was redrawn by the program TreeView (ver. 1.6.6) (Page 1996).

show cases where languages have changed more or less since the common ancestor.

Trees generated using these three approaches in fact show strikingly similar patterns. Before considering these subgroupings, however, there are several points to note which refer to both Neighbour-Joining and Maximum-Likelihood trees. First, the program produces diagrams which look rather different from conventional family trees, because they are unrooted. These diagrams look like stars, with each group identifiable because

the members appear together, and branch off together from the centre. This subgrouping is emphasised in the tree in Fig. 1 by labelling each arc separately.

These unrooted trees reflect the acceptance in biology of the hypothesis that all species ultimately derive from a single common ancestor: that is, monogenesis in terms of species is the usual assumption. Of course, using these diagrams in linguistics does not necessitate an acceptance of monogenesis or the assumption that all extant languages share a single common ancestor: preliminary work reported in McMahon, Lohr and McMahon (1999) indicates that when non-Indo-European languages are included, the whole Indo-European group forms a separate group branching away from the centre of the diagram, and that within this larger group, subgroups are visible as before. However, these star diagrams can also be converted into a more conventional family tree, and indeed this is very typically the procedure followed in biology, where any species can be selected as the root. Precisely the same conversion can be implemented linguistically, by selecting one language as the root, as illustrated in Fig. 2.

Fig. 2 shows Albanian as the outgroup used to root the tree. This does not imply that Albanian is nearer the common ancestor: any language could form the root in a diagram of this sort, but we have chosen Albanian for convenience, since it is not usually taken as part of a larger subgroup within Indo-European, so that selecting Albanian does not disrupt the form of any other subgroup. The process of rooting the tree simply produces an alternative, and more familiar visual representation of precisely the same set of relationships. The process is exactly analogous to that involved in converting a mobile from a starting state spread out on the floor, as with the star diagram, to hanging from the ceiling. Picking up any one of the 'leaves' of the mobile allows the other groups to hang down in their appropriate groups, and gives the appearance of a conventional family tree.

The important issue here is that these trees, and their Maximum-Likelihood equivalents, identify the subgroups which would typically be proposed for Indo-European. There were, in fact, a maximum 44 differences between the Neighbour-Joining and Maximum-Likelihood Kitch trees, from a total of 374 possible

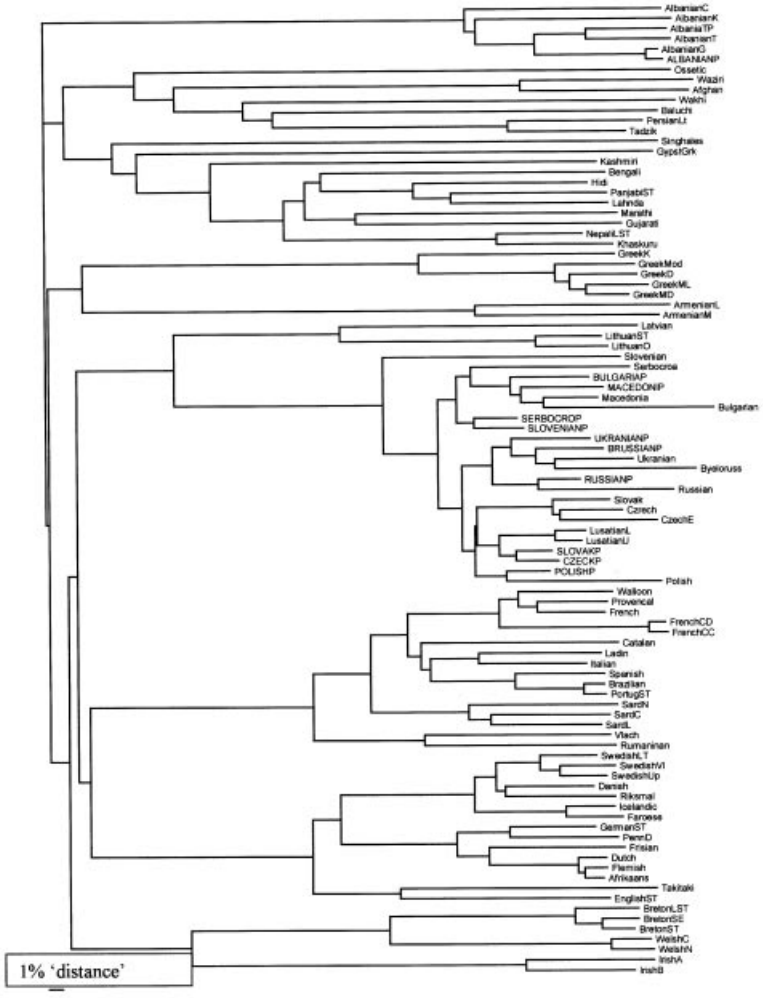

Figure 2. Pseudo-rooted tree, setting the Albanian group as an outgroup to root the rest of the tree. Based on a Neighbour-Joining tree calculated by the program Neighbour (Felsenstein 2001). Output tree redrawn in TreeView (Page 1996) as a rooted phylogram.

Figure 3. Radial unrooted tree based on a single output from the Fitch program (Felsenstein 2001), redrawn in TreeView (ver. 1.6.6) (Page 1996).

differences, meaning that, although there are some minor discrepancies (mainly concentrated within a specific Slavic subgroup), the trees are highly congruent, as can be seen from Fig. 3 and 4, the Maximum-Likelihood equivalent of Fig. 1 and 2, respectively. We shall therefore concentrate in the discussion below on the Neighbour-Joining trees, only.

