# *Chapter 16*

# Interdisciplinary Indiscipline? Can Phylogenetic Methods Meaningfully Be Applied to Language Data — and to Dating Language?

Paul Heggarty

A number of recent papers have sought to apply to language data various phylogenetic 'tree-drawing' techniques initially developed for uses outside linguistics. The reaction from many historical linguists, however, has typically been critical, if not outright hostile. This paper explores, and aims to explain, why it is that there has been such a long-running failure to reach a consensus between linguists and specialists from other disciplines, notably genetics and archaeology.

We consider linguists' fundamental concerns as to how non-linguists go about using language data; especially whether (and if so, how) one can meaningfully use such phylogenetic analyses on language data, interpret their results, and attempt to put dates on particular nodes in the trees. We look into certain aspects of the very nature of language that it is crucial to bear in mind in order to handle language data appropriately for these purposes, but which many linguists feel are not truly appreciated by non-linguists. These aspects are: language's inherent susceptibility to powerful external forces which vary tremendously through history; the nature of language data, and what it entails for how those data can meaningfully be compared and measured; and finally the nature of change and historical development in language, critical to how we are to interpret any parallels we observe between languages, not least for the purposes of dating.

It emerges, moreover, that these same characteristics of language change also challenge linguists' own 'established' dating of Proto-Indo-European by means of the so-called 'linguistic palaeontology', and that this issue is in truth much more open than Indo-Europeanist linguists generally admit.

**An inter-disciplinary problem – Why?**

Few articles in recent years have provoked such controversy in historical linguistics as Gray & Atkinson (2003) and Forster & Toth (2003). We do not propose here to give detailed or specific critiques of these, which are available elsewhere, but look at the broader problem behind the failure to achieve an interdisciplinary consensus on how phylogenetic methods, initially developed for applications outside linguistics, can be applied to language data, and particularly to dating languages.

The tone of commentaries from some historical linguistics circles on papers such as the above has ranged from the sceptical or bemused to the outright hostile, disparaging and dismissive, as if there were a few home truths that specialists from outside linguistics have yet to take fully on board. Their detailed criticisms revolve around two general aspects of language that are frequently poorly grasped.

The first is the nature of language as a **social phenomenon**. If languages can be so informative of history, then this is only because language in general, and all languages individually, are inherently susceptible to being moulded by the forces of history. What makes languages a function of the history of their speakers are certain social, cultural and political factors: contact, isolation, population size and density, conquest, and so on; and all the forces that influence those in turn. Such forces are themselves anything but stable over time; necessarily then, nor is their impact on language change and divergence. This is the key principle behind the rejection by mainstream linguistics of any dating mechanism that assumes that the rate of change is constant.

It has sometimes been claimed that Embleton (1986) 'rehabilitated' glottochronology to some extent. Despite her other improvements, however, her key innovation of 'borrowing rates' simply continues the original and flawed glottochronological assumption. For while she allows her borrowing factor $b$ to vary from one language pair to the next, for the dating method to work it still cannot be allowed to vary over time. Embleton (1986, 102–3) works with a borrowing

rate of '8.16% per millennium' for English from Danish, for instance. The problematic assumption remains, then: that change, specifically in this case borrowing under contact influence, is necessarily steady over time. This is plainly at odds with known historical and linguistic fact. There clearly never was even a single millennium throughout which Danish had any significant impact on English. A huge impact it did have, to be sure; but this was everything but constant 'per millennium'. It had precious little impact at all until the short age of Viking expansion, then a huge impact over a period of just some three centuries, when half of England was the Danelaw, before that intense influence then faded rapidly, to end pretty much abruptly and for good. Far from being an atypical case, this punctuality has been pretty much the normal scenario for contact influence between European languages, not least that of many of the Romance and Germanic languages on each other, (Norman) French and English being another case in point.

Language as a 'social animal' also determines the processes and patterns by which languages *diverge*, particularly into dialect continua by the wave model and other processes. Despite the lip-service paid to them, these are still all too often overlooked and their true nature misconstrued. Speciation in particular is a very poor analogy for these processes. And even if it were not, the idealized assumption that splits are punctual is more problematic in linguistics because its time-scales are so different to those of genetics: margins of error in estimating split-dates are consequently much greater. All this has led to confusion as to exactly what the abstract concept of a language 'split' actually corresponds to in real languages — yet this 'event' is exactly what phylogenetic models both aspire to put dates on, and use to calibrate their dating mechanisms in the first place.

The other main grounds on which linguists have rejected given applications of phylogenetic methods as plain 'bad science' have to do with the general approach to language **data** as input: how they are analyzed and processed, and what they are claimed to demonstrate.

The onus still appears to be on linguists to explain adequately these two key aspects of language to specialists from other disciplines. Unfortunately there is not space here to do so fully for both, and this paper has had to choose to focus on the second. We consider and illustrate a number of aspects of the very nature of language data and language change that have critical consequences for the use of phylogenetic methods. We shall also see, however, that among these same inescapable characteristics of language are some that also question the grounds on which some historical linguists have based their own estimates for the date of the first Proto-Indo-European split, specifically the claim that it must necessarily have occurred *after* the emergence of certain technologies such as wheeled vehicles.

**Applying phylogenetic methods to language: a multi-stage process**

Applying phylogenetic methods to language is a process that it is helpful to see as a series of discrete stages.

(a) The *encoding* stage: getting from real languages to some expression of the relationships between them in the form of numerical or state data, so that those data can then be used as input to phylogenetic methods.
(b) The *representation* stage: applying phylogenetic analysis methods to extract from those numerical and/or state data a signal that is converted into some useful form of representation, usually two-dimensional graphical ones such as trees or networks, which synthesize and 'collapse' what are often highly complex multi-dimensional relationships in the signal.
(c) The *interpretation* stage: assessing those tree and network representations to extract from them what they actually mean for real languages and their relationships through time.

As Embleton (1986, 3) makes clear, we first have to confirm that this whole process is viable, by testing it against cases for which we already know the answers, where we are fortunate enough to have copious historical knowledge available. Once we can trust the process, we can then progress to the ultimate point, to make use of it in cases where we do *not* have sufficient knowledge to find out the answers by other more traditional means.

