Introduction: the Indo-European debate and why it matters

On August 23, 2012, leading media outlets across much of the world announced that one of the major puzzles of human history had at long last been resolved: the origin of the Indo-European language family. "Family Tree of Languages Has Roots in Anatolia", proclaimed the *New York Times*'s sprawling front-section article.[1] The *Times* stressed the role of biologists in determining Indo-European origins, highlighting their use of "tools for drawing evolutionary family trees". On the same day, *Scientific American* showcased the same study, emphasizing instead its novel use of cartography: "Disease Maps Pinpoint Origin of Indo-European Languages".[2] Two days later, the BBC went so far as to announce that the "English Language 'Originated in Turkey'", illustrated with an image of the great lexicographer Samuel Johnson looking especially grumpy.[3] Similarly, Spanish, French, Italian, and Polish media outlets all saw the origins of their respective languages placed in Turkey as well.[4] One Russian blog went as far as to ask: "Are we all a little bit Turks?"[5] As the extra set of quotation marks in the BBC headline shows, such headlines were not meant to be taken literally, as English undoubtedly originated in England, just as Italian originated in Italy. The point was rather that the most

[1] nytimes.com/2012/08/24/science/indo-european-languages-originated-in-anatolia-analysis-suggests.html?_r=0

[2] scientificamerican.com/article/disease-maps-pinpoint-origin-of-indo-european-languages/

[3] bbc.co.uk/news/science-environment-19368988

[4] "El idioma español podría tener su origen en Turquía" [The Spanish language might have been originated in Turkey]: aprendemas.com/Noticias/html/N10747_F27082012.html

"Finalement, la langue française serait née en Turquie" [Finally, the French language was born in Turkey]: http://rue89.nouvelobs.com/2012/08/24/finalement-le-francais-serait-plutot-ne-en-turquie-234823

"L'origine turca dell'Italiano" [The Turkish origin of the Italian language]: http://oggiscienza.wordpress.com/2012/08/27/lorigine-turca-dellitaliano/

"Nasza mowa narodziła się w Turcji, nie w Rosji" [Our language was born in Turkey, not in Russia]: polskieradio.pl/23/3/Artykul/670988,Nasza-mowa-narodzila-sie-w-Turcji-nie-w-Rosji

[5] "Все мы немножко турки?" [Are we all a little bit Turks?]: http://echo.msk.ru/blog/ramera1/922907-echo/

distant ancestor of English, as well as of Spanish, French, Italian, Polish, Russian, and so on, supposedly emerged in what is now Turkey.

The study on which these articles were based – "Mapping the Origins and Expansion of the Indo-European Language Family" – appeared at the same time in *Science*, arguably the world's premier scientific journal. As is typical of a paper in the natural sciences, it was written by a team of scholars; although Remco Bouckaert is listed first, the actual leaders were Russell Gray and Quentin Atkinson, both at the School of Psychology at the University of Auckland. These scholars claim to have found "decisive support" for Indo-European origins among Neolithic farmers in central Anatolia some 8,000 to 9,500 years ago by using techniques of phylogenetic and phylogeographical analysis designed, respectively, to construct evolutionary family trees and trace the spread of viruses. Such methods, they claimed, allowed them to overturn the reigning hypothesis, which holds instead that the language family emerged among pastoral peoples living in the grasslands, or steppes, north of the Black Sea some 4,000 to 6,000 years ago. According to the *Science* study, an Indo-European homeland in Anatolia must now be regarded as orders of magnitude more likely than one in the steppe zone.

It might seem odd that this seemingly obscure topic would have attracted such intense media attention. But the Indo-European question, with its long, complex, and vexed heritage, is more intellectually significant than it might appear at first glance. No other language family comes close to the global significance of Indo-European, which encompasses the mother tongues spoken by almost half of humankind (Maps 1 and 2). Almost all languages found from central India through Iran and across Europe, and by extension through most of the Americas, descended from a single ancestral tongue spoken in some limited part of Eurasia thousands of years ago. That the daughter languages of this one variety of speech spread so widely has long intrigued scholars and has periodically engaged the public imagination as well. Indo-European studies, moreover, have often been ideologically misused by those seeking to establish their own pet theories of human history and cultural diversity. As a result, the announcement of a solution of the puzzle of Indo-European origins is indeed a newsworthy event.

