The origins, evolution, and diversity of human languages project description

By George Starostin and Tanmoy Bhattacharya

The evolutionary relationships of languages have been a lively field of research for nearly two centuries, ever since August Schleicher's genealogical trees for language families, or Dumont d'Urville's attempt to introduce a quantitative aspect into the comparison of "Oceanic" languages. These early roots actually predate Darwin's first tree of species evolution and show that the tradition of integrating linguistics into the field of natural science is nearly as old as linguistics itself.

Early work on phylogenetics in biology has been grounded in detailed expert descriptions of morphology. The advent of a deluge of molecular data and the relative simplicity of the mechanisms underlying sequence evolution has transformed molecular phylogenetics into a data-driven, computational science. In contrast to biological evolution, the detailed forces and mechanisms shaping language change over time remain much less understood. Nevertheless, it seems safe to assume that spoken language consists of a set of core features that are mainly uniparentally coinherited with occasional horizontal influences from other speech communities.

Given two or more different languages or language groups, we may quantify the probability that they share a common ancestor at a given depth of time—say, 100, 1,000, or 10,000 years back—and, also on a probabilistic basis, reconstruct certain features of that ancestor. Where biological evolution uses mutations in RNA or DNA bases and amino acids, proteins, core housekeeping genes, and regulatory networks, we operate with such linguistic data as phonemes, lexical roots and grammatical markers, core words, regular sound shifts, and other linguistic features. These results may then be integrated with anthropological evidence, which is also made easier within a general probabilistic framework.

Over time, as a language is passed from one generation to the next, its meanings can change, as can the way it is spoken. The processes by which meaning and pronunciation evolve seem to be generally independent, so that changes in semantics (meaning) are not precipitated by changes in pronunciation. Meaning thus contrasts both with phonetics, in which instrumental measurement of physical properties of articulation and acoustics is relatively straightforward, and with grammatical structure, for which there is general agreement on a number of basic units of analysis.

As linguistic communities grow and diverge, different mechanisms of change begin to apply to different communities, ultimately resulting in the enormous linguistic diversity that is observed today. This means that linguistic change may be studied on different levels—the "micro-level" at which particular phonetic, grammatical, and semantic changes take place, and the "macro-level", which generalizes and systematizes these changes over long periods of time. Linguistic research conducted at the Santa Fe Institute relates to both the "micro-" and the "macro-level", as well as their complicated interaction, and is represented by several subprojects whose results, when properly integrated, are expected to shed proper light on both *what* kinds of linguistic change humanity has gone through over the last 10,000 years or more, and *how* and *why* exactly it underwent that change.

Of these subprojects, "Quantifying Semantic Change" and "Quantifying Phonetic Change", coordinated by Tanmoy Bhattacharya in two small research groups, largely deal with "micro-level" change, whereas phonetic, lexical, and grammatical evolution over large periods of time is the primary concern of "Evolution of Human Languages", an international program led by Murray Gell-Mann and George Starostin. Brief descriptions of all these three lines of research are provided below.

Quantifying Semantic Change

The space of concepts expressible in any language is vast. This space is covered by individual words representing semantically tight neighborhoods of salient concepts. There has been much debate about whether semantic similarity (similar meanings) of concepts is shared across languages. On the one hand, all human beings belong to a single species characterized by, among other things, a shared set of cognitive abilities. On the other hand, the 6,000 or so extant human languages spoken by different societies in different environments across the globe are extremely diverse and may reflect accidents of history as well as adaptations to local environments. Thus, the question of the degree to which conceptual structures expressed in language are due to universal properties of human cognition, the particulars of cultural history, or the environment inhabited by a society, remains unresolved.

