
PAUL HEGGARTY

Scope, Context and Contribution. This paper well attests to Robert Sokal’s broad contribution to a cross-disciplinary human prehistory. For here, he and Jiangtian Chen team up with Merritt Ruhlen to add a linguistic perspective to a genetic one: specifically, to test whether the lineages of our languages might correlate, on a very ambitious worldwide scale, with those of our genes. “The implications . . . for the origin and history of our species would be great,” as they put it, and certainly the stakes could hardly be higher for a holistic understanding of our origins.

With such potential does come a commensurate temptation, however. Indeed, it must immediately be pointed out that the particular vision of language relationships and prehistory espoused here by Ruhlen is distinctly unorthodox and exceptionally controversial among historical linguists, such that grave concerns attend the “classification” that he contributes as the language “data” here. Not that this disqualifies the approach, however. Rather, in a sense it only reflects the ambition: to investigate potential correlations between linguistic and genetic lineages at the broadest possible level worldwide, and at the greatest time-depth possible. The very value of this paper is that it aspires to put to the test Ruhlen’s most controversial claims for deep language relationships. To do so, care is taken to keep a fairly tight rein on them, and to keep the method itself sufficiently distanced from them, without taking them as given (“If higher linguistic structure did exist, as claimed, we would expect . . .”).

A key result of this paper, in fact, is that it duly finds nothing to those claims, at least in the aspect assessed here: whether there emerges any clear correlation with genetics. Instead, the results incline toward the orthodox position as to what languages can and cannot reliably tell us of prehistory. In so doing, they render a service to all disciplines with a stake in understanding human prehistory. Notwithstanding the great advances in genetics since this paper was published, the latest generation of scholars in this cross-disciplinary field would do well to bear more in mind the lessons from this pioneering contribution.

1Dept. of Linguistics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. E-mail: Paul.Heggarty@gmail.com.

Copyright © 2013 Wayne State University Press, Detroit, Michigan 48201-1309.
Obviously, the task that Sokal and his colleagues set for themselves here is best understood in the context of its time. They refer specifically to Cavalli-Sforza et al. (1988) as having “reported not only a remarkable match between genetic and linguistic groupings but also a correspondence between linguistic superfamilies with major genetic clusters.” That claim had at once been very heavily disputed, of course, on two main counts: the putative superfamilies are in fact but linguistic fiction; and any apparent correlation is merely an artifact of the linguistic claims and the genetic clusters—both being effectively functions of the same determining factor, geography. Chen et al. (1995) specifically set out to address and test both of these concerns.

Another element of the context was a counterreaction, at times likewise overzealous, downplaying too far the reasonable default expectation that genetic and linguistic lineages should be expected to match at least in some circumstances. Such was what Sokal had already striven to demonstrate in “other geographically more limited studies” together with other co-authors, as cited here. Linguistic orthodoxy might even deem those illustrations more restrained and rather more sound than Chen et al. (1995) itself, but it alone has the merit of testing the potential correlation on a grander worldwide scale. And on that broad level it does suggest that to insist on “no intrinsic relation between genetics and language” is indeed overstated—even if those indications hold only on the level of certain language families well-established in linguistic orthodoxy, and not Ruhlen’s macro-family claims.

Ultimately, Sokal’s approaches would contribute to integrating comparative language data more fully into a cross-disciplinary prehistory. For as well as spurring and facilitating research into correlations with genetics, papers such as this helped push (historical) linguistics in a salutary new direction in methodology too. Quantitative and phylogenetic analysis techniques initially developed in the biological sciences can, in principle, also be applied to the process of “descent with modification” through which language lineages likewise diverge. Chen et al. (1995) stands as an early forerunner of a now significant new trend in research into language prehistory, illustrated by the cross-disciplinary research initiatives of Russell Gray and colleagues (e.g., Bouckaert et al. 2012; Gray and Jordan 2000), Don Ringe and colleagues (2002), and McMahon et al. (2005), among others.

