Recent studies in language prehistory with special reference to South Asia.

Taraka Rama K.

1 Introduction

The idea that languages and genes have correlated evolutionary paths was observed at least since Darwin (1871). The recent advances in genetics and comparative linguistics have allowed researchers to make bold claims about the correlation between linguistic history and the speakers’ history. The tools developed in genetic studies allow researchers to test specific hypotheses about population prehistory and the prehistory of the languages spoken by those populations.

One such original study is the study of Cavalli-Sforza et al. (1988), who attempt to combine evidence from genes, language, and archaeology to make findings about the history of modern humans. There is also another strand of work by Bolhuis et al. (2014) – who mainly work with generative linguistics – claiming that there is an abstract language organ present in the human mind that allows infants to learn the language from milieu on an astonishingly rapid rate. A third strand of work summarized by Heggarty (2014) attempts to look into the prehistory of a language population(s) through the evidence obtained from archaeology.

Akin to idea of modern human migrations from Africa (Stoneking, 2006; Vigilant et al., 1991) there is the idea of a putative homeland for reconstructed proto-languages in historical linguistics. Again, the posited homeland for a language family can either be modified or discarded from extra-linguistic evidence. An example of one such study concerns the putative homeland of Austro-Asiatic language family (Sidwell and Blench, 2011). Closely related to homelands is the problem of dating the migratory movements of a human population and its subsequent divergence that can be tracked in terms of linguistic diversity (Wichmann et al., 2010).

There is also some research working the reverse way. It consists of using the accumulated findings from genetics such as Neandertal intermixing with humans and looking at the Hominin anatomical findings to make conclusions about the origins of human language itself. This research consists of looking at the Hominin phylogeny and employing the most parsimonious explanation for human language origin (Dediu and Levinson, 2014; Dediu and Levinson, 2013).

Proposals for long-range family relationships are treated with caution in historical linguistics. Dene-Yeniseian language connection is a demonstrated example of a long-range relationship. Scott and O’Rourke (2010) is a study which uses evidence from physical anthropology to substantiate the long-distance proposal.

2 Genetic diversity and prehistory

2.1 Early studies and problems

There are at least two early studies that combine genetic history with that of linguistic and archaeological history: Sokal (1988) and Cavalli-Sforza et al. (1988).

Sokal (1988) followed a relatively simple procedure when compared to the recent Bayesian methods employed for phylogenetic inference (Ronquist and Huelsenbeck, 2003). Sokal compared three different distances – genetic, geographic, and linguistic
– in the form of matrix correlations. The individual datapoints were clustered beforehand so that the comparisons are made within systems. Genetic distance is computed as the mean absolute difference between two allele frequencies. Spatial distances are computed as the great-circle distances. Linguistic distances follow the notorious classification of Ruhlen (1991).

Sokal performed all possible correlations between the distance matrices. Partial correlations are conducted to negate the effect of one factor while testing the correlation between the other two. Sokal noticed that there are strong partial correlations between genetic distance and linguistic distance as well as geographic distance – if the other distance is kept constant. Sokal observed that more systems show stronger and significant genetic–geographical partial correlations than genetic–language partial correlations. Overall, the paper concludes by pointing out that there is correlation between language and genes even after accounting for geographical factor.

Now, I will move to review some studies on language prehistory that involve genetic studies.

2.2 Genetic studies: A world-wide scenario

Pakendorf (2014) reviews three different angles of prehistory investigation that forms the recent interdisciplinary field:

- Coevolution of language and genes.
- Prehistoric contact and its effect on language change and evolution.
- Demographic history of language families and its effect on the prehistory of the language speakers.

The first angle of investigation is concerned whether linguistic boundaries also act as barriers to genetic flow (admixture). There are two models in this investigation: branch-split model and isolation-by-distance model. The branch-split model hypothesizes that languages and genes coevolve in a successive split fashion and subsequent isolation. The isolation-by-distance hypothesis is characterized by an inverse relation between increasing linguistic distance and decreasing genetic affinity (Cavalli-Sforza et al., 1988; Sokal, 1988). The past studies, undertaken from this angle, suffer from misinformed linguistic classifications.

