Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-021-00991-8 OPEN

Check for updates

Foundations of the Age-Area Hypothesis

Matthew J. Baker[1]✉

A useful tool in understanding the roots of the world geography of culture is the Age-Area-Hypothesis. The Age-Area Hypothesis (AAH) asserts that the point of geographical origin of a group of related cultures is most likely where the culture speaking the most divergent language is located. In spite of its widespread, multidisciplinary application, the hypothesis remains imprecisely stated, and has no theoretical underpinnings. This paper describes a model of the AAH based on an economic theory of mass migrations. The theory leads to a family of measures of cultural divergence, which can be referred to as Dyen divergence measures. One measure is used to develop an Age-Area Theorem, which links linguistic divergence and likelihood of geographical origin. The theory allows for computation of the likelihood different locations are origin points for a group of related cultures, and can be applied recursively to yield probabilities of different historical migratory paths. The theory yields an Occam's-razor-like result: migratory paths that are the simplest are also the most likely; a key principle of the AAH. The paper concludes with an application to the geographical origins of the peoples speaking Semitic languages.

[1] Hunter College and the Graduate Center, CUNY, New York, NY, USA. ✉email: matthew.baker@hunter.cuny.edu

Introduction

Much cross-disciplinary effort has been devoted to understanding how the world-wide geographical distribution of cultures came to be. This research program has revealed deep connections between culture, genetics, and language (Cavalli-Sforza and Cavalli-Sforza, 1995). Linguistic, historical, archeological, and genetic evidence have been applied in concert to shed light into the origins of Indo-European peoples (Atkinson and Gray, 2003; Mallory, 1997; Renfrew, 1987), the peopling of Africa (Ehret, 2002; Holden, 2006), the settlement of the South Pacific (Greenhill and Gray, 2005), and the spread of agriculture around the world (Diamond and Bellwood, 2003; Holden, 2006), to name but a few prominent examples.[1]

Understanding the evolution of culture is fundamental in cultural anthropology, and a branch of this research, exemplified by Mace et al. (2005), brings to bear sophisticated phylogenetic techniques adapted from computational biology and historical linguistics to study cultural evolution in cross-cultural data. Some recent work makes data available on all of these aspects of culture to facilitate such study. Kirby et al. (2016) combines a variety of information on geography, culture, and language in a unified cross-cultural data set to aid analysis of culture. Even economists have taken an active interest in cultural origins, as ancient cultural practices are seemingly important in understanding economic outcomes in the present (Alesina et al., 2005; Ashraf and Galor, 2013; Spolaore and Wacziarg, 2013). As this literature has evolved, some economists have investigated the sources of ethnic diversity itself (Michalopoulus, 2012). Others have studied the evolution of culture over time (Giuliano and Nunn, 2018; Lowes et al., 2017).

A common thread running through analyses of cultural evolution is the need to understand the geography of culture. Bayesian phylogeography, originally developed for modeling geographical dispersal of viruses (Lemey et al., 2009), has also been applied to analysis of linguistic dispersals (Bouckaert et al., 2012, 2018). The method treats locations as states in a discrete-state, continuous time model of drift (Lemey et al., 2005), also a method used to analyze the evolution of language over time (Forster and Renfrew, 2006). Other recent research blends phylogenetic linguistic analysis with models of geographical dispersal. For example, Currie et al. (2013), compare different expansion scenarios for the Bantu peoples, coupling phylogenetic linguistic analysis with a Brownian-motion-based model of population dispersal.

Extension of techniques for analyzing change over time to geography have thus deepened understanding of the timing and nature of historical population dispersals. Older, more heuristic methods for deducing population expansions are still in common usage, however, and one might wonder how they contrast with and complement newer methods. Perhaps the oldest and most well-known approach for locating linguistic homelands derives from the so-called Age-Area Hypothesis (AAH). Commonly known as the center-of-gravity principle, but also known as the genetic diversity principle or Sapir's principle—the ideas underlying the AAH have proven to be powerful, flexible tools in understanding population dispersals. The AAH posits that the homeland of a group of related languages is likely where the constituent languages are most diverse, or divergent.[2] The AAH has proven to be an indispensable in filling in gaps in the historical relationships between cultures, and has often been invoked to buttress archeological, historical, and genetic evidence.

Indeed, applications of the AAH abound, and often usage of the idea is so second-nature that it is applied without explicit reference. Atkinson and Gray (2003), following Renfrew (1987) and Dolgopolsky (1988), couple computational linguistics with the AAH to suggest that the Indo-European languages originated in Anatolia, not as is sometimes argued, on the steppes of Siberia. Ruhlen (1994) uses the AAH and also provides overviews of debates about the origins of Na-Dene Native American cultures, the Bantu expansion in Africa (also see Ehret, 2002; Grollemund et al., 2015), and the peopling of the South Pacific. Ehret (2002) makes extensive and efficient use of the AAH in his sweeping account of how and when the cultures of Africa came to occupy their current locations. The AAH has exerted a significant impact on our understanding of ancient world history.

The aim of this paper is to develop a firm theoretical underpinning for the AAH, which the AAH currently lacks. The lack of theoretical basis for the AAH generates doubt as to its actual meaning, creating ambiguities, which can lead to outright contradictions between researchers who purport to be using the same methods. The approach of the paper is to describe the underlying process leading to population movements, and then show how migratory expansions that emanate from locations with more divergent languages are simpler in that fewer distinct historical events, and fewer model parameters are required to explain the expansion, which in turn implies a larger likelihood value. The approach leads to measures of divergence between cultures and a likelihood-based interpretation of the Age-Area Hypothesis. One can view the paper as an attempt to construct a rigorous model of migration and linguistic dispersal from the principles embodied in the earlier works of Sapir (1916) and Dyen (1956).

Overview of the literature

The origins of the hypothesis. One of the earliest applications of the AAH is Sapir (1916), which discusses the origins of Native American peoples speaking languages belonging to the Na-Dene family.[3] The Na-Dene language speakers cluster into three distinct geographical regions: (1) the northern Pacific coast of Alaska and Canada, and (2) the coast of Northern California and Oregon, and (3) the Southwestern United States. Navajo is the Na-Dene language with the largest number of current speakers. Athabaskan-speaking peoples such as the Dogrib and Chipewyan of Northern Canada also belong to this group.

While any one of the three geographical clusters identified by Sapir might be the place of origin of the entire language family, Sapir (1916) argued the geographical point of origin of the Na-Dene cultures was the Northwestern Pacific. His argument was based on the fact that Athabaskan peoples collectively form a branch of the Na-Dene linguistic family, while the cultures of the Northwest form separate branches of the tree on their own:

> The argument for the northern provenience of the Athabaskan tribes is clinched by a ... linguistic fact, namely that the Athabaskan dialects form one of the three major divisions of the Na-Dene stock, the other two being Haida and Tlingit. The fact that the latter are spoken in the northwest coast area so emphatically locates the historical center of gravity of the stock in the north that it becomes completely impossible to think of the Athabaskan tribes as having spread north from California or the southwest. (Sapir, 1916, p. 81–82)

Sapir's argument for a Northwestern Pacific origin point is based on the languages spoken by the Haida and Tlingit cultures of the northwest Pacific coast comprising distinct branches of the Na-Dene linguistic tree. Hence, they branched off from the rest of the cultures at an earlier time. His argument does not rely on any characteristics of the constituent cultures of the tree, or of the present geographical arrangement of these cultures. Moreover, his argument does not mention any ethnographic or archeological evidence; it only uses phylogenetic relationships between cultures implied by language relatedness to infer past geographical distribution and spread. Others, notably Dyen (1956), have pushed these ideas a bit further and posited that phylogenetic relationships suggest a probabilistic interpretation of possible past geographical movement. Dyen (1956, p. 623), in citing Sapir's passage, writes "the probabilities of different reconstructed migrations are in inverse relation to the number of language movements that is required." (Dyen, 1956, p. 613) Sapir and Dyen are superimposing phylogeny upon geography, assuming that the phylogenetic relationship between languages implies something about the timing and direction of population movements determining the current geographical distribution of the cultures.

Recent literature. The insight that linguistics provides a precise description of the timing of population splits and migrations is now a critical tool in the analysis of historical population dispersion.[4] But linguistic relationships themselves say nothing about geography, or the direction of population movements. While Bayesian phylogeographic methods (Bouckaert et al., 2012, 2018) seem to produce patterns consistent with Sapir's vision of the dispersal process, the question remains: Why couldn't people have migrated from central Canada towards the pacific northwest coast in the distant past, and then later from central Canada to the American southwest?

As it happens, population dispersal narratives often contradict the AAH. As one example, Kitchen et al. (2009) use sophisticated Bayesian methods to construct a phylogenetic tree of the Semitic languages, finding that the Akkadian language is the most divergent of the Semitic grouping. Yet Kitchen et al. (2009) then argue that the Semitic cultures likely originated in the Levant, not 1000 miles to the east where Akkadian was spoken. While Kitchen et al. (2009) offer a plausible explanation for the contradiction, it is hard to see why in this case inconsistency with the AAH is tolerated, and because of the lack of a theoretical structure for the AAH, it is also hard to see what has gone wrong. This prompts a general question: How likely must contrary evidence be to overturn the predictions of the AAH? The methods developed in this paper provide a probabilistic assessment as to how convincing historical, archeological, or other evidence would have to be to outweigh the AAH. Further discussion of this apparent anomaly is taken up in section "Origins of Semitic".

Definitional confusion is another unfortunate consequence of the lack of theoretical underpinnings for the AAH. Consider the discussion of the origins of Native American peoples speaking Algic languages in Wichmann et al. (2010, p. 78). They note that some researchers, notably Sapir (1916), hypothesize that these peoples originated in the American West, as the most divergent languages of the group, including Blackfoot, Arapaho, and Cheyenne, are spoken there. In fact, Sapir (1916) noted that the even-more-distantly related languages of Wiyot and Yoruk suggest even deeper origins on the west coast of America in California. In almost exact opposition to this position, Wichmann et al. (2010) suggest the Algic peoples originated in the eastern United States, in New York-New England, a conclusion supported by other research on the geographic origins of Algonquian languages (for example, Siebert (1967)). Shockingly, arguments for both origin stories are based on the AAH! A little reflection reveals this is the result of confusion over the definition of the AAH. Sapir (1916) defines the AAH in terms of *maximum linguistic divergence*, looking for the language group that is most different from all the others in the group, while Wichmann et al. (2010) define the AAH in terms of *maximum linguistic diversity*, i.e., the place where the density of different dialects is the highest.[5] The key questions then become: What is the correct way of stating the AAH? And what measure of divergence or dissimilarity best reflects the essence of the AAH?

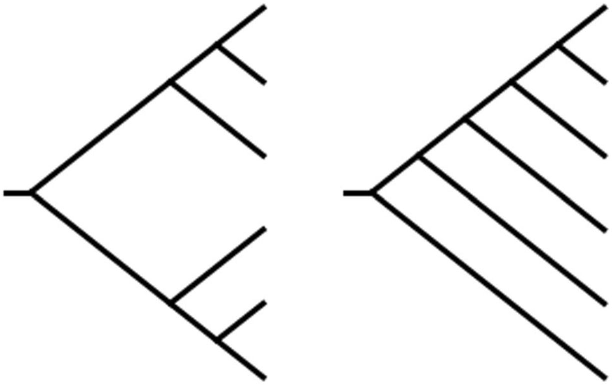


Fig. 1 Balanced (left) and unbalanced phylogenetic trees. The balanced tree on the left-hand side of the figure has equal numbers of terminal nodes along each branch, unlike the unbalanced tree on the right.

The role of a theory of migration. This discussion suggests why what could be called an economic theory of migration—that is, a theory in which decisions about migration and movement are made rationally by populations in a way consistent with what is observed—might be useful in formulating a rigorous foundation for the AAH. For example (Currie et al., 2013) write: "large-scale migrations of human populations are thought to be a major feature of human history during the Holocene."(Currie et al., 2013, p. 1). Others (Neureiter et al., 2021) have found that Bayesian phylogeographic methods work best when following an "expansionary" vision of the migratory process.

These ideas suggest that a model purpose-built for analyzing large-scale population movements might lend insight into the structure and patterning of population dispersals. The model developed here is based on simple economic principles and a model of mass migration. The micro foundations explain how mass population movements might occur in an ongoing fashion from occupied locations to new locations at irregular time intervals. The theory suggests that these movements are well-described by a Poisson-exponential model of the timing of migratory events. The model also shows how probabilities of the locations of different cultures being origin points relates to the Occam's-razor-like idea that simpler migration narratives are more likely. Migratory paths that originate at deeper points in a phylogenetic tree can be thought of as simpler in a precise probabilistic fashion under the model, in that inclusion of such paths in a migratory history also results in a model with fewer parameters, which in turn generates a greater likelihood. These ideas also suggest a way to define divergence or dissimilarity measures that are consistent with the AAH, some of which seem to have been anticipated by Dyen (1956). These ideas can be developed into an Age-Area Theorem, which makes explicit the assumptions under which a culture that is more linguistically divergent from the others in the stock is also more likely to reside at the point of origin of the stock.

