

Languages and Genes

Bernard Comrie

*University of California Santa Barbara and
Max Planck Institute for Evolutionary Anthropology*

Contact e-mail address: comrie@eva.mpg.de

1. Preamble

This paper is being distributed as preliminary material for the conference Language and Genes to be held at the University of California Santa Barbara September 8-10, 2006 - for further information, see the conference web site <http://www.linguistics.ucsb.edu/projects/Languages-and-Genes/>.

My aim in the paper is to set out some of the issues that strike me as particularly important in comparing work in linguistics and genetics, with reference also to other disciplines such as archeology and cultural anthropology, in reconstructing various aspects of prehistoric human population history. Some disclaimers are in order.

Although I have tried to balance the contribution of different disciplines, inevitably the paper is written from the perspective of a linguist. It may well be that some of the things I say about other disciplines will not be accepted by all, or even most, practitioners of those disciplines, but I see this as being potentially a strength of the discussions to take place at the conference, where the contributions of different disciplines need to be confronted as well as reconciled. Indeed, not all of the claims I will make will be accepted by all linguists, so that many of the points are more identification of issues for consideration than necessarily firm statements of undoubted truth. In other words, even within the discipline of linguistics, the paper is to some extent written from my own perspective.

My aim throughout has been to set out problems, or at least discussion questions, rather than to identify solutions - this last being the aim of our conference!

Given the "position-paper" nature of the paper, I have not included bibliographical references, though I will be happy to provide references on demand for the various claims I make. I have deliberately not written this as a "publishable paper", not even as the first draft of one, but rather to stimulate discussion.

To avoid terminological confusion, I will throughout use the term "genetic" only in reference to genetics, and use the term "genealogical" to refer to the historical relationships among languages. (I apologize in advance for any inadvertent slips in adhering to this terminology.)

2. Introduction

The reason for comparing linguistics and genetics with regard to human population history is that both disciplines make important claims with regard to both facts about human population prehistory and the validity of methods used in unraveling this question. To the extent that the results achieved by each discipline are reliable internal to that

discipline, they should be compatible with the results obtainable from the other discipline. But here an important factor has to be taken into account. Genes are transmitted biologically, from parents to their biological offspring. Languages are transmitted culturally, i.e. human children acquire their language (or languages) by interacting with the speech community in which they grow up. To the extent that the speech community is constituted primarily by the child's parents or by others speaking the same language as the child's parents, there will be agreement between biological and cultural inheritance. To the extent that these differ, there is the potential for differences. A crucial question that arises in considering this question in more detail is the nature of these differences and the likelihood of their occurring, and in considering the question of likelihood there is, of course, no guarantee that this probability has been the same at all times and in all places. There are several reasons why the cultural transmission of language might not match the biological transmission of genes, including at least the following.

The child may fail to acquire the language of one or both parents, either through the willful non-transmission of this language (as can happen in the case of stigmatized languages) or through separation from one of both parents, e.g. being adopted by speakers of another language. Note that if the two parents speak different languages, the situation becomes even more complicated - the child may grow up bilingual, speaking the languages of both parents, or may grow up speaking only one of those languages, as well as varying intermediate degrees of dominance of one language over the other. In addition, if the speech community as a whole has other languages in addition to the child's parents', the child may grow up also speaking other languages in addition to his or her parents, including languages spoken by people who do not bear any close genetic similarity to the child.

In addition to discrepancies between genetic and linguistic ancestry arising from acquisition of language by the child, a second possibility is language shift (in the genetic literature often referred to as: language replacement) among the adult community, either at a community level or in the individual, i.e. the abandonment of the ancestral language in favor of some other language. While language shift is often a process that occurs over several generations of intervening bilingualism, it can also take place quite abruptly, and I have myself encountered examples of families where grandparents and grandchildren are effectively unable to communicate linguistically - the grandparents speak the "heritage" language, and have little or no competence in the other language; the parents speak both languages, but have decided to transmit only the non-heritage language to their children, who grow up essentially monolingual in this language.

A crucial question is how frequent situations occur in which such discrepancies between linguistic and genetic heritage arises. The early work on comparison between genetic and linguistic classifications of human populations tended to emphasize the parallelism between the two sets of transmission, somewhat to the consternation of many linguists. More recent work has pointed to a number of instances where there is clearly lack of such parallelism. But it still remains to be investigated empirically to what extent we find correlation, taking care to pay due attention to the reliability of both the linguistic and the population genetic classifications that are proposed. From the linguistic side, there are a number of factors that need to be taken into account, some of which also bring up aspects of cultural anthropology, i.e. of the cultural setting within which biological and linguistic interactions take place.

