Are You My Mother…Tongue?

The story of the Tower of Babel is the creationist's version of the origin of language diversity: Man, in one of his many hubristic moments decides to build a tower to Heaven. God, realizing that communication is the key to completing any massive public works project, foils the plan by replacing the single common language of the workers by many different languages, thereby making impossible their cooperation, not to mention the scheduling of car pools and the organization of a softball team.

Is there a true "mother tongue" to which all existent modern languages can trace their origin? More precisely, what is the genealogical tree or phylogeny of language? It is these sorts of questions, the ones that look to tell a story of a branching journey of the development of languages that is the goal of SFI's Evolution of Human Languages (EHL) Project, funded by the John D. and Catherine T. MacArthur Foundation, and spearheaded by SFI Distinguished Fellow Murray Gell-Mann and Russian Academy of Sciences Member (and frequent SFI visitor) Sergei Starostin. The third leader of the project is Dr. Merritt Ruhlen from Stanford University, author of the monographs "Guide to the World's Languages" and "The Origin of Languages".

The EHL project falls squarely within the discipline of comparative linguistics. The last two hundred years or so of the subject have been devoted to the clarification of the most elementary stages of linguistic organization, an effort which has resulted in a partitioning of the roughly six thousand attested languages into several hundred more fundamental "language families", each of which implies the existence of a single language ancestor for its family members.

The standard methodology used to show relatedness involves the identification of a set of phonetic similarities between the words in the respective ***basic vocabularies*** (eg., words for body parts, numerals, natural phenomena etc.). This is the sort of comparison that supports the existence of a common Germanic language able to account for the English "hundred" and the German "hundert", or uses the Italian "cento" and the French "cent" as evidence for an older ancestral Romance language - actually attested as Latin. The reconstructed protolanguages are then grouped together into families of the next level, in our case forming the so called Indo-European family. It is estimated that its proto-language was spoken (in a homeland that is still a matter of dispute) some six or seven thousand years ago. A number of other universally recognized families have similar "time depths". Although many comparative linguists maintain that further classification is impossible because too many changes impede comparison and reconstruction, a few bold scholars go further to find superfamilies composed of several such families, with protolanguages spoken thousands of years earlier. Instead of comparing modern languages they use the reconstructed protolanguages that are naturally closer to each other than their modern descendants.

This is the so-called *step-by-step reconstruction*, a technique due to the Russian school of comparative linguistics first used in the construction of Eurasiatic proto-language. After several decades of research the evidence for macrofamilies became overwhelming, and there are many indications that even those can be further grouped together suggesting the existence at some point in time of a single common ancestor.

These achievements are, in the words of Starostin, "pre-science", insofar as they are obtained without mathematical tools. However, it is in the search for deeper levels of organization, and in the investigation of temporal considerations, that the tools of mathematics and statistics truly come to the forefront, marking a transition from pre-science to science for comparative linguistics, and the starting point of the discipline of “lexicostatistics” or "glottochronology", originally started in the US by Maurice Swadesh. It is in this domain that SFI is making a big contribution.

It is fitting that Gell-Mann is the person leading this search. The son of the founder of the Arthur Gell-Mann School for Languages (which taught English to immigrants and other languages to Americans), Gell-Mann has been interested in etymologies and language sound systems since childhood.

In essence, what Gell-Mann and Starostin seek is the linguistic equivalent of Gell-Mann’s Nobel Prize winning “Eightfold Way”, his insightful 1950’s reorganization of what was then a “zoo of particles” (over one hundred of them) thought to be the fundamental constituents of nuclear matter. By grouping them according to certain approximate symmetry conditions, and creating new mathematical techniques for their study, Gell-Mann was able to see this apparent confusion of particle types as parts of a more coherent whole. The reorganization suggested a new, more basic fundamental particle which Gell-Mann named the quark, as responsible for this "zoo", and in so doing, he brought our understanding of the story of matter closer to the beginning of time. By fusing his great love of language with his scientific proclivities he has found what seems to be a promising approach toward the search for the Mother Tongue.

At the heart of the problem is estimating the rate at which languages change, as measured by the changes that occur in the basic vocabulary as it passes from generation to generation, passed on like genes of “cultural DNA”. The basic principles underlying the model formulation are that language requires stability to ensure communication between generations, but that nevertheless there is inevitable information drift, resulting in changes during transmission. The latter takes place via a mechanism of ***replacement*** which occurs either through ***borrowing*** or through ***synonymic shift***. Replacement by borrowing occurs when a word is replaced by its foreign equivalent: an example is “mountain”, borrowed from French to English, supplanting the old English “berg”. Replacement by synonym occurs within a language when a word drifts to a new, but nearly equivalent meaning. An example of this is the current usage of the word “kill”, which has its origins in the Germanic word for torture. Keeping in mind the genetic model, these sorts of language mutations are akin to horizontal and vertical replacement (transmission) in genetics which result in the evolution of a particular genetic sequence.

