2.3
Deep time and first settlement: What, if anything, can linguistics tell us?

Paul Heggarty

1. Deep time and first settlement

Chapters 1.3 and 3.4 in this book survey what linguistics can and does usefully say on the Andes–Amazonia divide. This chapter bears a sober message also on what it *can't*. It is equally needed, however – and we shall shortly see why – for the avoidance of any doubts across the disciplines, on this touchstone of misconceptions between them.

This chapter's starting point is the same contrast on which Chapters 1.3 and 3.4 are structured: the opposing concepts of a language family, diverging out of a single origin, and a linguistic area, formed by languages converging (partially!) 'out of different origins'. Yet that formulation already raises a nagging question: 'But didn't *all* human languages ultimately start from the same origin, perhaps even long before human expansion out of Africa?' In South America particularly, is it not possible that only a small founder population originally crossed the isthmus of Panama, speaking just one language? In that case, there would originally have been no linguistic divide along the Andes–Amazonia frontier. Or does linguistics tell us that multiple different 'ethno-linguistic' groups entered South America and dispersed by different routes through the continent, establishing a linguistic Andes–Amazonia divide from the very first?

All of this is in fact quite possible. But linguistics – despite many speculative attempts and claims – is simply not able to bear on the first settlement of even this last of the continents to be colonized by *homo sapiens*. There is no real linguistic foundation to the speculative claims, schemas and deep-time 'language' entities that have sometimes been entertained. They are not some 'best guess' that we can go on at this level, even if 'controversial'. They offer nothing valid to go on at all. So one could just simply end the discussion here, were it not for a grave and ongoing interdisciplinary problem.

For over three decades now, many researchers outside linguistics, notably in genetics, *have* listened to one siren song of a purportedly linguistic framework on first settlement, and within it a potential early Andes–Amazonia divide. Greenberg's (1987) *Language in the Americas* interprets certain language data as constituting evidence that all languages of South America (and most of North and Central America) can be proven to descend from a single source, 'Proto-Amerind'. For the Andes–Amazonia question in particular, within Greenberg's purported 'Amerind' family are also his purported sub-branches, which risk being taken to support such a divide. One of those branches, indeed, he names specifically 'Andean'.

From the first, linguists have retorted, and repeatedly demonstrated, that Greenberg's 'data' provide no such *evidence* at all, as we shall see in the next part of this chapter. Linguists, then, immediately saw through the methodological deception of Greenberg's 'mass comparison' approach – or 'megalo-comparison', as Matisoff (1990) dubbed it. Frustratingly, though, many scholars in other disciplines did succumb to the temptation of a grandiose, 'big picture' pigeon-holing of all indigenous populations of the Americas, not least where it provided helpful myths upon which they could build. In genetics particularly, broad-scale publications on the indigenous Americas still routinely identify and group their genetic samples by Greenberg's constructs. Even high-profile recent papers as Reich et al. (2012), Rasmussen et al. (2014) and Moreno-Mayar, Potter et al. (2018), all published in *Nature*, use Greenberg's purported 'Andean', 'Equatorial-Tucanoan', 'Northern Amerind' and 'Central Amerind' categories, for example.

These are not the big-picture reference points that many geneticists imagine, but mere faces in the fire. They are subjective interpretations proposed by one scholar, and decried as vacuous by the rest of the discipline. Even these second-tier branches in Greenberg's schema are not valid language families. What coherence they may have is on a different level, obvious from the very names Greenberg gave them. *Andean*, *Equatorial*, *Northern*, *Central* – these are essentially just *geographical* groupings. For the challenge of working out whether linguistics aligns with the Andes–Amazonia divide, the first two are especially circular: purported linguistic entities, but actually geographical ones. If geneticists, then, find parallels in their own data, that is no support for the linguistic claims, but for the known relationship frequently found between genetics and *geography*. It is frustrating how many genetics papers could actually make considerably *more* of their findings, if only they switched to standard, meaningful language classifications, such as Campbell (1997) for the Americas, or the worldwide *Glottolog* freely available online (Hammarström et al. 2019: https://glottolog.org). It goes without saying that there is no trace of Greenberg's chimeras in those standard classifications.

But how come linguistics can say so little of deep time? Chapter 1.2 set out how the discipline can be particularly valuable in South America, where the historical record is so shallow. That is because well beyond just the five centuries or so of history here, linguistics works at its highest level of detail and confidence back a few millennia more. What dictates this timeframe, over which linguistics is most

applicable, are the typical natural rates of *change* in language. Change, and thus language divergence, happen fast enough that they give high resolution over this timescale of just a few millennia. One cannot have it both ways, however. For that natural rate is so fast that after much longer periods, so many changes have built up that one can no longer see through them all to whatever the original, deepest linguistic signal may have been.