Mere claims to scientific progress, however, do not guarantee the discovery of truth; nor do advanced mathematical approaches always generate accurate or even serviceable results. If such promises are to be realized, the data must be solid and the assumptions must accord with reality. Yet in regard to "Mapping the Origins" – and more generally the research program that generated it – neither stipulation holds true. Here, incorrect and in some cases incoherent linguistic information is fed into complex equations, systematically corrupting the results; as the common adage puts it, "garbage in/garbage out". The cartography underlying the phylogeographic component of "Mapping the

Origins" is no better than its linguistics. When it comes to issues of historical geography, the model produces incorrect results at virtually every turn, consistently contradicting the empirical record. It is thus not merely that the approach fails, but rather that it fails spectacularly. But the problems go deeper, as even ideal data would have yielded untenable conclusions in this case. Key here are the authors' erroneous and unexamined suppositions about language differentiation, distribution, and expansion. In actuality, linguistic evolution is only vaguely analogous to organic evolution, and hence cannot be analyzed with the same techniques, whereas language-group expansion has virtually nothing in common with the spread of viruses. The distribution of human languages, it turns out, has been heavily molded by historical contingencies, events that have not, and probably cannot, be factored into the biological model employed by the authors. Given both the faulty data and the demonstrably incorrect assumptions, the model used in "Mapping the Origins" can never deliver what it promises.

Although our opposition to a single *Science* article might strike some readers as excessive, much larger issues are at stake. Significantly, the vast majority of Indo-European specialists, whether linguists or archeologists, reject the Anatolian theory of Indo-European origins and remain skeptical of the phylogenetic and phylogeographical methods employed by the Gray–Atkinson approach. This near unanimity of opinion brings up a crucial yet largely unexamined issue. Most of the publications that acclaimed this study also maintain that it is essential to recognize the consensus of the scientific community on the fundamental issues of the day, such as climate change. It would be unthinkable, for example, for the *New York Times, Scientific American*, or the BBC to showcase in such a manner a single study purporting to debunk global warming based on techniques rejected by the vast majority of climatologists. Yet when it comes to matters of historical linguistics, scientific consensus appears to count for little, if anything. Why this should be the case is itself an interesting and important topic that demands investigation.

Unfortunately, such lack of respect for the scientific consensus in linguistics appears to be relatively widespread. As David Pesetsky emphasized in his plenary address at the 2013 annual meeting of the Linguistics Society of America, the "absence of [. . .] responsible publications [. . .] in high profile journals" is not limited to historical linguistics, as it extends as well to generative syntax and other linguistic subfields (Pesetsky 2013). Articles seeking to discredit generative grammar, arguably the theoretical heart of modern linguistics, that have appeared recently in major scientific journals often misconstrue the key concepts of the generative approach, such as hierarchical structure and linguistic universals. (Notably, several of these papers were co-authored by scholars who contributed to "Mapping the Origins"). This is not, however, to argue that works seeking to discredit such

fields as generative grammar are somehow inadmissible, much less to insist that scholars from other disciplines should avoid treading in this domain. To the contrary, all key scientific concepts must remain perennially open to challenge and dissent. Such challenges, however, have the responsibility of fairly and adequately grappling with the ideas that they are trying to dismantle. Unfortunately, in the field of linguistics this is often not the case. As Pesetsky further argues, the lack of publications in the broader scientific field "that *do* engage the field of generative syntax and accurately represent its results [. . .] probably reflects nothing more than the complete ignorance on the part of editors and reviewers that there even *is* a field and a set of results that should be engaged". If true, such ignorance needs to be remedied, and we hope that this book is a contribution in that direction.