While the replacement by borrowing is unpredictable, replacement by synonym seems to follow a standard model of genetic drift, the mechanism which many believe is responsible in biology for the species diversity we see today. In the context of language this model provides a means by which the times of language divergence can be estimated. At work here is an implicit assumption of a regular process of change which Starostin likens to the measurable rate of isotope decay that makes carbon-dating the exact science that it is today. The glottochronological version of carbon-dating suggests that one word of basic vocabulary is replaced roughly every 200 to 300 years or about 5 over a millennium. The original model assigns approximately the same probability of replacement to each word in the basic vocabulary. This model is quite naïve and Gell-Mann is leading an effort directed toward tuning the model using more realistic estimates of replacement probabilities of individual words.

Current techniques appear to reliably reconstruct the “proto-languages” in use six to seven thousand years ago. In addition, there is striking evidence for the existence of about ten “superfamilies” responsible for all languages in use today. The analysis reveals some interesting family relations, for example, indicating that northeast Asian languages such as Korean and Japanese are closer to European languages that southeast Asian languages (eg., Chinese).

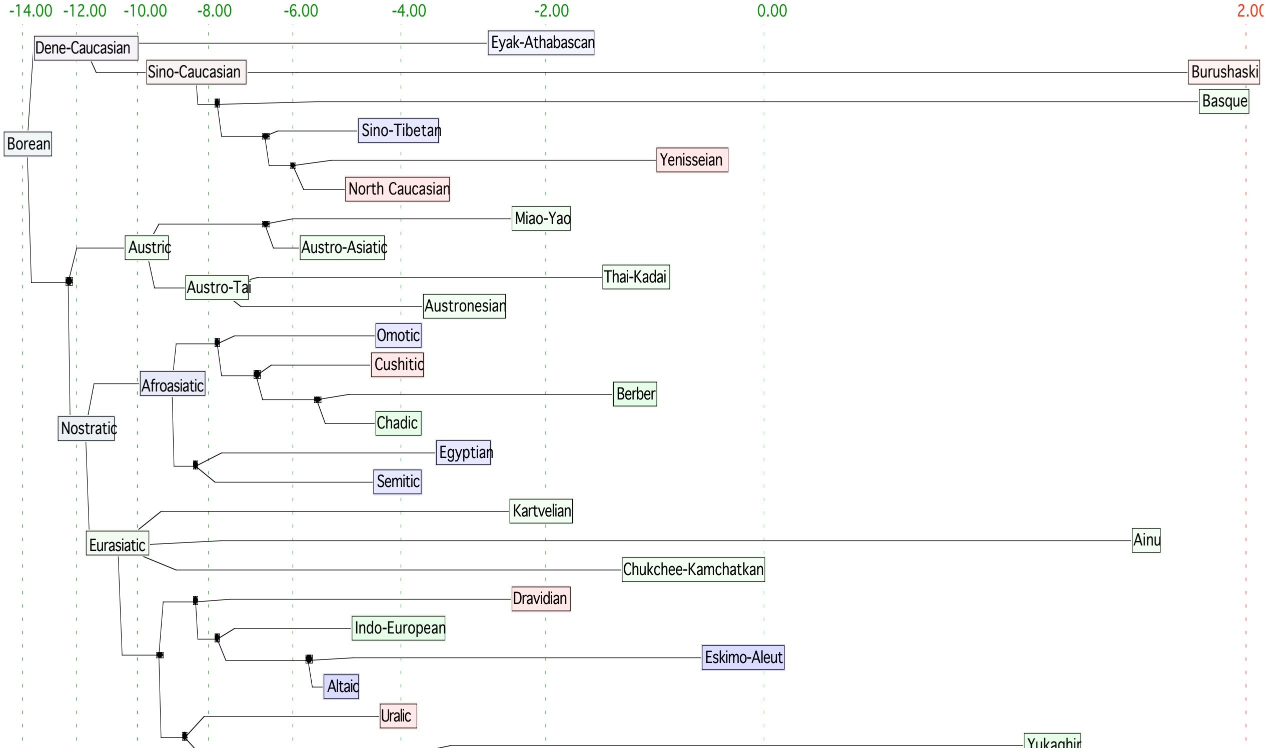
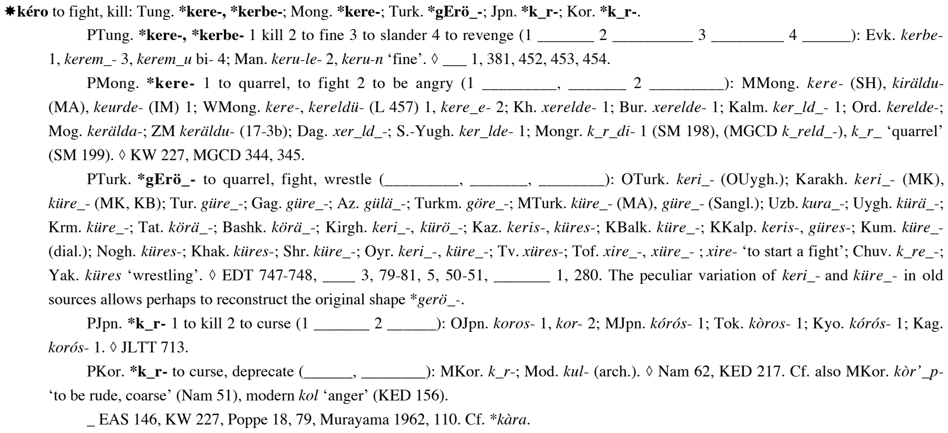