Early methods such as Sokal’s UPGMA have been progressively complemented and by now largely superseded by other analyses, especially Bayesian approaches and network as well as “family tree” models. Nonetheless, Sokal’s contributions rank among those that most inspired this new general trend, and Chen et al. (1995) has stood the test of time better than many successors. The trend has also generated much controversy and even notoriety for a number of papers (e.g., Atkinson 2011; Forster and Toth 2003) invalidated by fundamental linguistic errors and misconceptions, typically for lack of suitable linguist co-authors and necessary methodological restraint.

Here, Sokal and colleagues were able largely to avoid such pitfalls and still make valuable progress on some of the broadest questions in a cross-disciplinary
human prehistory, thanks to their approach from a position suitably informed of the key issues and correspondingly cautious in methodology. They explicitly raise and summarize the various complicating and confounding factors that can, in principle, “diminish a strict correspondence between genetics and language,” and more realistically ask “whether, despite all these complicating factors, a significant correspondence exists . . . on a worldwide basis.” This inspires rather more confidence in their eventual result that does find at least some indications of correlations for reliably established language families. Similarly, that the authors are “not convinced of the legitimacy of presenting intraspecific population data in the form of estimated phylogenetic trees” echoes ongoing debates in historical linguistics too (see Heggarty et al. 2010).

**Linguistic Concerns.** As foreshadowed above, to understand the nature of this paper’s approach and the significance of its findings entails recognizing that in linguistics the pattern that it proposes to test—especially the “unequal branch length” vision—is not at all the orthodox vision within the discipline. Ruhlen’s hand seems clearly to have shaped how linguistics is presented here, from the very first sentences—with which almost all others in the discipline would immediately take issue. For them, it is to jump the gun to imagine that “The ultimate goal of historical linguistics is a comprehensive, hierarchical, and phylogenetic classification of all the world’s languages” and that “Such a classification would necessarily unite all current language families in a single high-level taxon, assuming a single origin for extant languages.” This rather naively misrepresents what historical linguistics can and does aspire to. Whether all human language lineages might stem ultimately from a single “monogenesis” is by no means a necessary assumption, nor one that we shall ever be able to demonstrate. And even if one were to take that leap of faith that there must once have been some single ancestral “Proto-World” language, Ruhlen here is further assuming that reconstructing back to it is feasible methodologically. Again, for almost all historical linguists, that is a pipe-dream. Language transforms so rapidly and so completely over time that beyond a given time-depth, linguistic prehistory and relationships become irrecoverable from the chance background patterns. Granted, there is some leeway in debating precisely when our ability to discern such relationships fades out entirely, but it is generally set at around ten millennia or so, and certainly not at any multiple of that, as Ruhlen imagines.

Certain other sections of this text, also presumably Ruhlen’s, take further linguistic liberties, not the least in statements (as if of fact) that run counter to orthodoxy in the discipline. Several references to “families,” “cognates,” and putative “patterns” would by no means be recognized as such by orthodox historical linguistics. Chen et al.’s Figure 1, meanwhile, reveals how erratic and unrepresentative is the sample that Ruhlen feels sufficient to represent the many thousands of languages around the world, and hundreds of independent lineages in standard classifications. In various cases, just a single language is taken to
stand in not only for its entire family, but also for some far greater “macro-
family” claim. “Abxaz” alone represents Ruhlen’s “Caucasian” macro-family,
whereas another single taxon stands in for the entirety of Australia (and its own
supposedly eponymous macro-family). Moreover, unlike all other taxa (e.g.,
“French”), this “Pama-Nyungan” is not a specific language at all, but a
hypothetical large family, not widely accepted as related to other families of
(“Top End”) Australia in any case.

Readers should be under no illusions, then, as to the scale of the departure
from orthodoxy that the language “data” presented here actually represent.
Ruhlen’s “linguistic evidence” is, for orthodoxy in the discipline, just a
subjective interpretation on a methodology flatly rejected (see, e.g., Campbell
and Poser 2008; Ringe 2000). Of the 17 linguistic phyla listed at the top of page
599, more than half are routinely dismissed as fictional by most specialists in the
languages concerned. This applies even to some relationships at the shallow end
of the time-depth scale, and thus to corresponding parts of the equal-branch-
length analyses too. Examples include “Khoisan” and “Altaic,” against which
specialist opinion has only hardened further in the years since this paper was
published.