The second angle of investigation focuses on the undocumented prehistoric contact between two different language groups which can be revealed from genetic analysis. In contrast to the macro-scale correlation studies, this line of investigation focuses on language group specific hypotheses concerning the extant spatial distribution of language speakers. Identifying the reasons behind the spread of click consonants in Bantu languages is an example of such a study. The reason for the borrowing is attributed to maternal ancestors (female speakers) belonging to click consonant language groups (Bostoen and Sands, 2012). There are similar studies for Khoisan populations (Pickrell et al., 2012) placed alongside the Kx’a language family speakers that shows a deep genetic split but have different linguistic affinities.

Another study from Burkina Faso (Barbieri et al., 2011) showed that languages belonging to two vastly different phyla from Niger-Congo language family – Mande and Gur – have no within-population difference in the maternal lineage whereas a significant difference shows up in the paternal lineage. This evidence is in line with the patrilocal marriage pattern observed in these language groups and the subsequent spread of grammatical patterns of maternal language group into the child language.

The third angle of investigation uses language family histories to infer the language family’s prehistory. This is done by analyzing the lexical cognate data for members of a language family and inferring a phylogenetic tree through a off-the-shelf phylogenetic software and test the different possible hypotheses of homelands. Some examples of such studies for different language families are:

- **Austronesian**: Gray and Jordan (2000), Jordan et al. (2009).

\[1\] Traditional language families are called phyla and subgroups are labeled as language families.

\[2\] Significance testing is performed by means of Mantel test.

\[3\] Note that some linguistic experts dispute the evidence that was used to establish Khoisan languages since the family was postulated from the evidence based on shared click consonants and not demonstrated through the comparative method.
• **Tupian**: Walker et al. (2012).
• **Dene–Yeniseian**: Sicoli and Holton (2014) (Also uses structural features).
• **Melanesia**: Hunley et al. (2008).
• **Southeast Asia**: Donohue and Denham (2011) and Denham and Donohue (2012).

There are also studies which employ structural features for inferring phylogenies (Dunn et al., 2008) to distinguish Papuan languages vs. Oceanic languages.

Finally, there are some genetic studies which were performed on a geographical micro-scale for language groups from rest of the world.
• **Native America**: Reich et al. (2012), Amorim et al. (2013) (Only for South America)
• **Indo-European**: Balanovsky et al. (2013) and Burlak (2014).
• **Bantu languages**: Quintana-Murci et al. (2008).

Now, I will review some literature on genetic studies in South Asia and their relation to different linguistic families spoken in India.

### 2.3 Genetic studies in South Asia

Tamang et al. (2012a) and Tamang et al. (2012b) describe the three-dimensional view on peopling of South Asia – social stratification, language affiliation, and genetic affiliation – when trying to reconstruct the history of the peopling of South Asia. The authors begin their review by introducing two scenarios of human spread from Africa. One possible route is the southern coastal route which went along the rim of the Arabian sea and then into the Andaman and Nicobar islands (Thangaraj et al., 2005), finally terminating in Australia (Pugach et al., 2013). The authors also account for the recent Muslim and Siddi (Bantu language speakers) migrations by doing a genetic study regarding the gene flow in the aftermath of different migrations. The Muslim migrations did not leave a large genetic imprint but left a strong imprint on the culture through conquest and conversions.

The authors also perform another study on two populations who speak Indo-Aryan and Dravidian languages but with a plausible Austro-Asiatic descent. Their study reveals that the speakers are actually of Austro-Asiatic origin but the whole community underwent a language shift.

The authors also take a stab at the origin of Austro-Asiatic populations and note that their genetic study suggests a spread of the populations from a South-East Asian homeland. This study seems to be in congruence with the recent view on the original divergence of Austro-Asiatic populations (Sidwell and Blench, 2011).

The scholars also summarize the work on likely origins of the Andaman Islanders who are divided into two distinct tribes: Great Andaman and Little Andaman Islanders. Little Andaman Islanders are divided into Onge, Sentinelese, and Jarawa groups. With evidence from phylogenetic analysis, the scholars propose that the Onge and Andamanese are the descendants of the early group of humans who migrated out of Africa about 65,000 BP. Their studies reveal that the Andaman populations shared a common genetic affinity with Ancient South Indian populations (ASI) and also were genetically isolated for a long time.

In continuing study from Redd et al. (2002), Pugach et al. (2013) examine the hypotheses regarding the common origin of Papua New Guinea and Australian aboriginals. The common origin hypothesis implies that the ancestral group of this two populations reached Sahul (the undivided landmass of New Guinea and Australia) about 45,000 years ago. Further, this ancestral group split from the ancestral Eurasian population about 62,000–75,000 years ago before the migration out of Africa. The migration route followed by this population is a Southern route along the Arabian sea.

