Journal of
Phylogenetics & Evolutionary Biology

Bast, J Phylogen Evolution Biol 2015, 3:3
DOI: 10.4172/2329-9002.1000e115

Editorial Open Access

Time Calibration of Linguistic Phylograms: A Molecular Clock for Historical Linguistics

Felix Bast*

Centre for Plant Sciences, Central University of Punjab, Mansa Road, Bathinda, 151001, Punjab, India

Abstract

Evolutionary patterns of languages and organisms have surprising similarity, as Darwin famously captured by what he termed as "Curious Parallelism." While traditional comparative and historical linguistic methods such as detailed analysis of cognate correspondences reveal similarities between languages and group them into linguistic families like how Carl Linnaeus grouped organisms into taxonomical hierarchies based on overall similarity-an approach known as phenetic clustering, this will not help to answer such as "when did Proto-Dravidian split to Proto-South Dravidian and Proto-South-Central Dravidian?" and so on. Conventional methods for dating linguistic trees such as glottochronology are severely flawed such that these have now been largely discredited. Proposed in this invited editorial is the direct extension of the molecular clock hypothesis and time-calibration techniques of molecular phylogenetics to the field of phylolinguistics. For 'calibration checkpoints', ancient dated texts, as well as dated and reliable historical information (such as Cro-Magnon migration to Europe, etc.), can be employed. Also deliberated here is a call to make use of Maximum Parsimony-based approaches for the ancient character-state reconstruction, for reconstructing long-lost languages.

Keywords: Phylogenetics; Linguistics; Parsimony; Cognate; Glottochronology; Glottoclock

Languages are a lot like organisms; new languages (as well as accents, dialects) form when geographically separated from other languages like how new species forms (allopatric speciation), and evolve spatially and temporally like how organisms evolve. Like families of biological taxa, languages too have linguistic families, and languages within these families are connected via ancestor-descendant relationship like how biological taxa are related within the taxonomical families. Like organisms evolve by mutations occurring in DNA molecules slowly over time, languages too evolve slowly by innovations. Consider ancient Indian language Sanskrit, Indians did not decide overnight to stop using Sanskrit-the language they have been comfortably using more than a millennium, and form a number of daughter languages, like Hindi, Bengali, Punjabi, Marathi, Gujarati, etc. Due to vast geographical space and reduced intermixing between distantly located speakers, the original Sanskrit slowly acquired many innovations unique to the regions- analogous to geographic clines in biology, which started diverging first as mere "accents", later as 'dialects', and over the course of many hundred years, as distinct languages. Consider Yiddish or Dutch, these had been regarded as mere dialects of German merely a century ago. Pre-requirements for organic speciation and linguistic lineage split (new language formation) are also remarkably similar; intra-population variability should be minimum while interpopulation variability should be maximum, and to achieve this, immigration should be small (emigration have no effect). Genes from one species to other species sometimes jump, the so called Horizontal Gene Transfer-which is common for bacterial species, and languages too have an analogous phenomenon, the so-called "borrowings"/"loanwords" (examples include English word Shampoo, borrowed from Hindi, and English words Mango, Copra and Coir, borrowed from Malayalam). Plants of different species sometimes cross-fertilize to form hybrids, so as two or more languages mix along the border to form stable language creoles as well as its simplified version, pidgins.

Organic evolution can be defined as the change in allele frequencies of a population over time while linguistic evolution is the change in linguistic units over time, for example, lexicons. Consider the English word 'omnibus' which is now considered archaic and rarely ever used in everyday communication, but used to be the principal word for what is now called 'bus.' Or consider this sentence: "My house is painting", which now makes no sense and could even be ridiculed by the conservative prescriptivists of our time; but it made perfect sense merely a century ago when the prescriptivists of yore ridiculed syntactic innovations such as 'my house is being painted'-which is now universally adopted as standard. In other words, the frequency of the usage of these linguistic forms, be it lexicons in the first example, or syntactic forms in the second example, have evolved over time, what we can refer as the linguistic evolution [Figure 1].

It was Charles Darwin who first deliberated the so-called "Curious Parallelism" between the evolution of organisms and languages 14 [1] Stephen Jay Gould and Richard Lewontin-through their concept of spandrel [2] and Noam Chomsky-through his much contested concept of Universal Grammar [3] either directly or indirectly regarded languages as a phenotypic by-product of some other characteristic rather than the direct product of adaptive evolution. Language can also be viewed as Richard Dawkins' concept of "extended phenotype"-the entire repertoire of the effects of genes outside the body that may influence the chances of its replication [4]. Genes such as *FOXP2* have now been discovered to be associated with human linguistic cognition [5]. Of course mastering a language and *lingua francas* - which are culturally transmitted replicators [6] help in surviving and transmission of genes in a human population, as it aid to communicate, social grooming/gossiping and courtship.