The need for the different disciplines to work together in this 'new synthesis' is clear from how these different stages implicate specialists from the respective fields. Encoding, certainly, cannot be successful without a deep understanding of the nature of language data and change. Interpretation, likewise, needs an awareness of how real languages behave through time, and of the processes and scenarios by which they diverge and converge. Yet it is equally crucial for interpretation, and for the representation stage that precedes it, to understand in detail precisely how the phylogenetic methods work. This calls for a particularly sound grasp of statistics, significance, and the various alternative configurations possible in the methods that can effect important adjustments to the trees and networks they produce, suited to a range of different applications and data sets.

And what of dating, and its place in this process? Dating is not just an interpretation, but an extra stage in that it adds further assumptions, about rates of change over time. Glottochronology's assumption of a constant rate of change in core vocabulary has been demonstrated to be fundamentally flawed; it remains to be established whether Gray & Atkinson's calibration and rate-smoothing can give us a viable alternative, or whether their assumptions too are invalid.

**The nature of language data**

Linguists have objected to what they characterize as an undisciplined, unscientific approach to using language data, ignoring key characteristics that limit the suitability of those data as input to phylogenetic methods designed for what are actually very different data from other disciplines. A tacit assumption is often made that language offers data that can be used just like many data from the natural sciences.

The most fundamental point to bear in mind here is that language is not a creation of the natural world, but of the human mind. There may exist some high-level 'universal grammar' type constraints, but these are not of concern at the level of the language data used as input to phylogenetic methods. In such data, the human mind can be perfectly happy with gradations, irregularities, and a whole host of complex and indirect relationships. As any linguist who has sought to quantify data from any field of language knows all too well, real language data are by no means always amenable to clear-cut representation either in numerical terms, or as oppositions between two or more discrete states all 'equally different' from each other. Analyses and measurements of language data, even in a single feature, are typically nothing like as precise, unequivocal and reliable as a carbon-14 reading, for instance. True, carbon datings are not quite as simple as was first thought, and it was soon realized that they need to be adjusted in order to take certain other environmental variables into account. Yet the point, and the qualitative difference with language data, remains: these data too are nonetheless such that they can be measured and processed to a degree of accuracy and reliability *sufficient to achieve a consensus among specialists* as to the robustness of the results thus 'corrected'.

Since language data are often not reliably clear-cut, it follows too that *changes* in language from one state to another are not discrete either, but eminently gradual questions of degree. All these difficulties increase exponentially when one tries to match up multiple different systems in different languages in order to quantify precisely how different they are to each other.

Of course, language data will always need to be simplified to some extent to make them suitable for numerical processing. Yet on many occasions this simplification is taken so far, in order to fit data into far too rigid a mathematically-led model, as to render the whole process meaningless. The data end up so distorted and correspond so indirectly to the linguistic reality that as 'features' or quantifications they misrepresent it completely, and are actually worse than no data at all. For a stark illustration, see the critiques in Heggarty *et al*. (2005) and Heggarty (forthcoming, §2.2.1) of Nerbonne *et al.*'s (1999) dialect phylogenies based on comparing word strings by 'edit distance', on the model of applications in the natural sciences. For computational convenience their simplistic adaptation ends up treating the least change in sound quality as a 'mutation' rated as twice as significant as the loss of the whole sound. Every sound is effectively rated as more similar to silence than to any other speech sound. Take the example of RP English *mother* against Standard German *Mutter*, where there are certain differences in the pronunciations of the last three of the four sounds: ['mʌðə] vs ['mʊtɐ]. A 'mutation' approach like that of Nerbonne *et al.* would end up measuring these words as *more* different to each other than either is to a sound sequence such as [im]. Such results are all but meaningless as measures and representations of the linguistic reality — not that this comes as a surprise, because an analysis model in terms of mutations and edit distance so completely misrepresents the nature of language differences and processes.

The term 'mutation' is often heard misapplied by proponents of tree-drawing models too, presumably because it fits with the applications in genetics that they were originally designed for. The superficial analogy is a dangerously misleading one. Language data are by no means all made up of discrete slots and states, all equally different from one another. Sounds, affixes, words, meanings, grammatical contexts, and no end of other linguistic variables, are not carbon isotopes, nor the distinct A, C, G and T bases of DNA. Likewise, the real linguistic processes that bring about change and determine our language data at any given time are not at all sudden 'mutations' or discrete switches from one state to another, but typically subtle, gradual shifts in the relationships between multiple different variables on various levels.

Linguists' countless objections to particular lexicostatistical data are all just instances of this general principle as it applies to lexical semantics. The types of change that result in the meaning-cognacy mismatches that form lexicostatistical data are not 'mutations', the sudden complete 'loss' or 'replacement' of a given word. The cognate is often not lost at all in

the language, but just drifts to a very closely related meaning slot, as with German *Hund* (*dog*) vs English *hound*. Had Swadesh (1952) not drawn up his list in English in the first place, it would be a close call between *small* and *little* as the word that best corresponds to the intended meaning-slot. Such debatable calls in multiple languages at the same time can have serious repercussions for the 'matches' between them. How should English really be matched up with Danish, for instance, which has singular *lille* but a plural *små*? Indeed this case is even more complex than meets the eye, and only serves to illustrate how even a yes/no cognacy judgement between two superficially similar-looking forms can be far from a clear-cut question. According to the *Oxford English Dictionary*'s (*c.* 2000–) etymologies, the true relationship between *small* and *små* is 'doubtful'; and while in principle *little* and *lille* actually go back to *different* roots, they also raise strong suspicions of contact influence.

Contact and borrowing are critical problems for lexicostatistics, whose proponents take highly inconsistent approaches to them — another instance of a rather undisciplined approach to using and interpreting language data, even by linguists. Gray & Atkinson (2003), at least, cannot be faulted for relying on data analyzed not by themselves but by linguists, albeit advocates of lexicostatistics: Dyen *et al.* (1992). Yet this hardly means that all linguists are convinced of how meaningful and reliable such data are for producing precise phylogenies, given the troublesome nature of language data. For the *small* meaning slot, for instance, Dyen's data and cognacy choices put English not with any other Germanic language, but alone with Slavic. For the *Oxford English Dictionary* (*OED*), however, the relationship with Slavic is 'somewhat uncertain'. And even if the Slavic words *are* cognates with *small*, it is unclear whether this modern meaning match might actually be the result of parallel meaning shifts that happened *independently* in each branch, rather than the preservation of the original form-to-meaning relationship. Indeed, in the *OED*'s earliest records for *small*, in Old English it originally had a more specific sense of *narrow*, as still in German *schmal* (one form that we can say is definitely cognate).