All disciplines and individual methods can have their limits. Much of the archaeological discussion in this book mentions the paucity of the Early Holocene archaeological record across the Andes–Amazonia divide, particularly in Amazonia (see Chapters 1.1 and 2.1). Radiocarbon dating offers a more specific methodological analogy. For ultimately the decay of carbon-14 isotope leaves so little left that by 50,000 BP the method comes up against its intrinsic limits. In comparative linguistics, even though different aspects of language change at different paces, beyond a certain time-depth limit, *none* of the signals that can firmly establish *family relatedness* survives enough.

The natural pace of language change – and thus signal 'decay' – is so fast that any surviving traces of an ultimate common origin progressively fade. An absolute cut-off date is hard to pin down, for it depends on a number of variables, but the contexts in South America are very far from the ideal. Unlike in Eurasia, where ancient texts gave a head start of up to four millennia into the past, in South America no (decipherable) ancient writing existed that could take back the starting point for the decay of the linguistic signal here. And the crucial comparative data needed have been decimated by the irrecoverable extinction of so many indigenous language lineages after 1492 (see Dixon and Aikhenvald 1999, 19), driven not least by the pandemics unleashed by Old World pathogens, with their devastating demographic consequences (Chapter 1.1). Estimates therefore vary, but there is consensus that certainly by a time-depth of ten millennia or so, so little trace is left of whatever deep origins a language may have had that the signal starts to become indistinguishable from the background level of resemblances between languages that are inevitable by statistical chance. Even in South America, and even assuming the most recent timeframe postulated for first settlement, that lies already significantly beyond this ceiling on the ability of linguistics to recover the past. So there is no real prospect of recovering anything much of linguistic patterns at the time of the first peopling of the Andes or of Amazonia. In fact, as noted in Chapter 3.4, it turns out that linguistics has not been able to establish any language families in South America that might approach ten millennia. The families that are detected here began to diverge much more recently, and Chapter 3.4 finds a significant contrast between the Andes and Amazonia in just how recently.

We can now return to our original puzzle: 'But didn't *all* human languages ultimately start from the same origin?' All that is missing is just a key qualification to any language classification. This qualification is so intrinsic to historical linguistics that it is generally just left tacit – but with understandably misleading consequences for other disciplines. Linguistic texts (including the various chapters here)

normally simply state that given languages are not related to each other, and that their lineages are independent of each other in origin. The tacit, missing qualification is that they are unrelated and independent *as far back as linguistic methodology can detect relationships of common descent at all*. Such statements are understood to hold for all practical intents and purposes, or more precisely, for any attempt to use linguistics to contribute to understanding prehistory. To that end, 'unrelated' means only that two languages (or families) do not go back to the same origin at least *within* the last ten millennia or so. For otherwise, that relationship should be detectable – although in South America the visibility limit may be even shallower here, given the unfavourable contexts described above.

Full details on the inapplicability of linguistics to the question of first settlement of the Americas can be found in Goddard and Campbell (1994). Wider discussions oriented for non-linguists are Heggarty and Renfrew (2014a, 25–8), or specifically for South America, Heggarty and Renfrew (2014b, 1347–51).

Beyond the question of first settlement, this chapter has three remaining tasks. Section 2 below justifies the rejection of the methodology behind Greenberg's 'Amerind', 'Andean', 'Equatorial' and such like. The lessons there then serve also in section 3, to row back from various other speculative, deep-time claims for deep relationships of common language origin, specifically across the Andes–Amazonia divide. Finally, section 4 looks at attempts to uncover deep language relationships through correspondences not in specific sounds and meanings, but in more general and abstract characteristics of language structure. Again, we explore the limitations that necessarily attend those ambitions.

2. What is so wrong with Greenberg's 'Amerind', 'Andean' and 'Equatorial'?

For disciplines other than linguistics, it can be disconcerting to see the vehemence with which linguists have rejected Greenberg's 'Amerind', especially when it so temptingly offers the deep-time, big-picture perspective that suits others' deep-time research purposes so well. What could really be so invalid with the method Greenberg employed? Does it not appear, on the surface at least, reminiscent of how historical linguists usually seem to establish language relatedness: by comparing words from different languages in similar meanings? And if enough words look sufficiently similar, then do they not demonstrate that those languages are related? Didn't Greenberg just take this to a new level, the entire continental scale of the Americas, daring to perceive links that narrower, regionalized studies had simply failed to notice until then?

