«Evolution of Human Languages»:

current state of affairs (03.2014)

Contents:

I. Currently active members of the project.

George Starostin	*Primary affiliation*: Senior researcher, Center for Comparative Studies, Russian State University for the Humanities (Moscow). *Web info*: http://ivka.rsuh.ru/article.html?id=80197 *Publications*: http://rggu.academia.edu/GeorgeStarostin *Research interests:* Methodology of historical linguistics; long- vs. short-range linguistic comparison; history and classification of African languages; history of the Chinese language; comparative and historical linguistics of various language families (Indo-European, Altaic, Yeniseian, Dravidian, etc.).
Ilia Peiros	*Primary affiliation*: Visiting researcher, Santa Fe Institute. Formerly, professor of linguistics at the University of Melbourne. *Publications*: http://orlabs.oclc.org/identities/lccn-n97-4759 *Research interests:* Genetic and areal language relationships in Southeast Asia; history and classification of Sino-Tibetan, Austronesian, Austroasiatic languages; macro- and micro-families of the Americas; methodology of historical linguistics.
Sergei Nikolayev	*Primary affiliation*: Senior researcher, Institute of Slavic Studies, Russian Academy of Sciences (Moscow / Novosibirsk). *Web info / publications list (in Russian)*: http://www.inslav.ru/index.php?option-=com_content&view=article&id=358:2010-06-09-18-14-01 *Research interests:* Comparative Indo-European and Slavic studies; internal and external genetic relations of North Caucasian languages; internal and external genetic relations of North American languages (Na-Dene; Algic; Mosan).
Alexei Kassian	*Primary affiliation*: Researcher, Institute of Linguistics, Russian Academy of Sciences (Moscow). *Additional affiliation*: Professor, Center for Comparative Studies, Russian State University for the Humanities (Moscow). *Web info*: http://ivka.rsuh.ru/article.html?id=91107. *Publications*: http://iling-ran.academia.edu/AlexeiKassian. *Research interests*: Methodology of linguistic classification; Indo-European and Anatolian historical linguistics; the Nostratic hypothesis; internal and external genetic relations of North Caucasian languages; Na-Dene languages and historical linguistics.

Mikhail Zhivlov	*Primary affiliation*: Professor, Center for Comparative Studies, Russian State University for the Humanities (Moscow). *Additional affiliation*: Researcher, Institute of Linguistics, Russian Academy of Sciences (Moscow). *Publications and info*: http://rggu.academia.edu/MikhailZhivlov. *Research interests*: Comparative Indo-European studies; history and classification of the Uralic languages; Nostratic linguistics; internal and external genetic relations of North American language families (Hokan, Penutian).
Timothy Usher	*Primary affiliation*: Research fellow, The Rosetta Project. *Web info*: https://sites.google.com/site/newguineaworld/. *Research interests*: Theory and methodology of long-range comparison; genetic and areal relationships in the «Indo-Pacific» region; linguistic reconstruction and classification for non-Austronesian languages of New Guinea.
John Bengtson	*Primary affiliation*: Formally retired. Vice-president of The Association for the Study of Language in Prehistory (ASLIP). *Web info*: http://jdbengt.net/. *Publications*: http://jdbengt.net/biblio.htm. *Research interests*: Theory and methodology of long-range comparison; «global etymologies» and their interpretation; the Dene-Caucasian hypothesis; history and genetic affiliation of the Basque language.

II. Linguistic experts associated with the project (past members; consultants; occasional collaborators, etc.).

† Sergei Starostin **(1953-2005).** *Center of Comparative Studies, RSUH, Moscow.*

Former co-director of the EHL project (2000-2005) on the Russian side. Works on general theory, methodology of historical linguistics and lexicostatistics; Indo-European, North Caucasian, Yeniseian, Sino-Tibetan, Altaic language families; Nostratic and Sino-Caucasian hypotheses. Primary author of two major etymological dictionaries («A North Caucasian Etymological Dictionary», co-authored with S. Nikolayev; «Etymological Dictionary of the Altaic Languages», co-authored with A. Dybo and O. Mudrak). Creator of the STARLING linguistic software, now used as the main computer tool for EHL. Author of the «Tower of Babel» site, where most of the EHL databases are stored and published.

Anna Dybo. *Institute of Linguistics, Moscow; Center of Comparative Studies, RSUH, Moscow.*

Russia's leading expert on the history, classification, dialectology of Turkic languages. Works on general theory and methodology of historical linguistics; Proto-Turkic reconstruction; Altaic and Nostratic hypotheses; areal connections between Turkic languages and other linguistic units of Central Asia. **Official EHL participant from 2001 to 2010**; beyond that point, continues to provide consultation services and occasional collaboration on issues related to language classification in Eurasia.