Linguists' protestations that there is something particularly awkward about language data are often met with scepticism from specialists in other disciplines. Geneticists may retort that in trees based on language data, the phylogenetic signal is typically extremely strong by the usual standards of genetics, and can be frustrated at linguists' hesitation in venturing claims about real language histories on the basis of such clear phylogenies.

In fact there is truth in the claims from both sides, but importantly because each focuses on a very different aspect of what constitutes 'messy' data. The messiness in genetics is not so much in getting from the DNA to an encoding of it; it is simply that the picture the DNA gives is messy in the sense that the phylogenetic signal it contains is often weak, and can be hard to interpret conclusively. In linguistics, the opposite holds: the messiness in language data is not so much in the picture we get once we have encoded the data, it's in deciding what is a meaningful encoding of them in the first place. The signal may be stronger, but we are less confident that it actually does necessarily mean what we would like it to mean (such as a period of shared ancestry). Linguists' experience leaves them acutely conscious that language data are inherently open to multiple different analyses and interpretations, and it is this that has undermined and frustrated efforts to come up with workable encoding methods. Those proposed so far, such as lexicostatistics, force on language data what we know to be a seriously simplistic characterization, and we are left in doubt as to whether this encoding really is a particularly meaningful representation of the actual relationships between the real languages that we are hoping it will help us investigate.

**What does an apparent correspondence mean?**

For the particular task of working out the phylogeny (and from that the datings?) inherent in a set of data features, it goes without saying that however sophisticated and potentially useful an analysis tool may be, unless correspondences in the data features really *are* valid indicators of common origin, its output can have no meaning as a true genealogy.

At times some linguists too have not been immune to the temptation of granting themselves great leeway with their data and their interpretations of them, Greenberg's (1987) 'multilateral comparison' technique being the most notorious case. Abandoning due scientific rigour may offer a way of claiming discoveries novel and significant enough as to make big waves outside one's discipline; it is also a way to question if not discredit one's work and reputation amongst colleagues in the know.

How is one to know how much phylogeny can safely be read into a particular correspondence that at first sight may seem particularly striking? For a parallel in archaeology, take the issue of whether other civilizations or cultures had contact with and/or settled South America before Columbus. Granted, certain parallels can be found in various cultural attributes such as pottery technologies and styles, or the construction of reed boats, but they are hardly enough

to convince us of a necessary link with Oceania or Egypt; independent and parallel development offers an alternative explanation. Or in biology, as McMahon & McMahon (1995, 158) point out, just because a mammalian mole and a marsupial mole have many morphological similarities does not mean it is correct to assume they are closely related genealogically.

There are a number of inherent characteristics of language that conspire to produce a great many apparent correspondences of this type in language data, arguably to a degree far more misleading than in other disciplines. Certainly, a linguist's training is a long experience of constantly realizing how much more complex language data are than meets the untrained eye. A non-specialist — in some cases, even a linguist who is not a specialist in the particular languages concerned — cannot simply 'eyeball' the data and reliably tell apart which correspondences do or do not necessarily go back to a time of shared history before a split. To have any value, phylogenies drawn from feature-based comparisons *must* be based on features that are powerfully diagnostic of shared origin, as opposed to the many others which, despite appearances, do not necessarily mean anything of the sort. Such are the complexities and pitfalls of language data that the only way to be sure is through informed, painstaking research.

Four key characteristics of language data to consider here are that many features of language:

- tend to proceed along the same typical steps or *pathways of change* in any language;
- offer precious *few alternative states*;
- involve simply *re-interpreting* in new ways *existing resources* already available to a language;
- are not fully independent of each other, because of the *systemic* nature of language structure.

Among other problems they raise, these characteristics entail that changes often occur *independently* and *in parallel* in different languages, especially when they are tipped into a relatively sudden whole-system change. Superficial correspondences in features of language that are subject to these characteristics are therefore of very limited significance as evidence of a period of common history during which they arose, since they fail to exclude convincingly chance and systemic reasons as alternative explanations. We shall now look at each characteristic in more depth.

**Typical paths of change**

A common misapprehension about language change is that the changes are by their nature random, on the model of mutations. This is correct only in part. It is true that, other than in known cases of direct contact influence, we usually cannot explain why in any particular language a given change occurred when it did, nor why that change did while other possible ones did not. However, the changes a language undergoes are not so much random as just a randomly 'selected' subset of a larger number of well-known, typical steps of change. These changes *do* follow clear types, and contrast with a vast number of others which in principle are highly unnatural, and in practice do not occur. Indeed it is only this predictable aspect of language change that enables us to 'reverse' it so as to make reconstruction possible at all.

It is essentially random, for example, whether a language allows words to end in the [l] sound; and if so, whether that sound in this position does or does not change during a particular period. What is not random is that if its pronunciation *does* change, then it is overwhelmingly likely to change along a particular known path, for there are very few natural directions for a word-final [l] sound to change in. The first step is typically from clear [l] as in German, to dark [ɬ] as in most varieties of English, and independently in European Portuguese. Next this dark [ɬ] typically turns into [w] or [u], as occurred independently in Cockney, Glaswegian, Brazilian Portuguese, Polish and some varieties of Bulgarian. This can then even change to [o], as happened in French (*cheval~chevaux*) and in Serbo-Croat (*Belgrad→Beograd*). That a sound change like this occurs in two or more language varieties is no necessary indicator whatsoever that it happened during a period of common history, for the innovation so often occurs independently anyway, as all these known cases show.

There are similar 'typical paths of change' in the grammatical system too. In *I saw a cat today; the cat was black*, reversing the articles would be a grammatical mistake. This is because English 'grammaticalizes' definiteness: i.e. it is compulsory to make the *a* vs *the* distinction where appropriate. Other languages, including Latin and most Slav languages, simply do not bother to make this distinction, and have no words to perform the functions of English *a* vs *the*. The historical process of acquiring or 'grammaticalizing' such a distinction from a previous state of not having it has occurred repeatedly and independently in many European languages: in modern Romance, Germanic and Bulgarian, for instance. Again, definiteness marking is no sound evidence for historical relationships.