That beguiling sell has been unmasked by a rollcall of prominent figures in comparative and historical linguistics, ever since Greenberg's *Language in the Americas* first appeared. Outside linguistics, however, their publications remain less

known than Greenberg's work itself. So this section will attempt to warn off unsuspecting disciplines in the terms perhaps clearest to them, by setting out just how invalid is Greenberg's entire methodology. (It is not at all, of course, how historical linguistics actually goes about establishing relationships of linguistic descent.)

For a start, there is immediate methodological concern with Greenberg's cavalier approach to the data. He reassures that 'the method of multilateral comparison is so powerful that it will give reliable results even with the poorest of materials' (Greenberg 1987, 29). In fact, so great is the power of the method that it can *always* be made to give positive results, that is, to find large numbers of 'matches' between any desired language families in the world (see below). Greenberg took this 'power', moreover, specifically to exonerate using 'the poorest of materials'. As Adelaar (1989, 252) observes of the data quality for Quechua, for example: 'the number of erroneous forms probably exceeds that of the correct forms'. It can even be unclear which languages the 'data' are supposedly from. Experts in Quechua are rightly bemused by Greenberg's multiple references to a so-called 'Huanacucho' dialect. As Adelaar (1989, 252) puts it: 'Is this to be interpreted as the Ayacucho dialect, spoken by more than a million people and not mentioned even once … or is it the undocumented (and probably hypothetical) Huamachuco dialect …?'

We focus here only on the single most basic methodological issue, which can be seen grossly in statistical terms. Necessarily, lookalike words can just happen. Spanish *mucho* and English *much* do not in fact come from the same source, and they resemble each other purely by chance. They are evidence of nothing, in this case. So in order to use apparent similarities in sound and meaning to prove that languages are related, it is crucial to exclude statistically that they could be lookalikes just by chance. (One also needs to exclude other sources of lookalikes: sound symbolism like *shush*, near-universal nursery words like *mama* and, above all, loanwords. Greenberg makes no real attempt to exclude any of these.)

This, indeed, is where lies the most fundamental error of all in Greenberg's 'multilateral comparison' methodology. For in the name of big-picture scale, Greenberg so relaxes the criteria for a match, on all levels, that the statistical effect, far from excluding chance, is exactly the opposite: opening the floodgates so widely that 'matches' are statistically *guaranteed*. His 'method' is a machine for generating false positives, as follows.

Firstly, matches are drawn not between individual pairs of languages A and B, but between any two languages within large pools of languages. For 'Amerind', the pool effectively extends to the vast majority of indigenous languages in the Americas. Moreover, the small subsets of languages in which 'matches' are reported vary hugely from word to word. This multiplies enormously the probability of finding lookalikes by chance.

Secondly, on the level of sound, the criteria are likewise far too lax. As Goddard (1987, 657) points out, for Greenberg 'acceptable similarity … is often a match of only a single consonant', citing examples such as **mye:w* 'road' matched with *ma* 'go', or **-sit-* with *ʔas* for 'foot'. Greenberg abandons any requirement for regular,

recurrent patterns, makes free recourse to misleading spellings, and in any case, as Adelaar (1989, 252) observes, 'most examples are erroneous (e.g. Quechua *ruk* "to see" …, presumably meant to represent the verb *riku-*)'. Again, this methodological laxity hugely raises the probability of chance lookalikes.

Thirdly, on the level of meaning, comparison is made not between one word in language A and one in language B, but between potentially dozens of words with even the faintest semantic connection (and across any of hundreds of languages). Greenberg reports 'matches' between words that mean variously *night*, *excrement* and *grass*; or between *back*, *wing*, *shoulder*, *hand*, *buttocks* and *behind* (Goddard 1987, 657). If the desired sound string in *bitter* in one language is not found in *bitter* in another language, then a match is accepted also with sounds in *to rot*, *sour*, *sweet*, *ripe*, *spleen* or *gall*, while sounds in *body* can match with any of *belly*, *heart, skin, meat, be greasy, fat, deer*, and so on (Campbell 1988, 600). This too multiplies the pool of possible words for any match, and with it the probability of finding lookalikes by chance.