Figure1: Genealogical tree of the languages of the Old World (with time-scale in millennia), obtained on the basis of comparing lists of the thirty-five most stable words in various language families of Northern Africa and Eurasia. The margin of error at the deepest points in the tree is on the order of one a millennium.

The theoretical (i.e., model building) component of the EHL program is paired with (if not made possible by) a huge empirical component. Starostin is also a co-author of the recently completed Etymological Dictionary of Altaic Languages which gives a comparative study of the Altaic languages.



✱**kéro** to fight, kill: Tung. ***kere-, *kerbe-**; Mong. ***kere-**; Turk. ***gErö_-**; Jpn. ***k_r-**; Kor. ***k_r-**.

PTung. ***kere-, *kerbe-** 1 kill 2 to fine 3 to slander 4 to revenge (1 _______ 2 __________ 3 _________ 4 ______): Evk. *kerbe-* 1, *kerem_-* 3, *kerem_u* bi- 4; Man. *keru-le-* 2, *keru-n* 'fine'. ◊ ___ 1, 381, 452, 453, 454.

PMong. ***kere-** 1 to quarrel, to fight 2 to be angry (1 _________, _______ 2 _________): MMong. *kere-* (SH), *kiräldu-* (MA), *keurde-* (IM) 1; WMong. *kere-*, *kereldü-* (L 457) 1, *kere_e-* 2; Kh. *xerelde-* 1; Bur. *xerelde-* 1; Kalm. *ker_ld_-* 1; Ord. *kerelde-*; Mog. *kerälda-*; ZM *keräldu-* (17-3b); Dag. *xer_ld_-*; S.-Yugh. *ker_lde-* 1; Mongr. *k_r_di-* 1 (SM 198), (MGCD *k_reld_-*), *k_r_* 'quarrel' (SM 199). ◊ KW 227, MGCD 344, 345.

PTurk. ***gErö_-** to quarrel, fight, wrestle (_________, _______, ________): OTurk. *keri_-* (OUygh.); Karakh. *keri_-* (MK), *küre_-* (MK, KB); Tur. *güre_-*; Gag. *güre_-*; Az. *gülä_-*; Turkm. *göre_-*; MTurk. *küre_-* (MA), *güre_-* (Sangl.); Uzb. *kura_-*; Uygh. *kürä_-*; Krm. *küre_-*; Tat. *körä_-*; Bashk. *körä_-*; Kirgh. *keri_-*, *kürö_-*; Kaz. *keris-*, *küres-*; KBalk. *küre_-*; KKalp. *keris-*, *güres-*; Kum. *küre_-* (dial.); Nogh. *küres-*; Khak. *küres-*; Shr. *küre_-*; Oyr. *keri_-*, *küre_-*; Tv. *xüres-*; Tof. *xire_-*, *xüre_-* ; *xire-* 'to start a fight'; Chuv. *k_re_-*; Yak. *küres* 'wrestling'. ◊ EDT 747-748, ____ 3, 79-81, 5, 50-51, _______ 1, 280. The peculiar variation of *keri_-* and *küre_-* in old sources allows perhaps to reconstruct the original shape **gerö_-*.