Parallelism between genetic and linguistic classifications is to be expected when one is dealing with populations that are for the most part sealed off from one another, interacting neither biologically nor linguistically. Populations of this kind are presented, to some extent, by the major nation states, and to some extent their precursors, for instance in Europe, where there are factors leading to much more frequent marriage within the speech community than across language boundaries. But it would be misleading to generalize this pattern to the world as a whole, or to history as a whole. Indeed, at the opposite extreme we know of societies, for instance among the indigenous inhabitants of Amazonia and Australia, characterized by the practice of "linguistic exogamy", i.e. a requirement that one take as one's marriage partner a speaker of a language different from one's own, a practice that guarantees a discrepancy between linguistic and genetic heritages, but which is also one way of guarding against too close inbreeding, especially in the case of very small speech communities; indeed, violation of the linguistic exogamy taboo in such societies is viewed in much the same way as incest. I emphasize that the determination of the extent to which mate selection across linguistic boundaries takes or took place in any location or at any time is an empirical issue, and one that deserves to be investigated without prejudice. One will find numerous claims made in the literature that are often either unsubstantiated, or only substantiated with respect to particular communities. (To take an example that I myself believe, but that I am not sure has ever been really substantiated: Linguists often maintain that most of the world's population is and has always, or at least for a long time, been at least bilingual. If true, this would provide an easy conduit for discrepancies between linguistic and genetic classifications to arise.) Crucial to this enterprise is consideration of the social circumstances in which the relevant genetic and linguistic transmission took place. Ideally, one would welcome an investigation of the likely genetic and linguistic outcomes of each kind of encounter and interaction between different populations and languages.

3. Tree structures

Representations in terms of tree structures have played an important role in both linguistic and genetic accounts of human prehistory, and it is therefore important to consider the similarities and differences in the interpretation of such representations in the two disciplines.

I will start here with genetics, and more specifically with the study of mitochondrial DNA (mt-DNA) and of non-recombinant Y-chromosomal DNA (Y-chr DNA) in the individual (returning later to the consideration of the community). Investigation of mt-DNA and Y-chr DNA constitutes, or at least such is my impression, the basis of the bulk of current work in population genetics. They have the advantage of monoparental transmission - mt-DNA is inherited only from the mother (but to both sons and daughters), while Y-chr DNA is inherited only from the father (and only by sons). This means that one can draw a tree starting from a common maternal-line ancestor or from a common paternal-line ancestor, charting the different mutations that take place through history, each mutation defining a bifurcation in the tree structure, until one comes to that ancestor's latest descendants at the end nodes of the tree structure. In other words, each individual can be traced back to a single node in the previous generation, and in the generation before that, and so on indefinitely, although different individuals may be traceable back to the same node. The trees for each of mt-DNA and Y-chr DNA thus have a "parthenogenetic" structure.

Of course, in reality, there are two sets of complications that one needs to take into account in population genetic studies. First, the model as presented applies to non-recombinant DNA. With recombinant DNA, which is inherited randomly from either parent, then unless one wishes simply to trace the history of a particular gene, the resulting picture will be much more complex, with some genetic material inherited from the mother, some from the father, and one generation further back: some from the paternal grandfather, some from the maternal grandfather, some from the maternal grandmother, some from the maternal grandfather, and so on. In other words, in terms of recombinant DNA, an individual is a mixture of components from different ancestors, which calls for a different kind of model than the tree model. Note that even if we stick to non-recombinant DNA, but include both mt-DNA and Y-chr DNA, we will get different facets of an individual's genetic history, since the maternal and paternal lines can in principle be completely different from one another, not only in the trivial sense that the individuals are different, but in the for us more relevant sense that they may belong to (earlier) different populations.

Second, the tree representation provides a model of the ancestry of an individual, but not of a population, since different individuals in a population will have different ancestries. This provides an even more direct link to some of the problems with a literal interpretation of the tree model in linguistics to be discussed below. In particular, if we study the genetics of the individual, whether in terms of non-recombinant or recombinant DNA, it makes no sense to talk of "horizontal transmission" - all genetic material is inherited vertically from one's ancestors. But in talking of the genetic profile of a population, one can consider genetic material coming into the population from outside, as a result of selecting mates from outside the population. Much of the recent work on phylogenetic methods has concentrated, almost if not entirely, on vertical transmission, and only recently - at least according to my impression - is the relevance of horizontal transmission being seriously integrated into the relevant formal models, although this is an important aspect for those dealing with the history of language.