**A limited number of states and re-interpreting existing resources in a language**

Moreover, whether a language does or does not grammaticalize definiteness is a variable with only two

basic states (though the details are actually much more complex), either of which is entirely natural for a language. So this simple feature alone can hardly be taken as a reliable or statistically significant indicator of the relatedness of any two languages. Likewise for the *position* of definite articles: either in front of the noun they refer to, or after it (or both). French and English share the former, Romanian and Danish (with some complications) the latter, but these correspondences do not in themselves prove that they arose during a period of shared history — indeed, again we know full well that the *opposite* genealogical relationships hold between these languages.

Furthermore, with the exception of the borrowing of a specific loanword, in any change (including many other contact-induced changes such as calques) the new form or arrangement can only come about by using in a new way resources *already available* in that language. Normally this means that there is only a very limited range of existing resources with a meaning suitable to be put to the new use. For indefinite articles, a typical source is the numeral *one*, while for definite articles it is demonstratives like *this* and *that*. Not only have multiple languages grammaticalized definiteness; in most cases they have even grammaticalized exactly the 'same' words in this use, again entirely independently.

This means also that superficial correspondences can be even more deceptive when languages that change in parallel are ones that have previously diverged from a common origin, for they can thus end up not just with similar systems but with specific *forms* that are ultimately related too, such as English *a(n)* and French *un*. Even this form-to-meaning correspondence is no evidence that they actually arose *as articles* before the languages split; again, on the contrary, we know they did not.

**Whole-system change**

Many instances of language change are not isolated and independent of each other, but have knock-on and cumulative effects that together can help tip a language into a more general transformation, as when a language switches the whole character of its sound or grammatical system from one type to another. This typically occurs as per a 'punctuated equilibrium' model (see Dixon 1997), in bursts of faster change in that area of language structure, separated by periods of relative system stability, or at least much slower change. Here we have another factor causing instability in rates of change through history, only this time not an external but essentially a language-internal one.

What is more, wholesale system change in one area of language structure may accelerate change in other areas too. Generalized *phonetic* attrition of word endings, for instance, can accelerate the collapse of *grammatical* case systems (or vice versa), as in the development of Vulgar Latin into the early stages of the Western Romance languages. And by causing originally distinct words all to end up pronounced the same, *phonetic* attrition can also accelerate *lexical* replacements, in order to disambiguate the new homonyms.

Again, it is with languages related at an earlier stage that such whole-system changes can be particularly misleading. Related languages not only share many form-to-meaning correspondences, but they are also likely still to be at a similar stage in many typological characteristics, all just 'waiting' for some trigger to set them off along the next typical stretch of accelerated cyclical change. Despite all their immediate ancestors being highly inflectional and fusional languages, for instance, most modern Indo-European languages have become much less so, all but entirely independently of each other down each branch.

**Dating by linguistic palaeontology?**

One proposed linguistic method for locating and dating the origin of languages is the so-called 'linguistic palaeontology', the one on which the proposed date of *c.* 6500 BP is largely founded. To start with, it is well to remind readers that linguistic palaeontology as a method has never attained any status as linguistic orthodoxy as a dating mechanism, for just like glottochronology, it reposes on assumptions about language change in lexical semantics that the majority of historical linguists do not consider tenable.

For Dixon (1997, 49), for example: 'What has always filled me with wonder is the assurance with which many historical linguists assign a date to their reconstructed proto-language … it does seem to be a house of cards'. Sims-Williams (1998, 510), meanwhile, considers linguistic palaeontology specifically in the case of Indo-European, only to find that its arguments rely on 'unprovable assumptions about the absolute chronology of the prehistoric IE sound changes'. He is bound to conclude that

> To sum up, then, there seems to be no reliable way of establishing the upper limit for the breakup of PIE. If Renfrew were able convince his fellow archaeologists that the first farmers were the only possible bearers of PIE, then philologists could probably explain away all the shared vocabulary that has seemed to imply later phases of civilization.

How is it, though, that linguists could 'explain away'

shared vocabulary like the *axle* or the *wheel*? Again, it is typical paths of change that provide the alternative explanations. We have looked at examples of these paths, and of the many resulting cases of independent parallel change, so far only in the grammatical and sound systems of languages; it is often overlooked that in lexical semantics too there are a number of typical paths along which word-to-meaning relationships shift over time. Almost all changes fall into one of a number of types well known to linguists: extension, specialization, pejoration, and so on.

Underlying all of these is that when one word comes to replace another in a particular meaning, that new word (unless it is a loanword) has not simply been generated out of nowhere, but was a word that either already existed in the language in a very similar meaning, or was coined from an original word by adding an affix. The word for *tomorrow* in various languages (including German and Spanish) has independently come by extension from the existing word for *morning*; likewise the *money* word has been repeatedly taken from *silver* (French, Latin American Spanish); *drink* from *take*; and so on. This matters not just for the lexical data for many phylogenetic studies, but also for linguistic palaeontology.

Firstly as a means of locating the Indo-European homeland, the key weakness of the method is that peoples who move into new areas frequently apply their existing lexical roots to any similar-looking species in the new environment, so a Proto-Indo-European root for a species name such as *beech (tree)* is by no means conclusive for locating the PIE environment.

For dating, meanwhile, linguistic palaeontologists appeal to the roots that it has been possible to reconstruct for certain technologies, in particular wheeled transport, as proof that the PIE split cannot be dated before those cultural developments. Again though, this is by no means the only, inescapable conclusion from such language data. Once more, the spanner in the works is language's tendency to base new meanings, lexical as well as grammatical, on new interpretations of resources already present within the language, simply extending them to new senses.

First consider a modern example: the new sense acquired by English *mouse* in computing. That the German equivalent is *Maus* and the Dutch *muis* is most plausibly attributable to contact of some sort, on the model of the initial meaning extension in English (quite how best to analyze the contact process is debatable). This new *sense* was acquired first in one language, English, then 'borrowed' into others, at least fifteen hundred years after the languages 'split'. Yet since the two languages still have cognates that are phonetically similar, the result is indistinguishable from shared origin.