PJpn. ***k_r-** 1 to kill 2 to curse (1 _______ 2 ______): OJpn. *koros-* 1, *kor-* 2; MJpn. *kórós-* 1; Tok. *kòros-* 1; Kyo. *kórós-* 1; Kag. *korós-* 1. ◊ JLTT 713.

PKor. ***k_r-** to curse, deprecate (______, ________): MKor. *k_r-*; Mod. *kul-* (arch.). ◊ Nam 62, KED 217. Cf. also MKor. *kòr'_p-* 'to be rude, coarse' (Nam 51), modern *kol* 'anger' (KED 156).

_ EAS 146, KW 227, Poppe 18, 79, Murayama 1962, 110. Cf. **kàra*.

An example of an etymology and a small excerpt from the Eurasiatic database: the common Altaic root for "fight, kill", with a phonological reconstruction and a detailed account of its descendants in Turkic, Mongolian, Tungus-Manchu, Korean and Japanese".

The print component of the project is important, but the process of comparison and modeling is primarily focused on the development and management of a growing collection of on-line language databases. At the head of the database effort is Starostin, whose software package "STARLING" is designed specifically for linguistic database management (see http://starling.rinet.ru). The number of on-line language databases is increasing steadily. Of great current interest is the effort to digitize all the languages of New Guinea, an effort which will go a long way toward the reconstruction of the Indo-Pacific proto-language.

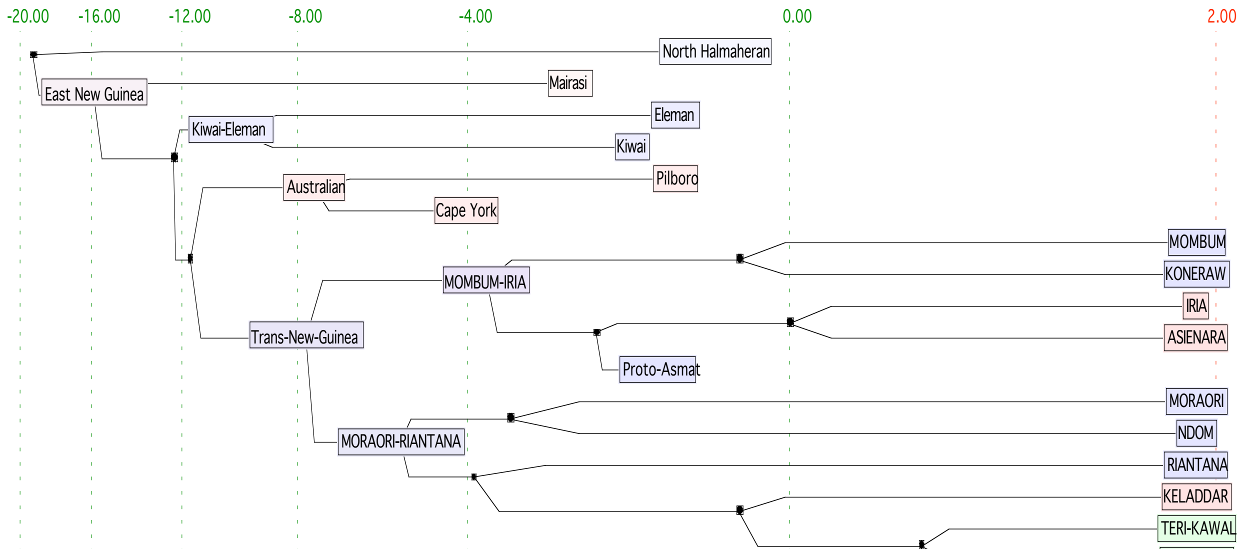


Figure 2: The genealogical tree of several language families of New Guinea and Australia (with time-scale in millennia) also obtained on the basis of the evolution of the thirty-five most stable words. The classification is far from complete, since most of the languages are not yet processed in a proper way; however, it gives an idea of the time distance and level of divergence of languages in this part of the world.

The installation at SFI of the entire STARLING project (software, webserver, etc.) is one of the major directions of current work in the EHL project. This, in concert with the mathematical modeling effort, defines the EHL project as another cornerstone in SFI's work at the scientific frontier. Our generation is bearing witness to a long overdue mathematicization of the life and social sciences, a modern updating of the Tower of Babel tale in which through the ever-broadening mediation by the universal language of number, scientific knowledge is growing via a renewed unification across disciplines. SFI's work to find the Mother Tongue is yet another instance of the progress propelled by the rewriting of sciences in the Mother Tongue of mathematics.