Turning now to language, there is a model, the so-called family tree model, which corresponds exactly to the "parthenogenetic" model proposed above, and we can illustrate it with the example of English within the Indo-European language family. English is a member of the West Germanic sub-branch of the Germanic branch of the Indo-European family. This means that English is descended from an ancestor language, Proto-West-Germanic, and at this time-depth only from that ancestor. (There are other languages descended from the same ancestor, such as German.) Going further back in-time, Proto-West-Germanic is descended from Proto-Germanic, which is thus the sole ancestor at that time-depth of English. (Again, Proto-Germanic has other descendants, such as Proto-North-Germanic, the ancestor of Icelandic and Swedish, among others.) And going back still further, Proto-Germanic is a descendant of Proto-Indo-European, which at that time-depth is again the sole ancestor of English, although again Proto-Indo-European has a number of other descendants (such as Proto-Slavic, with as one of its descendants Proto-East Slavic, with as one of its descendants Russian). The structure of such a family tree exactly matches that of the tree model of the history of non-recombinant DNA, which is of course one of the reasons why it has been so obvious to compare the linguistic and genetic trees and see to what extent they match, to what extent they differ.

At a basic level, the interpretation of a family tree in linguistics is exactly the same as in the study of mt-DNA or Y-chr DNA in genetics. As one proceeds forwards in time, a bifurcation in the tree structure indicates an innovation (or set of innovations) that is shared by one of the branches, setting it off from the other(s), i.e. each branch is defined by an innovation or set of innovations, comparable to a mutation in genetics. It is crucial to bear in mind that an inherited feature or set of inherited features does not define a branch. To take one of the parts of the family tree discussed in the previous paragraphs, the Germanic branch is defined (inter alia) by a set of changes that radically altered part of the consonant system of inherited Indo-European material, for instance shifting inherited *p* to *f* (compare English *father* with Latin *pater*, English *fish* with Latin *piscis*, or English *foot* with Latin *pes*.)

Although the family tree model forms the basis of much of comparative-historical linguistics, it necessarily involves a simplification of the actual historical facts, and it is important to consider some aspects of this simplification in evaluating the model and comparing its validity with that of similar models in other areas, including of course genetics.

First, the family tree model gives the impression of a clean break, whereby at a particular point in time one branch splits off from the rest of the tree. We know from the study of ongoing language change, as well as from the detailed results of historical-comparative linguistics, that this is a gross oversimplification. I will take an illustrative example from the division of the Slavic languages (i.e. of Proto-Slavic) into the three main divisions of that branch of Indo-European, East Slavic, West Slavic, and South Slavic. Each of these branches is characterized by a set of innovations not shared by the other two branches. For instance, in East Slavic words with the original structure *CarC* (where *C* stands for any consonant) show up with the structure *CoroC*, e.g. *gorod* 'city', whereas in South Slavic languages like Serbian the result is *CraC*, e.g. *grad* (compare the Russian and Serbian city names *Belgorod* and *Beograd* 'Belgrade', both literally 'white city'). However, when one looks at the distribution of word-initial *o* and *e* across the Slavic languages, and indeed across different words in the different Slavic languages, the pattern of distribution cuts across the three-way division, and has been shown to reflect a sound change considerably earlier than those dividing Slavic into three sub-branches; intervening population movements meant that the original distribution of the *o/e* distinction was overlain by a larger number of other of later innovations, so that although it actually reflects an earlier innovation, it is not considered criterial in constructing the family tree of the Slavic languages.

Conversely, even as a language is breaking up, it is possible for innovations to affect part of the overall area, including some, but not all of what will end up as separate languages. In extreme cases, one may even have one change starting from one end of the geographical area, another starting from the other end, and overlapping in the middle, so that one ends up with peripheral languages sharing either only change A or only change B, but an intermediate area sharing both changes - leading to a contradiction if one tries to draw a family tree, since the intermediate area shares innovations with each of A and B that are not shared by A and B. A somewhat simpler example is provided again by the break-up of Slavic. At around the time that Slavic was breaking up, a change took place that led to the loss of the weak vowels corresponding to Indo-European short *i* and *u*. Although this change is common to nearly all of Slavic, including languages from all three sub-branches, it clearly spread across Slavic after the break-up of the common

language. Indeed, these weak vowels are still present in the earliest texts in what is indisputably East Slavic (Old Russian), disappearing only during the first period of written records. One peripheral dialect of Old Russian, Old Novgorod, seems to have escaped the change altogether, until this dialect disappeared under the influence of its neighbors.

Second, it is well known that languages not only inherit material vertically, through the chain of ancestor languages, but also horizontally, through borrowing from other languages. For instance, English is a West Germanic language, but during its history it has borrowed substantially from two non-West Germanic languages. As a result of the Viking settlement of large parts of England during the Old English period, English borrowed heavily from the North Germanic language of the Vikings. And as a result of the Norman Conquest, English borrowed heavily from French, a member of the Italic branch of the Indo-European family. I return in the next section to more detailed consideration of the relevance of borrowing to establishing genealogical relatedness among languages, but in the case of English we may note that the horizontal transmission has been primarily of vocabulary, but includes some quite basic words in the language; for instance, *take* is of North Germanic origin, *very* of French origin.