New cultural items are well-known to be eminently susceptible to contact influence, indeed none more so than technological terminology for transport (examples are legion with seafaring terms borrowed between the languages of maritime western European peoples). There is every reason to suppose that terms for wheeled transport could indeed have spread through some form of contact — and this too could have been just by calquing new senses and meaning extensions, not necessarily borrowing new words. An entirely plausible linguistic scenario would run as follows. Many centuries after Proto-Indo-European first split and began diverging into dialects and languages, an existing Indo-European root for some meaning like *pole* or *post* acquires the additional sense of *axle*, initially in one dialect or language whose speakers are the first to use the new technology. Other early Indo-European dialects or closely related languages still have cognate forms either identical or instantly recognizable as just dialectal variants of this *pole* word, that now also means *axle* in one dialect. So there simply is no 'new word' to borrow, only a new sense — *axle* as well as *pole* — to extend their existing cognate lexemes to. As the innovation spreads, every dialect starts using its own variant of the *pole* word also to mean *axle*. Another typical path of language change in lexical semantics, specialization, may well then narrow the meaning of this term to *axle* only, with the original *pole* meaning taken over by a near-synonym (*stick*, *post*, etc.) to disambiguate the now unhelpfully broad meaning of the original *pole/axle* word.

When one recalls how change in lexical semantics typically proceeds, the supposed proofs of linguistic palaeontology soon begin to seem highly speculative. *Bronze* might have referred originally perhaps to a colour, or to the rock that was first used as bronze ore, or to a number of other related meanings. Linguistic palaeontologists themselves are very happy to envisage such well-known types of semantic changes when it suits them (i.e. when they happen to have been able to reconstruct forms that look like possible more remote cognates), as per Beekes's (1995, 37) suggestion: '"Gold", *$\acute{g}^h$(o)l(H)- … perhaps derived from the word for "yellow"'. Indeed Beekes (1995, 37) not only takes care to specify two forms for *wheel* *Hrot-$h_2$ and *$k^we$-$k^wl$-os*, but immediately goes on to explain 'The root *Hret- meant "walk", *kWel- "turning"'. We know that these words are derivations of some sort. Even if we have indications in the data of how far back the derivation happened in the history of particular sub-branches of Indo-European, or indeed back at the Proto-Indo-European stage, we do not know the pre-

cise meaning the derivation had at the time. Indeed, it is indicative that Beekes's example languages that preserved reflexes of *$Hrot\text{-}h_2$ vary in the precise meanings: *wheel*, *wagon*, and '*wheel, circle*'. And just as they can borrow and calque long after they split, languages can also make derivations that are still cognate even after many centuries of divergence.

Indo-Europeanists certainly did excellent detective work to reconstruct the words. But reconstruction takes us back only to assumed *forms* (in the linguistic sense of a word seen only on the level of sound), not necessarily their *exact meanings* in the proto-language. Indeed, reconstruction in so many cases is specifically compelled to allow explicitly for indirect meaning matches. We can reconstruct an ancient *form* whose modern reflexes now occupy the meaning slot of a current technological term; but this does not of itself necessarily entail that the proto-form already had that technical *sense* at the time of the proto-language, nor thereby that the language split can only have happened after the technological innovation. With Proto-Indo-European we are working at such a depth in time that the exact meanings of the reconstructed forms are by no means certain, particularly in areas of the lexicon subject to very significant technological and cultural changes over the millennia.

Nor are our reconstructed forms reliable to any great phonetic precision either, at such a great remove. While reconstructions may appear fairly sound to a more abstract phonemic level, precious little *phonetic* detail at all is absolutely certain. What this means is that we are unlikely indeed to have enough phonetic resolution to distinguish clear signals of either one of the two possible scenarios by which early Indo-European languages could have ended up with terms for new technologies that now appear shared, even though those meanings arose only long after the initial PIE split. Firstly, as discussed so far, there could have been a calque only of the extension of the meaning of a particular word, whose cognate forms were still clearly recognizable as such in the early Indo-European languages, despite some regular sound changes since an initial PIE split possibly many centuries earlier. Alternatively, a wordform itself could have been borrowed, either from an outside language, or from one early Indo-European language in which it still survived while others had lost its cognates. In each Indo-European language that borrowed this word, it would undergo different phonetic adaptations, and these are largely determined by precisely the same characteristics of its individual sound system that lie behind the historic sound changes that differentiated it from its sister languages in the first place. When modern Romance languages borrow a word from a language either within or outside the family, whatever the precise pronunciation of any r-sounds in the lender language, each borrowing language will adapt it to pronounce it precisely as per its normal modern reflex of Latin r-sounds: a uvular fricative in French, a tap in Spanish, an [h] sound in many accents of Brazilian Portuguese, and so on. That is, much phonetic adaptation mimics precisely the effects of regular sound changes.

Whichever of these two scenarios happened, the output can look confusingly similar. For the natural phonetic adaptation of a loanword into a borrower language's phonological system can further muddy the waters, not least at the time depth of PIE. And in highlighting this, again the scenario we propose hardly has to appeal to some one-off that can be dismissed by linguistic palaeontologists as relatively unlikely to have actually happened. Quite on the contrary, linguists are bound to recognize that such phonetic adaptation of loanwords is an entirely automatic process, and absolutely what we should expect by default.

Only if a language has undergone particularly significant change in the phonological status of the sounds concerned, or lost all traces of the cognate root in any related meanings, is a borrowing likely to enter in a different form to an existing cognate. Between languages that are still closely related, most cognate roots, and derivations from them, remain immediately recognizable as such, wherever any sort of meaning connection can be made. Alongside the *mouse* words in Germanic, Romance words for *train* illustrate an even more involved mix of borrowing, meaning extension, regular sound change and phonetic adaptation. Moreover, these are languages which we have good reason to imagine have been diverging at perhaps an abnormally rapid pace over the tumultuous last two millennia. So we can well imagine similar borrowing processes occurring between early Indo-European languages for at least as long after they first began to diverge.