Oleg Mudrak. *Institute of Linguistics, Moscow; Center of Comparative Studies, RSUH, Moscow.*

Expert on languages and linguistics of various national minorities of the former USSR and neighboring territories: Turkic, Mongolic, Tungusic, «Paleo-Siberian» (Chukchee-Kamchatkan, Yukaghir, Nivkh, Eskimo-Aleut). Works on historical Turkic, Altaic, Chukchee-Kamchatkan, Eskimo-Aleut linguistics; genetic and areal connections between the various linguistic units of Siberia and the Far East. **Official EHL participant from 2001 to 2010**; beyond that point, continues to provide consultation services and occasional collaboration on issues related to language classification in Eurasia.

Alexander Militarev. *Institute for Oriental and Classical Studies, RSUH, Moscow.*

Russia's leading expert on Semitic and Afro-Asiatic languages and linguistics. Numerous works in these fields, including issues of linguistic and cultural reconstruction, classification, lexicostatistics, areal connections, as well as interdisciplinary integration of data accumulated through linguistic reconstruction. Co-author of the first two volumes of «Semitic Etymological Dictionary» (together with L. Kogan). **Official EHL participant from 2001 to 2010**; beyond that point, provides consultation services and occasional collaboration on various issues related to Afro-Asiatic linguistics.

Olga Stolbova. *Institute of Oriental Studies, Russian Academy of Sciences, Moscow.*

Expert on Chadic and general Afro-Asiatic languages and linguistics. Works on historical reconstruction of Proto-Chadic and its external (genetic and areal) connections. Author of the on-going monograph series «Chadic lexical database». **Official EHL participant from 2001 to 2010**; beyond that point, provides consultation services and occasional collaboration on various issues related to Afro-Asiatic linguistics in general and its Chadic subbranch in particular.

Vladimir Dybo. *Center of Comparative Studies, RSUH, Moscow*; *Institute of Slavic Studies.*

Russia's leading expert on comparative Indo-European and Slavic studies. Numerous works on Indo-European reconstruction, Slavic philology and dialectology, typology of prosodic systems in languages of the world, Nostratic linguistics, theory and methodology of short-range and long-range comparison. **Official EHL participant from 2001 to 2005**; beyond that point, provides occasional consultations on issues related to Indo-European linguistics.

Václav Blažek. *Masaryk University, Brno*.

Expert on Indo-European, Afro-Asiatic, Nilo-Saharan linguistics. Multiple works on historical linguistics, focusing primarily on the production of comparative lexical lists and lexicostatistics. Provides frequent consultation services for EHL members; has collaborated with some EHL participants (e. g., Alexander Militarev) on a number of projects, usually concerning the internal and external relations of Afro-Asiatic languages.

Valentin Vydrin. *Institut national des langues et civilisations orientales, Paris*.

Expert on African linguistics, in particular the Mande language family. Works on synchronic description, various aspects of historical reconstruction, lexicostatistics and glottochronology of the Mande languages. Provides consultation services and collaborates with EHL participants on issues of Niger-Congo linguistic classification and handling of African linguistic data in general.

Merritt Ruhlen. *Department of Anthropological Sciences, Stanford University*.

Author of numerous studies on theory and methodology of long-range comparison, historical linguistics of Native American languages, specific hypotheses of distant genetic relationship (Amerind, Dene-Caucasian, etc.). **Official EHL participant from 2001 to 2006**; continues to collaborate with EHL participants on multiple issues.

III. General description of EHL's goals and major lines of research

(***slightly modified from the original description here***: http://ehl.santafe.edu/intro1.htm)

There are currently more than 6000 languages on our planet, some spoken by millions, some by only a few dozen people. The primary goal of EHL, an international linguistic project launched in 2001, 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, it would be natural to suppose (although 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 «superfamilies» and so on, as far as one can penetrate. The methodology for such a classification is rooted in the traditional field of comparative-historical linguistics and has recently been supplemented with statistical and cladistic methods, borrowed from other branches of natural and anthropological science.

Most existing classifications, however, do not look behind some 150-200 language families that are relatively easy to discern. This restriction has its 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 are usually expected to lose almost all superficially common features after some 6,000-7,000 years of independent development.

Nevertheless, despite widespread skepticism and reluctance to tackle the problem, there is a number of scholars who believe that these obstacles are not altogether insurmountable. Certain lines of research, emerging in the middle of the 20th century, have appeared to indicate that **geographically larger and chronologically deeper genetic groupings are not only possible, but indeed quite plausible**. It can be shown, quite realistically, that the majority of the world's language families can be classified into roughly a dozen «extra-large» groupings, or macro-families. 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 to formally reconstruct unattested language stages, so-called *proto-languages*, based on observing and describing the laws of linguistic change that bind together 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. As an example, modern English, Finnish, and Turkish may have very little in common (and what little there is, is practically indistinguishable from chance), but their respective reconstructed ancestors – Proto-Indo-European, Proto-Uralic and Proto-Altaic – appear to have more common traits and common vocabulary. This means that it is possible, in theory and on practice, to extend the time perspective and reconstruct even earlier stages of human language.
>
> 2) Where a detailed reconstruction of the proto-language is impossible to achieve (for example, due to insufficient data) or requires more time and effort than can be spared, 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. According to EHL research, out of all known 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 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:

(A) 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 more than a decade, 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 the stage of preparation.