To address our primary question, it is necessary to develop an empirical method to characterize the space of lexical meanings. Our project arrives at such a measure by noting that translations uncover the alternate ways that languages partition meanings into words. Many words have more than one meaning, or sense, to the extent that word senses can be individuated. Words gain meanings when speakers extend their use to similar meanings; words lose meanings when another word is extended to the first word's meaning, and the first word is replaced in that meaning. To the extent that words in transition account for many of the polysemies (multiple meanings of a single word form) revealed in cross-language translations, the frequency of polysemy found across an unbiased sample of languages can provide a measure of semantic similarity among word meanings. Translation is, therefore, used in our project to define a class of polysemies in core vocabulary, and to identify instances of such polysemies from a phylogenetically and geographically distributed set of languages chosen according to the methods of typology and universals research.

Several cross-linguistic surveys of lexical polysemy, and its potential for understanding semantic shift, have been carried out. The domains surveyed include: body part terms, cardinal direction terms, perception verbs, concepts associated with fire, and color metaphors. Our work has added a new dimension to the existing body of research by providing a comprehensive mathematical method using a systematically stratified global sample of languages to measure degrees of similarity.

Universal structures in lexical semantics greatly aids the reconstruction of human phylogeny (human evolutionary processes) using linguistic data. Much progress has been made in reconstructing the phylogenies of word forms from known cognates in various languages, thanks to the ability to measure phonetic similarity and our knowledge of the processes of sound change. However, the relationship between semantic similarity and semantic shift is still poorly understood. The standard view in historical linguistics is that any meaning can change to any other meaning, and that no constraint is imposed on what meanings can be compared to detect cognates. It is, however, generally accepted among historical linguists that language change is gradual, and that words in transition from having one meaning to being extended to another meaning should be polysemous. If this is true, then the weights on different links may reflect the probabilities that words in transition over these links will be captured in "snapshots" by language translation at any time. It may be reasonable to assume that such semantic shifts can be modeled as diffusion in the conceptual space, or along a universal polysemy network. A full "state-process" model of language structure and change would be required to model the entire semantic shift process. This model would be calibrated, in part, by comparing polysemy in the present with historical records, or with phylogenetic reconstructions of semantic shift. Nevertheless, the universality revealed by our study of polysemy networks across the world's languages is an important input to methods of inferring which words derive from a common ancestry.

Quantifying Phonetic Change

Concerted evolution is normally used to describe parallel changes at different sites in a genome, but it is also observed in languages where a specific phoneme (sound) changes to the same other phoneme in many words in the lexicon—a phenomenon known as regular sound change.

Linguists have long recognized concerted change that affects copies of the same sound (or phoneme) appearing in different words as a central feature of linguistic evolution. A well-known example is the *p>f sound change in the Germanic languages wherein an older Indo-European p sound was replaced by an f sound, such as in *pater>father, or *pes, *pedis>foot (linguistic convention is to use the ''>'' symbol to indicate a transition from one sound to another, and here the * symbol denotes a reconstructed ancestral form). These multiple instances of one phoneme changing to the same other phoneme yield regular sound correspondences between pairs or groups of languages. Linguists have proposed several explanations for the regularity of changes grounded in a number of basic processes, including speech production, perception, and cognition.

Can events of concerted change be detected statistically in sequence data, and do they improve the characterization of evolution and the inference of evolutionary histories? Although previous researchers working in a linguistic setting have used the concept of regular changes to build algorithms for automatically inferring cognacy (linguistic ancestry), our project was the first to build a probabilistic description of concerted change. This places concerted evolution in a statistical setting that allows for formal hypothesis testing about the nature and rates of concerted changes. For example, the question of how many parallel changes are required to be recognized as an instance of concerted change is naturally dealt with in one of our models: the statistical signature of concerted or regular change is that the multiple parallel events are more probable if treated as a single coordinated change than as a collection of independent changes.