The resulting model allows explanation of some other features of linguistic phylogenies. Many linguistic trees exhibit imbalance.[6] The left-hand side of Fig. 1 shows a balanced phylogenetic tree, while the right-hand side of Fig. 1 shows an unbalanced phylogenetic tree. Many linguistic phylogenies resemble the right-hand figure. As Aldous (2001) notes, this is

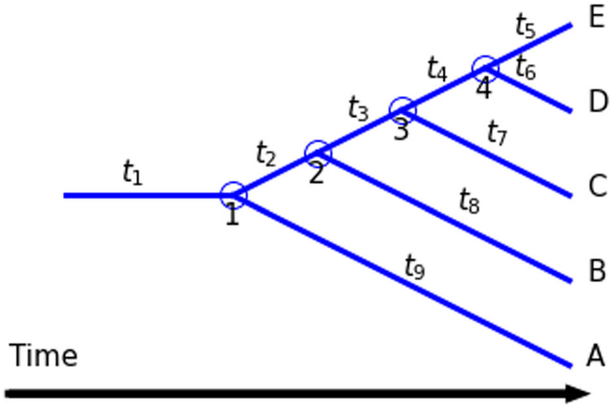


Fig. 2 A phylogenetic tree displaying the relationship between five cultures. The tree shows the relationship between five hypothetical cultures. The most time-distant split was between A and the common ancestor of B, C, D, and E.

a difficulty for modeling relatedness between cultures because it means simple models of branching such as a pure birth processes are not able to accurately model the data, even when considering the possibility that branches go extinct (Holman, 2009). Accordingly, it has been suggested (Gray et al., 2013) that rapid expansion and change could lead to an imbalanced structure, which might be a better way to model the sort of dispersal patterns consistent with population dispersals and linguistic drift (Neureiter et al., 2021). The model in this paper is based on a moving "propensity to migrate," which in turn is based upon a theory of local population growth and resource depletion. Trees like those on the right of Fig. 1 are more likely in a well-defined sense in the model: when presented with a choice between a tree structure to explain the relationship between a group of cultures, the right-hand picture of Fig. 1 would be preferred.

The model also allows for probabilistic assessment of the likelihood of alternative population dispersal narratives, and how these narratives change as additional evidence is included. In the application in section "Origins of emitic", the model is applied to some theories of ancient population dispersal of the Semitic peoples. The application shows how an important piece of the puzzle is where exactly one includes the ancient and relatively unknown Eblaite language on the Semitic tree.

Finally, consider the comment made by Greenhill and Gray (2005), who discuss the peopling of the South Pacific and also present a detailed discussion of quantitative methods in historical linguistics. They develop statistical tests comparing different hypotheses for how the South Pacific came to be settled. They write the following in describing the need for formal modeling and associated hypothesis tests in resolving disputes about migratory routes:

> ...many expansion scenarios are little more than plausible narratives. A common feature of these narratives is the assertion that a particular line of evidence (archeological, linguistic, or genetic) is 'consistent with' the scenario. 'Consistent with' covers a multitude of sins. Rigorous tests require a measure of exactly how well the data matches the proposed scenario. They also require an explicit evaluation of alternative hypotheses. ...a framework for the rigorous evaluation of these hypotheses is clearly desirable. (Greenhill and Gray, 2005, p. 31)

This statement could easily have been written in describing the reason for the current paper, and as a justification for taking building a formal model around the AAH.

Problem description

It is helpful to have a working example that fixes ideas and identifies the issues that a theory should address. Consider the phylogenetic tree in Fig. 2. The tree describes the degree of linguistic relatedness between a group of hypothetical cultures, and while it is drawn as a rooted tree, this aspect of the tree is inessential. What is important is that the tree captures the process driving cultural diversification as a series of times at which a single culture split in two, coming to occupy simultaneously a new and the old geographical location.[7] Some segment of the population moved to a new location, at which time the languages spoken by the cultures began to drift apart, allowing the phylogenetic tree to assume its structure. Culture A speaks the most divergent language, meaning it and the other cultures B, C, D and E—more correctly, the common ancestor of $B, C, D,$ and E—geographically divided at the most distant point in time. D and E are the most closely related cultures.

The AAH leads one to posit the geographical origins of this group of peoples is at A's current location.[8] Recursive application of the AAH would lead one to a most likely description of the migrations producing the current phylogenetic relationship between cultures and their geographical distribution. After originating at A, there was then a migration from A's location to B's location, then from B's to C's location, and then from C's to either D's or E's location.

The tree in Fig. 2 imposes constraints on the time sequencing of splits. It cannot be that a migration from D's location to E's preceded a migration from D's to C's location. This is inconsistent with the observed linguistic drift and implied timing of splits. The tree does not, however, impose directional constraints; it is possible, for example, that an initial migratory episode from C's to A's location occurred, followed by one from C's to B's location, which was then followed by a migration from C's location to D's or E's location. This alternative scenario yields a different homeland for the peoples, and is fully consistent with the phylogeny in Fig. 2.

A hypothetical geography coinciding with the phylogeny, along with some migratory events consistent with the tree timing, is described in Fig. 3. The AAH asserts that the right-hand sequence on the figure is a less-likely migratory history than that on the left. Why? The example shows the limited usefulness of positing a minimal number of moves to explain migrations, or even reliance on minimum distance paths. The two paths both require four distinct population movements. There also is not much difference in the total physical distance traversed, so physical distance is not of much use in intuiting the most likely dispersal. The AAH seemingly appeals to a certain kind of simplicity in migratory movements in suggesting a sequence of events - the events on the left-hand side of Fig. 3 seems to require one continuing expansion, while the events on the right-hand side require three separate expansions: an initial migration from A's to C's location, followed by another from A's location to B's. A third expansion then explains how the last two groups came to be at their positions in a way consistent with the Phylogeny: an expansion starts from C's location and goes to D's, and then from D's to E's location. Why is the left-hand narrative more likely?

The example presented in Figs. 2 and 3 can actually be viewed as a simplified representation of the opposing sides of the debate about the origins of many culture groups and even loosely applies to Sapir's observations about Na-Dene speakers. Consider also

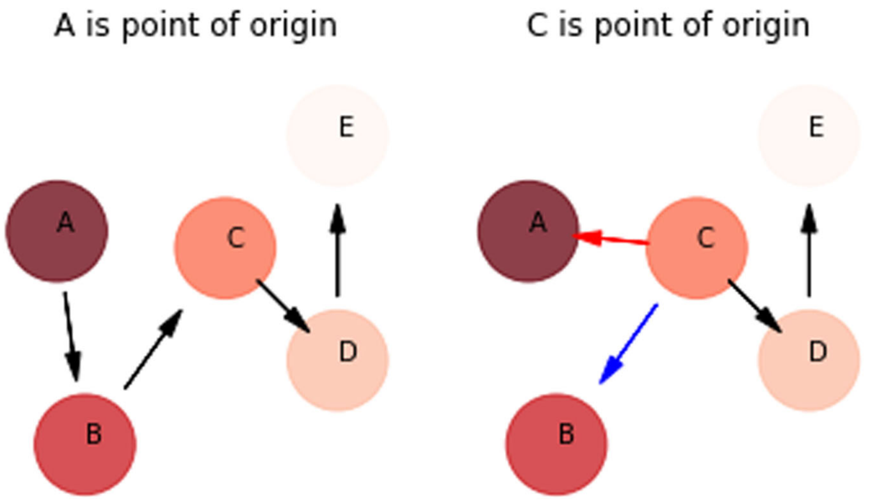


Fig. 3 Potential migratory routes consistent with the phylogenetic tree in Fig. 2. A single migratory process starting at A is depicted on the left, while three different processes starting from C are shown on the right.

the debate over the origins of the Afroasiatic or Afrasan languages and cultures, a linguistic group with a wide geographical distribution covering the Middle East and Northern Africa that includes Semitic and Arabic languages, ancient Egyptian, and a variety of languages spoken in the Ethiopian highlands. Diamond and Bellwood (2003) suggest that this language group originated in the Levant, while Ehret et al. (2004) suggests that this "generally abandoned" view "...[fails] to engage the five decades of Afroasiatic scholarship that rebutted this idea in the first place. This extensive, well-grounded linguistic research places the Afroasiatic homeland in the southeastern Sahara or adjacent Horn of Africa..." (Ehret et al., 2004, p. 1680). Roughly speaking, this argument is about the whether the right-hand (Diamond and Bellwood, 2003) or left-hand side (Ehret et al., 2004) of Fig. 3 is the more likely dispersion scenario.

The example of Figs. 2 and 3 is also in a sense emblematic of the recurring tension in geolocating the origins of peoples. Is it more likely that peoples originated at the center of their current geographical distribution, or current center of population, with the most distant relative, or somewhere else? If one believes migratory events to be rare, and wishes to conserve them in explaining historical migrations, what sort of model would reflect this concern? How might one characterize migratory parsimony in a meaningful mathematical and probabilistic way?

Theoretical framework

This section presents a model of migratory events, and as this aspect of the model reflects the ramifications of the theory, the discussion of the microeconomic foundations of the model is postponed until this part of theory is clear.

The development of the theory begins by building up a phylogenetic tree from its constituent migratory events. These events, which constitute a node on the phylogenetic tree combined with a directional arrow, as shown in Figs. 2 and 3, can be collected into migratory chains. Migratory chains can then be grouped into a migratory history. In the end, a history with fewer chains will be simpler, and simpler histories will have higher likelihood under the model, while at the same time being associated with languages that are more divergent.

The model begins by assuming that a phylogenetic tree describing the relationships between a group of cultures is known. In the simplest version of the model, only the structure of the tree and hence the sequence of population splits need be known, not the branch lengths or exact timing of the splits. The phylogenetic tree is assumed to be full, rooted, and binary, which implies there is a single origin node and root.[9] All nodes excepting the root have a single parent node, all interior nodes have exactly two children, and all terminal nodes have zero children. A binary tree with $K+1$ terminal nodes (sometimes called taxa or leaves) has K internal nodes. The tree in Fig. 2, for example, has five terminal nodes/taxa and four internal nodes.

To link phylogeny with geography, assume that each terminal node coincides with a physical location. Each location/terminal node then carries a label $c_i, i \in 1, 2, 3, \ldots, K+1$ and a location l_i, so one may refer to the "culture" c_i currently at "location" l_i.[10] The locations of the terminal nodes on Fig. 2 are mapped on Fig. 3. A branch connected to nodes of the tree is associated with a *migratory event* E_k. That is, a migratory event E_k is a tuple consisting of an internal node $k \in K$, an elapsed time t_k, a starting location-culture pair i, and an ending location-culture pair i':

$$E_k = \{t_k, (c_i, l_i), (c_{i'}, l_{i'})\}$$

On Fig. 2, one might then consider the migration from A's location to B's location as $E_4 = \{t_4, (A, l_A), (B, l_B)\}$, which indicates after t_4 units of time elapsed, a migratory event from the current location of culture A to the current location of culture B occurred. The spatial representation of this event is as shown on the left-hand part of Fig. 3. The migratory event is a movement of a fraction of the population at the first location to the second, so that now the original population occupies *both* the initial and the terminal location. Now that the population has split into two sub-populations, the sub-populations begin to change independently of one another, and the languages spoken by A and B drift apart as reflected in the phylogenetic tree.

An important part of the model are null migratory events:

$$E_k^o = \{t_k, (l_1, c_1), ()\}$$

The null event E_k^o means that proceeding from node k of the tree, t_k time elapsed, and no migratory event occurred. Null migratory events are of interest because they are a useful way to characterize the terminal nodes of the tree.

Migratory chains. A *migratory chain* is a sequence of *migratory events* which form a directed path both through space and along the phylogenetic tree, coupled with a parameterization of the process describing the times and spatial movements of the process. The parameterization of the chain is referred to as θ, which will describe a probability distribution over the timing of migratory events, and perhaps spatial aspects of migration. Thus, a migratory chain consists of the tuple: $C = \{\{E_m\}_{m=1}^{n}, \theta\}$. In the simplest version of the model, the parameterization of a migratory chain consists of some probabilistic model describing the timing of migratory events, where the spatial model just specifies that any spatial movement to an as-yet unoccupied location is equally likely.