The only inkling of all this lies in the note that “A distance of 5 denoted the
limit of difference between two languages that most linguists claim can be
discovered by comparative methods.” In other words, all of Ruhlen’s numbers
above 5 are, for most linguists, vacuous.

Not that his co-authors are unaware of this, for the text is at pains to clarify
repeatedly how Ruhlen’s linguistic distances were arrived at: “The linguist on
our team (M. Ruhlen) estimated linguistic distances”; “These distances were
subjective estimates”; “the designation of interphyletic distances in both . . .
classifications was subjective”; “Ruhlen believes that the reliability within each
phylum is fairly high” (my emphasis in each case). Such “subjectivity” and
“belief” sit uneasily with the scientific method, of course, and macro-family
claims have remained dogged by conflicting interpretations in which almost all
permutations of deep relationships between putative macro-families have been
claimed, self-contradictory and inconsistent with each other.

It is equally crucial that the authors duly acknowledge that “Ruhlen might
have been influenced by the existing genetic tree of human populations,
especially at the highest level (i.e., the African/non-African separation).” This
specific mention of a geographical contrast highlights the same circularity that
most undermined Cavalli-Sforza et al. (1988). For many of, the very parallels that
“macro-family” classifications rely on are seen in more orthodox analyses as
going back not to a common origin but to convergence between languages in
contact with each other within the same region. Effectively then, popular
“macro-family” hypotheses were arrived at in large part by (mis)interpreting
known geographical patterns in languages as if they were evidence of related-
ness instead. This is so intrinsic in constructs such as “Altaiic” (and the very
names of “Amerind” and “Eurasiatic”) that even when studies like this one seek
to hold geography constant, they cannot truly achieve that because the suppos-
edly purely linguistic classification is in part still a function of the geography that shaped ideas about it in the first place.

Finally, once the result emerges—no correlation between that classification and the genetic data—circularity returns in the attempt to explain away this inconvenient result for Ruhlen. To hope that “alternative higher classifications” or “Choosing different values . . . might yield higher . . . correlations” would be a very questionable statistical and methodological procedure. Fishing for whatever combination might happen to throw up the match desired a priori seems a clear case of the disciplines “building on each other’s myths” (Renfrew 1987: 287).

Key Findings. With all these objections to Ruhlen’s linguistics, readers might be forgiven for fearing that the research here might be largely invalidated. Yet they detract less than it might seem, precisely because this paper was effectively designed specifically to test Ruhlen’s classificatory vision against genetic data. Indeed, if anything, one detects a suitably tight leash put on Ruhlen’s putative higher-level classification by his co-authors.

And the results returned are, of course, all essentially negative for that classification, and highly consistent with orthodox linguistic analyses instead. “This study finds no genetic support for the proposed higher linguistic structure. On the contrary, the linguistic distance matrix without the claimed higher structure shows stronger correlation with the genetic data . . . . The present data do not support a Eurasian superphylum.”

Very little here is surprising for established linguistic scholarship, which would predict results of just this form. Because it sees no support for Ruhlen’s claimed higher structure on linguistic grounds in any case, it is no surprise that genetics yields no trace of a pattern that never existed in the first place. As for the lower-level families that linguistic consensus establishes certainly do exist, it is perfectly reasonable for the authors to assume that for some of them at least, the contexts through which their speaker populations lived through prehistory would have left some degree of surviving correlation with their genetic lineages.

In the handful of cases in which a correlation turns out to be present but in the negative, the authors adduce an explanation that seems plausible in cases like the Basque one to which they refer (pg. 608). Nonetheless, many linguists would in any case assume that a range of such effects may have arisen as artifacts of the subjectivity inherent in the relationship claims, and the thin and erratic language sampling.

On general methodological questions, the authors contrast the respective merits of their alternative forms of representing their results as dendrograms or as distances. Again, with welcome frankness, they first observe how different and inconsistent are their respective outputs, and thus must seek to explain this. In doing so, they hit upon a number of points still very much in debate within historical linguistics too. Remarking that “the consensus between language trees
and genetic trees is... so low as to make the trees incomparable,” they suggest that the problem may lie with the genetic trees not being “sufficiently reliable.” Few linguists would be convinced by their presumption that the language trees were more reliable, however, because even the equal-branch-length tree includes various nodes that (pace Ruhlen) still do not in fact “reflect established language classification.”