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

(B) Lexicostatistical testing of both traditional and new theories of language relationship.

In the fall of 2011, so as to emphasize the tremendous importance of lexicostatistics in determining the proper historical relations between languages, EHL has launched, as one of its sub-projects, the construction of the **Global Lexicostatistical Database** (GLD). The GLD intends to eventually host properly assembled and annotated Swadesh wordlists for the majority of the world's languages, as well as for proto-languages, reconstructed on various chronological levels, based on rigorous methodological procedures.

(C) 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 managed to take into account all of the necessary factors responsible for historical evolution, making «manual» handling of the data irreplaceable. Nevertheless, EHL stll 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.

Apart from 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.

Brief historical note:
The Evolution of Human Language project was originally founded in 2001, due to 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.

IV. Up-to-date results / achievements of EHL research

A reasonably detailed account of EHL's then-current perspective on the linguistic prehistory of the world was published in 2009 by three of the project's principal contributors (*Murray Gell-Mann, Ilia Peiros, George Starostin*. Distant Language Relationship: The Current Perspective. Journal of Language Relationship, v. 1, pp. 13-30). The paper briefly introduces some of the crucial methodological aspects of comparative-historical linguistics in general and long-range comparison in particular, and then proceeds to list the major hypotheses on macrofamily allocation of the languages of Eurasia.

Since then, significant process has been achieved not only in refining the linguistic classification of the languages of Eurasia (as described in Gell-Mann, Peiros, Starostin 2009), but also in expanding EHL's vision to other continents, primarily Africa and Native America, and, to a more limited extent, New Guinea and Australia. Below we will attempt to concisely summarize the principal «working hypotheses» on the deep-level classification of the world's languages, as per the overall state of EHL research circa mid-2013.

Methodologically, current EHL classification models tend to rely on comparative lexical lists that contain 50 words with – on average – the most stable «Swadesh meanings» (*ashes, bird, black, blood, bone, claw/nail/, die, dog, drink, dry, ear, eat, egg, eye, fire, foot, hair, hand, head, hear, heart, horn, I, kill, leaf, louse, meat, moon, mouth, name, new, night, nose, not, one, rain, smoke, star, stone, sun, tail, thou, tongue, tooth, tree, two, water, we, what, who*), reconstructed for a variety of low-level (not older than 2,000-3,000 years) language groups. Deeper level genealogical trees are then constructed, based on lexicostatistical analysis of the material. Datings are assigned based on the glottochronological method, with important modifications introduced by S. Starostin in the 1980s.

The trees vary in quality, depending on the selected method of «cognate scoring» (etymological analysis; phonetic similarity; manual vs. automatic procedures, etc.). Usually, tree nodes that are glottochronologically dated beyond 10,000-12,000 BP should be deemed unreliable on their own. Long-range hypotheses that go beyond this time point, yet are still cautiously endorsed by EHL, are based on additional data from comparative wordlists, but, since the methodology for historical analysis of such wordlists is not particularly robust, should be viewed as highly speculative and preliminary.

A. **Eurasia**.

(1) Lexical and morphological evidence for the **Nostratic** hypothesis continues to grow. This macrofamily is currently regarded as consisting of a tripartite «core»:

(a) *Indo-European* languages;
(b) *Uralic* languages;
(c) *Altaic* languages,
their separation tentatively dated to 10,000 – 12,000 BP.

Reliable evidence connecting these three families with two other previously suggested branches of Nostratic — Dravidian and Kartvelian languages — is much more scarce. Direct comparison of basic lexics reveals only occasional points of intersection; however, the etymological method, to some extent, confirms these connections. It has been suggested by some EHL participants to reserve the old term «Nostratic» for the «core» grouping and use J. Greenberg's term **Eurasiatic** for the larger grouping that would include:

(a) Nostratic (= Indo-European, Uralic, Altaic);
(b) Dravidian;
(c) Kartvelian,
their separation tentatively dated to 12,000 – 14,000 BP.