The field of phylogenetic linguistics, or phylolinguistics as often referred, finds its root back to the landmark publication by Cavalli-Sforza et al. [7] which directly compared human genetic and linguistic trees. In my last editorial in this journal, I have deliberated some of the pivotal developments in this field, with several examples of the application of computational phylogenetic methods in historical linguistics [8,9]. Phylolinguistics have not only corroborated existence of a number of language families of the world but also intricate evolutionary patterns of these languages that are easily missed in traditional comparative linguistic methods. For example, we now

***Corresponding author:** Felix Bast, Centre for Plant Sciences, Central University of Punjab, Mansa Road, Bathinda, 151001, Punjab, India
E-mail: felix.bast@gmail.com

Received August 21, 2015; **Accepted** August 24, 2015; **Published** August 31, 2015

Citation: Bast F (2015) Time Calibration of Linguistic Phylograms: A Molecular Clock for Historical Linguistics. J Phylogen Evolution Biol 3: e115. doi:10.4172/2329-9002.1000e115

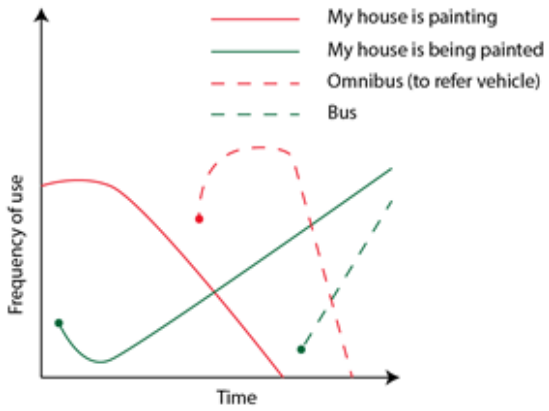

Figure 1: Examples of language evolution via changes in syntactic (continuous lines) and lexical (dashed lines) usage frequencies. Solid dots represent "linguistic innovations," that might correspond to scientific inventions, for example, the invention of the omnibus transportation system in 1824.

know that English is phylolinguistically related to Flemish and Dutch, all these three languages are related to German in similar way as Scandinavian languages are related to German as well (Icelandic, Norwegian, Swedish, Danish, Faroese)- all these languages including German and Yiddish are now thought to have a common ancestor, the so-called "Proto-Germanic" language of Western Europe. Romance languages (Latin and its descendants; Spanish, Portuguese, French, Italian, and Romanian) belong to Italic language family. The common ancestor of Italic and Proto-Germanic languages is also the ancestor of Sanskrit and Iranian language families, what is now called "Proto Indo-European" language [10] Indo-European is one of the most extensive linguistic families of the world, which encompasses the vast majority of languages of Eurasia. Now the question is how long did these languages divulge from Proto Indo-European to form new languages, the so-called 'cladogenesis'.

Before the entrance of phylogenetics, historical linguists solely depended on comparative methods for the reconstruction of ancient languages, but it had a number of shortcomings. For example, Proto-Germanic language is not directly attested by any coherent surviving texts and to infer linguistic attributes of this extinct language one need to compare surviving texts of its descendants; either extant or extinct languages. Extinct languages can also be utilized if any surviving texts are available for the analysis. Comparative method makes use of linguistic correspondences between languages and is analogous to phenetic (taximetric) clustering methods used in phylogenetics such as Neighbor-Joining, that compare organisms based on overall similarity of observable traits. Comparative linguistics help to reconstruct ancestral state, but not the date at which the language split happened.

Come to the era of phylolinguistics, which uses celebrated methods of molecular phylogenetics [11] to analyze linguistic "cognate" (=homologous) datasets. Phylolinguistics keep on enlightening the field of historical linguistics to divulge cryptic evolutionary relationships of languages in the similar fashion how molecular phylogenetics help in divulging evolutionary relationships of organisms [for example 12-17]. Molecular phylogenetics is such a robust tool that can do much more than revealing the evolutionary relationships; it can predict the ancestral states and can calculate the time since divergence- the time calibration of phylograms. However, attempts for the extension of the time calibration techniques in phylolinguistics have been scanty, except an extension of penalized-likelihood rate smoothing in relaxed framework attempted for Indo-European family [18]. I propose here an easy method for such an extension for the enormous benefit to the field of historical linguistics.