A null migratory chain can be defined using a single null migratory event: $C^o = \{\{E^o\}, \theta_o\}$. This is a migratory chain starting from some location from which no migration is observed. These null chains characterize locations that are occupied and "passed through" as part of a continuing mass migration, but then do not produce additional mass migrations themselves. Indeed, one interesting feature of many discussions of migration and cultural evolution is that ancient population movements figure into the discussion, but it is often never explained why some migrations apparently continue on to multiple locations, while others peter out after a single movement.

The propensity to migrate. The model posits that each migratory event carries a propensity to migrate with it, which leaves the initial location for the new location whenever a population splits. This means that the migratory chain and its migration propensity moves along with to the most recently occupied location as the migratory chain progresses. This view of the migratory process means that mass migrations are unlike other stochastic processes describing population dispersion, such as a Yule/pure-birth process, as migration is more likely to occur and continue from newly occupied locations than from previously occupied locations, as is seemingly the case in expansionary population dispersals. In a branching process, each split of a population results in two new populations, each of which could split again, and so on. The migratory process is best thought of as the result of a split in which part of the population leaves for a new location and carries a similar potential for future splitting along with it. Any subsequent out migration from the original location is governed by a new chain with a new parameterization. What sort of economic process organically produces such an effect should be explained and is one of the key roles played by the micro foundations of the model presented in section "Microeconomic foundations."

Migratory histories. A *migratory history* is a collection of migratory chains $H = \{C_j\}_{j=1}^{N}$ that together comprise a minimal spanning of the phylogenetic tree, so that every branch of the tree is traversed by a single migratory chain. Hypothesizing that a location is the geographical point of origin of all the constituent cultures of the tree amounts to positing some sequence of migratory chains that minimally span the tree, with the deepest chain starting emanating from the origin culture-location. Each culture-location combination (l_i, c_i) represented by a terminal node of the tree can be associated with a set of migratory histories $\mathcal{H}_i$ that start at the location.

A migratory history has some count of the total number of non-null migratory chains required to span the tree, referred to as $N(H)$. On Fig. 3, the history described by the sequence of events on the left has $N(H) = 1$, while for that on the right, $N(H) = 3$. Define a counting function for the number of migratory events in a migratory chain as $n(C)$, which counts the number of non-null migratory events spanned by the chain. The sole chain in the scenario on the left of Fig. 3 has $n(C_{ABCDE}) = 4$, while the three chains on the right-hand side of the figure give $n(C_{CA}) = 1$, $n(C_{CB}) = 1$, and $N(C_{CDE}) = 2$.[11]

The basic assumptions characterizing migratory chains and histories are then as follows.

Assumptions 1 Migratory Events, Chains, and Histories

1. Each migratory chain occupies a single location at any given point in time.
2. When a migratory event occurs, the propensity to migrate for the chain moves to the new location, and a new chain starts at the origin location.
3. Migratory chains move from their current locations to a new location at random times according to a chain-specific probability density function.

The only element that has not been discussed in the above list of assumptions is 3). This assumption implies that a migratory history with fewer chains introduces fewer parameters into the model, and under some basic assumptions this translates into a larger likelihood, as discussed in the next section.

Likelihood and a Poisson-exponential model. The building blocks of the previous section can be used to create a likelihood associated with any particular history, which then leads to measures of divergence that allow a precise statement of the AAH. These measures of divergence are referred to as "Dyen Divergence measures" after Dyen (1956). A Dyen Divergence measure is not a conventional distance measure, in that it does not measure how distant one culture is on average from other cultures according to some pairwise comparison of characteristics, as is used in, for example, Wichmann et al. (2010). A Dyen Divergence measure captures how dissimilar a culture is from the rest of the component cultures in a Phylogenetic tree by relating dissimilarity to a probabilistic model of the tree. The Dyen measure also reflects the relative chances that any particular culture/location is the point of origin of the group of related cultures.

The intuition underlying the AAH suggests that simpler migratory histories should be more likely, and in the model simpler histories are those that are comprised of fewer migratory chains. In this way, each migratory chain can be associated with a one-time historical occurrence, such as a discovery of a means of exploiting a new resource, and one might prefer to use fewer such exceptional events to explain the entire language family expansion if possible. Operationally, as each chain carries its own parameterized distribution, a history with fewer chains also has fewer parameters, grouping more observations together, and is therefore simpler in that it has a smaller number of parameters.[12] When a Poisson likelihood is optimized over this smaller number of parameters a larger likelihood results; this is also often true of an exponential likelihood. Thus, histories with smaller N that group together larger values of n, have higher probabilities. In addition to explaining the propensity to migrate, the microeconomic foundations for the model also suggest that a Poisson-exponential model is a good way to characterize migratory chains.

Consider a migratory chain with a number of constituent events n occurring over a time T, where as a migratory event occurs, the chain leaves the old location and occupies the new one. A Poisson likelihood comprises a simple model of these n migratory events spread out over time T:

$$L = \frac{(\lambda T)^n e^{-nT}}{n!} \tag{1}$$

The value of λ that maximizes (1) is $\lambda^* = \frac{n}{T}$, and substituting this back into (1) gives a chain likelihood

$$L = \frac{n^n e^{-n}}{n!} \tag{2}$$

the critical feature of the function (2) is that the kernel $\frac{n^n}{n!}$ is convex, as is elaborated on further below. As the concentrated likelihood in (2) does not depend upon T, it is a suitable model of a migratory chain when the number of branching events are known, but the timing of these events is not.[13] By contrast, if branch lengths are known, an alternative is to use an exponential model for n events occurring over a total time T:

$$L = \lambda^n e^{-nT} \tag{3}$$

The exponential likelihood in (3) has concentrated value:

$$L = \left(\frac{n}{T}\right)^n e^{-n} \tag{4}$$

If a history is comprised of N chains, its likelihood is:

$$L_H = \prod_{j=1}^{N} L_j$$

where L_j is given by either (1) or (3).

The migratory histories described in Figs. 2 and 3 can be used to illustrate the ideas. The left-hand migratory history on Fig. 3 requires an initial migratory chain to start at A, which then proceeds from location A to B, from B to C, and then finally to D or E. However, each time the migratory chain proceeds to a new location, by Assumption 1, a new migratory chain starts in its place. In the example in Figs. 2 and 3, most of these new chains are null chains—they never create any new migratory events and only lead to terminal nodes. The probability of observing this sequence of events, then, can be written by combining the densities of the component chains of history H_A, which can be written as a combination of the migratory chain C_{ABCDE} along with a collection of null chains needed to span the tree:[14]

$$L_A = \text{Prob}(H_A) = P(C_{ABCDE})P(C_A^o)P(C_B^o)P(C_C^o)P(C_D^o)$$

Using the Poisson distribution to parameterize the timing of migratory events, L_A is:

$$L_A = \frac{(\lambda_1 T)^4 e^{-\lambda_1 T}}{4!}\frac{(\lambda_A t_A)^0 e^{-\lambda_A t_A}}{0!}\frac{(\lambda_B t_B)^0 e^{-\lambda_3 t_B}}{0!}\frac{(\lambda_C t_C)^0 e^{-\lambda_C t_C}}{0!}\frac{(\lambda_D t_D)^0 e^{-\lambda_D t_D}}{0!} \tag{5}$$

Equation (5) simplifies to:

$$L_A = \frac{(\lambda_1 T)^4 e^{-\lambda_1 T}}{4!} e^{-\lambda_A t_A} e^{-\lambda_B t_B} e^{-\lambda_C t_C} e^{-\lambda_D t_D} \tag{6}$$

The log-likelihood associated with equation (6) is:

$$\ln L_A = 4 \ln(\lambda_1 T) - 4\lambda_1 T - \ln(4!) - \lambda_A t_A - \lambda_B t_B - \lambda_C t_C - \lambda_D t_D \tag{7}$$

The log-likelihood in (7) is maximized with $\lambda_A = \lambda_B = \lambda_C = \lambda_D = 0$ - since these chains are all null, the maximum likelihood estimate of the rate parameter for the timing of migratory events along these chains is zero. In contrast, $\lambda_1^* = \frac{4}{T}$. Substituting this and other optimal values back into the objective function gives the (concentrated) likelihood L_A as:

$$L_A = \frac{4^4 e^{-4}}{4!} \tag{8}$$

Equation (8) yields a probability for migratory history A, but also embodies the concept of simplicity. The back story for (8) is simple in that only one non-degenerate migratory chain is needed to explain the whole tree, given that the migratory history starts at A. Contrast this with the case in which C is posited to be the origin point. To maintain consistency with the splitting events of the phylogenetic tree, the requirements are: (1) a chain starting at C's location leading to A's, (2) another migratory chain starting at C's location going to the location of B, and then (3) a migratory chain that starts at C's location proceeding to D's (or E's) and then finally to E's (or D's). Degenerate chains start at location C and D (or E) when each chain moves on from these locations. The likelihood for history H_C is then:

$$L_C = \text{Prob}(H_C) = P(C_{CA})P(C_{CB})P(C_{CDE})P(C_D^o)P(C_C^o)$$

Again parameterizing each chain using the Poisson distribution gives likelihood:

$$L_C = \frac{(\lambda_1(t_4+t_A)^1 e^{-\lambda_1(t_4+t_A)}}{1!}\frac{(\lambda_2(t_3+t_B))^1 e^{-\lambda_B(t_3+t_B)}}{1!}\frac{(\lambda_3(t_2+t_1+t_E))^2 e^{-\lambda_3(t_2+t_1+t_E)}}{2!} \times \frac{(\lambda_C t_C)^0 e^{-\lambda_C t_C}}{0!}\frac{(\lambda_D t_D)^0 e^{-\lambda_D t_D}}{0!} \tag{9}$$

Maximizing L_C in (9) with respect to λ_1, λ_2 and λ_3, noting that $\lambda_C = \lambda_D = 0$, and substituting the result back into (9) gives:

$$L_C = \frac{1^1 e^{-}}{1!}\frac{1^1 e^{-1}}{1!}\frac{2^2 e^{-2}}{2!} = \frac{2^2 e^{-4}}{2!} \tag{10}$$

This forms a basis for a probabilistic comparison of the chances each of these two locations were the point of origin of the tree.[15] The likelihood that A is the point of origin relative to C is $L_A/(L_A + L_C)$. Since we have $L_A \propto \frac{4^4}{4!}$ and $L_C \propto \frac{2^2}{2!}$, the relative probability A is the point of origin is:

$$\frac{\frac{4^4}{4!}}{\frac{4^4}{4!} + \frac{2^2}{2!}} = 84\%$$

The result that A is a relatively more likely point of origin owes to the kernel:

$$z(n) = \frac{n^n}{n!} \tag{11}$$

which is convex in n and increases more rapidly than any polynomial in n. This "extreme" convexity,[16] which is due to the Poisson-exponential distribution, implies that histories that require fewer non-degenerate parameters, or equivalently, that lump migratory events into fewer, longer chains are more likely. The kernel (11) also means that if one could potentially increase the length of a longer chain at the expense of a shorter chain, one would increase the likelihood. That is, if one has two histories with lengths n_1 and n_2, where $n_1 > n_2$, and one could rearrange things to subtract a migratory event from the second chain and add it to the first, this will increase the likelihood. This is because, using the Poisson likelihood kernel:

$$\frac{(n_1+1)^{n_1+1}}{(n_1+1)!}\frac{(n_2-1)^{n_2-1}}{(n_2-1)!} > \frac{n_1^{n_1}}{n_1!}\frac{n_2^{n_2}}{n_2!} \quad \rightarrow \quad \left(\frac{n_1+1}{n_1}\right)^{n_1} > \left(\frac{n_2}{n_2-1}\right)^{n_2-1}.$$

Such rules of thumb—fewer chains and longer chains should be formed if possible—help one find a history with maximum divergence or likelihood for each culture location pair, which is an important part of the Age-Area Theorem presented below. This is discussed after first defining a divergence measure which leverages these ideas:

Dyen divergence. A simple measure of divergence that replicates the properties of (11) can now be created. This measure forges a link between probability and a measure of how divergent a particular culture is from the rest of the tree.