The next step is to consider how good this tree really is: although a tree may be the best found by the program on a particular run, the program may be testing several hundred thousand possible trees, within which the data may support five hundred different trees approximately equally well, so that this one has potentially been

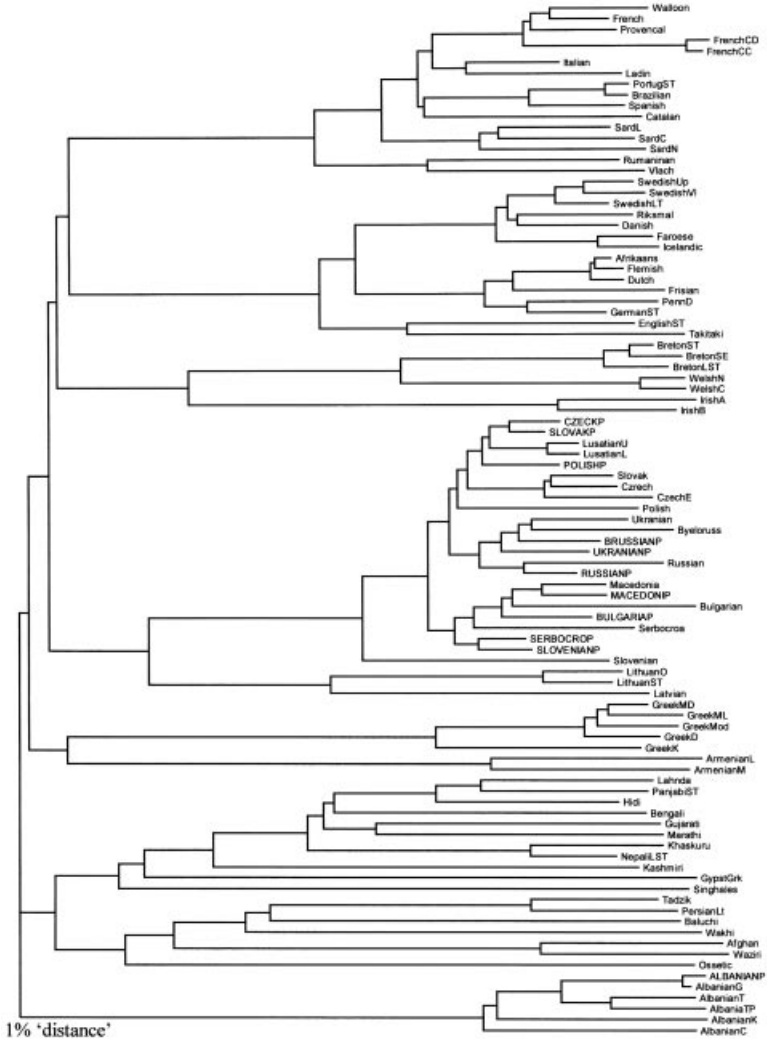

Figure 4. Same tree as in 3, redrawn with Albanian as an outgroup; comparison with Fig. 2 shows very few differences.

chosen essentially at random from within that class. There is also the (admittedly unlikely) possibility that the program may not have found any very good trees, so that the trees in Fig. 1–4 above might be the best only of that rather inadequate sample, rather than absolutely the optimal trees. To produce objective judgements here, it is necessary to test statistically how much confidence we can have in any particular branching, and therefore cumulatively in the tree as a whole. The main statistical method here is bootstrapping, which essentially means resampling the data sequentially to test the robustness of any given part of the tree. Strictly, this resampling should also involve replacement: 5% of the data (= ten meanings, for a 200-meaning list) would be removed at random, then those empty spaces are refilled by randomly choosing ten meanings from the same data set (this means that the same ten meanings could conceivably be replaced, or ten times the same meaning included, although both those possibilities are rather remote). Using these altered lists, the tree-finding algorithm is run again; and this procedure is repeated, perhaps 100 or 1000 times depending on the projected size of the final confidence intervals. This mechanism is rather time consuming, because ten meanings must be removed by hand each time, then the data are resampled, and then rerun; in the case of Maximum-Likelihood runs, this involves essentially one full day's work per iteration, which may explain why our bootstrapping results currently apply only to the Neighbour-Joining method. The consensus tree for bootstrap iterations performed to date is given in Fig. 5, and shows very strong congruence with the Neighbour-Joining and Maximum-Likelihood diagrams from Fig. 2 and 4 above.

The number at each node shows how many of the ten bootstrap iterations performed to date support that particular bifurcation. It is obvious that the subfamily branches are extremely robust, with Romance, Germanic and the other subgroups being found absolutely routinely. However, in different iterations, a small number of languages below the subfamily level do seem to have a less stable position: the Slavic languages are particularly prone to changing places. Nonetheless, these languages do consistently stay within their subgroup. Our working hypothesis is that these

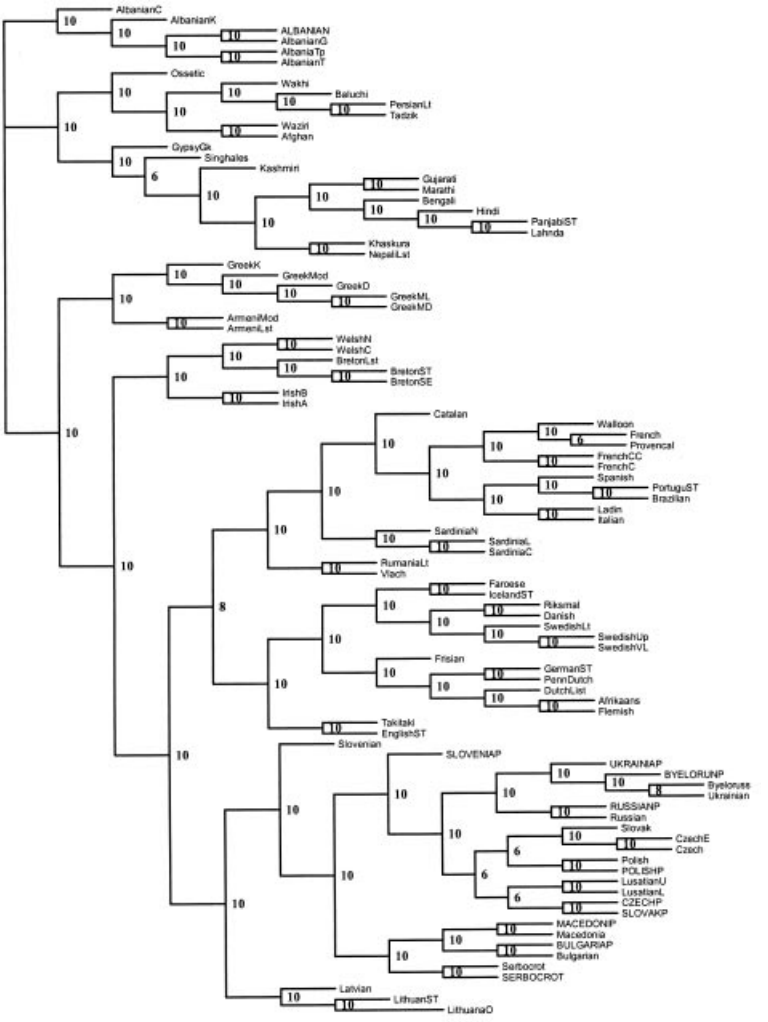

Figure 5. Consensus tree of ten bootstrap iterations using Neighbour (Felsenstein 2001). Numbers at nodes represent the number of iterations (out of ten) in which the branch to the immediate left of the number is supported: so, French, Provençal and Walloon form a group in all ten iterations, but French and Provençal form a group within this in only six iterations.