Under these criteria, pronouncing 'matches' becomes utterly subjective, and turns into a self-fulfilling prophesy. Critics have repeatedly shown how the combined result of these relaxed criteria is that multilateral comparison can produce 'matches' between any languages selected at random (see Campbell 1988, for example, on Finnish with Greenberg's 'Penutian'). Or for a new illustration, take some colour terms in English and compare them with Cuzco Quechua: /ɹɛd/ with /puka/; /gɾiːn/ with /q'umiɾ/; and /jɛləʊ/ with /q'iʎu/. None appear to match (and they are indeed all unrelated). But if we relax all our criteria, then we can instead propose 'matches' between Ayacucho ('Huanacucho'?) Quechua *jellu* (in Spanish spelling) and *yellow*, between *(j)omer* and *emer(ald)*, and even between *(p)uca* and *ochre*. If this seems fanciful nonsense, then of course it is – and it matches the impression one has as an informed linguist perusing much of the supposed 'data' in Greenberg's *Language in the Americas*.

In short, wherever one might wish to find false positives, multilateral comparison can oblige. There is a great deal more that is wrong, invalid and beguiling in Greenberg's approach than can be said here. (And there is far more to the methodology of historical linguistics than just comparing across languages the phonetic forms of their words for the same meanings.) Further dismantling of Greenberg's chimera of a big-picture linguistic prehistory of the Americas can be found, *inter alia*, in Campbell (1988), Adelaar (1989), Matisoff (1990), McMahon and McMahon (1995) and Campbell and Poser (2008).

3. Other linguistic misreadings on an Andes–Amazonia divide

Here is also the place to forewarn of certain other, not dissimilar dangers for the linguistic assessment of an Andes–Amazonia divide. In older linguistic literature, one finds a series of speculative hypotheses that would link individual languages

(or families) from different sides of that divide, in claiming to detect between them a signal of a deep, long-range relationship of common descent. One such suggestion is that the Uro family of the Andes, for instance, is related to the Pano family of Amazonia (Fabre 1995). Chapter 4.2 in this book takes up that particular speculation, based on similarly lax methodological criteria to Greenberg, and illustrates, in the detail of that case too, just how poor the methodology behind it really is.

In other cases, supposed shared linguistic origins between the Andes and Amazonia result from a straight misunderstanding between the disciplines, a confusion across the fundamental contrast in linguistics between language divergence and convergence (see Chapters 1.3 and 3.4). Hornborg (2005, 605, endnote 49), for instance, reports that 'Torero (2002, 488–92) suggests that Puquina … and Uru … both share an Arawakan derivation'. But Puquina and Uru are not related to each other in any case, so they cannot be derived in common from Arawak. And what Torero actually refers to here is just contact and convergence, not common 'derivation' or origin in Arawak. (On the Puquina case, see also Chapter 4.1.) Linguists themselves, like Torero in this case, sometimes muddy the waters, by talking loosely in terms of a 'contact relationship', when the ambiguous term 'relationship' is best reserved uniquely for common ancestry within a language family.

Many a misconception about language relationships goes back to this same general error. Certain linguistic parallels are often misread as evidence of a supposed deep-time language family and divergence event, when the linguistic signal concerned in fact results from and attests to convergence processes instead, often much more recent. One such discredited claim is that by Büttner (1983) for a supposed 'Quechumara' *family* uniting Quechua and Aymara, when the parallels he identifies were actually the result of intense convergence (Mannheim 1991; Torero 2002). Yet despite two decades of dismissal by linguists of the Andes, when Diamond and Bellwood (2003, Figure 3) applied to South America the hypothesis that major world language *families* were spread by farming, they nonetheless invoked the chimera 'Quechumara' *non*-family as if in support.

Claims for such 'deep' relationships pepper the older linguistic literature, particularly during and around the 1960s. At that time, enthusiasm remained fresh for staking bold, far-reaching claims upon all too superficial comparisons of just minimal lists of words. The consensus methodology of comparative linguistics had not yet been applied to many indigenous language families of South America (even for Quechua, not until the mid-1960s). As those rigorous analyses did gather pace over subsequent decades, almost all of the old claims duly fell by the wayside. Very few hypotheses of common descent of languages of the Andes and of Amazonia are even entertained today, and only where a more solid case has been made for a potential connection. See Chapter 4.1 for a case-study.