Definition 1 *Dyen Divergence*. Define a function which collects the number of non-degenerate chains in a migratory history, and the number of events in each of its constituent chains as follows:

$$d(H) = -N(H)\ln(2\pi) - \sum_{j=1}^{N(H)} \ln n(C_j)$$

Let $\mathcal{H}_i$ denote the set of migratory histories emanating from culture-location pair i. Define the **Dyen Divergence** of culture-location i as the maximum value of $d(H)$ for the culture-location pair:

$$D_i = \max\{d(H_{1i}), d(H_{2i}), \dots d(H_{Li})\}, \quad \forall \quad \{H_{li}\}_{l=1}^{L} \in \mathcal{H}_i$$

The Dyen divergence for culture-location pair A from the previous example requires considering two possible histories, which are almost trivially different: a history with migratory sequence $ABCDE$, and one with sequence $ABCED$. Based on the tree structure alone, one cannot distinguish between these two histories, but can select either in forming the divergence measure for A:

$$D_A = -\ln 2\pi - \ln 4 \approx -3.22$$

As $N(H_A) = 1$, and $n(C_{H_A,1}) = 4$. By contrast, D_C, corresponding to the right-hand of Fig. 2, also produces two possible histories; with it once again being impossible based on the tree alone to distinguish between CA, CB, CDE and CA, CB, CED. Still, these histories produce the same value for the Dyen divergence measure, which is:

$$D_C = -3\ln 2\pi - \ln 1 - \ln 1 - \ln 2 \approx -6.206$$

Since $D_A > D_C$, A is more divergent that C. One can also see from this example how the divergence measure exploits concavity of the natural log function to preferentially treat longer migratory chains. If a chain of length 4 can be formed, instead of two chains of length 2, the Dyen divergence measure gives the longer chain a higher score, because $\ln(4) < \ln(2) + \ln(2)$, or $-\ln(4) > -\ln(2) - \ln(2)$.

Age-Area Theorem. The Age-Area Theorem links the divergence measure to a probabilistic model, which allows one to say a more divergent culture is more likely a point of origin of the group of related cultures.

Theorem 1 (Age-Area Theorem). *Define Dyen Divergence as in definition 1, and suppose that mass migrations along a migratory chain occur at times described by a Poisson model. Then*:

$$D_i \geq D_{i'} \Longleftrightarrow L_i \geq L_{i'}$$

and if

$$i_D^* = \arg\max_i \left[D_1, D_2, \dots D_i, \dots, D_K\right]$$

and

$$i_K^* = \arg\max_i \left[L_1, L_2, \dots, L_i, \dots, L_K\right]$$

Then $i_D^* = k_K^*$—the most likely point of origin has the largest Dyen divergence measure.

Proof. The probabilistic model suggests that the likelihood associated with a particular location being the point of origin can be written as the product of a group of concentrated Poisson likelihoods:

$$L_k = \Pi_{i=1}^{N_k} \frac{n_i^{n_i} e^{-n_i}}{n_i!} \tag{12}$$

by Stirling's lemma, $n! \approx (2\pi n) n^n e^{-n}$, so the right hand side of (12) is approximately:

$$L_k \approx \Pi_{i=1}^{N_k} (2\pi n_i)^{-\frac{1}{2}} \tag{13}$$

Since monotone transforms of functions preserve order, square L_k in equation (13) and take the logarithm to get $d(H_k)$. Since $d(H_k)$ is a monotone transform of L_k, it follows that whenever $D_i > D_j$, $L_i > L_j$, $i, j \in K$ to the same degree of accuracy as the Stirling approximation.

The distance measure attaches significance to the number of events that proceed from a given split, and in this way creates larger values for the deepest routes of the tree. This is very much in line with the spirit of Sapir's idea that one should, in determining the homeland of a language, focus on the deepest roots of the tree; or, as Sapir says, should focus not on "all the dialects of the stock, but rather on the basis of its major divisions."[17] (Sapir, 1949, p. 455)

Example. Consider first the left-hand side of Fig. 4, which is drawn taking culture E's location as the point of origin. On the figure dashed lines represent null chains, while solid lines represent chains composed of at least one migratory event. The red solid line represents the migratory chain C_{EDABC}, with the red dashed lines representing null migratory chains. The chain starts with an initial migratory event in which a sub-population left E and went to D, with the chain continuing on to A, then B and finally C. To complete a history assuming E is the point of origin, another, more recent chain, written C_{EFGIH} is also needed, and this is shown in green on the figure. Sometime after the migration that started with a movement to D, a new chain started with a migration to F that then moved on to I, then G, then H. By the Age-Area Theorem, this migratory history will produce the largest likelihood associated with E being the point of origin of the constituent cultures of the tree, as this history has grouped all migratory events into the smallest possible number of chains given the event started at point E.[18]

The associated likelihood reflecting the two non-degenerate chains comprising this history is:

$$L_E^* = \frac{4^4 e^{-4}}{4!} \frac{4^4 e^{-4}}{4!} = \left(\frac{4^4}{4!}\right)^2 e^{-8}$$

The Dyen Divergence measure associated with this history is:

$$D_E = -2\ln(2\pi) - 2\ln(4) = -6.44$$

Suppose instead that location A is posited to be the point of origin, as shown on the right of Fig. 4. The history with the smallest number of long chains requires an initial migratory chain to produce the opposite branch of the tree, C_{CEFIGH}. This migration left A for location E, then moved onto F, then I, then, G, then H. Sometime later, a migration from A to D occurred—a migratory chain with only a single migratory event. Then, a third migratory episode leaving A for B and then C occurred. These three non-degenerate migratory chains combine to give likelihood:

$$L_A^* = \frac{5^5 e^{-5}}{5!} \frac{1^1 e^{-1}}{1!} \frac{2^2 e^{-2}}{2!} = \frac{5^5}{5!} \frac{2^2}{2!} e^{-8}$$

With Dyen divergence:

$$-3\ln(2\pi) - \ln(5) - \ln(1) - \ln(2) = -7.81$$

So, E is more divergent and also the more likely point of origin for the tree of these two possibilities. However, if one were to consider location D as the origin, one has:

$$-2\ln(2\pi) - \ln(5) - \ln(3) = -6.38$$

Thus, the location currently occupied by D is the most likely point of origin of the phylogeny absent other information. Like the history emanating from E, the history starting from D requires only two chains, but improves upon that of E by building a 5-event migratory chain and a 3-event migratory chain, whereas E required two 4-event chains.

Exponential distribution and known branch lengths. The method outlined in the proof of the Age-Area Theorem can be applied to create alternative measures of divergence which may be useful in situations in which more information about the structure of the tree is available. A leading situation is when branch

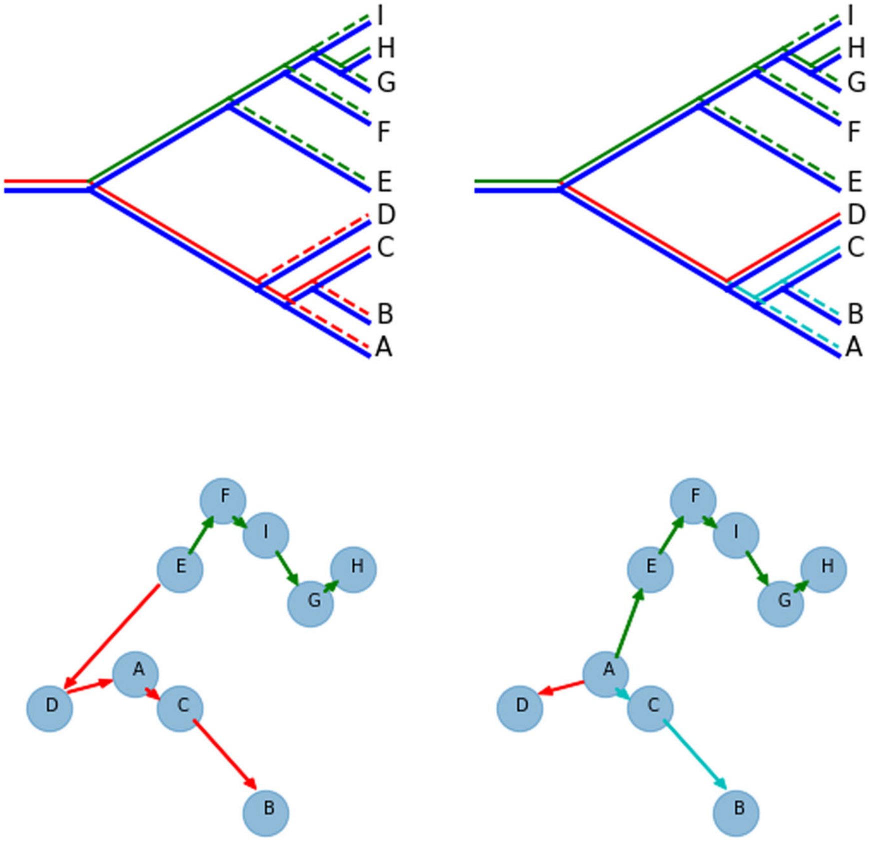


Fig. 4 Two hypothetical migratory histories through a given phylogenetic tree. The picture on the left takes location E as the point of origin, while that on the right takes point A as the origin. Arrow colors on the top and bottom elements of the figure show branching events and population movements.

lengths are known (i.e., divergence times are known). Then, an exponential model may be used for the timing of migratory events. The exponential density associated with n migratory events occurring over a total time span T is:

$$\lambda^n e^{-\lambda T} \qquad (14)$$

The value of λ that maximizes (14) is: $\lambda^* = \frac{n}{T}$. The resulting likelihood for a history with N chains under the exponential model is then:

$$L = \Pi_{j=1}^{N}\left(\frac{n_j}{T_j}\right)^{n_j} e^{-n_j} \qquad (15)$$

where T_j in (15) is the total time length spanned by the migratory chain. A Dyen divergence measure can be formed based on this likelihood embodies the key ideas of the model. All migratory histories will eventually have a term that amounts to e^{-K} because of the K splits needed to explain K interior nodes, and since this is common to all likelihoods, it can be dropped. Also, let $T_j = T\alpha_j$, where α_j is the fraction of the total time depth of the tree elapsed in the chain. Then form the relative likelihood:

$$L \propto \Pi_{m=1}^{N}\alpha_m^{n_m} \qquad (16)$$

So an exponential Dyen divergence measure for the migratory history from culture-location pair k can be formed by taking logs of (16):

$$D_e = \sum_{j=1}^{N} n_j(\ln n_j - \ln \alpha_j) \qquad (17)$$

The divergence measure in (17) emphasizes some additional features of the tree. The convex function $n\ln n$, which forms the first part of the measure in (17), means longer chains increase divergence, as was true of the Poisson-distribution-based measure used in the Age-Area Theorem. However, D_e includes parameters α which captures the fraction of tree time spanned by a migratory chain. Since $\alpha \in (0, 1)$, the smaller is α, the greater the divergence. This means that D_e emphasizes chains which lump a large number of migratory events into a short-time-period chain. This phenomenon plays a prominent role in the literature (Atkinson et al., 2008), and migratory episodes with rapid expansions often figure prominently in migration narratives, such as the peopling of the South Pacific (Gray and Jordan, 2000; Greenhill and Gray, 2005). The exponential model, as it allows for explicit inclusion of timing in splits, therefore admits a way to view more recent and perhaps more rapid expansions as a component of a different migratory episode than events far-removed in time.

A consequence of the way in which α enters D_e in (17) is that it allows the possibility that the history with the longest chain is not necessarily the most likely point of origin. Consider the two possible migratory histories depicted on Fig. 5. On the left-hand side of the figure the migratory history runs from the location of the most divergent culture, A, through B to C to D then finally E's location. The solid blue line running along the tree depicts the sole migratory chain needed to capture this sequence of events, with dashed lines representing null migratory chains that start as the chain moves on. The Poisson likelihood for this tree is $\frac{4^4}{4!}$, with Dyen divergence measure -3.22. The right-hand side of the figure depicts an alternative migratory history, where there was first a migration from B to A, which produced no additional migratory events. Then, a migratory chain followed, running from B to C to D to E. Under the Poisson model, the likelihood of this sequence of events is $\frac{1^1}{1!}\frac{3^3}{3!}$, with corresponding Dyen divergence measure – 4.774. Therefore the Poisson model would suggest that A is the more likely point of origin.

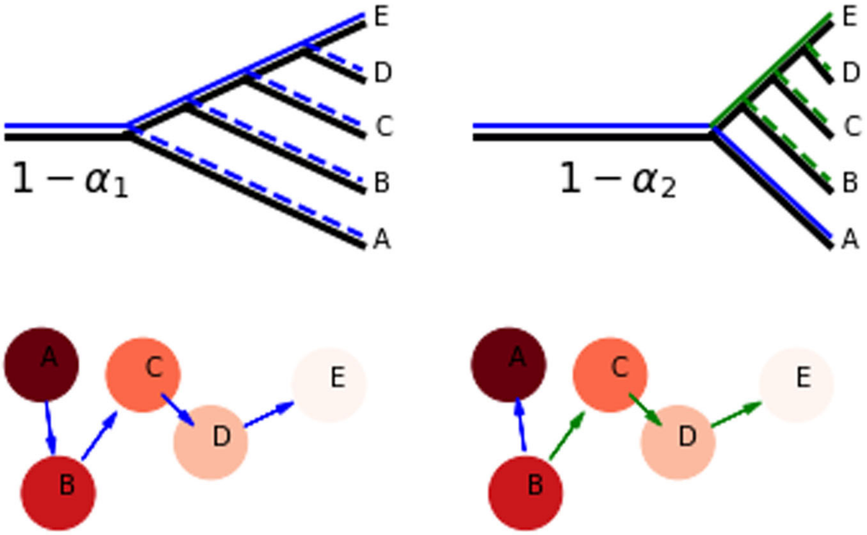


Fig. 5 Different migratory histories for a tree with the same basic structure. On the right side of the figure, migratory events occur in a more compressed time frame.