relatively unstable cases are also those which are not consistent as between Neighbour-Joining and Maximum-Likelihood trees, and that the ultimate cause of this instability is either borrowing from geographically and genetically related languages here, or alterations to the rate of change, for reasons other than contact. On the other hand, the difficulty in the case of these Slavic languages may be altogether more mundane and less generally interesting: Dyen, Kruskal and Black include the relevant lists on their website, but exclude them from their (1992) monograph, noting that, 'Eleven Slavic lists taken from Fodor (1961) were ultimately excluded because they seemed as a group to have higher percentages within Slavic than Slavic lists from other sources do. Apparently his lists have a tendency to favor cognate Slavic forms' (1992: 19). We return to the vexed question of borrowing in the following section.

4.5. Composition of meaning lists

The results from the previous two sections can be summarised in two main points. First, the subfamily structure of Indo-European, in terms of the conventional daughter groups like Romance, Germanic, Slavic, Celtic and so on, emerges extremely robustly from the various PHYLIP programs of Felsenstein (2001), whether these use a Neighbour-Joining or a Maximum-Likelihood approach to the computation. Second, those subgroups also emerge robustly when the data are subjected to bootstrapping by resampling, though this confirmatory work is still proceeding. It is true that none of this really tells us anything new; but arguably, this should be seen as a positive fact rather than a negative one, since we would hope not to find serious deviations from generally accepted results when working with a family which has been as carefully researched as Indo-European, using traditional methods which are generally regarded as reliable and sound. Nonetheless, there remains a potential concern, in that the trees drawn and selected in these experiments, however robust they may be, depend absolutely on the cognacy judgements and the meaning-list composition which underlie them. That is, it is hardly surprising that the tree-drawing algorithms 'find'

the conventional Indo-European subgroups, since the cognacy judgements on which the distance matrices are based have essentially built in a predisposition to producing exactly those groupings. Those judgements of whether particular forms are cognate or not, and therefore reflect common ancestry or not, are independent of the construction and testing of the trees; but this is not necessarily a point in our favour, since we are simply accepting existing judgements, which by their nature are subjective and based on the individual linguist's knowledge and experience of certain languages – the very antithesis of the characteristics we wish to claim for our objective, neutral and quantifiable methods.

There are two points to be made here. First, the Comparative Method by its nature is experience-based and subjective to some degree, and if we are to retain it, as we may very well wish to given its positive results to date, then we may simply have to work with that. What we can provide is a set of tests which can be applied after the fact to test the robustness of the classifications arrived at on that basis. An appropriate analogy here might involve medicine: a typical General Practitioner will become experienced through having seen multitudes of patients, and will achieve diagnoses, not by using medical textbooks and journal articles, but largely by intuition, based on that accumulated experience. However, when a GP arrives at a diagnosis, she will wish to send the patient to hospital for appropriate tests, which will provide a definite answer – the scientific confirmation of the initial, essentially arts-type evaluation.[2] Of course, the hospital staff would not be doing those particular tests if the GP had not asked them to; so both are, in opposite proportions, relying on a combination of scientific knowledge, and past experience, with the more subjective elements less able to interfere in the later stages of the process. In exactly the same way, we hope to provide a means of confirming the more subjective initial hypotheses of relatedness which arise from the Comparative Method or from the use of meaning lists, or of identifying areas within the data where that subjectivity may have got out of hand, by accident or design, and which may require re-evaluation. On the

[2] Thanks to Ricardo Bermúdez-Otero for suggesting this analogy.

other hand, it may be that the programs we are using, even when run on such intuitive, judgement-based and potentially degenerate data, can identify interesting patterns which we might not otherwise have been able to discern, perhaps thereby speaking in favour of both the initial judgements and our computational tools. One illustration of this type arises in connection with the composition of meaning lists.

We have already seen that there are reasons for choosing two hundred meanings rather than more or fewer, but of course there are infinitely many possible 200-meaning lists, and we must be prepared to consider the composition of our particular list in detail. Our calculations immediately show that not all the two hundred meanings are behaving in exactly the same way. That fact emerges incontrovertibly from the bootstrapping analyses; although those are still preliminary, they do give slightly different results for particular languages depending on the meanings which are excluded, or sampled out, on each run. If each meaning were contributing absolutely equally and identically to the results, all the runs would have precisely the same outcome; and this would indicate that any random set of the two hundred meanings could be omitted, so long as sufficient remained to keep the analysis statistically robust. This would be an important result because one very well known objection to the use of meaning lists for glottochronology lies in the difficulty, and perhaps sheer impossibility in some cases, of translating all the items on the list into certain other languages. Swadesh chose the items on the 100- and 200-meaning lists for their relative cultural neutrality; but there are still well reported problems with some meanings, which also differ depending on the target language (see Hoijer 1956 for Navajo).

As we shall see below, further investigations show that not all meanings are changing at the same rate; though this is an equally important and interesting result, since identifying meanings which are contributing to the overall analysis in divergent ways should allow us to predict the impact omitting particular meanings is likely to have, where these cannot be translated into a particular language. The discovery that meanings change at different rates has already been illustrated in a preliminary way by, for instance, Kruskal, Dyen and Black (1971) and Pagel (2000); but what has

not been demonstrated so far, though it has surely been suspected (see for example Clackson 1994: 26ff), is that these differential rates of change might have an impact on the shape of the resulting trees.

However, although bootstrapping certainly provides different outcomes depending on the meanings omitted, this could still be a random effect, spread across the list in such a way that we might reasonably expect the minor changes to cancel one another out. More importantly, there would then be very little action we could reasonably take to guard against such a contribution of different meanings: any items omitted would have an effect on the recovered tree, but we might expect the magnitude of that change to be relatively minor, so long as the number of meanings omitted was quite small. On the other hand, the whole problem becomes potentially much more serious if it turns out that there is a disproportionate effect on the outcome of omitting a particular class of meanings, or some members of that class. In other words, a random effect would not be correctable, but would not be very important, either; one might even suggest including a minor numerical correction to adjust for this factor, in cases where the absence of meanings was known. However, a systematic effect, involving particular meanings, is both more serious, but also potentially resolvable, if we can pinpoint exactly which are the meanings causing the problem.