Only one significant case *has* been made with a methodology that is fairly orthodox: by Rodrigues (2009), for a hypothetical 'Jê–Tupi–Carib'. But the data invoked are extremely sparse, and this proposal remains firmly outside standard

classification. Tellingly, older speculative proposals had claimed to relate Tupí to Arawak instead, and Jê and Carib to Panoan. So mutual incompatibility alone entails that a majority of such claims must inevitably be wrong – if not indeed all of them. And for our purposes, even if Rodrigues were right, this would only reinforce the Andes–Amazonia divide, for even his vast 'Jê–Tupi–Carib' would obey it.

4. Alternative linguistic signals on deep prehistory?

We remain with the limitation, then, that beyond a threshold of ten millennia or so, we cannot trace language relationships back any further through sound-to-meaning correspondences. But might some other type of language take their place, to push back the threshold deeper into prehistory? In particular, one current in linguistics looks hopefully to structural characteristics – of the type discussed in Chapter 1.2. As an example, how does a language put together the basic components in a sentence, particularly the main verb, its subject and object? English follows the order svo, but most languages in South America use sov (https://sails.clld.org/parameters/NP2#5). Such fundamental contrasts in how languages structure their grammatical systems have long been taken to define fundamental 'types' of language, as in some sense intrinsic, deep-seated characteristics of a language. With that, might they also be unusually stable over long time-periods, and thus potential indicators of language relationships deeper than even ten millennia or so?

Nichols (1992) marked the first major attempt to identify which structural features might be so stable. More systematic and wider-scale research is now possible thanks to major comparative databases such as the *World Atlas of Language Structures Online* (Dryer and Haspelmath 2013b, http://wals.info), the *South American Indigenous Language Structures* database (SAILS) (Muysken, Hammarström, Krasnoukhova et al. 2014, the data source for Chapter 3.4), and the *GramBank* database now nearing completion (Harald Hammarström, personal communication). For all their value for research in linguistic typology, however, the aspiration to use these databases to demonstrate deep language relationships still faces existential challenges. Each abstract, structural criterion allows of only a small set of possible answers, often just two: does a language have nasal vowels or not, for example, or does it put the adjective before a noun, or after? With so few options to choose from, hundreds if not thousands of languages around the world, irrespective of whether they are related or not, necessarily share the values they have on such criteria. These characteristics thus offer little statistical power to exclude chance as an explanation for the parallels. Moreover, many structural characteristics are not fully independent of each other in any case, further reducing their diagnostic power.

Recall too, from Chapter 1.2, that many structural features are well known to pattern geographically. That is, they are susceptible to convergence between neighbouring languages, irrespective of whether they are related to each other or

not. Attempts are made to try to 'control for geography' statistically, to hone down to parallels that might result from deep relatedness instead, but they generally fail to convince. Indeed, the search for the methodological holy grail of structural characteristics that are deeply stable is proving increasingly frustrating, even to its followers. Many of the candidates in fact turn out to be considerably *less* stable than even sound-to-meaning correspondences in core vocabulary (Greenhill et al. 2017). Meanwhile, there are good grounds to consider 'deep' features actually to be positively unreliable as indicators of language relatedness. Stable they may be, but in almost the opposite sense. When speakers switch to another language, not least of a totally different family, the 'deepest' characteristics of their original native tongue can be precisely the ones they retain. That is, they carry those characteristics over into how they speak the new language that they are ('imperfectly') learning. Far from keeping in step with deep relatedness, then, these characteristics intrude into unrelated languages (see Heggarty 2017, 169–71). South America itself provides plenty of examples. Many languages distinguish two forms of the pronoun *we*. Cuzco Quechua, for instance, uses *nuqayku* for *I + you* (inclusive *we*), but *nuqakuna* for *I + other(s), not you* (exclusive *we*). This structural characteristic is precisely the one that Nichols (1992, 209) ranks as the 'most stable' of all those she analyses worldwide. Yet within even the shallow Quechua family, while Cuzco Quechua does make a distinction, Ecuador Quechua does not.

In short, no deep-time language relationship has ever been proven on the basis of structural characteristics. Nor, given the considerations above, is it ever likely to be. The optimism for 'deep' characteristics always comes back up against the reality, that it falls foul of the basic opposition that has always defined and demarcated comparative linguistics into two complementary fields (Heggarty 2017, 140–3). Historical linguistics employs those concrete, sound-based forms of language data that *are* amenable to proving language relationships. Language universals and typology studies the more abstract, structural characteristics that have so much to say on aspects of language *other* than relatedness. New structural databases like SAILS are a great advance in many ways, as we shall see for our Andes–Amazonia question in Chapter 3.4. But they are unlikely to prove any new, deep families on either side of the Andes–Amazonia divide, or spanning it.