Usefully, the genetic and linguistic phenomena share fundamental properties relevant to their statistical characterization. Phonemes are the units of sound that make up words and distinguish one word from another, just as the four nucleotide bases (A, C, T, G) make up DNA gene sequences or the 20 amino acids make up protein sequences. The number of distinct sounds in a language varies greatly, but somewhere around 30–60 phonemes are commonly sufficient to describe the range of distinctive sounds in a language’s words. Collections of words can therefore be thought of as providing phonemic “sequence information”, that might be informative as to the history, rate, and patterns of concerted evolutionary change in language, and in a manner analogous to sequences of DNA.

Collections of words can therefore be thought of as providing phonemic ‘‘sequence information’’ that might be informative as to the history, rate, and patterns of concerted evolutionary change in language, and in a manner analogous to sequences of DNA.

We have developed a general statistical model that can detect concerted changes in aligned sequence data and, in one study of regular sound changes in the Turkic language family, we have used this method to create a map of the linguistic ancestry that infers the widely-accepted historical timings of linguistic change, without embedding prior knowledge of these dates into the model.

In future pursuits, we hope to apply the same model and methods to illumine those regions of linguistic evolution that are yet unknown.

Evolution of Human Languages

According to the latest calculations, the number of languages currently spoken on Earth exceeds 7,000 distinct units, not counting innumerable dialectal varieties. The degrees of linguistic diversity attested in the sound, grammar, and lexical systems of all these units are staggering—yet much, if not most, of this diversity upon close scrutiny turns out to be relatively recent. Over the past two hundred years, dozens, sometimes hundreds of languages have been successfully clustered together by scientists into "families", such as Indo-European, Uralic, Austronesian, or Niger-Congo; within each such family, all of its units are tied together by recurrent patterns of linguistic development, and are traced back to reconstructed ancestral states called "protolanguages". However, linguistic change tends to be rapid, and few of these reconstructions deal with time depths larger than 5,000–6,000 years—a serious discrepancy with the age of human language as such, the lower time limits on which are usually defined as at least 50,000 years (and probably vastly longer).

The primary goal of the international program known as EHL (Evolution of Human Language) is to work out a detailed historical classification of these languages, organizing them into a genealogical tree similar to the accepted classification of biological species. Since all representatives of the species Homo sapiens presumably share a common origin, and the only known genetic changes related to our linguistic ability predate the origin of modern humans, it would be natural to suppose—although extremely hard to prove—that all or most known human languages also go back to some common source. The only way to proceed here is “bottoms up”: classifying attested languages and dialects into groups, groups into families, families into “macro ” or “super” families and so on, as far as one can penetrate using comparative-historical and cladistic methodology. Most existing classifications, however, do not look beyond the 300–400 language families that are relatively easy to discern. This restriction has natural reasons: languages must have been spoken and constantly evolving for at least 40,000 years (and quite probably more), while any two languages separated from a common source usually lose almost all superficially common features after some 6,000–7,000 years.

Nevertheless, despite widespread skepticism and reluctance to tackle the problem, there are a number of scholars who believe that these obstacles are not insurmountable. Research has been going on over the past several decades that appears to indicate that larger genetic groupings are not only possible, but indeed quite plausible. It can be shown that most of the world's language families can be classified into roughly a dozen large groupings, or macrofamilies. Two sorts of evidence can be used for this purpose:

1) The science of historical linguistics has developed a very powerful tool—the comparative method—that allows the reconstruction of unattested language stages, so-called proto-languages, based on systematic comparison of their present day descendants. With the gradual accumulation of this data over the past 200 years, it has become evident that, while modern languages may vary significantly, protolanguages in many cases tend to be much more similar to one other. Thus, modern English, Finnish, and Turkish may have very little in common (and what little there is, is practically indistinguishable from chance), but their respective ancestors—Proto-Indo-European, Proto-Uralic and Proto-Altaic—appear to have many more common traits and common vocabulary items. This means that it is possible, in theory and with practice, to extend the time perspective and reconstruct even earlier stages of human language. In fact, much of this research has already been conducted.