Their explanations also rightly touch upon the “considerable question whether a dendrogram is an appropriate representation for the populations of the world,” and venture that “In view of the continuum of variation and the numerous interconnections resulting from gene flow, distances may be a far better representation than trees.” Similar questions are—or at least should be—at the heart of linguistic prehistory too, not least in its interaction with genetics. Likewise, languages are hardly immune to processes of interaction and convergence, although there lies here another point worth adding to Chen et al.’s (1995) summary of cross-disciplinary issues, because to this day it remains all too often overlooked.

For in a crucial sense, genetic and linguistic distances and classifications typically do not compare like with like at all. Indeed, while genetic measures normally do intrinsically reflect the impacts of contact and admixture, linguistic ones based on language families (as here) effectively do the reverse. Those classifications are achieved in the first place precisely by abstracting away from any contact influences, deliberately in order to isolate only those parallels that go back to common origin instead. Ultimately, of course, one reaches time-depths at which it is no longer possible to tell the two apart with any confidence—the orthodox position that whatever deeper linguistic relationships there once may have been, they are irrecoverable.

Chasing after such shadows still distracts all too many linguists themselves, to the detriment of the other half of the linguistic story, to be found not in diverging family trees, but in the linguistic data on the convergence of populations. Yet the study of those patterns worldwide still lacks any principled cross-disciplinary approach for tying them in with evidence for parallel convergent processes impacting on the archaeological and genetic records.

Progress and Continuing Relevance. Much else has changed since this paper was published. Modern analyses are far more detailed and sophisticated than was possible for Chen et al. (1995). While they calculated geographical distances, for instance, “as geodesic (great-circle) distances,” alternative models are now available that are more sensitive to landscape, topography and travel times, or that follow entirely different approaches in any case, such as Bouckaert et al.’s (2012) “phylogeography.” Much more revolutionary still have been the technological advances toward both full-genome and ancient DNA analyses, banishing Chen et al.’s resignation before the basic problem that “genetic information for human populations is still far from complete.” It is true that certain ethical, political, and simple cost factors still conspire to impose limits and distortions on
genetic sampling of population diversity worldwide, but possibilities are fast opening up to rerun studies such as Chen et al. (1995) on what are now vastly richer genetic databases.

And as they emerge, how are these authors’ predictions faring? They wondered whether “other or more copious genetic data may still reveal evidence for structure among the phyla” after all. Yet recent attempts (e.g., Reich et al. 2009) continue to show much complexity and still no conclusive case for a general genetics-linguistics correlation at the deepest worldwide level—let alone one uniquely compatible with claimed macro-families rather than established classifications. And where Chen et al. (1995) once saw a limitation in genetic data, by now it lies more in language, inherently an object for essentially qualitative analysis, far less amenable to rapid technological advances. However spectacular the progress in recovering DNA from samples as old as the Denisova hominins, there is no prospect of bringing back human speech from such time-depths.

This deep frustration that language relationships cannot be traced remotely so deeply goes back ultimately to the simple idea that Chen et al. (1995) themselves touch upon here: our linguistic and genetic lineages do indeed diverge “at different rates.” Perhaps this explains why geneticists, in particular, have grasped at such controversial linguistic claims that may seem to be “all we have to go on” at the much deeper time-scales to which their genetic data can be taken. Not least for geneticists, then, does Sokal’s contribution here remain so significant: a first cautionary tale, from their own data, that such spectacular linguistic claims are in truth nothing to go on at all.

Finally, how are we to interpret the essential cross-disciplinary finding of this study? Some correlations can be found between languages and genes, but far from universally. Sokal would surely not have taken this as a cue to throw up our hands, as if the lack of a simplistic correlation condemned the whole enterprise as a hopeless task. Rather, it is precisely the contrasting patterns of where our linguistic and genetic lineages do match, and where they do not, that we should mine as rich seams of data on our prehistory.

Received and revision accepted for publication 9 October 2012.

Literature Cited