The issue of the total number of separate branches in the Nostratic / Euroasiatic macrofamily remains open. There is significant evidence that the Eskimo-Aleut languages also belong to «core» Nostratic, as well as impressive lexical isoglosses between «core» Nostratic and Chukchee-Kamchatkan languages. However, Chukchee-Kamchatkan languages also show non-accidental similarities to other, decidedly non-Nostratic, languages both in the Old World (Nivkh) and the New World (Algic), and it has been so far impossible to reach a proper consensus among EHL members as to their classification. It is quite likely that Chukchee-Kamchatkan languages are either an old Nostratic offshoot with a heavy «Beringian» substrate (see below), or, vice versa, a member of the «Beringian» family that has been significantly influenced by Nostratic languages (such as Eskimo-Aleut), arriving in Northeast Eurasia at a later date.

Comparison of basic lexicon also shows that the Yukaghir languages probably belong to the Nostratic «core», confirming the old hypothesis of a particular proximity between Yukaghir and the Uralic languages, better explained in genetic than areal terms. However, some of the EHL-affiliated specialists prefer to affiliate Yukaghir with «Beringian» languages (Nivkh, Algic, and possibly Chukchee-Kamchatkan), so the issue remains controversial even within EHL.

(2) In contrast, the constituency and internal classification of the **Dene-Caucasian** («Sino-Caucasian», according to S. Starostin's original terminology) macrofamily seems to have been established quite firmly. This macrofamily tentatively consists of the following branches:

I. «West Dene-Caucasian»:
    I.A. Vasco-Caucasian:
        (a) Basque;
        (b) North Caucasian.
    I.B. Burusho-Yeniseian:
        (a) Burushaski;
        (b) Yeniseian.
II. «East Dene-Caucasian»:
    (a) Sino-Tibetan;
    (b) Na-Dene.

The approximate age of Dene-Caucasian is comparable to that of «Narrow Nostratic», i. e. estimated around 10,000 - 12,000 BP.

The Dene-Caucasian hypothesis is partially compatible with the hypothesis of a Dene-Yeniseian relationship, originally proposed by M. Ruhlen and recently revived and popularized by E. Vajda. However, EHL members generally agree that the proximity between Na-Dene and Yeniseian languages has been somewhat exaggerated by Vajda, and that both families should rather be regarded in a broader Dene-Caucasian context, where the closest relative of Yeniseian is the Burushaski language, and the closest relative of Na-Dene are the Sino-Tibetan languages.

The Na-Dene family itself, according to old models of classification (E. Sapir, etc.), consists of three branches: the large Eyak-Athapaskan family and two remote outliers — Tlingit and the extinct Haida. Most specialists in these languages today, such as M. Krauss and J. Leer, accept the relationship between Eyak-Athapaskan and Tlingit, but not between either of them and Haida. Lexicostatistical analysis of Na-Dene data corroborates that position: there is too little evidence to regard Haida as a certified member of this family. However, a broader «Dene-Caucasian» affiliation for Haida remains an open possibility, to be explored further.

(3) The **Afro-Asiatic** macrofamily remains largely unchallenged. Analysis of 50-item wordlists by G. Starostin shows that, much like Euroasiatic, Afro-Asiatic should be viewed as consisting of a «core» of four families (Semitic, Berber, Egyptian, Chadic) that disintegrated around 8,000-10,000 BP, and a «periphery» that almost certainly includes Cushitic languages and, with much less certainty, the Omotic languages of Ethiopia. This «broader» version of Afro-Asiatic, like the broader Eurasiatic, is tentatively dated to around 12,000-10,000 BP.

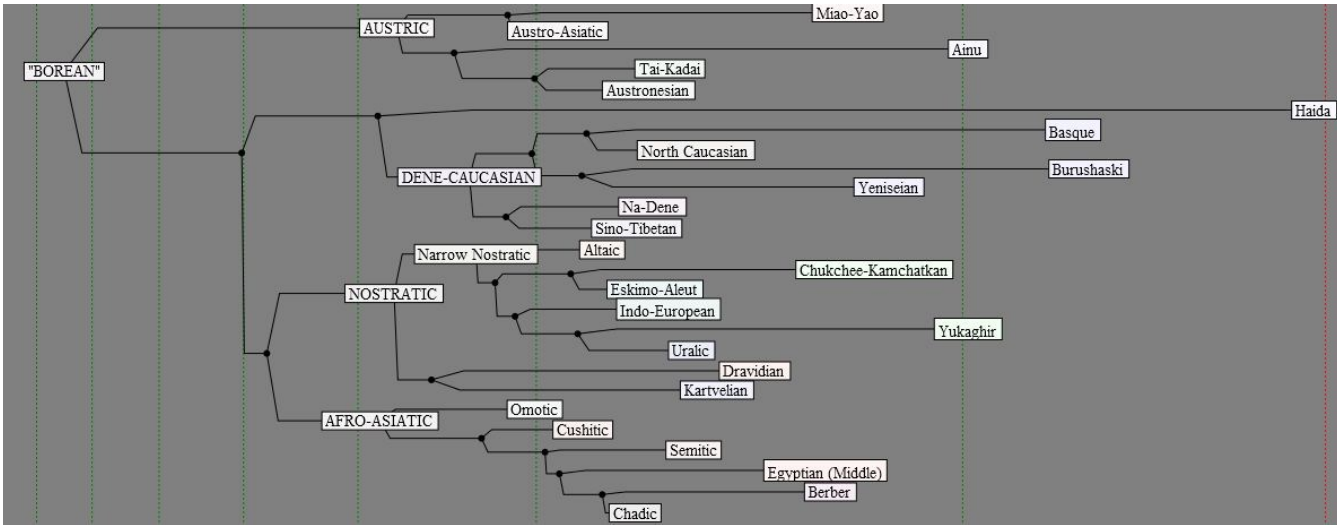
(4) Status of the fourth hypothetical macrofamily, **Austric**, remains somewhat unclear. According to 50-item wordlist lexicostatistics, its four potential branches hint very strongly at two binary groupings, each of them disintegrating around 9,000-10,000 BP:

(a) Austronesian + Tai-Kadai (= «Austro-Tai»);
(b) Austro-Asiatic + Hmong-Mien /= Miao-Yao/ (= «Austro-Miao»).

Lexicostatistical evidence for a nearest common ancestor for both of these groupings, however, remains very scarce, and the hypothesis is visually better supported by a set of potential etymological parallels, assembled by S. Starostin and I. Peiros. It remains unclear to what degree these parallels are truly reliable. If it is confirmed that the currently attested similarities are non-accidental, the Austric macrofamily would still be extremely deep chronologically, comparable to «broad» Eurasiatic or even older (e. g. disintegrating not later than 14,000 BP).

It has been convincingly shown by EHL associates V. Blažek and J. Bengtson that, out of several competing hypotheses that try to suggest different genetic affiliations for the isolated **Ainu** language of Japan, the Austric connection is the likeliest one. This is somewhat, though not definitively, corroborated by lexicostatistical analysis, which suggests that Ainu could be a separate branch of Austric, somewhere «in between» Austro-Tai and Austro-Miao. However, Ainu also shows a distinct layer of «Siberian» lexics, particularly lexical isoglosses with Nivkh; due to this, some EHL associates (O. Mudrak) prefer to group this language together with Nivkh, Yukaghir, and Chukchee-Kamchatkan (see above).

The overall classificatory picture for the majority of linguistic families of Eurasia currently looks as follows (see below for further comments on the upper-level «Borean» node):