In assessing the differential contribution of particular meanings, we draw on Lohr (1999). Lohr's thesis surveyed different methods in comparative linguistics, and she was particularly concerned with the question of how to define the 'best' meaning lists in an objective way. She developed two scales on which meanings can vary: these rely on the relative reconstructability and the retentiveness of certain meanings.

First, Lohr considered four reconstructed protolanguages for four different families: Proto-Indo-European (Buck 1949), Proto-Afroasiatic (Ehret 1995), Proto-Austronesian (Zorc 1995) and Proto-Sino-Tibetan (Luce 1981). She collected lists of meanings which could be reconstructed for two, three and four of these protolanguages; the argument is that 'such meanings are likely to be relatively basic, universal and stable, since they reflect cultures of

several millennia ago, cross at least two cultures, and were able to be reconstructed from descendant languages' (Lohr 1999: 54). Lohr found 61 meanings which were reconstructible for all four protolanguages, 196 meanings shared by three protolanguages and 281 meanings shared by two protolanguages.

Second, Lohr estimated the retentiveness of a set of these meanings, for Indo-European, only. She traced the histories of a range of meanings in Buck's dictionary, with the addition of some particles, numerals and pronouns, for the language groupings and time periods in (5).

(5) Lohr (1999): time periods for calculating rates of replacement:
Proto-Indo-European to Classical Greek, Sanskrit and Latin (approximately 2.5 millennia in each case)
PIE to Proto-Germanic (2.7)
PIE to Old Church Slavonic (3.7)
Classical Greek to Modern Greek (2.5)
Proto-Germanic to Proto-West-Germanic (0.7)
Vulgar Latin to French, Italian and Rumanian (1.8 in each case)
Proto-Germanic to Danish (2.3)
Proto-West-Germanic to English and German (1.6 in each case)
Old Church Slavonic to Serbo-Croat and Russian (1.3 in each case).

Essentially, Lohr was interested in the number of replacements which each meaning underwent during each time period, that is, the number of times a different form is documented with that same meaning. The number of replacements per meaning was calculated for each of these time periods; Lohr then simply totalled the number of replacement events per meaning for the combined time period of 31 millennia. Consequently, the replacement rate is expressed as the number of millennia one might expect to wait for a replacement event for that meaning; these rates are shown in (6) below. Note that, although one might argue over the time periods which Lohr allocates to the PIE to Latin, Greek and Sanskrit intervals, shortening or extending these specific periods would not affect the relative results. Since PIE is reconstructed, rather few replacements can be

projected for these periods, since evidence is lacking; and consequently, although changing the projected date for PIE would alter the overall number of millennia in the calculations, it would not affect the number of replacements or the relative ranking of rates for different meanings. The idea of more and less retentive meanings is not a new one, and Starostin (2000), for instance, discusses a 55-word list of highly persistent meanings from the Swadesh 100- and 200-word lists, though Starostin and Lohr have constructed their lists by different means.

(6) Lohr (1999): average Indo-European replacement rates per meaning

Visible replacements over time period	Replacement rate in millennia	Number of meanings
0	infinite retentiveness	17
1	31.3	40
2	15.7	20
3	10.4	25
4	7.8	37
5	6.3	44
6	5.2	53
7	4.5	56
8 or more	3.9 or lower	138

We selected two extreme samples from the Dyen, Kruskal and Black (1992) database, for reasons of comparability, to assess whether there would be a difference in the trees generated using these particular groups of meanings. We therefore chose thirty meanings which scored as high as possible on Lohr's indices of reconstructability and retentiveness: all were reconstructable for at least three protolanguages, and had no more than three replacements (indeed, 23 of these forms had only one replacement or none at all). This high-reconstructability–high-retentiveness class is shown in (7). Conversely, we selected 23 meanings which were reconstructable for only two protolanguages, and which had eight or more visible replacements in the 31.3-millennium total sample: this class is given in (8).

(7) High-reconstructability–high-retentiveness

four	name	three	two
foot	to give	long	salt
sun	other	to sleep	to come
day	to eat	not	thin
five	I	ear	mother
new	night	one	to spit
to stand	star	thou	tongue
tooth	wind		

(8) Low-reconstructability–low-retentiveness

grass	mouth	stone	heavy
year	bird	near	smooth
wing	man	neck	tail
to walk	back	to flow	left (hand)
to pull	to push	river	rope
straight	to think	to throw	

The numbers of meanings in these two opposing categories are different because it turned out that six of the meanings in the high-reconstructability–high-retentiveness class were totally uninformative, being cognate for all subgroups of Indo-European. There are some few cases where these meanings are omitted for individual languages in Dyen, Kruskal and Black (1992), reflecting language-specific changes affecting those items; but none of these changes is shared by a whole subfamily, and there is therefore no salient information for tree-drawing purposes, since a unique innovation (or loss, for that matter) affecting only a single language is consistent with any possible tree structure.

One issue emerging from this comparison of classes differentiated for retentiveness and reconstructability involves the Germanic group. When only the low-reconstructability–low-retentiveness meanings are used, as shown in Fig. 6, Frisian appears as a sister of a group containing Afrikaans, Flemish and Dutch (as it does, indeed, in the tree selected on the basis of the whole 200-meaning list).

However, when the high-reconstructability–high-retentiveness sublist is used, as in Fig. 7, Frisian is related to these languages only at a deeper level, and the tree indicates an earlier split of Frisian as against the rest of the West-Germanic group.

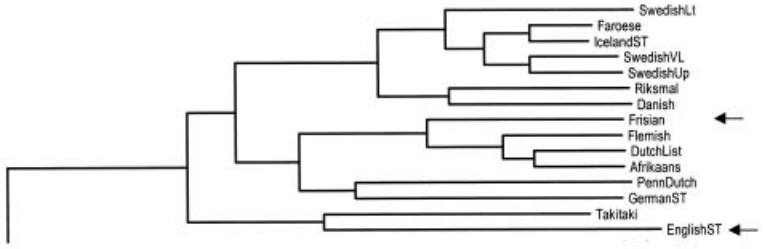

Figure 6. Close-up of the Germanic branch from the Neighbour-Joining tree based on the least retentive meanings.