2) Where a detailed reconstruction of the proto-language is impossible to achieve (e.g., because of insufficient data) or requires more time and effort than can be devoted to the task, it is still possible to build somewhat weaker models of language evolution based on a combination of manual and automatic analysis of limited corpora of data. Of all types of linguistic data that can be used for historical purposes, it is the so-called "basic lexicon" that generally persists the longest over time. Focusing our attention on the comparison of small groups of words, such as the Swadesh wordlist, and tracing their evolution on micro- and micro-levels, reduces the amount of "noise" (such as due to borrowings, from which no language is free) and helps strengthen the case for many proposals of long-range relationship.

Based on these theoretical considerations, the particular work that goes on within EHL is being carried out in three main directions:

I. Reconstruction of proto-languages and compilation of computerized etymological dictionaries (databases) in accordance with the traditional comparative method. A large set of such databases has already been open to public access for a long time and is gradually being enlarged as more data become available and more analytical work is performed on various language families. The set currently includes data on comparative Indo-European, Uralic, Altaic, Dravidian, North Caucasian, Yeniseian, Sino-Tibetan, Indo-European, Austroasiatic, Chukchee-Kamchatkan, Eskimo, Semitic, and several families collectively known as Khoisan languages. Many more databases, in particular those on specific language families of Africa and America, are in preparation.

EHL also adopts a generally tolerant position towards attempts to prepare etymological databases for those deep-level macrofamilies whose daughter proto-languages have been already reconstructed to general satisfaction. Currently, databases for three major macrofamilies of the Old World already exist: viz., for the Eurasiatic (Nostratic), Sino-Caucasian, and Afroasiatic macrofamilies. Exploration of macrofamily connections for Africa, America, and the Pacific region is also well on the way.

II. Lexicostatistical testing of both traditional and new theories of language relationship. To emphasize the central role that lexicostatistics has played in determining the proper historical relations between languages, EHL has currently launched, as one of its subprojects, the construction of the Global Lexicostatistical Database (GLD) that will contain properly assembled and annotated Swadesh wordlists for the majority of the world's languages, as well as for reconstructed proto-languages on all levels, based on rigorous methodological procedures.

III. Procedures for automatic data handling. An important issue in historical linguistics is the amount of subjectivity on the part of the researcher when hypotheses on unattested ancestral stages of languages are concerned. According to the collective opinion of historical linguists working within EHL, none of the existing models and algorithms that have been proposed for language classification purposes have yet managed to take into account all of the necessary factors responsible for historical evolution, making “manual” handling of the data irreplaceable. Nevertheless, EHL still sees the elaboration of such models as an integral part of the project. Improved, more elaborate, algorithms of automatic classification, and even reconstruction, are being worked on within the EHL team; EHL participants also exchange data and experience with several other working groups conducting research in the same direction.

Besides its theoretical goals, one of the major purposes of EHL is to provide specialists and enthusiasts around the world with as much information on the history of language(s) as possible. To that purpose, all of the databases, as soon as they reach “usable” shape, are made public. EHL provides wordlists and etymologies for many languages and language families that are poorly known and data on which is almost impossible to find in any kind of open access system. EHL participants have also scanned, recognized, and converted to database format some of the major existing etymological dictionaries, such as Pokorny’s Indo-European etymological dictionary.

The Evolution of Human Language project was originally founded in 2001, with the joint efforts of Murray Gell-Mann, Sergei Starostin (1953-2005), and Merritt Ruhlen, a generous grant from the John D. & Catherine T. MacArthur Foundation, and plenty of support from the Santa Fe Institute. Back then, the experience of the EHL team did not extend significantly beyond professional work on several large families of the Old World and their prehistorical connections. Today, the EHL team is integrating data from all of the world’s major and minor language stocks in order to push our knowledge of linguistic prehistory as far back as possible. Once the assembled data have been properly organized and their analysis, combining sound traditional methodology with modern cladistic methods, completed, EHL’s classification aspires to become a solid reference model for linguists, historians, anthropologists, geneticists and everyone even remotely interested in human prehistory.