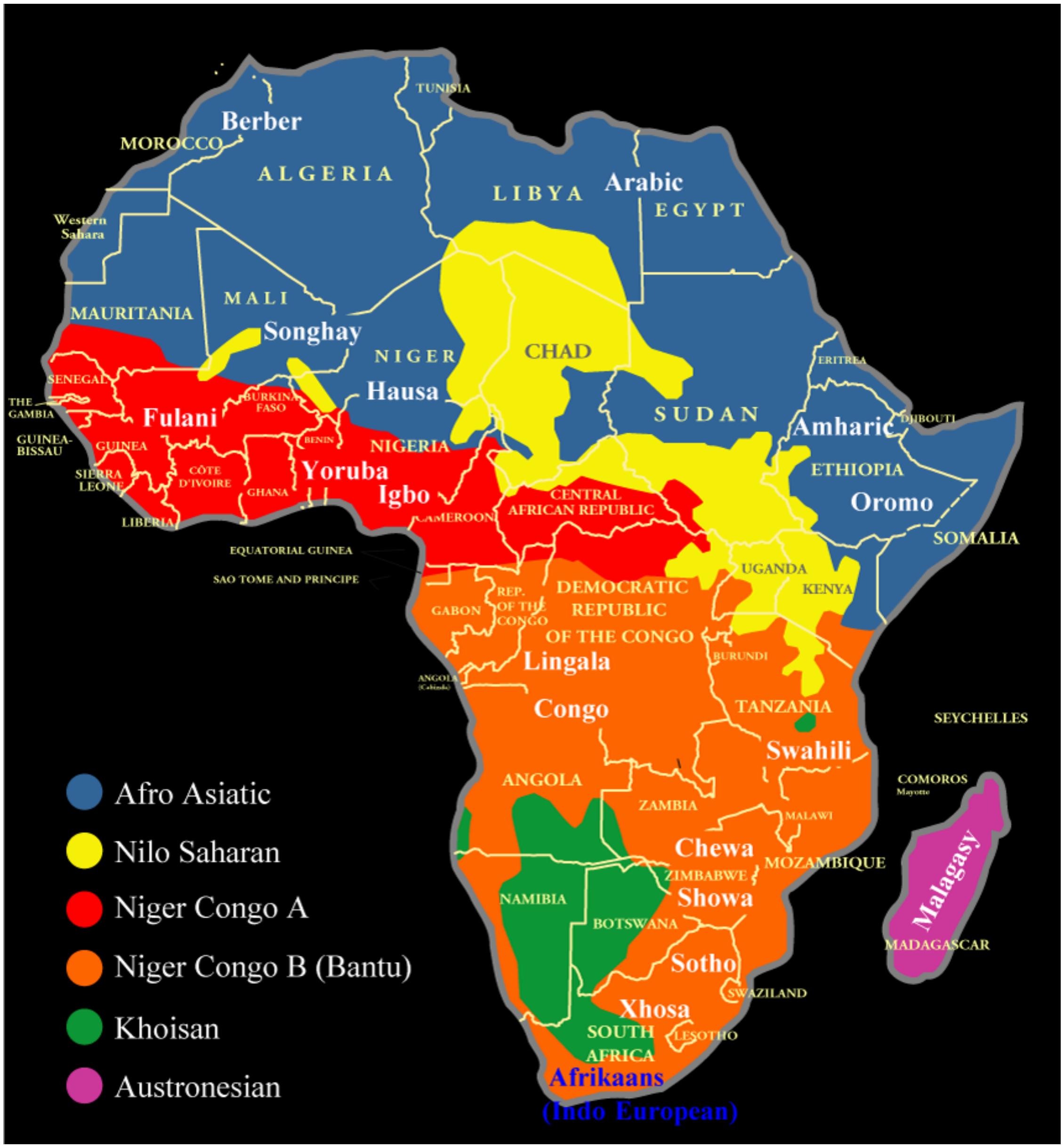


B. Africa.

The «standard model» of linguistic classification for Africa remains essentially based on seminal work by Josef Greenberg, who had grouped nearly all of the continent's languages into four macro-families: **Afro-Asiatic**, **Nilo-Saharan**, **Niger-Kordofanian** (= **Niger-Congo**), and **Khoisan**. This classification was based almost exclusively on the unreliable methodology of «mass comparison», and requires a thorough revision, which is currently carried out by G. Starostin, with help from several EHL associates. Research carried out from 2001 to 2013 has allowed to formulate the following conclusions (some of them still liable to further amendments):

(1) The small **Khoisan** (formerly = **Bushman-Hottentot**) macrofamily of so-called «click languages» is reliably reinterpreted as a typologically similar agglomeration of two distinct families: **Peripheral Khoisan** (= **Juu-Taa**, consisting of a number of small Bushman languages) and **Khoe-Kwadi-Sandawe**, which, among others, includes all «Hottentot» languages (such as Nama) and, as a very distant outlier, the Tanzanian isolate of Sandawe. Disintegration of both of these families, according to preliminary lexicostatistics, must have taken place around 8,000 - 10,000 BP.

The issue of whether Peripheral Khoisan and Khoe-Kwadi-Sandawe can be considered nearest closest relatives remains open: lexicostatistical evidence for such a connection is exceedingly small (limited to just a few pronouns and occasional lexical similarities), and etymological evidence is, in most cases, very difficult to distinguish from traces of later areal contacts. Additionally, Khoe-Kwadi-Sandawe shows certain transparent typological and morphemic connections to languages of Central and Northern Africa, most often to Bantu (Niger-Congo) and Chadic (Afro-Asiatic). It may be speculated that the relation of these languages to Peripheral Khoisan has been obscured by heave «adstrate» influence on the part of archaic migrations from the North of Africa by people speaking Afro-Asiatic languages.



Greenberg's «standard model» for Africa.

Additionally, yet another isolate of Tanzania, the **Hadza** language, traditionally grouped with Khoisan because of the heavy presence of «click» phonemes in its sound inventory, has been definitively shown not to be Khoisan — the amount of similarities that Hadza shares with Afro-Asiatic languages is at least equal to, and possibly exceeds similarities that it shares with different varieties of «Khoisan».

(2) The much larger **Nilo-Saharan** language family has been subjected to the full procedure of «preliminary lexicostatistics», also including some elements of etymological analysis. EHL's current views on this taxon are as follows:

(a) The two major parts of the «core» of Nilo-Saharan are the (almost universally recognized, but with various models of internal classification) **East Sudanic** language family (which includes such well-known branches as Nilotic and Nubian, as well as Surmic, Daju, Tama, etc.) and the **Central Sudanic** language family (including Moru-Ma'di, Sara-Bongo-Bagirmi and other languages). Disintegration of both of these is tentatively dated to around 8,000-6,000 BP.