A similar shift in the position of English can be observed from these Germanic subtrees. In the full 200-word-list tree, English appears (along with Takitaki, an English-based creole) as a relatively deep, distant sister of the whole Germanic group; these results are mirrored in the tree in Fig. 6 for the low-reconstructability–low-retentiveness meanings. However, when we work with the high-reconstructability–high-retentiveness meanings, as in Fig. 7, we find that English migrates into the West-Germanic subgroup, with close affinities to both German, and the Flemish–Afrikaans–Dutch cluster.

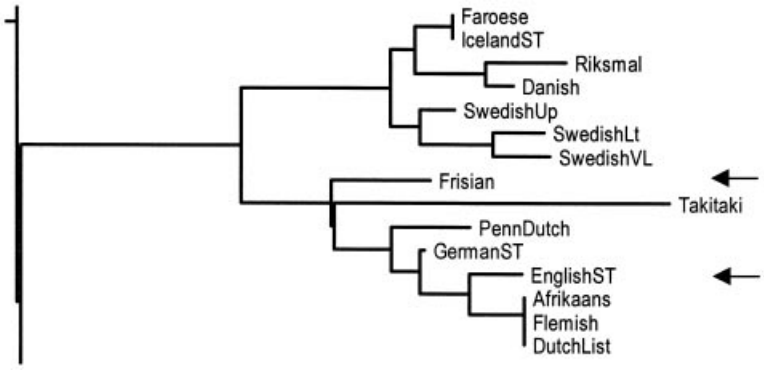

Figure 7. Close-up of the Germanic branch from the Neighbour-Joining tree based on the most retentive meanings. Compare the positions of Frisian and English in this tree with Fig. 6.

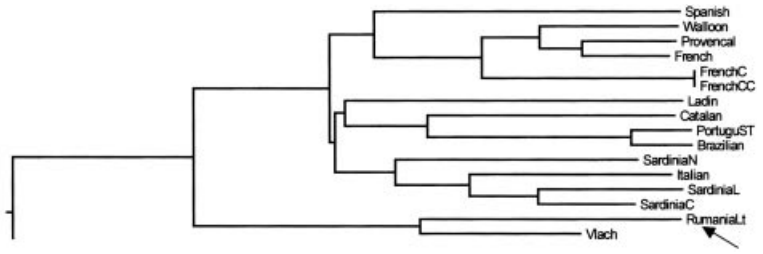

Figure 8. Close-up of the Romance branch from the Neighbour-Joining tree based on the least retentive meanings.

Turning to a different subfamily, we find similar results for Rumanian. When we use only the low-reconstructability–low-retentiveness meanings, as in Fig. 8, Rumanian is marginal to the Romance group.

On the other hand, in Fig. 9, calculated with the high-reconstructability–high-retentiveness meanings, Rumanian is much more integrated within Romance, forming a subgroup with Ladin and Sardinian.

It is not possible to discuss this at any length here, but our hypothesis is that we are seeing a 'signature' of borrowing in these trees: in other words, although there is a tendency to assume that loans have been filtered out of lexical comparisons, this may not always be the case, where borrowings are particularly hard to detect,

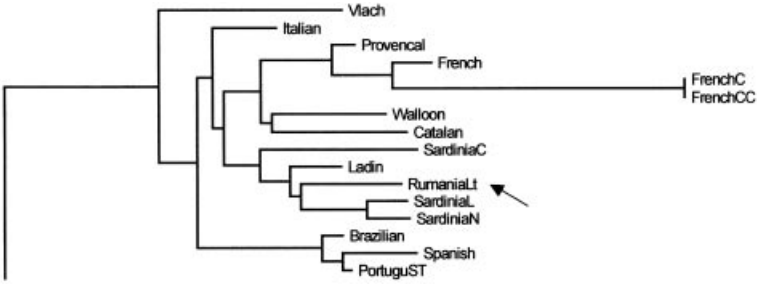

Figure 9. Close-up of the Romance branch from the Neighbour-Joining tree based on the most retentive meanings. Compare the position of Rumanian with that in Fig. 8.

or from closely related languages, or both. This hypothesis could be confirmed by using a particularly retentive set of meanings as a control, to see whether this triggers a change in position of the language(s) which have been borrowing. Conversely, the relocation of the language in question using the less retentive meanings, or the whole list, should indicate approximately the source of the loans. In the case of Frisian, we have heavy borrowing from Dutch; and for English, although the trees themselves are less informative, we know that there is considerable early borrowing from North Germanic (Embleton 1986: 100–1 finds fifteen Dutch-to-Frisian loans in the Swadesh 200-word list, and sixteen from North Germanic to English). English does not move into the North-Germanic group altogether; but this is quite possibly because of the additional, extensive borrowing from Romance, which is effectively pulling English towards the margins of the Germanic subfamily altogether (again, Embleton 1986: 100 reports twelve borrowings from French into English in the 200-word list). In the case of Rumanian, there is also well documented and extensive borrowing from Slavic languages, which again may account for these shifts in the tree depending on the relative retentiveness of the meanings used. Having said this, the general structure of the Romance subfamily is much more volatile than was the case for Germanic, and Rumanian is by no means the only language to shift within the tree between Fig. 8 and 9, though it is one of those most radically affected. Romance, in other words, is potentially less robustly treelike in its internal relationships than Germanic. General patterns of shift like this may indicate continued contact between daughter languages after their initial separation.

5. Future methodological developments

The preliminary results reported above provide an initial indication that standard trees for Indo-European, which cohere with results from the Comparative Method, can be produced using biological tree-drawing programs, which in turn operate on distance matrices derived from standard cognacy judgements based on 200-meaning lists. This is true for both the Neighbour-Joining and Maximum-Likelihood approaches; and bootstrapping shows that these trees

are also extremely robust. So far, we have made very few real methodological innovations, and have used only familiar Indo-European data, thus remaining firmly within the first two of Embleton's (1986) three points.