Under the exponential model the left-hand history has a likelihood of 4^4 after normalizing the overall length of the tree to unity, while the right hand side of the tree has likelihood $\left(\frac{3}{\alpha_i}\right)^3$. The exponential Dyen divergence for the left-hand side of Fig. 5 is $4\ln 4 = 5.54$, while it is $3\ln 3 - 3\ln\alpha$ for the right-hand side. For α small enough—less than $\frac{3}{4^{\frac{4}{3}}} = 0.472$ to be exact, the scenario on the right-hand side of Fig. 5 is more likely.

It is also possible to blend the Poisson and the exponential model if some branches are of known length, while others are not. For example, if the tree is unrooted, one may have no information on the time leading up to the initial population split; one knows only that it occurred. Therefore, one could use a Poisson likelihood for the first split in the tree, followed by exponential branchings for each subsequent split.

Geographical distances. One might wish to allow other factors besides the order and timing of population splits to influence the likelihood of different migratory histories. An important case is inclusion of physical distance as a component of the history so that both the structure of the tree and the length of hypothesized routes impact likelihood. If nothing else, geographical distance could be used to break ties in cases in which the tree offers no guidance. On Fig. 5, the tree itself does not suggest whether either chain should end with a migration from D to E or E to D. However, the geography depicted on the lower part of Fig. 5 suggests that the more natural (i.e., shorter) geographical path might be from C to D, and then from D to E.

A simple model of distance that both makes intuitive sense and allows one to simply add an additional term onto either the Poisson or the exponential Dyen divergence can be described as follows. Suppose that when a mass migration occurs and a culture-location pair emits an out-migrating sub-population, a random direction is selected from a uniform distribution around the circle, and then the distance traveled in that direction is governed by an exponential distribution. Let the migration from location i to i' traverse a distance $d_{i,i'}$, and let the parameter describing the density of the distance be denoted by $\mu_{i,i'}$, so that the parameter is specific to the migratory event. The probability of observing a move from i to i' then depends only on the distance between the two points, and can be written as:

$$P(d_{i,i'}) = \mu_{i,i'} e^{-\mu d_{i,i'}} \tag{18}$$

Maximizing (18) with respect to $\mu_{i,i'}$ gives $\mu^*_{i,i'} = \frac{1}{d_{i,i'}}$, so the concentrated likelihood of observing the jump is:

$$L(d_{i,i'}) = \frac{1}{d_{i,i'}} e^{-1} \tag{19}$$

Collecting all such terms associated with a migratory history, the likelihood of observing the distances associated with the history is:

$$L^D_H = e^{-(K-1)} \prod_{(i,i')\in H} \frac{1}{d_{i,i'}} \tag{20}$$

The $e^{-(K-1)}$ term in the above is a result of the fact that if there are K terminal nodes, there are $K-1$ internal nodes of the tree, and each migratory event associated with an internal node adds e^{-1} to the likelihood. As this is true of all migratory histories, one can write:

$$\ln L^D_H \propto -\sum_{(i,i')\in H} \ln d_{i,i'} \tag{21}$$

That is, a simple way of favoring paths requiring shorter distances is to include the negative of the sum of the log distances traversed. One advantage of this approach is that it remains agnostic about how physical distance is reckoned. For example, one could use a distance metric that weights east-west and north-south distances differently.

If one includes geographical distances in such a way so that smaller geographical jumps are more likely, the result would be something like the expansionary model of Neureiter et al. (2021), and this would also produce an effect similar to the geographical grouping of divergent dialogs. One might also allow that jumps that traverse different terrain or water are more difficult. Such a modification would allow for saturation of an already-populated island before expansion to a new island, which might also result in a region where the most divergent language is near where the languages of the phylogeny are most diverse, as appears to be the case for the Austronesian languages, where the most divergent languages are found on Taiwan, also a region of great linguistic diversity, suggesting a Taiwanese origin for this language family (Blust, 1984, 1999).

Microeconomic foundations

Two crucial assumptions of the model are (1) migratory events occur according to a Poisson-exponential distribution, and (2) the propensity to migrate leaves the origin location and moves to the new location with the departing sub-population. How can these points of departure for the model be justified? This section shows how exponential migration timings occur in a model in which population growth is dictated by a stochastic logistic growth model combined with a critical population level, that, when achieved, precipitates a resource crash. This crash generates a local superabundance of population,[19] which makes migration for a segment of the population desirable. If the crash can be thought of as irreversible, as might occur if a distinct resource is completely used up beyond the point of recovery, then the singular nature of the migratory chain can be justified. More broadly, the spirit of the model is consistent with the idea that a new innovation or approach to resource exploitation is discovered, but that there is some possibility that eventually a innovation or resource-specific crash will occur, which generates a superabundance of population and allows for a large segment of the population to leave. The idea is to justify the one-time nature of the event propelling population forward, but also to (eventually) generate a superabundance of population, which creates a large-scale migration from the current location to a new location.

Suppose that there is a discrete set of habitable locations that are not yet occupied.[20] Each location has some given carrying capacity K. Carrying capacity can be exhausted. Exhaustion of the carrying capacity is a discrete event and occurs if and when the population ever reaches some upper barrier B at the location.

In fact, when $B \geq K$, the carrying capacity is exhausted and falls to a much lower level for a generation, represented by κK, $\kappa < 1$. Once the event has passed, the carrying capacity does not revert to its old level, but instead a new K, B combination is randomly drawn for the location. The crucial thing is that the sudden crash that occurs when population reaches B generates a local superabundance of population, creating the potential for a sizeable fraction of the local population to wish to migrate to a new location.[21]

The population growth process is modeled as follows. Suppose food output per capita depends upon a resource level and also the current number of inhabitants at a location. That is, net per capita output during generation t is

$$y_t = f(p_t, \epsilon_t(\Delta); \theta_t) \tag{22}$$

where p is population, e is a random error, Δ is the time length of a generation, and θ is a collection of parameters. All income is devoted to production of children, and the number of children created by each individual is proportional to income. Hence:

$$p_{t+\Delta} = p_t f(p_t, \epsilon_t(\Delta); \theta_t) \tag{23}$$

A first-order Taylor expansion of (23), and then parameterization and normalization yields:

$$p_{t+\Delta} = \Delta p_t \left[1 + r\left(1 - \frac{p_t}{K}\right) + \sigma(\epsilon_{t+\Delta} - \epsilon_t)\right] \tag{24}$$

The expression in (24) can be rewritten as:

$$\frac{p_{t+\Delta} - p_t}{\Delta} = p_t\left[r\left(1 - \frac{p_t}{K}\right) + \sigma(\epsilon_{t+\Delta} - \epsilon_t)\right] \tag{25}$$

Letting Δ approach zero, and assuming that ϵ is governed by a standard Brownian motion, (23) can be rewritten as a stochastic differential equation:

$$dp = p\left(1 - \frac{p}{K}\right) + \sigma p dz \tag{26}$$

Equation (26) parameterizes the model as a stochastic logistic population growth model. The crucial property of (26) for the model is that it is mean-reverting. Ricciardi et al. (1999) show that mean-reverting processes like (26) have time-independent, initial-condition-independent stationary distributions. Further, Ricciardi et al. (1999) and Nobile et al. (1985) also show that *the existence of a time-independent steady state-density implies that the first passage time to a “large”’barrier is approximately exponential*. That is, if $g(B, t|p_0)$ denotes the distribution of the first passage time to a barrier B given initial population p_0, then:

$$g(B, t|p_o) \sim \frac{1}{t_1(B|p_0)} \exp\left(-\frac{t}{t_1(B|p_0)}\right) \tag{27}$$

Where $t_1(B, p_0)$ is the mean first passage time corresponding with the distribution. So, if attaining the barrier B is associated with the timing of a migratory event, one might expect the timing of migratory events to be approximately exponential.

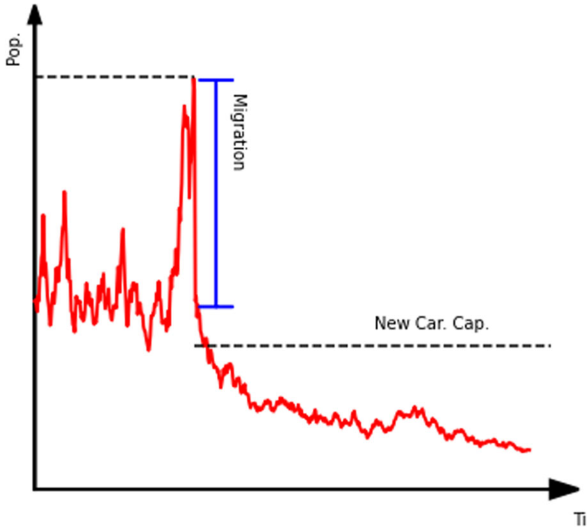


Fig. 6 An illustration of the stochastic population process. When the upper population barrier is attained, a resource crash occurs. The resulting population superabundance triggers a mass migration.

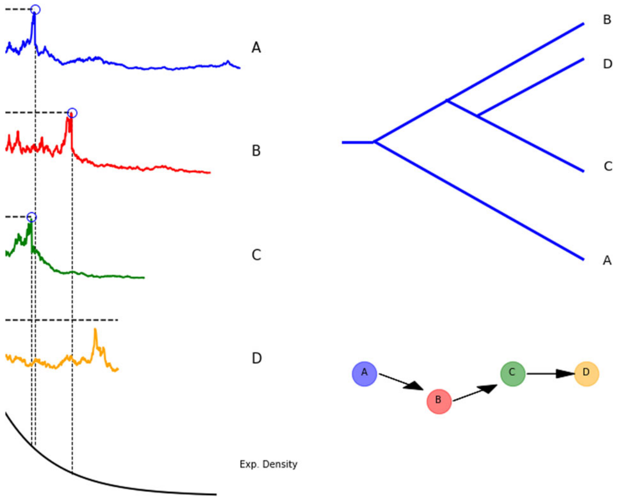


Fig. 7 An illustration of the formation of a phylogenetic tree and its underlying geography. The left-hand side shows times at which the upper population barrier is reached, prompting a segment of the population to leave.

Figure 6 illustrates the idea. The figure shows the population process moving along through time at a high carrying capacity. At some point, the carrying capacity is exhausted when the critical population level shown by the dashed line is reached, and a segment of the population leaves. The new carrying capacity in the location is then determined at the end of the generation; in the figure this new capacity is evidently lower.

Figure 7 takes the idea a bit further and shows how the process might play out sequentially over four different locations. In each area, the process hits a barrier at a random, approximately exponential time, triggering a local superabundance of population. Over time, the result is a series of jumps to new locations at exponentially distributed times. The length of the dashed lines corresponds with the lengths of the internal branches of the phylogenetic tree.

Some additional details round out the model. Since population is B at the barrier, expected per capita income if some segment of the population moves from the current location (24) is:

$$y_t = 1 + r\left(1 - \frac{B - m}{\kappa K}\right) \tag{28}$$

Suppose moving involves a per capita cost of c. Then, if m people move to a new location, expected income and utility among the migratory group is:

$$\tilde{y}_t = 1 + r\left(1 - \frac{m}{K}\right) - c \tag{29}$$

The arbitrage condition $y_t = \tilde{y}_t$ determines m^*, the size of the migratory group:

$$m^* = \frac{Br - Kc\kappa}{r(\kappa + 1)} \tag{30}$$

When the barrier is achieved, a migratory group of size m^* leaves for a new location. This population of m^* also plays the role of p_o in the distribution (27) as the migratory chain moves forward. To ensure that all of this group leaves for a single location, rather than fanning out, one might add in an additional assumption that a minimum migration size on the order of m^* is required to guarantee success.