(b) Further lexicostatistical comparison shows that a genetic relationship between East and Central Sudanic is quite likely, based on numerous similarities in their basic lexicon that show signs of pattern behaviour (although a proper set of regular correspondences and sound laws has not yet been assembled). Verification of this hypothesis runs into the problem of serious typological discrepancies between the basic root shapes in East and Central Sudanic, possibly due to heavy substrate influence on the latter by certain extinct languages of Central Africa (pigmies?).
Nevertheless, if the hypothesis is verified, genetic relationship between East and Central Sudanic, tentatively dated to around 12,000BP, would already confirm the bulk of Greenberg's Nilo-Saharan hypothesis, since, in between them, these two large families cover about 80% of the attested «Nilo-Saharan» languages.

(c) Lexicostatistical or proper etymological confirmation for a «Nilo-Saharan» affiliation of such African groups / families, as Saharan (including Kanuri and Tubu), Fur, Koman, Kuliak, and Songhay, has not been reached so far. Since most of these taxa are relatively small, and their ancestral states can only be reconstructed, on the average, to a time period of 2,000-3,000 years, it is possible that their genetic status has been obscured by significant accumulations of phonetic or lexical change. In this respect, it is not likely that much progress will be made on their «classificatory accomodation» before other, more global, issues of African classification have been settled, such as the nature of relationship between East and Central Sudanic.

(3) The huge **Niger-Congo** family, consisting of altogether more than a thousand languages, has not yet been properly elaborated by the EHL project. However, some preliminary work on Niger-Congo lexicostatistics seems to yield results that are in general agreement with the present day Africanist consensus: namely, that Niger-Congo as a whole is a certified genetic reality, but some of the languages and groups, originally included in that family by Greenberg, may have been erroneously included in this macrofamily.
The «core» of Niger-Congo includes the entire huge **Benue-Congo** family (including Bantu), as well as such smaller families in West Africa as **Mande**, **Kru**, **Dogon**, and, most likely, *parts* of the linguistic units usually known as **Atlantic**, **Gur**, **Adamawa**, **Ubangi**, and **Kwa**. Other parts of these groups, however, may have been integrated in them by mistake, due to millennia of areal convergence. These complicated issues remain to be investigated further.
Altogether, «core Niger-Congo» may be very tentatively dated to a period of 12,000-10,000 BP (e. g. approximately the same time as a hypothetical «core Nilo-Saharan», consisting of East and Central Sudanic).

(4) On **Afro-Asiatic** languages, the only large African macrofamily that is also represented outside of Africa (with its Semitic branch), see above, in the section on Eurasia.

C. America.

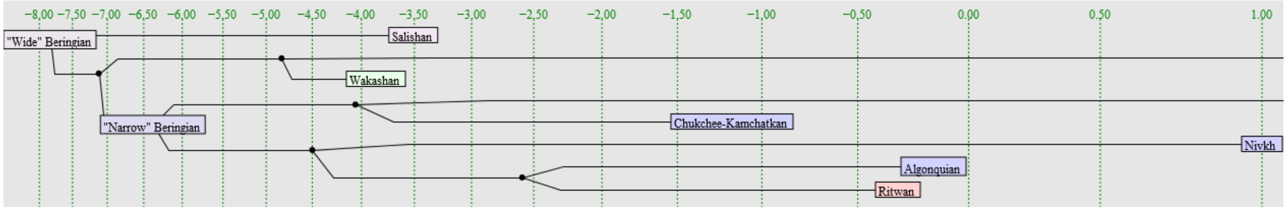
One of EHL's major goals is to verify J. Greenberg's highly controversial hypothesis that the absolute majority of Native American languages go back to a single common ancestor, «Proto-Amerind» (*Language in the Americas*, 1987). Most EHL experts agree that Greenberg's «exceptions» to this hypothesis, the Na-Dene languages and the Eskimo-Aleut languages, do indeed have their closest relatives in the Old rather than the New World (Na-Dene languages are grouped with the rest of Dene-Caucasian, and Eskimo-Aleut are most likely to be grouped with Eurasiatic, see above). However, there is, as of yet, no consensus on the status of the much more complex «Amerind» hypothesis.

To a certain extent, EHL members stand in opposition to «mainstream» specialists on historical linguistics in the Americas (such as Lyle Campbell, William Poser, and others) in that they are willing to accept the «Amerind» hypothesis as a serious proposal, backed by lexical and grammatical evidence that is currently insufficient to prove the hypothesis «beyond a reasonable doubt», but justifies further research on the issue. A proper «disproval» of the Amerind hypothesis, on the other hand, would consist of showing that one or more sub-sections of «Amerind» share their nearest common ancestor not with other sub-sections of «Amerind», but with certain linguistic groupings of the Old World — which is also a possibility.

As of 2013, EHL members working on «Amerind» languages have assembled representative basic lexicon wordlists for almost all of the minor groupings included in Greenberg's classification, but proper lexicostatistical and etymological analysis has been so far extended only to a fraction of these wordlists. Primary attention was paid to languages of North and Central America, since these are usually better described, and the phonetic complexity of their root shapes, on the whole, indicates a generally more conservative historical behavior than in the case of their South American counterparts.