There are two points to make here. First, it is important to stress that producing familiar trees from familiar data still constitutes a step forward: the trees are in fact different, not in shape, but in statistical robustness and assured viability. This role of quantification is not always appreciated, however, by other linguists. For instance, Gray and Jordan (2000) have been working independently with a second biological tree-drawing program, this time using data from Austronesian languages; again, they produce results strongly in keeping with one model of population settlement and therefore language divergence. However, these results did not meet with universal approbation among linguists: Cysouw (2000) notes that:

> The authors are very keen to proclaim that their quantitative methods, which are taken from biology . . . are important, or even better than the methods used by linguists . . . The authors used 'an efficient computer algorithm' on the unpublished data from Blust's Austronesian Comparative Dictionary to build a language-tree of the Austronesian languages. As far as I can see, nothing new results from their analyses. There is a rather nice congruence between their tree and the tree as I knew it from the literature . . . The method is a nice addition to historical linguistics, but there is nothing really new. So, it seems to be possible to publish an article in *Nature* just by using the right computer program and forget that many years of research has been performed in linguistics to be able to perform these analyses . . .

It cannot be stressed too much that, in the initial stages of such investigations, researchers hope they will not find anything new. The point of these initial analyses is to show that biological tree-drawing programs provide sufficiently standard, sensible trees to make it worthwhile extending them in the future, to cases which cannot be resolved by the application of more traditional linguistic methods. Of course we must not forget that prior linguistic analysis is necessarily involved; these programs are only ever as good as the

data and analysis that underlie them, and it is a priority to ensure that those are sound and robust. However, the application of computational techniques to cognacy scores can also show relationships and patterns, as in our putative 'signatures' of borrowing, which might be suspected but cannot be demonstrated by non-quantificational methods. Furthermore, although further applications might be developed, for meaning-list comparisons and potentially for other aspects of the Comparative Method, it seems highly unlikely that any program could be designed to confirm Greenberg's (1987) results for Amerind, precisely because his criteria are unclear, and programming depends crucially on explicitness.

For the moment, there are three obvious ways forward for this kind of quantitative analysis in comparative linguistics, all of which are scheduled as part of our ongoing project. The first involves the extension of the meaning-list and family-tree approach outlined above to non-Indo-European languages, and our first step here will involve the collection of meaning lists for Quechua, Aymara and Jaqaru. Application of the tree-drawing programs in these cases will extend these analyses to Embleton's (1986: 3) third level, involving '. . . data where linguistic opinions have not yet been produced, have not yet been firmly established, or perhaps are even in conflict. In practice, this usually means application to non-Indo-European data . . .'. We shall also be comparing trees based on Lohr's (1999) maximally retentive meanings, with Starostin's (2000) list of persistent roots, and with random selections from larger word lists.

Returning to Embleton's (1986) first step, we are also interested in developing truly new models for language comparison; in particular, it is important to assess the possibilities for comparison in terms of areas of the grammar outside the lexicon, which have not so far been fully exploited in comparative linguistics. While it is true that meaning lists involve aspects of semantics and the lexicon, and that the Comparative Method is not restricted to lexical comparisons, but revolves around recurrent similarities of sound and meaning, these methods do require decisions to be made in advance on which are the salient features for a group. The same would be true of Ringe et al.'s character-based approach to Indo-European (Warnow, Ringe and Taylor 1996; Ringe, Warnow and Taylor

2002), and indeed of our own meaning-list-based work reported above, which equally involves compiling a list ahead of time, and is not therefore strictly speaking objective. The ideal complementary method would involve the objective, quantitative comparison of a non-lexical system; and Paul Heggarty is currently developing just such a model for phonetic comparison (Heggarty 2000a, b) in a separate strand of our project work.

Finally, we propose to explore issues of contact and borrowing further; and here the Andean languages are also important, in that all have borrowed from Spanish, but it is clear that they have also inter-borrowed. As noted above, it already appears that our trees may show a 'signature' of borrowing when we compare the results of the full 200-meaning list, or indeed the 23 low-retentiveness–low-reconstructibility meanings, with the thirty (or 24 informative) high-retentiveness–high-reconstructibility meanings. It may also be possible to see where the borrowing is coming from by the relative position of the language(s) in question in the different trees. This is particularly important because it distances our methods from alternative ways of dealing with borrowing. Many meaning-list approaches simply seem to assume that borrowing will be unproblematic, because the Swadesh lists, and variants of these, were chosen specifically to include basic, core vocabulary items, which are said to be more resistant to borrowing. However, this greater resistance to borrowing has never really been tested, and indeed it is not entirely clear how this testing could be accomplished. Even if we accept that the core vocabulary is less liable to borrowing, we cannot rule it out altogether, especially for language groups or periods where contact is particularly intense and perhaps prestige of one language over the other(s) very high. We have already seen that Embleton (1986) finds a whole range of inter-borrowings within the Indo-European languages, for the very Swadesh 200-meaning list which is intended as the epitome of borrowing resistance; Embleton's figures include, for example, twelve loans from French into English, sixteen from North Germanic into English, eleven from Danish into Norwegian, six from Hebrew into Yiddish, fifteen from Dutch into Frisian, nineteen from Danish into Faroese and five from Tolai into Tok Pisin, along with a whole series of smaller numbers.

The other alternative is to accept that borrowing will almost inevitably happen, but that it might be used productively to tell us more about the different types of relationship the languages we are considering contract with one another. This might entail attempting to derive some formula we might apply to the data to correct the effects of borrowing. This is Embleton's (1986) approach; but although it is an interesting one, it is also limited. Although Embleton argues that a figure for borrowing rate can be proposed, she also argues that this will be highly variable and therefore has to be calculated independently (though within normally certain specified limits, i.e., up to 30%) for each language pair. Effectively, this means we simply add up the borrowings already diagnosed, then extrapolate that total over a particular time period as the overall rate, which is then factored into the calculations. Where this does not help at all is in cases where there may or may not be borrowings. Our method may possibly identify just these cases, through the comparison of more retentive with less retentive meanings. This will not show exactly which meanings are being affected, although future work may be able to reveal this, with some effort, by removing individual groups of meanings or even single meanings to check the effect on the relative position of the languages in question.