Most linguists are probably of the opinion that, in at least most cases, it should be possible to assign a language to a single ancestry on the basis of vertical transmission, with the instances of horizontal transmission not serving to undermine that basic genealogical assignment. But there are some cases of languages, so-called mixed languages, where this seems much more problematic, and many linguists are prepared to consider that there are some languages that may not unequivocally be assignable to a single language family. A striking example is provided by Michif, spoken by a small community of the Canadian and US prairies of mixed Cree and French descent. Their language basically combines French noun phrases (including such idiosyncrasies as French genders) with Cree clause structure; the vocabulary is similarly divided, with nouns being overwhelmingly of French origin, verbs overwhelmingly of Cree origin. (Cree is a member of the Algonquian language family, with a syntactic structure very different from that of French or other Indo-European languages.) Another example is provided by creole languages, such as Haitian Creole, which, on at least one current hypothesis, combine a vocabulary that is largely of European origin (French, in the case of Haitian Creole) with a grammar that is largely of West African origin (Fongbe, in the case of Haitian Creole). Such extreme instances of language mixing are perhaps exceptional, though clearly possible, and seem to result from specific social circumstances: In the case of Michif, this was the emergence of a new ethnic group that considered itself neither French nor Cree, though descended from the coming together of French (and French-speaking) fathers and Cree (and Cree-speaking) mothers. In this case, the expected genetic correlates are clear, but in cases of more limited contact the precise correspondences between linguistic and genetic mixing will be more difficult to unravel. I return to borrowing in the next section.

4. Reconstructing history

Despite the similarities between the family tree models that play so important a role in genetics and in historical-comparative linguistics, it would be hard to imagine two more different methodologies preferred by the majority of practitioners of each of the two disciplines. In genetics, preference is given to the use of mathematical approaches that

can deal with huge amounts of data and construct trees that provide some kind of best fit to that data - the individual methods vary, but my general impression is that it is in general effectively impossible (given constraints on computing power and time) actually to seek out the best tree, rather different kinds of short-cuts can be taken to get an approximation. By contrast, linguists have proceeded primarily by the painstaking investigation of the history of individual forms, including both direct historical investigation (in texts from older periods) and comparison with related languages, in order to come up with detailed accounts of the histories of languages (e.g. the phonetic and other changes that separate Germanic languages from the rest of Indo-European), and of the etymologies of individual words. There are no doubt various reasons for the different approaches. For instance, the avoidance of mathematical methods by many historical linguists no doubt owes much to historical tradition - the methods were undeveloped when some of the major breakthroughs were made in historical-comparative linguistics in the nineteenth century - and to the fact that most linguists are simply not trained in the methods in question (although this situation is fortunately changing). In addition, the results of the painstaking research into individual linguistic changes of individual etymologies are valued in their own right as a contribution to the cultural history of the people speaking the language in question; think, for instance, of the importance assigned to etymology in the *Oxford English Dictionary*. To take an analogy: If one has established by careful investigation that the accused was at the scene of the crime, then one does not need to apply statistical methods to evaluate the probability of his having been there.

But there are, I think, other reasons that are more important in evaluating methodological strengths and weaknesses, and in what follows I will try to examine some of these, in particular as they relate to linguistics.

It will be recalled that in establishing a family tree in linguistics, crucial importance is assigned to common innovations. This means that it is not sufficient to establish similarities versus differences among languages, rather one must ascertain whether these similarities reflect common innovations (in which case they are criterial for establishing sub-groups of languages as reflected in a family tree) or common retentions from the ancestral state, in which case they are not. We can take as an example the shift of Indo-European *p* to Germanic *f* mentioned above. On the basis simply of similarities and differences, one could divide the Indo-European languages into two groups, those that have *f* in the relevant words, and those that have *p* in the relevant words. Crucially, however, the two different sets of languages have a very different status: Those that have *f* have undergone a change, so that the languages characterized by *f* have undergone a common innovation and therefore constitute a valid genealogical grouping; those that show *p* are simply those that did not undergo this change, and do not form a valid genealogical grouping - in fact, they comprise all the branches of Indo-European other than Germanic, so that we find *p* in languages from different branches such as Italic (Latin *pater*), Hellenic (Greek *patér*), Indo-Iranian (Sanskrit *pitr*), etc. The fact that there is a valid Germanic genealogical grouping but no valid "non-Germanic" grouping cannot be ascertained by inspection, but involves detailed investigation, taking into consideration such factors as the likelihood of different kinds of changes (the sound change from *p* to *f* is more likely than one from *f* to *p*). This has left linguists skeptical of approaches that apply mathematical methods to different states without taking into account which of the states is more likely to be the innovation and which ancestral.