The scholars examine the genetic evidence from Australian aborigines (AUA), New Guinea Highlanders (NGH), European (EUR), South Indian populations (DRA), Indian populations (IND), and Han Chinese in Beijing (CHB). They perform a divergence-time estimate for the populations and reaffirm the earlier estimates that CHB and EUR populations diverged about 43,000 years ago. The divergence times between AUA and NGH are estimated around 36,000 years, which is 9,000 years earlier than the consensus date of 45,000 years. The authors do not provide a satisfactory answer for this disagreement but move on to explain the possible signal for admixture found in Australian populations. The authors perform a maximum-likelihood based on an ADMIXTURE analysis for determining the Indian ancestry in Australian populations. The
analysis assumes that each person’s genome has $K$ hypothetical ancestors and tries to find the analysis which has the lowest error rate. They find that $K = 4$ has the lowest error rate and also shares the genetic component that occurs uniquely with highest frequency in Indian populations. This study was performed only on Australian aborigines, New Guinea Highlanders, European, and Indian populations.

To reaffirm the findings from this study, the scholars also performed another test by including African and Dravidian speakers (after removing other Indian language groups). This study revealed that the lowest error rate occurs at $K = 5$ and that the ancestry profile of Australian aborigines populations – not shared with New Guinea Highlanders – shows the highest similarity profile with Dravidian speakers. A divergence-time estimate of the similarity profile shows that the time period of divergence is around 4,230 years ago which agrees with the sudden change in the archaeological record in Australia: namely, change in stone tool technologies, processing of plants, and at least a morphological resemblance between dingo and Indian dog.

Rai et al. (2012) is a study which investigates the likely Indian origins of European Roma populations. Rai et al. analyze DNA samples from 57 linguistic groups and determine that the European Roma populations followed a migratory route through Afghanistan and into Europe starting around the 13th century. The authors report a founder population study on Roma population along with the northwestern Indian population to determine the age of the migration initiated by the founder population. The study yields a age estimate of 1405 $\pm$ 688 BP. This date is in conformity with the alternate hypothesis of van Driem that Roma people are the descendants of the original Northwestern populations who were recruited to defend the Hindu Northwestern region. The successful incursions of Ghazni raids triggered the migrations of the Roma ancestors. The study also finds that the descendant Roma population are closest to the Scheduled Tribes and Castes found in the Northwestern India.

Now I will list some fine-scale genetic studies that have been performed in the Indian subcontinent and investigate various hypotheses regarding the peopling of South Asia.

- **Austro-Asiatic**: Kumar and Reddy (2003), Kumar et al. (2007), Kumar et al. (2006).
- **South India (predominantly Dravidian)**: Kumar et al. (2012).
- **Caste studies**: Thanseem et al. (2006), Bamshad et al. (2001), Thangaraj et al. (2010), Sharma et al. (2012), Reddy et al. (2005).
- **Andaman islanders**: Thangaraj et al. (2005), Thangaraj et al. (2003).
- **Tibeto-Burman**: Gazi et al. (2013), Su et al. (2000), Chaubey et al. (2014), Kraaijenbrink et al. (2014), Van Driem (2014b).
- **Indo-Aryan invasion**: Kivisild et al. (1999).

### 3 Linguistic diversity and prehistory

Linguistic diversity, as defined by Nettle (1999), refers to one of the following: language diversity (language richness), phylogenetic diversity (number of different families), and typological diversity (amount of typological variation) in a unit geographical area. Linguistic diversity provides a quantification of the geographical and (indirectly) temporal spread of languages. Explaining the when, where, and why of linguistic diversity has been gaining the attention of scholars in the last few years. These questions have been addressed from different perspectives such as biology, language contact, ecology, and socio-linguistics.

#### 3.1 Diversity through non-historical linguistics

Gavin et al. (2013) is a recent survey of various efforts to explain the extant linguistic diversity in the world. The authors also outline a methodological and theoretical program for future research on linguistic diversity. The paper describes some caveats when employing correlation or regression studies in studying the causes behind linguistic diversity. Then, the authors propose the following four different processes as drivers of linguistic diversity.