This model of population growth and resource collapse requires a few parameter restrictions. For the size of the migratory group to be positive, parameters must obey:

$$Br > Kc\kappa \quad \rightarrow \quad \frac{Br}{K\kappa} > c \tag{31}$$

The model cannot just be a theory of mass migration, but also must be a theory able to explain why mass migration is rare. Parameter restrictions are required that ensure movements only occur when the rare event of population attaining the barrier occurs. When population is arbitrarily close to the barrier income at the new location is:

$$y_t = 1 + r\left(1 - \frac{B}{K}\right) \tag{32}$$

If a small segment of the population moved to a new location, expected income is:

$$\tilde{y}_t = 1 + r - c \tag{33}$$

For nobody to wish to move, it is required that $y_t > \tilde{y}_t$ whenever the barrier B has not been reached. This requires that the per capita cost of moving is sufficiently large:

$$c \geq \frac{rB}{K} \tag{34}$$

Conditions (31) and (34) bracket migration costs so that migrations only occur under rare conditions. Coupled with a migration cost, a theory with stochastic logistic population growth and a resource crash at a population barrier is consistent with a Poisson-exponential time sequencing of migratory events along a chain.

Computation

One can usually calculate a Dyen Divergence measure by hand and using simple rules of thumb about forming chains—indeed, that is the point of creating such measures of divergence—but in many circumstances it is useful to be in possession of a computational algorithm, as the case may be when it is not immediately obvious how to span a tree with the minimal number of migratory chains. Calculation is also difficult if one wishes to include other features besides the structure of the phylogeny in the calculation, such as geographical distance, as described in section "Geographical distances." Fortunately, Dyen Divergence measures or likelihoods can be computed by working recursively backwards through the tree as one would in a dynamic programming algorithm. The end result of this backwards iteration through the tree is a divergence measure or likelihood for each culture-location pair, reflecting the likelihood that each point in the phylogeny was the geographic point of origin of the entire tree.

The algorithm begins with enumeration of the interior nodes of the phylogenetic tree in depth-first order, so that nodes nearer to leaves carry lower indices, and nodes nearer leaves are traversed first. This allows a backwards traversal of the tree. Figure 2 follows this convention.

Given interior nodes $k = 1, 2, 3, \ldots, K$ and terminal nodes represented by location-culture pairs, (c_i, l_i), the recursion tracks the likelihood that culture-location pair i was the point of origin for all cultures after consideration of internal node k of the tree. Once $k = K$ is reached the end result is a vector of values representing the probability that each location was the beginning point of the migratory history of the entire tree. Needed for computation is a function which compiles the likelihood or divergence measure associated with a migratory chain, which generically depends upon state variables such as the current time length of the chain and the number of events in the chain. Let this function be denoted as $f(n, t)$, which could be (the log of) either (2) or (4), or either function's kernel.

State variables required for the calculation include the length of a migratory chain under consideration as of the traversal of node k, denoted t_{ik}, and the number of events in a chain, n_{ik}. A set-valued state variable r_{ik} is needed to keep track of nodes which have already been considered as part of i's value prior to the inclusion of k, where $r_{i0} = \{i\}$. Pursuant to keeping track of r_{ik}, it is also helpful to equip each internal node of the tree with a set of nodes reachable from k, called R_k. Let V_{ik} denote the likelihood or divergence measure value for culture-location pair i after consideration of node k, starting with $V_{i0} = 0$.

The algorithm changes character as new nodes become reachable when an internal node is considered, so it is therefore useful to have indicator variables capturing previous reachability. In particular, let $\mathbb{1}^r_{ik} = 1$ if $r_{ik} \neq \{i\}$ and $i \in R_k$, with $\mathbb{1}^r_{ik} = 0$ otherwise. The indicator therefore captures both whether a terminal node is reachable from an internal node and therefore should be updated when it is considered, and whether the terminal node was accessible from a previously considered interior node.

The recursion is defined by the following:

$$V_{ik+1} = V_{ik} + \mathbb{1}^r_{ik}\left[f(n_{ik}, t_{ik}) + V_{i^*k}\right] \tag{35}$$

where

$$i^* = \underset{i' \in \{R_k \backslash r_{ik}\}}{\arg\max}\left(f(n_{i'k} + 1, t_{i'k} + t_k) + V_{i'k}\right) \tag{36}$$

Equation (35) shows that the calculation depends on whether or not the node has previously considered other interior nodes. The maximization component of the algorithm in (36) occurs because when a new node is considered, a best next location for a given migratory path for each accessible terminal node must be chosen given location i is the starting point and now a migratory route has to be posited through some new location i'. The indicator variable captures the notion that when a node presents a culture-location pair with new options, the existing state variables have to be assembled into a chain likelihood or divergence measure. The state variables are updated/initialized according to:

$$n_{ik+1} = n_{i^*k} + 1, n_{i0} = 0 \qquad (37)$$

$$t_{ik+1} = t_{i^*k} + t_k, t_{i0} = t_i \qquad (38)$$

$$r_{ik+1} = R_k, r_{i0} = \{i\} \qquad (39)$$

Once the algorithm concludes with $k = K$, a final step is needed to collect the final values of the state variables:

$$V_{iK} = V_{ik} + f(n_{ik}, t_{ik}), \quad k = K \qquad (40)$$

As the algorithm can be difficult to picture until it is seen in operation, an extended example is worked in the supplemental appendix. In section "Origins of Semitic" these methods are applied to two competing theories of Semitic population dispersal.

Origins of Semitic

This section shows how the model might be used to attach a probabilistic interpretation to debates about the origins of a group of related peoples. The application is designed to frame debates about both the origins of a specific group of peoples—the Semitic peoples— but also to highlight how additional evidence shifts the likelihood of different origin narratives, even when additional evidence is imprecise. The intent in presenting the example is not to resolve the debate, but merely illustrate how the methods in this paper can be used to shape the debate in terms of probability.

The Semitic language family includes Hebrew, Arabic in its many versions, and ancient languages such as Amaraic and Akkadian. This language group has figured prominently in shaping both world history and the geopolitical landscape from ancient times into the present, hence understanding its origins and past is of paramount importance. As (Kitchen et al., 2009, p. 2703) point out, the cultures comprising Semitic created some of the earliest civilizations, three of the world's most important religions (Christianity, Islam, and Judaism), and also some of the first works of literature (for example, the Akkadian work *The Epic of Gilgamesh*).[22]

The early work on the origins of Semitic languages suggested that the point of origin of the Semitic family was in Arabia, but subsequent research argued for a northern origin for Semitic peoples in what is today Armenia (Grintz, 1962, Peters, 1919). The modern consensus has converged on an origin in the Levant, but as Kitchen et al. (2009) say, "Uncertainty about key details of this history persist despite extensive archeological, genetic, and linguistic studies of Semitic populations." (Kitchen et al., 2009, p. 2703)

Kitchen et al. (2009) apply state-of-the-art methods using comparative lexicons to fit a linguistic tree for the Semitic languages to shed further light on the origins of the constituent cultures of the group. The resulting tree is shown in Fig. 8.[23] The conclusion that Kitchen et al. (2009, p. 2708) reach is that linguistic and other evidence "...suggests a Semitic origin in the northeast Levant and a later movement of Akkadian eastward into Mesopotamia and Sumer," proposing later migrations out of this area into Ethiopia and Arabia, respectively. So, Kitchen et al. (2009) argue that the Levant is the point of origin for the group, in spite of Akkadian being the most divergent language of the family. Given the linguistic tree, this requires a set of distinct migratory chains emanating from the Levant to the locations of Akkadian, into Ethiopia, and into Arabia, using the vocabulary of this paper.

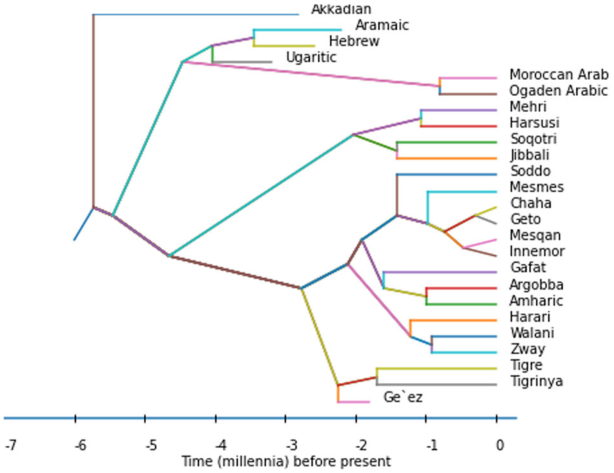


Fig. 8 A phylogenetic depiction of the Semitic Languages. The figure is constructed following the results of Kitchen et al. (2009).

Table 1 Semitic languages-cultures with divergence measures and probabilities of point of origin. The last column of the table is used in forming the map in Fig. 9.

Language/	Divergence		Probability	
Culture	Dyen	Exponential	Poisson	Exponential
Ge'ez	-24.9	-105.6	0.004	0.045
Tigrinya	-26.4	-105.7	0.001	0.041
Tigre	-26.4	-107.1	0.001	0.01
Zway	-29.6	-107.1	0.0	0.01
Walani	-28.1	-108.3	0.0	0.003
Harari	-25.1	-106.7	0.002	0.015
Amharic	-26.5	-106.5	0.0	0.019
Argobba	-26.5	-106.5	0.0	0.018
Gafat	-25.4	-105.9	0.001	0.032
Innemor	-28.5	-109.0	0.0	0.001
Mesqan	--30.0	-107.4	0.0	0.007
Geto	-30.0	-107.8	0.0	0.005
Chaha	-28.5	-109.0	0.0	0.001
Mesmes	-27.1	-107.8	0.0	0.005
Soddo	-25.5	-106.5	0.0	0.018
Jibbali	-24.6	-105.9	0.013	0.034
Soqotri	-24.6	-106.3	0.013	0.023
Harsusi	-24.6	-106.0	0.013	0.029
Mehri	-24.6	-105.5	0.013	0.049
Ogaden Arabic	-24.9	-106.1	0.033	0.028
Moroccan Arabic	-24.4	-108.4	0.033	0.003
Ugaritic	-24.7	-106.4	0.019	0.021
Hebrew	-25.9	-107.3	0.005	0.008
Aramaic	-25.9	-106.7	0.005	0.015
Akkadian	-23.3	-103.1	0.841	0.558

Using the tree in Fig. 8, divergence measures and distance measures for all members of the tree using the algorithm described in section "Computation" were computed. The results of these computations for two divergence measures and two likelihoods are presented in Table 1. The first column reports the Dyen divergence measure, which considers only the structure of the tree. The second column reports the exponential divergence measure, which considers the length of branches. The third column on Table 1 computes the Poisson probability for the location of each Semitic-speaking population being the point of origin, and the fourth column gives the exponential probability while

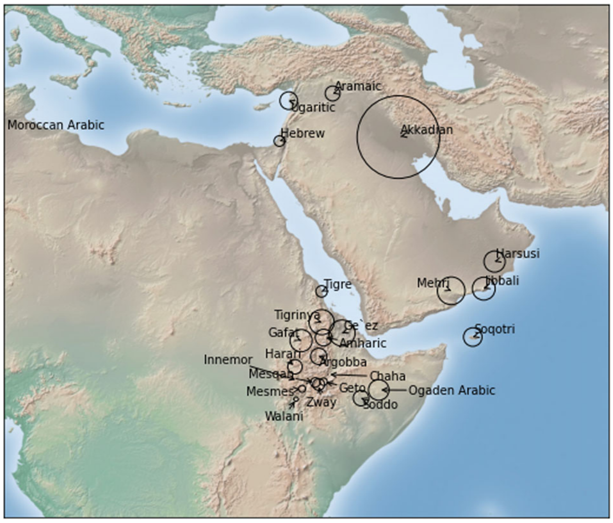


Fig. 9 Geography of the Semitic languages. Circles are proportional to the estimated probability of origin following the tree in Fig. 8.

also including geographical distance, as described in section "Geographical distances." From Table 1, the divergence measures and the exact probabilities all suggest that the most likely origin point for the Semitic cultures would seem to be where Akkadian was spoken if one relies solely on the evidence produced by the linguistic tree, and the probability of an Akkadian origin in the most expansive model is about 0.56. The cultures speaking languages in the Levant such as Ugaritic, Hebrew, and Aramaic, have much lower values, because if they are to be points of origin, more chains with fewer events in each are required to explain the current geographical distribution of the Semitic linguistic group.

The map in Fig. 9 depicts the exponential probabilities geographically. The circles drawn on the map in 9 are proportional to the last column of probabilities in Table 1. Again, the model suggests that the Akkadian location is the most likely origin point, but after this, Arabia or even the Ethiopian highlands appear to be just as likely a point of origin for the Semitic languages.

Based on this evidence, the arguments advanced by Kitchen et al. (2009) seem untenable given the structure of the tree they have constructed. However, that is not the whole story, as the authors discuss a variety of additional evidence in support of the hypothesis that Semitic-speaking peoples originated in the Levant, some of which can be folded into the model. One interesting component of this additional evidence is the place occupied in the tree by the long-extinct Eblaite language, for which no word list was available for the study. Eblaite, Kitchen et al. (2009) note, was likely a member of the Eastern Semitic languages along with Akkadian, spoken by a people living in a large area of the Levant, centered in what is now northern Syria.