In order to time-calibrate a phylogenetic tree we need to calculate the rate at which branches evolve, for which the molecular clock hypothesis (MCH) was postulated by Zuckerkandl and Linus Pauling in 1962 [19]. According to the original MCH, mutations are thought to accumulate linearly over time. Empirical prediction of molecular clock arose with Neutral Theory of Molecular Evolution proposed by Motoo Kimura, which holds that majority of mutations occurring in nature are evolutionarily neutral and are the result of random drift rather than Darwinian natural selection [20]. As majority of observed mutations are random, rate at which these mutations occur (nucleotide substitution rates) can be calculated to make the branch lengths of a phylogram proportional to the evolutionary time- an immediate extension of the MCH. However, rate of evolution might not be constant across all lineages and, therefore, strict molecular clock model was eventually replaced by a relaxed model that incorporates rate variation across lineages [21]. In order to time-calibrate a phylogenetic tree we would also need the so-called 'calibration checkpoints' -time constraints at the interior nodes of phylograms as imposed by the empirical data with known dates, for example radiometrically dated fossils or dated geological/human migratory events (such as drift of continents, Cro-Magnon migration to Europe etc.) [22].

What is proposed herein is the direct augmentation of MCH and time calibration methods to phylolinguistics. Rate of linguistic evolution can be calculated for each lineage in a phylolinguistic tree using morphological analyzes of recent history where we have excellent written record. For example, consider the following passage:

Her…Ælfred cyning … gefeaht wið ealne here, and hine geflymde, and him æfter rad oð þet geweorc, and þær sæt XIIII niht.

Which might make no sense in present English but it meant the following in Old English:

Here… King Alfred… fought against the whole army, and put it to flight, and rode after it to the fortress, and there he camped for thirteen nights.

The original passage is from a document called *Anglo-Saxon Chronicle* that has been historically dated to AD 878, which is written in what is now called Old English. A comparison between these two passages reveal that to change from "Her" in old English to "here" in modern English, as well as from "Here" in old English to "Army" in modern English, it took less than 1000 years. By analyzing a large dataset (cognates, phonemes, lexicons, syntax and so on), estimates of the rate of linguistic evolution in changes/unit time can be derived, which is analogous to the MCH, what we may now refer as "linguistic clocks." A similar concept called "glottoclocks" exists in glottochronology sub-discipline of historical and/or comparative linguistics. Glottochronology employs percentage of shared cognates between languages under the assumption of constant rates of lexical replacement, the glottoclocks. However, this have largely been discredited due to several major issues such as loss of discrete characters while summarizing cognate correspondences into percentage scores, and substantial borrowings of lexical attributes between languages makes the time estimates statistically inaccurate and unreliable. Linguistic clocks envisioned here do not condense the information in cognate datasets into percentage values and do not assume constant rate of evolution, but instead attempt to calibrate rate of lexical evolution at each linguistic lineages from the cognate datasets. For example, linguistic

evolutionary rates for a continuous population such as that at Haiti are expected to be lower than that of the discontinuous archipelago such as Papua New Guinea-the most linguistically diverse place on earth. In addition, we would also need to find reliable calibration checkpoints in the linguistic phylograms, for which ancient dated texts and reliable historical information can be used. Epic of Gilgamesh (ca 2100 BCE) written in extinct Semitic language of Akkadian (Afro-Asiatic language family) spoken in Mesopotamia, as a 'dated ancient text', and spread of agriculture from Anatolian Peninsula (Turkey) around 9000 years ago, as a 'reliable historical information', for instance. As the oldest written texts date from 27th or 28th Century BCE, we can supplement linguistic phylograms with dated ancient texts for the last 4900 years as calibration checkpoints. Of course, we have only a few surviving ancient texts (so as we have only a few dated fossils), and the writing itself was invented merely 5000 years ago-by then most of the extant linguistic families have already diverged from its last common ancestor. We could nevertheless effectively employ the surviving texts for the linguistic time calibration in the best available manner. This can be used to reconstruct within family time-trees (time calibrated phylogram)-such as divergence within Indo-Aryan language family, and between closely related families- such as divergence between Baltic and Slavic families ca 1200 BP (before present) or between Iranian and Indo-Aryan ca 3300 BP. At the same time, it might not have sufficient resolution when applied for between distantly related family time-trees, such as tree depicting the connection between Afro-Asiatic, Austronesian, Turkic, Dravidian and Uralic language families. For the latter, dated (either radiometric dating of archeological samples or time-calibrated molecular phylogeny of ancient DNA samples) migration events of human beings can be employed as calibration checkpoints. For example, migration of Cro-Magnons to Europe ca 32,000 BP, or migration of Paleo-Indians across Bering Land Bridge to North America ca 16,500 BP.