- **Neutral change**: This change is analogous to vertical transmission change in biology where hereditary units, genes, are transmitted from parents to children. Children learn new languages from their parents and environment. When a speech community is isolated from its parent community for a sufficiently long amount of time, the isolated community’s language undergoes sufficient neutral changes so
that it is no longer mutually intelligible with the parent community.

- **Movement and contact:** Although movement and contact are different factors, they are strongly interrelated drivers of language change. Physical and social barriers are also drivers of language diversification. As such, movement of population groups over long geographic distances followed by subsequent isolation caused languages to diversify. One example is the diversification of Vedic Sanskrit into middle Indo-Aryan languages that caused the western Indo-Aryan languages to become mutually unintelligible with eastern Indo-Aryan languages. Population movement can also replace the previously existing languages when entire populations shift from one language to another (usually living some trace of the shift). Contact can also result in language change through the process of borrowing parts of vocabulary or syntax from one language to another; or can result in the development of new languages (creolization).

- **Selection:** This driver of linguistic diversification has two senses. In one sense, it is the social factor that drives the leveling of linguistic varieties in a bounded mutually intelligible linguistic area. Factors such as age, gender, and social class also drive the linguistic diversification of a single language. In another sense, it is individual choice that can drive linguistic diversification. Individual choice can stem out of adoption or borrowing from a prestige language that can lead to upward movement in the social or economic status.

The article ends with a note on **linguistic complexity** – causal factors of linguistic disparity (also called typological diversity by Nettle 1999). The authors admit that socio-linguistic factors play a major role in the generation of linguistic variation. The factors which lead to the development of linguistic complexity are:

1. **Small group size:** Whether small group size causes greater linguistic change or not is debated, but there seems to be not much relation between size or spatial spread of language families to the subsistence type of a particular language family (Hammarström, 2010; Bowern, 2010).

2. **Low levels of contact:** There are studies which show that both grammatical as well lexical units can be borrowed in contact situations.

There are two other factors – large amounts of shared information, high stability – which are not discussed at all.

### 3.2 Diversity through historical linguistics

Now, I will try to link different prehistory studies in South Asia from a historical linguistic point of view (Heggarty, 2014). Historical linguists posit a reconstructed language (proto-language) through the application of the comparative method. The reality of these reconstructions are always questionable and need to fit with the stories told by other prehistoric disciplines such as genes and archaeology. Heggarty poses the following high-level questions regarding the reality of the reconstructed language families and the corresponding proto-languages.

- When was the proto-language spoken? This question refers to the age or time-depth of the proto-language. An extension would be
the time depth of various splits in the proto-language’s descendants.

– Language divergence: This technique is based on correlating the amount of divergence in a language group with its documented time-depth. This approach is at least in two forms. In one form, Holman et al. (2011) employs a regression model to predict the time depths of new families based on linguistic divergence. Another paper of Bouckaert et al. (2012) employs Bayesian phylogenetic techniques to date the proto-Indo-European to 9,000 years BP. In the case of Austro-Asiatic, Holman et al. give a date of 3600 years which is close to the beginning of the rice cultivation period in Mekong river – 4300 BP.

– Cultural reconstruction: This technique attempts to correlate the reconstructed proto-forms with technological or material advances recorded in archaeology. Presence of cultivation of specialized forms of agriculture or pastoralist related products are supposed to be interpreted as part of the culture of the proto-speakers.

• Where was the proto-language spoken? This question refers to the putative homeland of the language family.

– Linguistic paleontology has played a major role in narrowing down the plausible homelands of language families. Austro-Asiatic language family is a classic example where the reconstructed flora and fauna were used to propose a tropical, coastal homeland and reject any homeland in Southern China (Sidwell, 2010).

– Focus of diversity principle is another guiding principle in locating the homeland for a language family. The principle originates with Sapir (1916), who proposed that the geographical location with the largest number of deep branches is supposed to be the homeland. Wichmann et al. (2010) used lexical divergence as an index to triangulate the homelands of major language families of the world. The focus of diversity principle was also employed by the Austro-Asiatic scholars to propose that the Mekong river axis served as the original homeland for the family.

– Phylogeography is a technique which attempts to ground the language family’s phylogeny into the language areas so that a population movement model can explain conflicting hypotheses (Bouckaert et al., 2012).

• Why did the family come into existence? A partial answer to this question has been given by Gavin et al. (2013) in terms of environment, subsistence patterns, conquest, and prestige.