The second major problem is that provided by borrowing, or horizontal transmission, which is known to be a major factor in historical linguistics. Until recently, the mathematical methods applied to establish phylogenies have operated, at least for the most part, on the assumption that all transmission is vertical, an assumption that clearly makes sense with dealing with, for instance, data from mt-DNA or Y-chr DNA analysis. But with linguistic material, unless one can separate out borrowing from inheritance, there is a danger of the method confusing the two and assigning to vertical transmission what should in fact be assigned to horizontal transmission. The recent development and implementation of methods that do take account of horizontal transmission promises much in this area, even if linguists will continue (for partly independent reasons) to argue over whether a particular word is inherited or borrowed, even while perhaps agreeing on the genealogical affiliation of the language overall.

In this connection, it is worth stepping back a moment to consider more systematically the reasons why languages may share properties in common, given that common ancestry is only one of the possibilities. Basically, there are four main reasons. First, the property may be a language universal, i.e. a defining characteristic of language that is shared by all languages. For instance, all spoken languages have consonants, and finding that two languages have consonants would be no evidence whatsoever in favor of their common origin. Methodologically, there is no problem here; if all the languages in our sample share a particular property, then we have no basis on which to use this feature to divide them into groups. There is, however, a subtype of this first property that is potentially more insidious. In addition to absolute universals, characteristic of all languages, there are also properties that are widespread across the languages of the world, so-called universal tendencies, or general propensities of language, and here it can be difficult to decide whether similarities are universal tendencies or indicative of large genealogical groupings. A transparent example that has surfaced in recent discussion is the widespread distribution of words of the type *ma* in the meaning 'mother' across the languages of the world: Does this simply reflect the fact that this is one of the first syllables produced by babies, then interpreted as meaning 'mother' by adults, or does it reflect common inheritance from an ancestor language that just happened to have this sound sequence in this meaning? Opinions remain divided.

Second, languages might share a property because they have inherited it from their common ancestor. Third, they may share a property because one has borrowed it from the other (or both have borrowed it from some other source). And fourth, the similarity may be purely the result of chance. Distinguishing common inheritance from borrowing remains the most problematic case, to which I return below. Progress has, however, been made in distinguishing between chance and other possibilities, with particular regard to similarities among words, even if the method - the so-called shift method - has not yet been widely put into practice. Suppose one has comparative vocabulary lists for two languages, let us say for 100 semantic concepts. One goes through the lists, comparing the word in language A with the word in language B having the same meaning, and notes the percentage of words that seem similar (without as yet making any assumption as to why they are similar). One then shifts the B list one notch, so that the B word alongside A word 5 is no longer the word with the meaning of 5, but rather the one with the meaning 6. Again one counts the percentage of similar words across the two lists. One continues doing this - for reasons of computational feasibility, probably taking a sample of the 100 possible shifts, rather than trying to include all of them. If in general the pairing where the words correspond in meaning has a significantly higher percentage of similar forms

than the pairings where they do not correspond in meaning, then one has good reason to believe that the similarities are not due to chance. (A negative result does not necessarily point in the other direction, since there could be other explanations, e.g. that one is being so lax in identifying similarities that almost anything is treated as similar to almost anything else.) Note that this method enables one in principle to distinguish between chance and non-chance similarities. It does not, however, enable one to distinguish between non-chance similarities due to common inheritance and those due to borrowing.

Indeed, the crux is historical linguistics usually comes down to whether two similar corresponding forms in two languages are the result of common ancestry or borrowing, and it would certainly be nice to have a reliable method of distinguishing between these two situations. Unfortunately, no fool-proof method exists, although there are a number of factors that mitigate in one direction or the other. Basically, while there is probably no part of a language that cannot be borrowed, there are some parts of a language that are more likely to be borrowed than others, so that one can establish hierarchies of reliability of different sets of criteria. This has been done traditionally using non-mathematical methods, but could presumably be integrated into mathematical methods that would provide differential weighting to different kinds of criteria.

Some of the hierarchies of borrowability seem to be independent or largely independent of other considerations, in particular of social considerations. For instance, if one makes a distinction between basic vocabulary - names for concepts that are pretty much universal to the human condition - and cultural vocabulary - names for concepts that are specific to particular cultures - then it turns out that in general cultural vocabulary is more likely to be borrowed than basic vocabulary. So if we find two languages that share a lot of cultural vocabulary but little basic vocabulary, such as Arabic and Persian, then it is more likely to be the case that this is an instance of borrowing than of common inheritance, and indeed in this case we also have independent evidence that Persian has borrowed heavily from Arabic in the realm of cultural vocabulary. It should be emphasized that this is a hierarchical relation between basic and cultural vocabulary, and does not mean that basic vocabulary is never borrowed. While the English loans from French relate primarily to cultural vocabulary, they also include a handful of words from the basic vocabulary list like *mountain* and *round*.