How would the inclusion of Eblaite in the tree alter the picture of likely points of origin of Semitic speakers? For the sake of illustration, contrast the eastern Semitic hypothesis with another possibility: that Eblaite is in fact a closer relative to the Northwest Semitic languages like Aramaic, Hebrew, and Ugaritic, than it is to Akkadian. This alternative has, in fact, been suggested (Lipinski, 2001).

The ramifications for these two alternatives are depicted in tree form in Fig. 10. The left-hand side of the figure shows a tree where Akkadian and Eblaite are grouped together as Eastern Semitic languages. One can on the left-hand side of Fig. 10 see that the initial split in the tree has these peoples branching off together from all the other Semitic-speaking peoples. On the right-hand side of the figure, the alternative scenario is shown, where Eblaite is included with all other Semitic languages in the initial split.

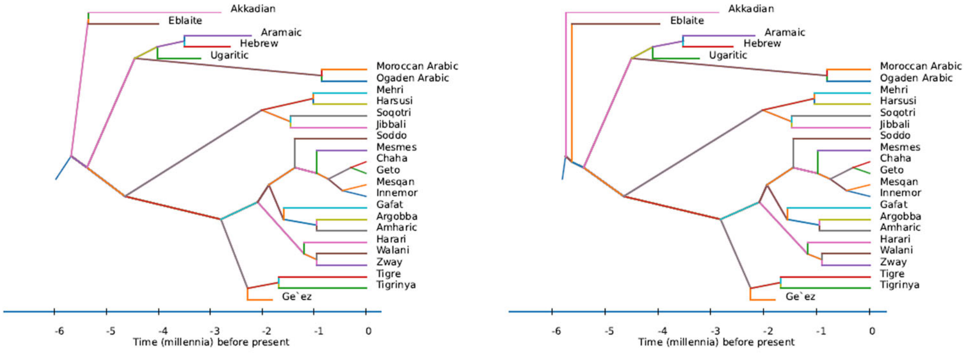
Table 2 Including Eblaite in the Semitic tree, either in a branch with Akkadian or grouped with all other Semitic languages as shown in Fig. 10.

Language/	Eastern Semitic		Northwestern Semitic	
Culture	Dyen	Poisson	Dyen	Poisson
Ge'ez	-25.6	0.005	-26.7	0.0
Tigrinya	-27.1	0.001	-28.2	0.0
Tigre	-27.1	0.0	-28.2	0.0
Zway	-30.2	0.0	-31.4	0.0
Walani	-28.8	0.0	-30.0	0.0
Harari	-25.8	0.001	-27.0	0.0
Amharic	-27.2	0.0	-28.3	0.0
Argobba	-27.2	0.0	-28.3	0.0
Gafat	-26.1	0.001	-27.2	0.0
Innemor	-29.2	0.0	-30.4	0.0
Mesqan	-30.7	0.0	-31.8	0.0
Geto	-30.7	0.0	-31.8	0.0
Chaha	-29.2	0.0	-30.4	0.0
Mesmes	-27.8	0.0	-28.9	0.0
Soddo	-26.2	0.0	-27.4	0.0
Jibbali	-25.3	0.006	-26.5	0.0
Soqotri	-25.3	0.005	-26.5	0.0
Harsusi	-25.3	0.005	-26.5	0.0
Mehri	-25.3	0.009	-26.5	0.0
Ogaden Arabic	-25.6	0.015	-26.7	0.0
Moroccan Arabic	-25.1	0.002	-26.3	0.0
Ugaritic	-25.4	0.112	-26.6	0.001
Hebrew	-26.6	0.007	-27.7	0.0
Aramaic	-26.6	0.011	-27.7	0.0
Eblaite	-24.8	0.725	-24.8	0.034
Akkadian	-25.1	0.094	-23.1	0.965

As no information about branch lengths for Eblaite are available, one can rely on the Dyen divergence measure and/or the Poisson likelihood to assess and compare the impact on the likelihood of the origin point of the group. Geographical distances can also be included in these computations, as described in section "Geographical distances", with results shown on Table 2. One can see from the table that the subtle difference in tree structure matters a great deal for pinpointing the point of origin of the entire phylogeny. Including Eblaite with Akkadian as the sole members of an eastern Semitic group shifts the most likely point of origin of the group as a whole to the location of the Eblaites in the Levant. This is because a separate branching of the tree—a distinct migratory chain—must be introduced to explain how both of these two peoples came to occupy their locations. Moreover, since situating the point of origin for subsequent migrations reduces the geographical distances traversed, Eblaite is more likely than Akkadian as a point of origin.

Alternatively, if Eblaite is included with the rest of the Semitic languages, the results in Table 2 suggest that the case for Akkadian as the point of origin is *strengthened*. This is because there now is another migratory event that can be attached to the long chain running from the location of the Akkadians, which increases the Akkadian divergence measure and Poisson probability. The geography of the two alternatives is shown in Fig. 11.

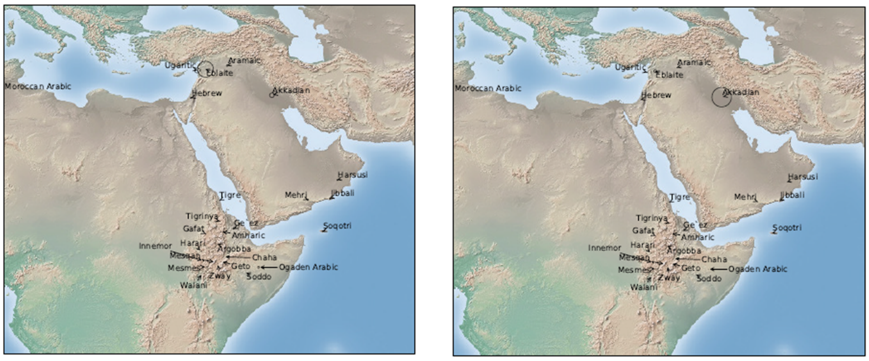
The purpose of this exercise is not to reach specific conclusions about where the Semitic peoples originated, but merely to show how one might deploy the model to assess the consequences of different sorts of assumptions about the structure of the tree. It bears mentioning that the tools presented in this paper might also be deployed in a different fashion. To wit, *if* one knew that Eblaite was near the point of origin of the Semitic peoples, *then* one might conclude it is most likely an eastern Semitic language like Akkadian.[24]



(a) Eastern Semitic

(b) Northwest Semitic

Fig. 10 Two alternatives for the inclusion of Eblaite in the Semitic language tree. The left-hand side of the figure modifies the Semitic language tree to include Eblaite and Akkadian together as part of an Eastern Semitic branch of the Semitic Tree (**a**), while the righthand side shows Eblaite branching off from Akkadian as part of a continuing expansion (**b**).



(a) Eastern Semitic

(b) Northwest Semitic

Fig. 11 Two theories of Eblaite. a Eastern Semitic, **b** Northwest Semitic. Circles are drawn in proportion to the likelihood of locations being the point of origin of the entire group.

Discussion and conclusions

This paper develops probabilistic and microeconomic foundations for a critical theoretical idea in piecing together the geographical dispersion of cultures: the Age-Area Hypothesis. It also develops related measures of linguistic divergence that can be used to compare potential geographical points of origin, and to assess the likelihood of alternative dispersal narratives. The model relies on the Poisson-exponential distribution in describing migratory events. This reliance can be justified with a theory of mass migration based on the stochastic arrival of a superabundance of population. This is because the time it takes a mean-reverting stochastic differential equation to reach a distant barrier is approximately exponential. The paper then forges a link between these singular events, in that fewer of them are required, meaning the parametric model of the tree has fewer parameters. This parametric simplicity is then translated into greater likelihood, through a tree-constructed distance measure.

A larger aim of this paper is the continuance of the project of grounding rules of thumb and other sorts of algorithms employed in the social sciences in probability theory and likelihood. Felsenstein (2004), for example, describes how tree-building algorithms and algorithms for inferring ancestral states in genetics initially developed using parsimony-based methods, but were then fashioned into probabilistic models enabling likelihood based methods, with important contributions from Felsenstein himself.

Probabilistic foundations are important for a variety of reasons. One such reason is suggested in the application to the Semitic languages in section “Origins of semitic”. As recent Bayesian phylogeographic research (Bouckaert et al., 2012, 2018) and other work (Currie et al., 2013) does, one can pair phylogenetic and geographic likelihoods in a joint model of the phylogeny and the geographic dispersal process.

Some recent research suggests that some patternings of population dispersals might be more likely than others. It has been argued (Neureiter et al., 2021) that an expansionary pattern seems to better capture population expansions than alternative models. However, this might lead one to wonder whether different visions of the migratory process could be created that could be blended in a more thorough analysis. Perhaps some of the ideas of this paper could be deployed to build other types of dispersal processes.

Blending geography and phylogeny does not just apply to tree-building —one major advantage of probability and likelihood is that it presents a straightforward way to include all sorts of different information—spatial and phylogenetic information can also be combined with prior information deriving from history or archeology as well. Different sorts of data, curated in places such as Kirby et al. (2016), for example, can be combined via likelihood in a joint model. In the end, it is hoped the paper will push forward the cross-disciplinary project of melding of culture, geography and history, especially as new and varied sorts of data are combined in more comprehensive analysis of cultural evolution.

Received: 2 August 2021; Accepted: 11 November 2021;
Published online: 03 December 2021

Notes

[1] An introduction to the role linguistic evidence plays in historical analysis, see Ostler (2006). A more detailed introduction to methods are Nichols (1992) or Nichols (1997).

[2] One must be careful in invoking "the" AAH, as there are possibilities for confusion with other, sometimes unrelated, ideas. What is referred to as the AAH in this paper should not be confounded with the Sapir-Whorf Hypothesis (Sapir, 1929; Whorf, 1956), which refers to "Linguistic relativity"—the idea that peoples' thinking is influenced by language structure. What might be called the Cultural Age-Area Hypothesis is the controversial idea that older cultural traits are likely to be more widely geographically distributed. See Graves et al. (1969).

[3] Trask (2000, p. 12) attributes the Age-Area Hypothesis (AAH) to the work of Latham (1851) and Sapir (1916).Trask (2000) further mentions the work of Mallory (1997) and Nichols (1997), as examples of applications and qualifications of the AAH. Dimmendaal (2011, p. 336) describes the AAH in part as the "principle of least effort," going on to note that "This principle probably was applied first by the scholars working on Amerindian Languages, e.g., Sapir (1916) and Dyen (1956)".

[4] Mace et al. (2005) provides an excellent overview of state-of-the-art methods for developing phylogenetic trees from languages. See also Nichols (1997) Atkinson and Gray (2003) or Kitchen et al. (2009). State-of-the-art methods are quite sophisticated and blend Markov-chain Monte Carlo sampling, Bayesian methods,and computational linguistics.

[5] In some instances, the most divergent languages are also found where languages are most diverse. This seems to be true of the Austronesian language family, where the most divergent languages and great diversity in languages are found on Taiwan, which supports a Taiwanese origin for the Austronesian family (Blust, 1984, 1999).

[6] In fact, tree imbalance is a common signature of phylogenetic trees in general. See Aldous (2001).

[7] A classic example is the migration of Germanic peoples into England; see Weale et al. (2002). This characterization of population movements leans on what Heggarty et al. (2010) refers to as a "splits" model, not a "wave" model, which produces something more akin to a continuum of dialects and could have different implications for the degree of relatedness between cultures.

[8] To be clear: this is one statement of the AAH, but others might state the principal differently. This is consistent with the usage in Sapir (1916), but not with Wichmann et al. (2010), who state the principal in terms of linguistic diversity being greatest near the geographic point of origin. The relationship between these two statements is discussed in section "Problem description."

[9] While no special knowledge of phylogenetic trees is needed for this paper, one could consult Jackson (2008) or Felsenstein (2004) for the basic ideas.

[10] One might assume l_i is something more concrete; for example, $l_i = (x_i, y_i)$ is composed of coordinates, which makes things easier to picture and simplifies relationships with cross-cultural data. Another possibility is that the locations are just a list of potentially habitable locations.

[11] For a shorthand notation for describing a chain, C with a subscript listing the locations visited in sequence by the chain is used.

[12] It is interesting to note that using small parameter sets is sometimes imposed during the estimation of models of geographical drift. For example, Lemey et al. (2009) use Bayesian stochastic search variable selection (BSSVS) to restrict the size of the parameter space in their model.

[13] The Poisson model also results from an exponential model in which the timing of events have been integrated out.

[14] While H_A is used to keep notation simple, this is with some abuse of notation, as there are other possible migratory histories that could arise from the culture/location A.