3.3 Linking genes, archaeology, and historical linguistics in South Asia

South Asia is a linguistically diverse region with six different language families and three isolates. Blench (2008) is a recent study that evaluates the linguistic prehistory of South Asia from genetic and archaeological perspectives. Blench takes up the case of each language family and evaluates the proposals for dating the spread of each language family and tries to fit the evidence given by genetic and archaeological studies.

I will concentrate on the Dravidian language group since that is the only language group that has been left out in the previous sections. Also, studies referring to the Dravidian homeland or the dating naturally includes Indo-Iranian group. I will survey the proposal for homelands, dating of the homelands (through linguistic archaeology; Southworth 2005) and attempt to reconcile these with the genetic studies’ findings since 2007.

Dravidian: The Dravidian language family is a well-researched language family with a reconstruction of Proto-Dravidian. Krishnamurti (2003) lists only 26 languages whereas the latest version of Ethnologue lists 83 languages. There is a need for fresh field work on these languages (or dialects). Fuller (2003) performs a cultural reconstruction of Proto-Dravidian populations by considering the reconstructed crop vocabulary based on the Dravidian etymological dictionary (Burrow and Emeneau, 1984). However, the dictionary suffers from two drawbacks: Reconstructions are not given in this
dictionary; there is a data collection bias towards literary languages.

Scholarly opinion, traditionally, stresses the geographical location of Brahui to determine the Dravidian homeland. This proposal argues for an eastern and southern spread of Dravidian languages from Northwestern India. However, Fuller (2007) argues for a peninsular homeland for Dravidian by invoking the parsimony principle. The proponents of the former homeland posit a agriculturalist expansion of Proto-Dravidian society. However, the reconstructed vocabulary of life-stock and crop vocabulary speaks a different story.

The other story would be that the original populations were hunter-gatherer populations but had technologies for dehusking, grinding and some form of storage (mesolithic culture) who transitioned into an agricultural society. If such is the case, then Brahui would just be a migration during the mesolithic stage of Proto-Dravidian society. The agricultural complex in South India is linked to the Southern Neolithic and is dated to about 4000 BP. This seems to fit with the recent study of Moorjani et al. (2013) claiming that the Indian population descended from a admixture of Ancestral South Indian and Ancestral North Indian population (related to West Eurasian). It is interesting to note that Moorjani et al. (2013) date the Indian population admixture to about 4,200-1,900 BP well after the first appearance of agriculture in South Asia (about 7,000 BP).

3.4 What needs to be done?

What remains to be investigated is the question regarding which languages were spoken during the time period between the Southern route of migration of first humans out-of-Africa and the first admixture between ANI and ASI populations? The dates from the comparative method might not give an answer since, dates are shallow – about 8,000 years BP – due to the inherent limit of the method itself. Hence, lexical data is not the best data for reconstructing the temporally deeper relations between languages or to fill the time gap left by recent genetic and archaeological studies.

For instance, Fuller (2003) provides a list of crop vocabulary which are neither of Indo-European nor of Dravidian origin. The words are not even of Austro-Asiatic affinity. This conundrum might be resolved by employing structural data for inferring a phylogenetic tree and then dating the internal nodes of the tree. Finally, the genetic studies in South Asia reject a Aryan invasion hypothesis in favor of admixture between Ancestral North Indian and Ancestral South Indian populations. However, one thing that misses from their rejection is the placement of the ancestral population’s migrations into South Asia. This missing link is also evident in linguistic studies in the form of words of unknown origin.

I will conclude with a note on Sino-Tibetan languages. The origin of Sino-Tibetan languages is a much debated issue. What is the position of Chinese in Sino-Tibetan languages (Van Driem, 2007)? van Driem (2014a) hypothesizes about the possible links between the peopling of Asia – north of Himalayas – and disputed language families whose time depth goes beyond historical linguistics’ ken.

Acknowledgements

I thank Søren Wichmann for the comments on an earlier draft. This draft was submitted to satisfy the requirements for a course on language prehistory.

References


Tim Denham and Mark Donohue. 2012. Reconnecting genes, languages and material culture in Island Southeast Asia: Aphorisms on geography and history. Language Dynamics and Change, 2:184–211.


Keith Hunley, Michael Dunn, Eva Lindström, Ger Reesink, Angela Terrill, Meghan E Healy, George


Mark Stoneking. 2006. Disentangling genes, geography, and language. *Languages and Genes*.