But in some cases differences in degree of borrowability can be sensitive to social factors, so here particular care must be taken. For instance, at one time it was widely believed that vocabulary, in particularly cultural vocabulary, is more likely to be borrowed than grammatical structure, and this does seem to be the case where a community basically goes on speaking the same language but incorporates elements from some other language. But under conditions of language shift, it is not unusual for the speakers carrying out the shift to make a near perfect job of shifting their vocabulary, certainly in form, while carrying over grammatical features of the heritage language into the new language, thus giving rise to a variety of the new language that has little lexical borrowing from the heritage language but shows considerable grammatical borrowing from that language. Hiberno-English, the variety of English spoken in Ireland as a result of language shift from Irish, has rather few lexical loans from Irish, but considerable syntactic influence.

Of the components of the grammar and lexicon of a language, the one that is least subject to borrowing is inflectional morphology, i.e. the way in which a word changes its form to express different grammatical meanings, such as the English plural in *-s* (*cat* - *cats*) or the

past tense in *-ed* (*walk - walked*). Again, borrowing of inflectional morphology is not excluded - varieties of Greek spoken in Turkey through the early twentieth century had in some cases borrowed substantial parts of Turkish verb conjugation, applied even to words of Greek origin, and even English has borrowed bits of Latin and Greek morphology, such as *criteria* as the plural of *criterion* - but it is relatively rare. As a result, two languages sharing substantially their inflectional morphology would probably be considered genealogically related even in the absence of much shared vocabulary; indeed I used precisely this criterion myself to show that a language of Highland New Guinea, Haruai, is related to one of its neighbors, Hagahai, rather than another of its neighbors, Kobon, although it shares roughly equal percentages of basic vocabulary with both, because Haruai and Hagahai inflectional morphology are virtually identical, while Kobon inflectional morphology is very different. (This implies that Haruai has borrowed a substantial amount of basic vocabulary from Kobon. A relevant factor seems to be the phenomenon of word taboo, which has led to a rather rapid rate of lexical replacement in Haruai.)

While some of the components of a language are universal - all languages have phonology, lexicon, and syntax, for instance - inflectional morphology is not, and there are languages like Vietnamese that have little or no inflectional morphology. Thus, although inflectional morphology is a good test of genealogical relatedness when available, in many cases it is simply not available, and significantly for a long time Vietnamese was thought to be related to Chinese until it was shown, on the basis of careful comparison of the most basic vocabulary, to be a member of the Austroasiatic language family. This is an important methodological lesson. In linguistics, one sometimes comes across the claim that the same methods should be applicable whatever languages one wants to compare historically. The example of inflectional morphology suggests that this is not the case, and there might well be instances of pairs of languages that are roughly equally closely related to one another but where we would easily show this in the case of the pair with rich inflectional morphology, but not, or at least not so easily, in the case of the pair with little or no inflectional morphology. For some reason this strikes some linguists as strange, or even unacceptable, although of course such restrictions are well known in other disciplines. In archeology, you can use carbon-14 dating if you want to date organic material, but if the material is inorganic you cannot.

One criterion that has played a major role in discussions of the reliability of cross-linguistic comparisons for historical-comparative purposes is the regularity of sound change. This generalization, the so-called Neogrammarian principle, says that a given sound always changes in the same way in the same phonetic environment. In the example of the shift of Indo-European *p* to Germanic *f* discussed above, this would mean, for instance, that we find the shift not just in the first word discussed, *father*, but also in other words that began with *p* in Proto-Indo-European, such as *fish* and *foot*. In particular phonetic environments, the change might not take place or a different change might take place, and indeed in the shift from Indo-European to Germanic *p* did not shift after *s*, as can be seen by comparing Latin *spuere* and English *spew*. But there cannot simply be random words that fail to undergo the shift. Thus, if we find a word like *paternal* that appears not to have undergone the shift, then this is an indication that the word is not related through common ancestry, but perhaps by borrowing (and indeed *paternal* is a borrowing, from French or medieval Latin).

The principle of the regularity of sound change has played an important role in placing etymology on a scientific foundation, and has avoided a number of pitfalls that would have followed from mere superficial comparison of lexical items. (For instance, despite the similarity of English *much* and its Spanish translation *mucho*, it can be shown that they cannot share a common ancestral form, the similarity between them being purely the result of chance.) In addition, it has succeeded in showing that words that do not look alike are nonetheless cognate (related by common ancestry) precisely because of the regular sound correspondences; despite the dissimilarity between Armenian *erku* 'two' and the corresponding words in other Indo-European languages (including English *two*), they can be shown to be regularly related. Nonetheless, there are some problems with the application of the criterion, and although most linguists continue to accept its importance, there are linguists who have denied its importance (for instance, those working in the Greenbergian approach to historical-comparative linguistics). It is therefore worth considering some of the problems.