The problems, however, go deeper still, as the Gray–Atkinson research program has gained traction not only with editors of high-profile scientific journals and the popular media, but they are beginning to make inroads into scholarship in neighboring disciplines as well. This broader appeal of the Gray–Atkinson approach is partly based on the fact that it claims to do much more than merely solve linguistic puzzles of the distant past. If confirmed, this research program would substantially discredit conventional research in historical linguistics, replacing its language-focused methods with those of evolutionary biology and epidemiology. Since languages differentiate, spread, and evolve just as species do, or so these authors contend, the same tools should be able to determine lines of descent and areas of origin for both. At the same time, they dismiss two key geo-historical methods of linguistic scholarship – comparative reconstruction and linguistic paleontology – as inadequately rigorous. Frustrated that historical linguists have been unable to definitely resolve some of their own central problems, the Gray–Atkinson team would reformulate the field on a supposedly more thoroughly scientific basis, employing cutting-edge bioinformatic methods and promising a computational revolution in knowledge. As a result, some scholars and journalists are now assuming the conclusions reached in "Mapping the Origins" as given and are basing further speculations upon them. For example, recent genetic research showing that the ancient Minoans were likely descended from Anatolian farmers has been extended by claims that "the Minoans may have spoken a proto-Indo-European language".[6] The only way to deduce this claim is by *assuming* that Neolithic Anatolian farmers spoke an Indo-European language, turning a highly questionable hypothesis into a foregone conclusion. (The claim is, moreover, linguistically nonsensical, as *Proto*-Indo-European was by definition a single language.)

[6] Tia Ghose, "Mysterious Minoans Were European, DNA Finds", *LiveScience*, May 14, 2013.

The broader drive to refashion the field of linguistics on a supposedly more scientific basis extends to the study of phonology and morphology. A number of scholars have recently sought to explain such purely linguistic attributes on the basis of adaptive pressures rooted in the physical environment. Such hypotheses include putative correlations between rich systems of case marking expressing fine spatial distinctions and the complex topography of the mountainous landscapes (languages of the Caucasus fit this correlation particularly well), and the supposed connection between the presence of nasal vowels and cold and damp climates (showcasing the fact that Modern French, based on the dialect of allegedly cold and damp Paris, has nasal vowels, whereas Spanish and Italian, mostly spoken in warmer and drier climes, do not). Such theories, while perhaps plausible for a small set of languages, fail to apply if a global language sample is used. A more recent article by anthropologist Caleb Everett, "Evidence for Direct Geographic Influences on Linguistic Sounds: The Case of Ejectives", purports to solve this problem by engaging in a large-scope typological study considering 567 languages. Here the author argues that the relatively rare type of sound called ejectives (discussed in more detail in Chapter 9), which involve the closure of the glottis to generate a dramatic burst of air when the oral closure is released, is linked to geographical elevation. Everett claims that such sounds are easier to make under conditions of low air pressure found at high elevations, and he further speculates that "ejective sounds may help to mitigate rates of water vapor loss through exhaled air". Unfortunately for Everett, these supposed causal mechanisms actually make no sense, and a careful examination of the distribution of languages with ejective sounds fails to reveal any actual correlation with altitude.[7] As Johanna Nichols puts it, "ejectives [. . .] can be found in mountain areas – not because harsh mountain geography deterministically causes languages to add harsh consonant series (!), but because isolation favors complexity" (2013: 38). In the end, supposedly scientific accounts of phonological and morphological properties such as Everett's tend to have the characteristics of just-so stories, providing superficially appealing explanations that fail to withstand scrutiny.

Other attempts to generate a supposedly scientific approach to the study of languages come from social scientists claiming to find correlations between linguistic features and various non-linguistic social or cultural traits. For example, a recent paper by three economists (Gay *et al*. 2013) links linguistic gender systems with female economic and political empowerment: women in countries with languages that make gender distinctions are supposedly less likely to participate in the labor market or in politics and have a reduced ability to obtain credit or own land. A somewhat similar paper by Boroditsky *et al*.