[15] One could and should consider the chances any of the points were the starting point, but for illustrative purposes, suppose that the origin point was known to be one of these two locations perhaps because of archeological or other historical evidence.

[16] For any $k \in (1, n-1)$, for functions like (11) it is true that: $h(n) > h(n-k)h(k)$.

[17] Thanks to an anonymous referee for directing attention to this quotation.

[18] Other histories will produce smaller or equal likelihoods. For example, a possible history with chain C_{EADBC}, where an initial migration from E to A is posited instead of E to D, along with chain C_{EFIGH}, and chains C_{AD}, and C_{ABC}, is also admissible. However, this history has four chains instead of two, and hence will produce a smaller likelihood. the migratory history in which the last two events of each chain are switched; i.e., $\{C_{EADCB}, C_{EGIHG}\}$, produces the same likelihood.

[19] Dow et al. (2009) employ a similar approach to explain the emergence of Agriculture in Southwest Asia; in that paper the circumstances generating the superabundance is different, but it plays a similar role in that it creates a singular event that in part explains why the events are rare and not ongoing.

[20] Or, if the locations are occupied, the costs of contesting the location could be folded into migration costs.

[21] A parable: a people currently occupy an island with a population of stylized birds. The population of birds extends across all islands in the area. While plump, the birds taste terrible, but by accident one day it is discovered that a spice on the islands makes the birds palatable, leading to an abundance of food. The population adjusts to the new diet of local agriculture and birds. One year, an unusually good agricultural crop pushes the human population to a new high. The additional human population strains the bird population, leading to a collapse from which the population cannot recover. Migration is costly and risky, but if everyone stays at the current location, there will be famine. Thus, a fraction of the population incurs the costs of migration and leaves for a new island, where the birds may be found and the process begins again.

[22] The rediscovery of this old epic poem in the early 20th century actually ignited some controversy over the origins of the Old Testament, as *Gilgamesh* contains a chapter about a flood. See Ziolkowski (2012). Hetzron (1997) is an overview of the Semitic languages and their relationships to one another.

[23] Kitchen et al. (2009) use word lists and Bayesian computational linguistics methods to fit the tree. While it is not apparent from Fig. 8, Kitchen et al. (2009) also produce confidence intervals for all branch lengths along the tree. In principle, one could, and probably should, use this information in constructing distance measures or likelihoods, for example, by repeatedly drawing branch lengths, computing distance measures, and then averaging results. This was not done to keep the exercise as transparent as possible.

[24] Additional information and applications, including a Python implementation of the ideas in this paper, can be found on the project sites: http://github.com/mbaker21231/instevo and http://github.com/mbaker21231/agearea.

References

Aldous DJ (2001) Stochastic models and descriptive statistics for phylogenetic trees, from yule to today. Stat Sci 16(1):23–34

Alesina A, Devleeschauwer A, Easterly W, Kerlat S, Wacziarg R (2005) Ethnic diversity and economic performance. J Econ Liter 43(3):762–800

Ashraf Q, Galor O (2013) The "out of africa" hypothesis, human genetic diversity, and comparative economic development. Am Econ Rev 103(1):1–46

Atkinson QD, Gray RD (2003) Language-tree divergence times support the anatolian theory of indo-european origin. Nature 2(426):435–439

Atkinson QD, Meade A, Venditti C, Greenhill SJ, Pagel M (2008) Languages evolve in punctuational bursts. Science 319(5863):pp. 588

Blust R (1984) The Austronesian Homeland: A Linguistic Perspective. Asian Perspect 26(1):45–67

Blust R (1999) Subgrouping, circularity, and extinction: some issues in Austronesian comparative linguistics. In: Zeitoun E, Jen-Kuei Li, P (eds), Selected papers from Eighth International Conference on Austronesian linguistics. Academica Sinica, Taipei. pp. 31–94

Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko AV, Drummond AJ, Gray RD, Suchard MA, Atkinson QD (2012) Mapping the origins and expansion of the Indo-European language family. Science 337(6097):957–960

Bouckaert R, Bowern C, Atkinson QD (2018) The origin and expansion of Pama-Nyungan languages across Australia. Nat: Ecol Evol 2:741–749

Bromham L, Hua X, Fitzpatrick TG, Greenhill SJ (2015) Rate of language evolution is affected by population size. Proc Natl Acad Sci 112(7):2097–2102

Cavalli-Sforza LL, Cavalli-Sforza F (1995) The great human diasporas: the history of diversity and evolution. Perseus Books, Cambridge, MA

Currie TE, Meade A, Guillon M, Mace R (2013) Cultural phylogeography of the Bantu Languages of sub-Saharan Africa. Proc R Soc B 280:20130695

Diamond J, Bellwood P (2003) Farmers and their languages: The first expansions. Science 300(5619):597–603

Dimmendaal GJ (2011) Historical Linguistics and the comparative study of african Languages. John Benjamins Publishing Company, Amsterdam and Philadelphia

Dolgopolsky A (1988) The indo-european homeland and lexical contacts of proto-indo-european with other languages. Mediterr Lang Rev 3:7–31

Dow GK, Reed CG, Olewiler N (2009) Climate reversals and the transition to agriculture. J Econ Growth 14(1):27–53

Dyen I (1956) Language distribution and migration theory. Language 32(4):611–626

Ehret C (2002) The civilizations of Africa: a history to 1800. University of Virginia Press, Charlottesville

Ehret C, Keita SOY, Newman P, Bellwood P (2004) The origins of afroasiatic. Science 306(5702):1680

Forster P, Renfrew C, editors (2006) Phylogenetic methods in the prehistory of languages. McDonald Institute, Cambridge

Felsenstein J (2004) Inferring Phylogenies. Sinuaer

Giuliano P, Nunn N (2018) Ancestral characteristics of modern populations. Econ Hist Dev Region 33(1):1–17

Graves TD, Graves NB, Kobrin MJ (1969) Historical inferences from Guttman scales: the return of age-area magic? Curr Anthropol 10(4):317–338

Gray RD, Atkinson QD, Greenhill SJ (2013) Phylogenetic models of language change: three new questions. In: Richerson PJ, Christiansen MH (eds), Cultural evolution: society, technology, language, and religion. MIT Press, Cambridge, MA. pp. 285–300

Gray RD, Jordan F (2000) Language trees support the express train sequence of austronesian expansion. Nature 405:1052–1055

Greenhill S, Gray RD (2005) Testing population dispersal hypotheses: Pacific settlment, phylogenetic trees, and austronesian languages. In: Mace R, Holden CJ, Shennnan S (eds), The evolution of cultural diversity: a phylogenetic approach, chapter 3. Left Coast Press. pages 31–65

Grintz JM (1962) On the original home of the semites. J Near East Stud 21(3):186–216

Grollemund R, Branford S, Bostoen K, Meade A, Venditti C, Pagel M (2015) Bantu expansion shows that habitat alters the route and pace of human dispersals. Proc Natl Acad Sci USA 112(43):13296–13301

Heggarty P, Maguire W, McMahon A (2010) Splits or waves? trees or webs? how divergence measures and network analysis can unravel language histories. Philos Trans R Soc B 1559(365):3829–3843

Hetzron R (editor) (1997). The Semitic Languages. Routledge, London and New York

Holden CJ (2006) The spread of languages, farming, and pastoralism in subsaharan africa. In Lipo CP, O'Brien MJ, Collard M, and Shennan SJ (eds). Mapping our ancestors: phylogenetic approaches in anthropology and prehistory. Aldine Transaction. pp. 249–268

Holman EW (2009) Do languages originate and become extinct at constant rates? Diachronica 27(2):214–225

Jackson MO (2008) Social and Economic Networks. Princeton University Press, Princeton and Cambridge

Kirby KR, Gray RD, Greenhill SJ, Jordan FM, Gomes-Ng S, Bibiko H-J, Blasi D, Botero CA, Bowern C, Ember CR, Leehr D, Low BS, McCarter J, Dival W, Gavin MC (2016) D-place: A global database of cultural, linguistic and environmental diversity. PLoS ONE 11(7):1–14

Kitchen A, Ehret C, Assefa S, Mulligan CJ (2009) Bayesian phylogenetic analysis of semitic languages identifies an early bronze age origin of semitic in the near east. Proc. Biol Sci 276(1668):2703–2710

Latham RG (1851) The Ethnology of the British Colonies and Dependencies. J. Van Voorst, London

Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009) Bayesian phylogeography finds its roots. PLoS Comput Biol 5(9):1–15

Lemey P, Salemi M, Vandamme A-M editors (2009) The phylogenetic handbook. Cambridge University Press, Cambridge and New York

Lipinski E (2001) Semitic Languages: outline of a comparative grammar, 2nd edn., volume 80 of Orientalia Lovaniensia Analecta. Peeters Publishers, Leuven, Belgian

Lowes S, Nunn N, Robinson JA, Weigel JL (2017) The evolution of culture and institutions: Evidence from the kuba kingdom. Econometrica 85(4):1065–91

Mace R, Holden CJ, Shennan S, editors (2005) The evolution of cultural diversity: a phylogenetic approach. Left Coase Press, Walnut Creek, CA

Mallory JP (1997) The homelands of the indo-europeans. In: Blench R, Spriggs M, (eds) Archaeology and language 1: theoretical and methodological orientations. Routledge, London. pp. 93–121

Michalopoulus S (2012) The origins of ethnolinguistic diversity. Am Econ Rev 102(4):1508–1539

Neureiter N, Ranacher P, Van Gijn R, Bickel B, Weibel R (2021) Can Bayesian phylogeography reconstruct migrations and expansions in linguistic evolution? R Soc Open Sci 8:201079

Nichols J (1992) Linguistic Diversity in space and time. University of Chicago Press, Chicago and London

Nichols J (1997) Modelling ancient population structures and movements in linguistics. Ann Rev Anthropol 26:359–384

Nobile AG, Ricciardi LM, Sacerdote L (1985) Exponential trends of first-passage-time densities for a class of diffusion processes with steady-state distribution. J Appl Probab 22(3):611–618

Ostler N (2006) Empires of the World: A Language History of the World. Harper, London and New York

Peters JP (1919) The home of the semites. J Am Orient Soc 39(1):243–260

Renfrew C (1987) Archaeology and Language: The Puzzle of Indo-European Origins. Cape, London

Ricciardi LM, Crescenzo AD, Giorno V, Nobile AG (1999) An outline of theoretical and algorithmic approaches to first passage time problems with applications to biological modeling. Math Jpn 50(2):247–322

Ruhlen M (1994) The Origin of Language: Tracing the Evolution of the Mother Tongue. John Wily and Sons, Inc., New York, New York

Sapir E (1916) Time Perspective in aboriginal American culture: A study in method. Number 19 in Geological Survey, Memoir 90, Anthropological Serie. Government Printing Bureau, Ottawa

Sapir E (1929) The status of linguistics as a science. Language 5(4):207–214

Sapir E (1949) Selected Writings of Edward Sapir In Language, Culture, and Personality. Mandelbaum, D (ed). University of California Press, Berkeley

Siebert F (1967) The original home of the proto-algonquian people. In: Dubois AD (ed), *Contributions to anthropology: linguistics I (Algonquian), bulletin No. 214, anthropological series no. 78*, Ottawa. National Museum of Canada, pp. 13–47

Spolaore E, Wacziarg R (2013) How deep are the roots of economic development? J Econ Liter 51(2):325–369

Swadesh M (1951) Lexico-statistic dating of prehistoric ethnic contacts: with special reference to north american indians and eskimos. Proc Am Philos Soc 96(4):452–463

Trask RL (2000) The Dictionary of Historical and Comparative Linguistics. Fitzroy Dearborn Publishers, Chicago and London

Weale ME, Weiss DA, Jager RF, Bradman N, Thomas MG (2002) Y chromosome evidence for anglo-saxon mass migration. Mol Biol Evol 19(7):1008–1021

Whorf B (1956) Language, thought, and reality: selected writings of Benjamin Whorf (edited by John B. Carroll). MIT Press, Cambridge, Massachusetts

Wichmann S, Müller A, Velupillai V (2010) Homelands of the world's language families: A quantitative approach. Diachronica 27:247–76

Ziolkowski T (2012) Gilgamesh among us: modern encounters with the ancient epic. Cornell University Press, Ithaca, New York

Acknowledgements

I would like to thank Jonathan Conning, Andrew Dickens, James Fenske, and seminar participants at Hunter College and Wesleyan University for helpful comments and suggestions.

Competing interests

The author declares no competing interests.

Ethical approval

This research did not require any ethical approval.

Informed consent

This research does not contain any studies with human participants performed by any of the authors

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-021-00991-8.

Correspondence and requests for materials should be addressed to Matthew J. Baker.