First, there are some items that simply violate the principle of regularity of sound change. In some cases, such exceptions can be explained through other principles. For instance, the expected past tense of the English verb *mow*, on the basis of regular sound change, would be *mew*, but in fact this form has been lost and the verb changed by analogy to fit the pattern of regular verbs, i.e. *mow* - *mowed* just like *walk* - *walked*. But in other cases, idiosyncratic factors seem to have intervened, so that one simply ends up with irregularities; for instance, the expected regular development of the third person singular feminine pronoun in English would have been not *she*, but rather *se* (pronounced the same as *see* or *sea*). Over a limited period of language change, the number of such irregularities is small, so that the overall principle of the regularity of sound change can still be used to constrain etymologies and avoid chance look-alikes. But over longer periods of time, the likelihood of irregularities increases, so that one would eventually reach a point where the number of irregularities is large enough to overwhelm the number of regularities.

Second, there is a problem in the other direction, namely the fact that when one language borrows a significant amount of vocabulary from another language over a limited period of time, it can do so by applying regular conversions to fit the alien phonology to its own, thus giving the appearance of regular correspondences. For instance, many languages that have borrowed heavily from Arabic show regular correspondences not as a result of sound changes but through having applied consistent modifications to adapt Arabic words to their phonological system. Malay, for example, lacks the sound *f*, and therefore regularly replaces Arabic *f* with *p*, e.g. *pikir* 'think' (from Arabic *fikr*), *pasal* 'chapter' (from Arabic *faSl*). Here, one must use other criteria (such as the fact that Arabic loans into Malay are typically cultural rather than basic vocabulary) in order to establish that these are loans rather than shared inheritance.

In reconstructing history, an overall account must of course not only establish common ancestry, but also try to date that common ancestry. In both archeology and genetics, such methods have been developed. In archeology, a number of processes have been identified that occur at a regular rate irrespective of other conditions. Or more accurately, a number of processes have been identified that have a constant probability of occurrence, which translates, when one has a large number of possible occurrences, into a regular rate of change. To take the best known, carbon-14 dating, there is a constant likelihood of carbon-14 undergoing radio-active decay, so that one can use this fact to determine the

age of organic material by measuring the percentage of carbon-14 that has decayed. (But see below for a caveat.) In genetics likewise, it is assumed that there is a constant probability of a particular gene mutating (different for different genes), so that the biological clock, just like the radioactive clock used by archeologists, can be used to establish chronologies. (Since one is actually dealing with probabilities, there is inevitably a certain range of error in the establishment of chronologies. In genetics, this range of error seems in general to be considerably greater than in archeology, and this does seem an important factor to take into account in evaluating the different chronologies. In addition, I sense an unease among some geneticists at some rates of mutation that have been proposed) The question therefore arises whether there is any comparable clock in linguistics.

Certainly there are many aspects of linguistic structure that will not serve, in particular general assessments of degree of similarity. French and Italian are both equally distant chronologically from their common ancestor Latin, but French has departed much more from Latin in almost every component of the grammar and vocabulary than has Italian. There is, however, one component that has been claimed to provide a clock analogous to the radioactive and biological clocks, although this claim remains one of the most controversial in historical linguistics. The hypothesis, usually known under the name "glottochronology", is that the basic vocabulary of a language is subject to a constant rate of replacement (or, more accurately, that the probability of replacement of basic vocabulary is constant). If this is true, then by comparing the percentage of shared vocabulary between two languages, one would be able to compute the date at which they separated. (It should be noted that this is distinct from the claim that basic vocabulary is less likely to be borrowed than many other components of the grammar/lexicon of a language. First, borrowing is only one form of replacement, the other being replacement from within the language. Second, for the application of glottochronology it is crucial that there should be lexical replacement, but at a fixed rate.) Different attitudes towards the reliability of glottochronology constitute one of the major battle lines within the community of historical linguistics. My own view is that the probability of lexical replacement in the basic vocabulary cannot be independent of social factors, on the basis of my work referred to above on Haruai, where the phenomenon of word taboo has led to a rapid rate of replacement even of basic lexicon, i.e. the probability of replacement of basic lexicon under such conditions is unusually high. (I would note that even some of the dating methods used elsewhere are not as completely impervious to environmental conditions as previously thought. For instance, carbon-14 dating relies not only on the constant rate of decay of carbon-14 (which seems to be incontrovertible), but also on the assumption that the proportion of carbon-14 in the atmosphere has remained constant over the Earth's history, or at least to the time-depth (around 30,000 years) to which this method is reliable; this second assumption is not only not necessarily true, but apparently false, a fact that has required some recalibration of dates given by carbon-14 dating.)

What this means for the chronology of historical events established by historical linguistics is that their dating is largely impressionistic, especially for those who reject the glottochronological hypothesis. This does not mean that anything goes - one can establish a range from very rapid to very slow attested change as a benchmark against which to measure claims about particular rates of change - but it does mean that most linguists are not comfortable providing chronologies parallel to those given by archeologists and geneticists.