In such cases where recognizable cognates still exist, rather than a true borrowing of the wordform itself, it is an entirely normal scenario for only the meaning extension to be borrowed, and such calques can be all but impossible to distinguish from true common origin at the depth of Proto-Indo-European. For the phonetic signal left in the data, calque processes are a shortcut that reproduces or 'inherits' in one fell swoop many of the relevant sound changes between the lender and borrower languages since they diverged, such as the two millennia of sound changes in the *mouse* word in the various Germanic languages.

Linguistic palaeontologists sometimes try to object that the *mouse* example is not pertinent, and

insist on a case-by-case scrutiny of their reconstructions, but this misses the point. It is not the details of a reconstructed Proto-Indo-European form that are in question, any more than they are for the Proto-Germanic form of *mouse*. What *is* in question is the series of assumptions that linguistic palaeontologists make: about what a given proto-form does or does not necessarily entail for the culture and date of the proto-language — by virtue, in their view, of the simple fact that they have been able to reconstruct it; and about how and when that form acquired the meaning that they now read for it. The Germanic *mouse* and Romance *train* words are just two examples of general and completely normal processes of borrowing and calque. The *mouse* example is relevant particularly as an illustration of how it is in the very nature of those processes that they conspire to render their output 'reconstructible' too, indeed indistinguishable from common origin. This applies *a fortiori* where all the output we have left to go on is whatever imperfect signal of it we can reconstruct many millennia later.

Advocates of linguistic palaeontology tend to present the alternative scenarios to theirs as unlikely one-offs that require a special justification; whereas in reality, in the specific case of coining words for new technology, those scenarios are precisely what we should *expect* to apply by default. All the processes we have cited as offering alternative explanations — derivation, borrowing, phonetic adaptation, extension, specialization, and so on — are the classic 'typical paths of change' in lexical semantics. Far more than that: the one class of words most subject to precisely those processes is none other than terms for new technologies. And the one scenario where it is most difficult to distinguish them from common origin is when they operate between closely related languages, still at a fairly early stage of divergence.

As soon as one abandons preconceived ideas about Proto-Indo-European necessarily being dated no earlier than 6500 BP, and envisages an alternative scenario in which terms for new technologies became necessary at a time when the family was already in the early stages of breaking up into a dialect continuum and/or still closely related languages, the boot is on the other foot. The terms for the new technology can hardly just be plucked out of nowhere: they *must* come either from borrowing (with automatic phonetic adaptation); or from the existing resources within the languages, whether by derivation or by some realignment between existing lexemes and related meanings. In both cases, the trace we would be left with now, many thousands of years down the line, will look temptingly reconstructible and dangerously indistinguishable from true common origin.

On whom does the onus of proof rest, then? Linguists who reject linguistic palaeontology as a reliable dating method are making no claims either way about what these particular language data — the terms for new technologies found in Proto-Indo-European — can be used to prove. They exclude neither scenario; on the contrary they recognize that the nature of the evidence is not such as to demonstrate conclusively either Renfrew's long chronology or the traditional shorter one. It is the linguistic palaeontologists who are making assumptions about what the mere fact that we can reconstruct a given lexeme means for the culture that spoke it. If they wish to use given reconstructions to argue for a particular chronology, the onus is on them to prove why their assumptions are valid. More specifically, the onus is those who hold that earlier dates are unacceptable to prove why the assumptions of linguistic palaeontology are so *uniquely* valid, and why other explanations — which moreover are perfectly in line with typical paths of change in coining terms for new technologies — are not plausible. For many historical linguists, linguistic palaeontologists have signally failed to provide such proofs, as attested by the above citations from Dixon and Sims-Williams. So while linguistic palaeontology can offer one possible historical scenario that can explain given patterns in the linguistic data, it has certainly not been able convincingly to exclude other linguistically and historically plausible scenarios that could have left us with exactly the same patterns in the data.

Now of course there may be other 'non-chronological' arguments in the linguistic data — in the network of relationships between the early Indo-European languages, for instance — which may not be neatly compatible with the geography of Renfrew's scenario. It is certainly not linguistic palaeontology, though, that can offer any conclusive linguistic evidence for why we may not put the first Indo-European split before the *wheel*, the *cart*, and the *horse*.

Quite how many lexemes and how much certainty in their reconstruction are enough to tip the balance between linguistic palaeontology's explanation for the shared terms, and the alternative scenario that the split came after the inventions, is another question open to our interpretation of language data and how languages change. Certainly, many linguists' objections to the earlier dates of the proposed long chronology tend to be expressed very subjectively: they are simply too 'hard to believe' for that writer's personal, unquantified impressions of how much change is feasible over how long a time-span. Yet if we have no certain dating mechanism, and no reliable quantifications of degrees of change over time, then there is no objective basis for deciding at which

particular date one crosses from the plausible to the 'hard to believe'. Maximum and minimum observed rates of change offer a possible yardstick for a *bracket* of plausible dates, as suggested for change in phonetics by Heggarty (2000). One of Gray & Atkinson's (2003) steps in the right direction is also to deal in terms of spans of dates. Both of those very different studies leave open broad spans of dates compatible with the Neolithic farming hypothesis.

With two scenarios imaginable, in weighing up the overall balance of probabilities between them we should do well to take into account other relevant questions. In particular, recalling how language is a function of social forces, a plausible scenario also needs to offer a sufficiently powerful socio-cultural factor to account convincingly for the astonishing spread of the Indo-European language family, and the completeness of its territorial domination, as Renfrew (1989, 124–31, 150–52) points out. Is 'elite dominance' really so realistic an explanation, when historical linguists know of endless examples of very powerful elites signally *failing* to impose their language: the Romans in Britain, Germanic tribes throughout the Romance area, the Normans in England, the Turks in the Balkans...

**Dangers in the data for feature-based phylogenetic methods**

The consequences of our four characteristics of language data are equally important for phylogenetic methods whose input consists of feature-based language data. For together they conspire to make even distantly related languages inevitably show parallels that in fact do *not* go back to their period of common descent, and so say nothing conclusive about *when* they split.

No analysis model magically produces a true phylogeny: they only convert patterns of correspondences in a data set into the tree or network diagrams that best fit those patterns. It goes without saying that for such an output to stand as a phylogeny, the features in the data much be such that correspondences in them do point unequivocally to shared innovation at a moment of common history before a split, rather than parallel change after it. It is crucial to take great care to identify how much diagnostic power can or cannot be read into a given feature: this is necessarily a task that requires much specific linguistic knowledge and experience.