Another application of phylogenetics is in the field of ancient character-state reconstruction of extinct languages. For example, consider Dravidian language family, we do not have any surviving texts of Proto-Dravidian language, which is the most recent common ancestor of Malayalam, Tamil, Kannada and Telugu. This language is conventionally been reconstructed using linguistic correspondences largely done by B. Krishnamurti [23]. However, the reconstruction suffers multiple deficiencies, including grammar and epoch determination. Also, the conventional reconstruction largely avoided Northern Dravidian languages such as Brahui, Malto, and Kurukh, and these are not derivable from the currently accepted morphotype of Proto-Dravidian. I propose a direct extension of the molecular phylogenetic method, Maximum Parsimony (MP), for the reconstruction of extinct languages like Proto-Dravidian. It is surprising that the historical linguists have not yet explored MP for the ancestral reconstruction till-date.

It is hopeful that the augmentation of phylogenetic methods as proposed herein, including time calibration, MCH and ancestral reconstruction using MP to the field of phylolinguistics will immensely be helpful for deriving time-trees of languages-a task never been attempted before, as well as for the reconstruction of extinct languages.

Acknowledgements

This study is supported by a grant-in-aid from ICSSR [Indian Council for Social Science Research) Grant No. 02/305/2014-15/RPR.

References

1. Darwin C (1859) On the origins of species by means of natural selection. Murray, London.
2. Gould SJ, Lewontin RC (1979) The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proceedings of the Royal Society of London B: Biological Sciences 205: 581-598.
3. Chomsky N, DiNozzi R (1972) Language and mind. Harcourt Brace Jovanovich New York.
4. Dawkins R (1999) The extended phenotype: The long reach of the gene. Oxford University Press.
5. Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, et al. (2002) Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418: 869-872.
6. Pagel M (2009) Human language as a culturally transmitted replicator. Nature Reviews Genetics 10: 405-415.
7. Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J (1988) Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. Proceedings of the National Academy of Sciences 85: 6002-6006.
8. Bast F (2015a) Phylogenetics: Tracing the Evolutionary Legacy of Organisms, Metastatic Clones, Bioactive Compounds and Languages. J Phylogen Evolution Biol 3:e112.
9. Bast F (2015) Tutorial on Phylogenetic Inference - 1. Resonance 20 :360-367.
10. Nakhleh L, Warnow T, Ringe D, Evans SN (2005) A comparison of phylogenetic reconstruction methods on an Indo-European dataset. Transactions of the Philological Society 103: 171-192.
11. Bast F (2013) Sequence Similarity Search, Multiple Sequence Alignment, Model Selection, Distance Matrix and Phylogeny Reconstruction. Nature Protocol Exchange.
12. Bast F, Kubota S, Okuda K (2015) Phylogeographic Assessment of Panmictic Monostroma Species from Kuroshio Coast, Japan Reveals Sympatric Speciation. J Appl Phycol. 27: 1725-1735.
13. Bast F, Bhushan S, John AA, Achankunju J, Panikkar NMV, et al. (2015) European Species of Sub aerial Green Alga *Trentepohlia annulata* (Trentepohliales, Ulvophyceae) Caused Blood Rain in Kerala, India. J Phylogen Evolution Biol 3:144.
14. Bast F, Bhushan S, John AA (2014) DNA barcoding of a new record of epi-endophytic green algae *Ulvella leptochaete* (Ulvellaceae, Chlorophyta) in India. Journal of Biosciences 39: 711-716.
15. Bast F, John AA, Bhushan S (2014) Strong Endemism of Bloom-Forming Tubular Ulva in Indian West Coast, with Description of Ulva paschima Sp. Nov. (Ulvales, Chlorophyta). PloS one 9: e109295.
16. Bast F, Bhushan S, John AA (2014) Morphological and molecular assessment of native carrageenophyte *Hypnea valentiae* (Cystocloniaceae, Gigartinales) in Indian Subcontinent. Phykos 44: 52-58.
17. Bast F, Rani P, Meena D (2014) Chloroplast DNA Phylogeography of Holy Basil (*Ocimum tenuiflorum*) in Indian Subcontinent. The Scientific World Journal.
18. Gray RD, Atkinson QD (2003) Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature 426: 435-439.
19. Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. Journal of theoretical biology 8: 357-366.
20. Kimura M (1984) The neutral theory of molecular evolution. Cambridge University Press.
21. Lepage T, Bryant D, Philippe H, Lartillot N (2007) A general comparison of relaxed molecular clock models. Molecular Biology and Evolution 24: 2669-2680.
22. Britton T, Anderson CL, Jacquet D, Lundqvist S, Bremer K (2007) Estimating divergence times in large phylogenetic trees. Systematic biology 56: 741-752.
23. Krishnamurti B (2003) The Dravidian Languages. Cambridge University Press.