[7] See our discussion found at http://languagesoftheworld.info/geolinguistics/ejectives-high-altitudes-grandiose-linguistic-hypotheses.html

(2003) reported a connection between linguistic gender and attributes ascribed to various inanimate objects. A study by behavioral economist Keith Chen from Yale University highlights a supposed correlation between tense marking and financial behavior: people whose native language makes few distinctions between the future and the present purportedly think differently about the future and therefore make different financial decisions than those whose language makes systematic temporal distinctions (Chen 2013).[8]

These and similar studies are reminiscent of the strong version of the Whorf–Sapir hypothesis, which claims that one's language determines how one thinks. Professional linguists and psychologists, however, long ago undermined such claims of linguistic determinism. As Steven Pinker puts it, "The cognitive revolution in psychology, which made the study of pure thought possible, and a number of studies showing the meager effects of language on concepts, appeared to kill the hypothesis by the 1990s, and I gave it an obituary in my book *The Language Instinct*" (2007: 124). In the end, Pinker justly observes that, "The Whorfian interpretation is a classic example of the fallacy of confusing correlation with causation" (2007: 127). John McWhorter echoes Pinker's sentiment in his recent book *The Language Hoax: Why the World Looks the Same in Any Language*: "Yet most would consider it a fair assessment that the work of [the "Neo-Whorfians"] has shown that language's effect on thought is distinctly subtle and, overall, minor. Not uninteresting – but nevertheless, minor" (2014: xiv).

Many of these problematic linguistic studies are also reminiscent of geographical determinism, a hoary idea abandoned by professional geographers generations ago that holds that the physical environment determines how people think and behave. In the early twentieth century, ambitious geographers argued that stark desert vistas automatically lead the human mind to monotheism (Semple 1911: 512); now their epigones are telling us that climate and altitude determine the sounds that we make. Not much has changed.

Yet, despite all the evidence to the contrary, such theorizing refuses to die, in large part because it has the semblance, although not the substance, of scientific reasoning. And as the digitization of linguistic information advances, the ability, and temptation, to advance such specious claims grows apace. As Mark Liberman perceptively noted in a *LanguageLog* blog post:[9]

> [M]any relevant linguistic and non-linguistic datasets are now pre-compiled and available for easy download, and the software needed for fitting various sorts of statistical models can easily be run on your laptop. So if you have a bright idea – maybe alcohol consumption correlates with phonotactic complexity? – really, it could – the chances are

[8] For our critique of this study, see languagesoftheworld.info/language-and-mind/you-save-what-you-speak.html
[9] http://languagelog.ldc.upenn.edu/nll/?p=4685

that you can test a model within a few hours. If it doesn't work out, there are plenty more to try – maybe coffee consumption helps to preserve morphological inflection?

Efforts to reformulate the study of language thus extend deep into the core practices of linguistic scholarship. But as Pesetsky put it in reference to recent articles on linguistic issues in high-profile scientific journals, "none of these papers [. . .] contain any linguistic theory, any linguistic analysis or any significant set of linguistic facts" (2013: 107).

The current push to reinvent linguistics on a supposedly more scientific basis is not the first. In the 1950s, the renowned psychologist B. F. Skinner (1957) argued that speech should be analyzed like any other behavior, in terms of stimuli and responses. Skinner's linguistics, like his psychology more generally, has not fared well in the intervening decades, and his key language concepts, such as that of the "mand", have all but vanished. Even at the time of publication, the Skinnerian approach never made much headway in linguistics, owing in part to Noam Chomsky's (1959) devastating critique. Chomsky's contrasting approach of generative grammar, which postulates an innate human proclivity for language acquisition based on universal principles and parameters, has instead emerged as the theoretical core of modern linguistics. It has done so for good reason, as it can account for the variety and complexity of language much better than any competing framework.