In comparing dating methods in linguistics and in the other disciplines, especially archeology, it should be noted that archeologists find themselves having to use different methods depending not only on what object they want to date (for instance, carbon-14 dating works with organic, but not with inorganic matter), but also on the time-depth they are interested in: Thus, carbon-14 dating is reliable for dates going back about 30,000 years BP, but drops off in reliability rapidly for dates earlier than that. Perhaps surprisingly, linguists who advocate glottochronology have not in general been willing to consider comparable limitations on this method, i.e. the usual assumption is that glottochronology is valid back into time without limit (or perhaps, that human language has not been around long enough to reach the limits on the method).

One question that inevitably arises out of the above, especially for those who are skeptical of methods like glottochronology, is: What is the limit back to which historical linguistics can reliably go in reconstructing the historical processes that have differentiated languages? This is another point on which different schools of historical linguists differ, with some believing that the methods will take us back to the dawn of human language, others that the methods take us back at best to around 10,000 years ago. In this connection, it is important to consider what leads to this proposed cut-off point of 10,000 years ago. In principle, given that every historical reconstruction involves some uncertainty, and that reconstructions based on reconstructions in order to go even further back in time are based on even more uncertainty, there must come a point at which that uncertainty becomes so great that we no longer have anything reliable that can be reconstructed. But this still leaves open the question of when we reach that point, with some historical linguists believing that this point is not reached before we get back to a common ancestor of all human languages, others believing that it stops well before that point, thus leaving it an open question whether or not all human languages have a common ancestor. I want to address further only one aspect of this problem.

When more circumspect historical linguistics propose the date of 10,000 years, it should be borne in mind that this is to be interpreted as "under the best available circumstances". Proto-Afroasiatic is probably to be dated at around that time or even a little earlier, with Proto-Indo-European having a shallower time-depth, probably around 6-7000 years ago (though some non-linguists, in particular, are willing to go with somewhat earlier dates, still within the 10,000 year window). Both Afroasiatic and Indo-European have a number of characteristics that enable one to go rather far back in time. First, some of the languages in the family are attested from thousands of years ago, meaning that for those languages the time-depth is actually considerably less. Second, both families have a dense internal tree structure, i.e. at different levels of the hierarchical structure there are multiple branches. For instance, in order to reconstruct Slavic we can take advantage of the diversity provided by its three branches, and as we go back to Indo-European we can take advantage of the diversity provided by the existence of around ten well documented branches. Where such factors are not available, the time-depth to which we can go is much less - in the case of a language isolate like Basque, for instance, comparison of the different dialects and of the earliest extensive texts (from the sixteenth century) leads to a much shallower time-depth.

Earlier, I emphasized the need to consider the social environment in which people lived in order to assess the most appropriate way of reconciling genetic and linguistic accounts of population histories. This is equally true in dealing with linguistic reconstructions. With respect to the grammar of a language, comparison with evidence of social structures

provided by archeology is not likely to be significant, but the situation is very different when we turn to vocabulary. Reconstructing vocabulary, in the sense of reconstructing pairs of forms and meanings, necessarily reconstructs sets of meanings, concepts that must have been present in the proto-community, if the reconstruction is correct. Even if linguistics is poor at assigning reliable chronologies to its reconstructions that can be checked against plausible archeological correlates, its ability to reconstruct vocabulary means that this part of the reconstruction, more especially its semantic-conceptual side, can be used in order to search for plausible archeological correlates among archeologically attested cultures. The method, sometimes known as linguistic paleontology, is not without its dangers. Like any other aspect of linguistic reconstruction, the reconstruction of proto-meanings is not fool-proof but necessarily involves some uncertainty. To take an obvious example, the fact that one can reconstruct a word for a crop or an animal that is now a domesticate does not mean that the equivalent in the proto-culture was domesticated. The method also only provides positive evidence; from the fact that a particular concept cannot be reconstructed, one cannot draw any firm conclusion - the proto-language may have had the concept, but the word may have been lost (i.e. replaced) in so many of the descendant languages that it can no longer be reconstructed.

5. Conclusions

In this position paper I have tried to set out some of the points of contact between work in historical linguistics and in genetics, with reference where relevant to archeology and anthropology. First, I considered possibilities of reconstructing human prehistory. The general line that I have taken here is that each discipline must reach its own conclusions, since there is no guarantee that, for instance, linguistic and population genetic classifications will coincide, but that these conclusions must then be integrated into a consistent historical account taking account also of the cultural setting in which the people lived. Second, I have tried to compare the different methods used in the different disciplines, noting where these reflect real differences in the objects of investigation that may preclude direct transfer of methods from one discipline to another, but also trying to identify those areas where such transfer is more promising. To the extent that I am right, it now only remains to do the work.