We close with an illustration, taken from Heggarty *et al.* (2005), of the dangers for otherwise valuable phylogenetic methods, when applied to selective feature-based language data. The phenogram in Figure 16.1 is the output of FITCH, one of the PHYLIP suite of programs (Felsenstein 2001), from input in the form of quantifications of phonetic similarity between the cognates of the numerals *one* to *ten* in a number of Romance varieties. These data and quantifications are effectively equivalent to a feature-based analysis of all the phonetic features responsible for the key differences between the pronunciations of these cognates in these languages, with all differences weighted against each other for their relative phonetic significance, and for their frequency in this data set.

The resulting phenogram immediately strikes linguists as 'wrong' at first glance. It is at odds with the phylogeny commonly proposed for Romance, which has an Iberian sub-branch in which Portuguese and Spanish are more closely related to each other than either is to French. Given the historical reality of the origins of Romance as a dialect continuum, that phylogeny is not necessarily accepted here without many provisos; but by the same token the aim here is certainly not to argue for the tree in Figure 16.1 either. Rather, it serves to illustrate just how easily a certain selection of language data can give a *similarity* or *correspondences* phenogram out of line with the true *phylogeny*. (There are more examples within the Spanish and Portuguese dialect tree structures here.)

Figure 16.1 is put forward here for illustrative purposes only, and there is no space here to go into the details of the methods used to produce the phenogram — for a fuller discussion see Heggarty *et al.* (2005). It should be pointed out that with bigger data sets the method does give trees more consistent with 'consensus' genealogies for Romance, Germanic and Indo-European. Moreover, its quantifications are closely in line with perceptions among linguists of the phonetic similarity between cognates. In fact this turns out to be true for these numerals too, as we shall now see on closer inspection, despite the different relationships between the languages in other fields.

Firstly, part of the branching in Figure 16.1 is due to Spanish having innovated in ways that neither French nor Portuguese have: changing certain vowels to diphthongs, for instance, as in *siete* vs *sept*, *sete*. A shared *retention* like this is of course completely uninformative as to branching — a reminder of how crucial it is, in order to arrive at a correct phylogeny, to identify which state of a feature is an innovation and which the original form. (See Landerman (1991) for a detailed discussion of this, but also Heggarty (2005) for a critique.)

Other features, however, warn that working this out is not always easy and unequivocal. From the original Latin cluster [kt] in *octō*, both French *huit* and Portuguese *oito* simplified to [t], while Spanish has ended up with [tʃ] in *ocho*. Here it is a more debatable call as to which is most innovative, since all languages have changed radically from the original [kt].

Moreover, even with shared retentions filtered out, there are still a host of features in which French and Portuguese together share what is indeed an *innovated* value, while Spanish has either a retention or a different innovation. Even in this tiny data set, further instances of French and Portuguese changing independently in parallel include: syllable-final nasalization (as in the numerals for *one* and *five*); original [$k^{(w)}$] before high front vowels becoming [s] in French *cinq* and Portuguese *cinco*, as opposed to [θ] in Spanish *cinco* (any similarity in modern spelling is irrelevant!); and devoicing and/or loss of unstressed final vowels, complete in French and well underway in European Portuguese (as in *four*, *five*, *seven*, *eight* and *nine*).

Taking this battery of sound changes together, the phenogram they produce no longer appears at all surprising. It so happens that of all the phonetic features in this small data sample, those that are consistent with a separate Iberian branch within Romance are outweighed by a number of other features in which French and Portuguese have shared retentions, and particularly parallel innovations, not found in Spanish. This serves as a warning of just how common parallel changes can be, and how many of a set of randomly chosen features can be entirely unreliable for representing, quantifying and diagnosing language relatedness.

Much depends, then, on one's particular selection of features. Given the arbitrariness of the form-to-meaning relationship in language, the numerals *one* to *ten* do form a *phonetically* random sample of dozens of phonetic differences, selected with no expectation of the 'accidental' result in Figure 16.1. When looking at very poorly attested languages there is a particular danger, then, for the selection is made for us by the limited records available. Where only a very small sample happens to have survived, nothing guarantees that these few data will necessarily contain enough diagnostically powerful features, nor a balanced set of them. This applies not just to phonetics: with grammatical features too, the limited subset contained in a random short text could easily be biased towards features in which that language has either innovated, or been particularly conservative while other related languages have changed in parallel. When branching structure is based on very few distinctive states, and on correspondences in features that are not diagnostic and actually go back to independent changes, phylogenetic methods can all too easily produce erroneous trees, and datings.

We can be confident in rejecting the tree in Figure 16.1 as a genealogy, and in accepting others as correct, only because we are able to confirm both which states are innovations, and which among these are shared ones predating a split as opposed independent parallel changes postdating it. This knowledge

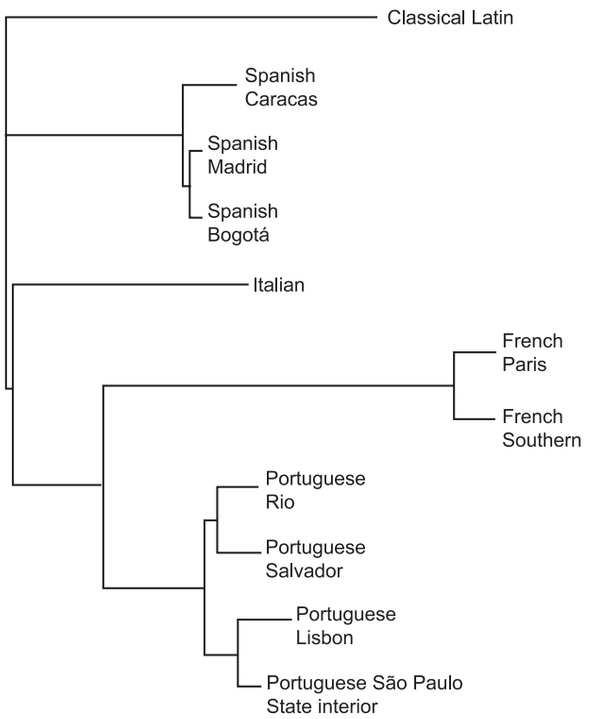


**Figure 16.1.** *Rooted* FITCH *tree for Romance varieties using the results for phonetic similarity for the numerals one to ten.*

can come only from a careful, linguistically informed analysis of the features; in this case supported by our good fortune in still having evidence of intermediate data, both through history and from a large number of other extant regional varieties. In cases where such information is lacking, and the data set contains too few features truly diagnostic of shared history, it is clearly methodologically unsound to trust any phylogenetic method, however refined, to uncover from such data the true historical tree, let alone date the nodes.