On the surface, the Gray–Atkinson approach to historical linguistics has little in common with B. F. Skinner's attempt to understand language acquisition on the basis of operant conditioning, or with studies that link specific sounds to specific features of the physical environment, or with works that purport to show how the grammar of a particular language determines the thought and behavior of the people who speak it. These various approaches, however, are similar insofar as they attempt to explain or at least comprehend crucial features of language on non-linguistic grounds or by applying non-linguistic methods. As a result, we contend, none of these approaches has much explanatory power, as language must be understood in its own terms. It is difficult to avoid the pitfalls of pseudo-scientific research, in short, when one engages in the non-linguistic study of language.

The seductive attractions of linguistic scientism

Owing to this broader intellectual environment, the present book should not be construed as a long-winded response to one rather short paper, even though most of it does focus on "Mapping the Origins" (also referred to here more formally as Bouckaert *et al.* 2012). Instead, we use this single article as a springboard for discussing far more fundamental issues than those pertaining to the origin of a particular language family. By scrutinizing the methods used,

the assumptions made, and the errors produced in this article and related works, we can gain insight into the actual nature of linguistic origination and change, allowing us to specify more precisely how languages diversify from each other and spread across landscapes. Doing so, in turn, can help clarify crucial processes in human history, as well as shed light on the nature of interdisciplinary research.

Ultimately, we seek to defend the established disciplinary methods used to investigate language change and other linguistic phenomena. Linguistics itself has long been a highly rigorous endeavor based on rational inquiry and empirical justification. Already a science in the broad sense of the term, it does not need to be reformulated to *look* more scientific. This is not to say that linguists should ignore new techniques pioneered in other disciplines or shun collaboration with scientists schooled in other disciplines. We do contend, however, that any such work should seek to complement rather than supplant the time-tested methods of linguistic research. These difficult and demanding linguistic techniques continue to produce new and important findings on human communication and indeed on the human condition itself. As such, they deserve to be cherished, not tossed away on the rubbish bin of intellectual history by overly eager, would-be paradigm shifters.

One of our key questions in regard to "Mapping the Origins" is why such deep and pervasive errors in research design and execution went unnoticed, whether by the authors themselves, brilliant and accomplished scientists as they are, by the peer reviewers at *Science*, or by the funding agencies that lavishly supported the research. Why were top science journalists, such as Nicholas Wade at the *New York Times*, so eager to trumpet such a deeply flawed study, giving only lip service to its critics? Although such questions elude simple answers, surely overweening faith in the methods of scientific research plays a central role. We are not concerned here with genuine science, an endeavor that we treasure and defend, but rather with *scientism*, or the uncritical belief in the universal applicability of specific scientific methods. In biological taxonomy, the new bioinformatic methods have indeed constituted a revolution in knowledge, but that is hardly a guarantee that they would do the same in linguistics. Insisting that they must, regardless of all evidence to the contrary, is not the reasoning of science, but rather that of scientism.

The appeal of linguistic scientism is easy to understand. The research techniques employed are themselves brilliant, requiring vast computational resources and mind-boggling mathematics. As Simon Greenhill and Russell Gray have noted, the number of possible family trees computed and weighed in "Mapping the Origins" exceeds the number of atoms in the universe (2012: 529). Bouckaert *et al.* (2012) assess probabilities through the ingenious techniques of Bayesian analysis, recently brought to wide public attention by statistician Nate Silver (2012). These authors produce impressive and sometimes

stunning visualizations of the branching and spreading patterns of language families. Their techniques, moreover, are fully generalizable, able to handle – or mishandle – any language grouping with equal facility. The new approach thus promises a great shortcut to knowledge, allowing one to dispense with the welter of empirical detail that traditional Indo-Europeanists, like students of other language families, must master. By the same token, most of the painstaking work of linguistic reconstruction is avoided. Once the model has been established and the inputs entered, the grinding toil of analysis is automated, carried out by extraordinarily capable computers.