Here again, we can perhaps learn from biology, which has many similar problems. Cann (2000: 1008) stresses that, 'There is a close connection between comparative linguistics and evolutionary biology. Both seek to account for the overall resemblance between entities that are now distinct; in both there are confounding cases of horizontal transfer of information; and both are bedevilled by spurious similarities that arise from convergence, parallelism or reversals in character states.' In other words, 'horizontal transfer', or borrowing, happens between populations just as it does between languages: people from outside the particular population will sometimes be introduced, just as words and sounds are borrowed from one language or language family into another. There is a mathematical model under development at the moment which may provide a means of identifying such genetic borrowing when it happens. In fact, this matrix or network model (Bandelt et al. 1995; Bandelt, Forster and Röhl 1999; Forster et al. 2001) is currently intended to filter out cases of convergent evolution, in other words cases where

the same feature has arisen more than once perhaps in the same sort of environment, but does not indicate common ancestry. However, network analysis is also directly applicable to the population rather than the species level, and hence must deal with inter-population mixing, an operation directly analogous to borrowing. The network analysis cannot be described in detail here, but a matrix for a population provides a composite picture of the relationships among a series of individuals, showing by the lengths of the branches how many mutational steps each individual is away from a common ancestor. The network actually includes all the most likely tree representations for this group, but the larger number of connections and dimensions allows for more than one possibility. Perhaps the most interesting aspect of network analysis is that, although the program is designed to draw networks, cases where there has been no borrowing or convergence will automatically be represented with the most likely tree. That is, the program involved draws a tree when the relationships are clear and tree-like, and a more complex network when the connections are more complex and show more interaction. The next step is to explore this further, to establish whether in fact this method can be used to detect linguistic cases with more and less inter-borrowing. It is ironic, but rather pleasing, that whereas borrowing between languages has been one of the single biggest obstacles to linguistic classification in the past, borrowing from biology may present one opportunity of resolving it, in the future.

Department of English Language and Linguistics
University of Sheffield
Firth Court
Western Bank
Sheffield
S10 2TN, UK
Email: April.McMahon@shef.ac.uk; R.McMahon@shef.ac.uk

References

ALLEN, W. S., 1953. 'Relationship in comparative linguistics', *Transactions of the Philological Society*, 52–108.

BAKKER, PETER, 1997. *'A Language of Our Own'. The Genesis of Michif – the Mixed Cree – French Language of the Canadian Métis*, New York: Oxford University Press.

BAKKER, PETER, 2000. 'Rapid language change: Creolization, intertwining, convergence', in Colin Renfrew, April McMahon and Larry Trask (eds.), *Time Depth in Historical Linguistics*, Cambridge: McDonald Institute for Archaeological Research, 585–620.

BAKKER, PETER & MOUS, M. (eds.), 1994. *Mixed Languages: 15 Case Studies in Language Intertwining*, Amsterdam: IFOTT.

BANDELT, H-J., FORSTER, P., SYKES, B. C. & RICHARDS, M. B., 1995. 'Mitochondrial portraits of human populations using median networks', *Genetics* 141: 743–753.

BANDELT, H-J., FORSTER, P. & RÖHL, A., 1999. 'Median-joining networks for inferring intraspecific phylogenies', *Molecular Biology and Evolution* 16, 37–48.

BARBUJANI, GUIDO, 1997. 'DNA variation and language affinities', *American Journal of Human Genetics* 61, 1011–1014.

BUCK, C. D., 1949. *A Dictionary of Selected Synonyms in the Principal Indo-European Languages*, Chicago: University of Chicago Press.

CAMPBELL, LYLE, 1988. Review of Greenberg (1987), *Language* 64, 591–615.

CAMPBELL, LYLE, 1998. *Historical Linguistics*, Edinburgh: Edinburgh University Press.

CANN, REBECCA L., 2000. 'Talking trees tell tales', *Nature* 405, 1008–1009.

CAVALLI-SFORZA, L. L., 2000. *Genes, Peoples and Languages*, London: Allen Lane/The Penguin Press.

CHOMSKY, NOAM & HALLE, MORRIS, 1968. *The Sound Pattern of English*, New York: Harper & Row.

CLACKSON, JAMES, 1994. *The Linguistic Relationship Between Armenian and Greek*, London: Blackwell (Publications of the Philological Society).

CYSOUW, MICHAEL, 2000. Message posted to the HISTLING discussion list. HISTLING@VM.SC.EDU.

DARWIN, CHARLES, 1996 [1859]. *The Origin of Species*, Oxford: Oxford University Press.

DURIE, MARK & ROSS, MALCOLM (eds.), 1996. *The Comparative Method Reviewed*, Oxford: Oxford University Press.

DYEN, ISIDORE, KRUSKAL, JOSEPH B. & BLACK, PAUL, 1992. 'An Indoeuropean classification: a lexicostatistical experiment', *Transactions of the American Philosophical Society 82, Part 5*. Data available at http://www.ldc.upenn.edu.

EHRET, CHRISTOPHER, 1995. *Reconstructing Proto-Afroasiatic (Proto-Afrasian). Vowels, Tone, Consonants and Vocabulary*, Berkeley: University of California Press.

EMBLETON, SHEILA M., 1986. *Statistics in Historical Linguistics*, Bochum: Brockmeyer.

FELSENSTEIN, J., 2001. *PHYLIP: Phylogeny Inference Package. Version 3.6*, Department of Genetics, University of Washington.

FODOR, I., 1961. 'The validity of glottochronology on the basis of the Slavonic languages', *Studia Slavica* 7, 295–346.

FORSTER, PETER, TORRONI, ANTONIO, RENFREW, COLIN & RÖHL, A., 2001. 'Phylogenetic star construction applied to Asian and Papuan mtDNA evolution', *Molecular Biology and Evolution* 18, 1864–1881.

FOX, ANTHONY, 1995. *Linguistic Reconstruction: An introduction to theory and method*, Oxford: Oxford University Press.

GAMKRELIDZE, T. V. & IVANOV, V. V., 1995. *Indo-European and the Indo-Europeans:*

A reconstruction and historical analysis of a proto-language and a proto-culture, Berlin: Mouton de Gruyter.

GRAY, RUSSELL D. & JORDAN, F. M., 2000. 'Language trees support the express-train sequence of Austronesian expansion', *Nature* 405, 1052–1055.

GREENBERG, JOSEPH H., 1987. *Language in the Americas*, Stanford: Stanford University Press.

GRIMES, J. E. & AGARD, F. B., 1959. 'Linguistic divergence in Romance', *Language* 35, 598–604.

HEGGARTY, PAUL, 2000a. *Quantification and Comparison in Language Structure: An exploration of new methodologies*. PhD thesis, University of Cambridge.

HEGGARTY, PAUL, 2000b. 'Quantifying change over time in phonetics', in Colin Renfrew, April McMahon and Larry Trask (eds.), *Time Depth in Historical Linguistics*, Cambridge: McDonald Institute for Archaeological Research, 531–562.

HEGGARTY, PAUL & MCMAHON, APRIL, 2002. 'How similar are sounds?' Paper presented at the 10th *Manchester Phonology Meeting*.