## Conclusions

So however applicable and promising phylogenetic methods might appear to be at first sight as tools for investigating language histories, even a cursory look at how language behaves in reality shows just how problematic their application to language data can be. The analysis of those data has to be alive to the myriad ways in which the nature of language, and how it changes, can so readily render language data immeasurably less straightforward than meets the eye. It is self-evident that unless the encodings we feed into phylogenetic methods are informed by such linguistic awareness, we cannot place any great faith in their output as meaningful enough to be of any real use.

The facets of language that we have looked at here, and more generally its inherent susceptibility to being moulded by external socio-cultural forces, conspire to make the attempt to use language data to extrapolate back into an unknown past an enterprise fraught with pitfalls. The unhappy history of methods proposed by linguists themselves serves as a cautionary tale: multilateral comparison and glottochronology raised grand edifices founded upon sweeping inferences from data that are in truth quite equivocal between various possible explanations. Linguistic palaeontology, too, has failed to convince many linguists that it escapes a similar trap. That the dating question remains more open than even many linguists like to admit is simply a function of the nature of language itself.

There is undoubted potential for linguistic purposes in phylogenetic methods, not least the more recent network-type representations that are more sensitive to the reality that relationships between languages by no means always go back to neatly branching histories (*Neighbor-Net* by Bryant & Moulton (2002), *Network* by Bandelt *et al.* (1999)). Long before we push phylogenetic methods so far as to hail them as miraculous new language-dating techniques, however, the first step in unlocking their potential is to ensure that their input, and thus also their output, is truly *linguistically* meaningful. Likewise for our historical interpretations of those outputs, which must accept the reality that certain configurations that we observe in surviving language data may go back to any of a number of different possible linguistic histories, all of them consistent with those same end results, as we saw for the shared terms for technological innovations like the axle. Telling which scenario among them was what actually happened is not a task to be taken lightly on the basis of questionable assumptions that we already know do not always hold in real language history. From all disciplines in the new synthesis, there is a need for closer co-operation, and for a careful and deeply linguistically-informed approach to the inherent complexities of language data, change, divergence and dating.

**References**


Bandelt, H.-J., P. Forster & A. Röhl, 1999. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution* 16, 37–48. Programme available at: http://www.fluxus-engineering.com.

Beekes, R.S.P., 1995. *Comparative Indo-European Linguistics: an Introduction.* Amsterdam: John Benjamins Publishing Company.

Bryant, D. & V. Moulton, 2002. NeighborNet: an agglomerative method for the construction of planar phylogenetic networks. *Proceedings of the Workshop in Algorithms for Bioinformatics*. Programme available at: http://www-ab.informatik.uni-tuebingen.de/software/jsplits/welcome_en.html.

Dixon, R.M.W., 1997. *The Rise and Fall of Languages.* Cambridge: Cambridge University Press.

Dyen, I., J.B. Kruskal & P. Black, 1992. An Indoeuropean classification: a lexicostatistical experiment. *Transactions of the American Philosophical Society* 82(5). Full data available at: http://www.ntu.edu.au/education/langs/ielex/HEADPAGE.html.

Embleton, S.M., 1986. *Statistics in Historical Linguistics*. Bochum: Brockmeyer.

Felsenstein, J., 2001. *PHYLIP: Phylogeny Inference Package. Vers. 3.6.* Department of Genetics, Univ. of Washington.

Forster, P. & A. Toth, 2003. Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European. *Proceedings of the National Academy of Sciences of the USA* 100(15), 9079–84.

Gray, R.D. & Q.D. Atkinson, 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426, 435–9.

Greenberg, J.H., 1987. *Language in the Americas*. Stanford (CA): Stanford University Press.

Heggarty, P.A., 2000. Quantifying change over time in phonetics, in *Time Depth in Historical Linguistics*, eds. C. Renfrew, A. McMahon & L. Trask. (Papers in the Prehistory of Languages.) Cambridge: McDonald Institute for Archaeological Research, 531–62.

Heggarty, P.A., 2005. Enigmas en el origen de las lenguas andinas: aplicando nuevas técnicas a las incógnitas por resolver. *Revista Andina*, 40, 9–57.

Heggarty, P.A., forthcoming. *Measured Language: From First Principles to New Techniques for Putting Numbers on Language Similarity*. Oxford: Blackwell.

Heggarty, P.A., A. McMahon & R. McMahon, 2005. From phonetic similarity to dialect classification: a principled approach, in *Perspectives on Variation*, eds. N. Delbecque, J. van der Auwera & D. Geeraerts. Amsterdam: Mouton de Gruyter, 43–91.

Landerman, P., 1991. *Quechua Dialects and their Classification*. Ann Arbor (MI): U.M.I. Dissertation Services.

McMahon, A.M.S. & R. McMahon, 1995. Linguistics, genetics and archaeology: internal and external evidence in the Amerind controversy. *Transactions of the Philological Society* 93(2), 125–225.

Nerbonne, J., with W. Heeringa & P. Kleiwig, 1999. Edit distance and dialect proximity, in *Time Warps, String Edits and Macromolecules: the Theory and Practice of Sequence Comparison*: eds. D. Sankoff & J. Kruskal. Stanford (CA): CSLI, v–xv.

*Oxford English Dictionary*, *c.* 2000– (web edition: www.oed.com) Oxford: Oxford University Press.

Renfrew, C., 1989. *Archaeology and Language*. London: Penguin.

Sims-Williams, P., 1998. Genetics, linguistics, and prehistory: thinking big and thinking straight. *Antiquity* 277, 505–27.

Swadesh, M., 1952. Lexico-statistical dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* 96, 452–63.