But untempered enthusiasm for conceptually elegant procedures can easily lead one astray. At the extreme, a project so conceptualized can become unmoored from the empirical world, and hence impervious to criticism. Unfortunately, "Mapping the Origins" – and the wider research program in which it belongs – seems to fall into this category. As our correspondences with the authors, as well as their responses to other critics, have revealed, they dismiss any error as a mere triviality, and wave away mistakes by the hundreds as temporary imperfections that will eventually be overcome as their model is fine-tuned. Sloppy and indeed unjustifiable language selection and language mapping are not regarded as significant, nor is faulty data tabulation. As a December 20, 2013 communication to the editors of *Science* reveals, the authors forgot to fix their data matrix after they removed thirteen languages from their original sample, resulting in "283 'empty' columns of zeros".[10] After repairing this flaw, the authors substituted a new version of the Supplementary Materials with a revised phylogenetic tree, although the publication date indicated by the *Science* website is still "24 August 2012".[11] Thus, the sheer fact of the substitution is concealed from the public, as if the original "results" never existed in the first place. Many of our criticisms are based on the originally published Supplementary Materials, but most hold for the revised version as well. We also indicate at several points in this book how the simple "tweaking" by the authors of the *Science* article changed their claims about the Indo-European (pre-)history. The biggest change is that the new tree places the divergence of the main Indo-European branches 500 years closer to the present. Yet, this new divergence date still deviates from the established consensus by some 2,500 years – not much of an improvement. As we show in the following pages, some of the results produced by this modification are not improvements at all, as they take the Gray–Atkinson model even further away from the established consensus in historical linguistics while making it even more vulnerable to counterevidence from the archeological record.

[10] *Science* 342: 1446 (sciencemag.org)
[11] http://www.sciencemag.org/content/suppl/2012/08/22/337.6097.957.DC1

But an even deeper problem stems from the authors excusing what turn out to be demonstrably false assumptions about language diversification and spread with the trivially true but misleading observation that science always demands simplification. In short, their approach seems to be unfalsifiable, and hence unscientific. When it comes to matters as complex and multifaceted as the origin and dispersion of language families, it turns out, massive shortcuts to knowledge necessarily fail, as the details matter. If one ignores or dismisses them, one cannot help but reproduce ignorance.

Unfortunately, such misleading scientific research is not limited to biologically inspired efforts to reinvent linguistics. In late 2013, the global scientific establishment was shocked when Nobel laureate Randy Schekman announced that he would no longer publish in the highly prestigious journals *Nature, Cell*, and *Science* in order to protest their tendency to showcase flashy but shoddy studies designed to attract the attention of journalists.[12] As Walter Russell Mead and his colleagues insightfully summarized the situation:

> The race for immediately exciting or actionable science has led to lots of "bunk science" that other scientists are often unable to later verify. The costs of this are enormous. It sends us down a lot of false alleys, fueling flashy but superficial headlines in the popular press and giving cover to bad politics. But most importantly it means resources and energies are [being] subtly but surely directed away from research that makes for less click-bait headlines but is nevertheless vitally important to innovation and scientific progress.[13]

In our view, the research program that produced "Mapping the Origins" exemplifies this problem, indicating a fundamental failure both in the vetting of articles at *Science* and in the reporting of such "scientific breakthroughs" by leading newspapers including the *New York Times*.

Such harsh charges, it is essential to note, are specific to the research methods used in "Mapping the Origins", not applying to computational linguistics more generally. When combined with rigorous linguistic analysis, computational linguistics shows great promise. Nor do we even seek to entirely discredit the bioinformatic approach; we may disagree with the manner in which it has been carried out, but we acknowledge its impressive computations, respect its audacity, and value its prompting of productive debate. Had the authors of the *Science* article in question framed their findings as suggestive rather than decisive, we would have shrugged it off as an intriguing if misguided effort. But by claiming to have resolved a crucial debate, they have crossed a line, veering from inadequately conceptualized science into a pernicious form of scientism that demands firm rebuttal.

[12] "Nobel winner declares boycott of top science journals". *The Guardian*, December 9, 2013: theguardian.com/science/2013/dec/09/nobel-winner-boycott-science-journals

[13] the-american-interest.com/blog/2013/12/14/nobel-prize-winner-criticizes-leading-science-journals/