HOIJER, H., 1956. 'Lexicostatistics: a critique', *Language* 32, 49–60.

KESSLER, BRETT, 2001. *The Significance of Word Lists*, Stanford: CSLI Publications.

KOERNER, KONRAD (ed.), 1983. *Linguistics and Evolutionary Theory: Three essays by August Schleicher, Ernst Haeckel & Wilhelm Bleek*, Benjamins: Amsterdam.

KROEBER, A. L. & CHRÉTIEN, C. D., 1937. 'Quantitative classification of Indo-European languages', *Language* 13, 83–103.

KRUSKAL, JOSEPH B., DYEN, ISIDORE & BLACK, PAUL, 1971. 'The vocabulary method of reconstructing family trees: innovations and large scale applications', in F. R. Hodson, D. G. Kendall and P. Tautu (eds.), *Mathematics in the Archaeological and Historical Sciences*, Edinburgh: Edinburgh University Press, 30–55.

LEIGH FERMOR, PATRICK, 1986. *Between the Woods and the Water*, London: Penguin.

LOHR, MARISA, 1999. *Methods for the Genetic Classification of Languages*. PhD thesis, University of Cambridge.

LUCE, G. H., 1981. *A Comparative Word-List of Old Burmese, Chinese and Tibetan*, London: School of Oriental and African Languages.

MATRAS, YARON, 2000. 'How predictable is contact-induced change in grammar?', in Colin Renfrew, April McMahon and Larry Trask (eds.) *Time Depth in Historical Linguistics*, Cambridge: McDonald Institute for Archaeological Research, 563–583.

MCMAHON, APRIL, LOHR, MARISA & MCMAHON, ROBERT, 1999. 'Family trees and favourite daughters', in Colin Renfrew and Daniel Nettle (eds.) *Nostratic: Examining a Linguistic Macrofamily*, Cambridge: McDonald Institute for Archaeological Research, 269–285.

MCMAHON, APRIL & MCMAHON, ROBERT, 1995. 'Linguistics, genetics and archaeology: internal and external evidence in the Amerind controversy', *Transactions of the Philological Society* 93, 125–225.

NICHOLS, JOHANNA, 1992. *Linguistic Diversity in Space and Time*, Chicago: University of Chicago Press.

PAGE, RODERIC D. M., 1996. 'TREEVIEW: An application to display phylogenetic trees on personal computers', *Computer Applications in the Biosciences* 12: 357–358.

PAGE, RODERIC D. M. & HOLMES, EDWARD C., 1998. *Molecular Evolution: A phylogenetic approach*, Oxford: Blackwell.

PAGEL, MARK, 2000. 'Maximum-likelihood models for glottochronology and for reconstructing linguistic phylogenies', in Colin Renfrew, April McMahon and

Larry Trask (eds.), *Time Depth in Historical Linguistics*, Cambridge: McDonald Institute for Archaeological Research, 189–207.

Poloni, E. S., Semino, O., Passarino, G., Santachiara-Benerecetti, A. S., Dupanloup, L., Langaney, A. & Excoffier, L., 1997. 'Human genetic affinities for Y-chromosome P49a,f/*Taq*I haplotypes show strong correspondence with linguistics', *American Journal of Human Genetics* 61, 1015–1035.

Renfrew, Colin, 1999. 'Reflections on the archaeology of human diversity', in Bryan Sykes (ed.), *The Human Inheritance: Genes, Language and Evolution*, Oxford: Oxford University Press, 1–32.

Renfrew, Colin & Nettle, Daniel (eds.), 1999. *Nostratic: Examining a Linguistic Macrofamily*, Cambridge: McDonald Institute for Archaeological Research.

Renfrew, Colin, McMahon, April & Trask, Larry (eds.), 2000. *Time Depth in Historical Linguistics*, 2 vols, Cambridge: McDonald Institute for Archaeological Research.

Ringe, Don, 1999. 'How hard is it to match CVC- roots?', *Transactions of the Philological Society* 97, 213–244.

Ringe, Don, Warnow, Tandy & Taylor, Ann, 2002. 'Indo-European and computational cladistics', *Transactions of the Philological Society* 100: 59–129.

Ross, Alan S. C., 1950. 'Philological probability problems', *Journal of the Royal Statistical Society Series B* 12, 19–59.

Ruhlen, Merritt, 1991. *A Guide to the World's Languages. Vol 1: Classification*, London: Edward Arnold.

Sims-williams, Patrick, 1998. 'Genetics, linguistics and prehistory: thinking big and thinking straight', *Antiquity* 72: 505–527.

Singh, Ishtla, 2000. *Pidgins and Creoles: An introduction*, London: Arnold.

Slaska, Natalia, 2002. 'Grafting contact onto family trees', Paper presented at the *Manchester Postgraduate Linguistics Conference*, March 2002.

Starostin, Sergei, 2000. 'Comparative-historical linguistics and lexicostatistics', in Colin Renfrew, April McMahon and Larry Trask (eds.), *Time Depth in Historical Linguistics*, Cambridge: McDonald Institute for Archaeological Research, 223–265.

Swadesh, Morris, 1952. 'Lexico-statistical dating of prehistoric ethnic contacts', *Proceedings of the American Philosophical Society* 96, 452–463.

Sykes, Bryan (ed.), 1999. *The Human Inheritance: Genes, Language and Evolution*, Oxford: Oxford University Press.

Thomason, Sarah G. (ed.), 1997. *Contact Languages: A wider perspective*, Amsterdam: Benjamins.

Thomason, Sarah G., 2001. *Language Contact: An introduction*, Edinburgh: Edinburgh University Press.

Thomason, Sarah G. & Kaufman, Terrence, 1988. *Language Contact, Creolization, and Genetic Linguistics*, Berkeley: University of California Press.

Tryon, D. T. (ed.), 1995. *Comparative Austronesian Dictionary: An Introduction to Austronesian Studies*, Berlin: Mouton.

Warnow, Tandy, Ringe, Donald & Taylor, Ann, 1996. 'Reconstructing the evolutionary history of natural languages,' *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, 314–322.

Zorc, R. D., 1995. 'A glossary of Austronesian reconstructions', in D. T. Tryon (ed.) *Comparative Austronesian Dictionary: An Introduction to Austronesian Studies*, Berlin: Mouton, Part 1, Fasc. 2, 1105–1197.