Two major hypotheses are currently being worked upon:

(1) S. Nikolayev is exploring the hypothesis of a «Beringian» language family, according to which three major groupings of North America (Algic, Salishan, and Wakashan) are genetically related to some of the «Siberian» languages, such as Nivkh and Chukchee-Kamchatkan. The novelty of this idea is in combining Algic / Salishan / Wakashan (formerly known as «Almosan», one of J. Greenberg's terms) with certain languages of the Old World in one grouping.



«Almosan» / «Beringian» relations according to Nikolayev's research, mid-2013.

However, it should also be remembered that Chukchee-Kamchatkan languages at least show very strong ties with Eurasiatic languages (see above), and that, accordingly, it has yet to be understood which of the «relationship signals» — from the Old World or the New World — for these languages is the stronger one, as the two classificatory hypotheses are mutually exclusive. As of now, EHL consensus on Nikolayev's hypothesis has not yet been established, although the quantity and quality of evidence on the connection between *Algic* and *Nivkh* (some of it has also been collected by O. Mudrak) seems quite impressive.

(2) I. Peiros has arrived at a positive relationship signal for several other families of North and Central America, suggesting a nearest common ancestor between **Mayan**, **Mixe-Zoque**, **Uto-Aztecan**, **Hokan**, and possibly several other small groupings as well. It is not excluded that a few of the South American groups belong in this taxon as well; most notably, Quechua languages seem to share a significant amount of potential cognates with some of the listed families. The taxon, which may be provisionally called «West Amerind», suggests an original date of disintegration circa 12,000-10,000 BP.

Neither of the two hypotheses explicitly disagrees with Greenberg's «Amerind» proposal (at most, Nikolayev's «Beringian» slightly expands «Amerind» so as to include Nivkh and Chukchee-Kamchatkan as well). If both turn out to be completely or even partially correct, the age of a super-hypothetical «Amerind» would highly likely predate the commonly accepted date of the initial settlement of the Americas (≈ 16,000 BP) or, at least, coincide with this settlement. At present, however, any discussion on «Amerind» as a whole can only be speculative, since even the required preliminary research on its constituents is far from complete.

D. New Guinea and Australia.

By area and population, New Guinea is the most linguistically complicated and taxonomically diverse region of the world., with classifications ranging from 20-25 families (Ethnologue) to as many as 60 (Foley 1986).

It has long been asserted that a large number of New Guinean families, including all families south of the Papuan cordillera, belong to an ancient macrofamily called Trans New Guinea (McElhanon and Voorhoeve 1970). This is not utterly different from Greenberg's (1971: 853) "Nuclear New Guinea," though with significant differences in detail. Neither of these proposals were backed with credible evidence, excepting similarities in the forms of the personal pronouns and their uses. These resemblances are so strong that there can be no question that *some* «Trans New Guinea» exists, and that many New Guinean subfamilies are mandatorily members thereof. However, there are also families which would not be included by this heuristic, but have traditionally been included anyway due primarily to typological resemblances (Wurm 1982).

Recent work by Andrew Pawley (2005, 2012) purports to reconstruct the proto-Trans New Guinea lexicon. The fundamental weakness of Pawley's approach is the use of a small number of widely separated individual languages, ignoring those in between. This has as much chance of succeeding as a reconstuction of «Eurasiatic» based upon French, Armenian, Finnish and Mongolian. While giving lip service to the comparative method, there is no way to build a credible historical phonological framework from only these descendants, and most (not all) of Pawley's proposed cognates, like McElhanon and Voorhoeve's and Greenberg's, can be shown to be invalid based upon family-level analysis.

EHL member Timothy Usher's approach is to reconstruct lower-level families to the extent possible given the varying depth and quality of resouces; these can range from, in the worst case, dozens to up to a thousand or more reconstructions. The addition of terms beyond the basic last can establish phonological correspondences which aren't sufficiently exemplified in a Swadesh term list when the lexical resemblance is relatively low; this in turn can confirm (or deny) the cognacy of terms which are counted as matches for the purposes of glottochronology. This has resulted in the division of Trans New Guinea sensu stricto (i.e. those families which are manditorily members) into under a dozen mid-level easily-defendable yet previously unrecognized subgroups, encompassing [NUMBER] languages.

In the far northwest of New Guinea and on the neighboring islands of Halmahera and Yapen, languages of the West Papuan phylum are spoken (q.v. Donohue 2002). Usher has found no connections between these and any other New Guinean families, although with the caveat that less investigation has been conducted in these families in recent years precisely for this reason.

Resemblances in personal pronouns suggest a relationship between West Papuan and Andamanese. Superficially, lexical comparison between Great Andaman vocabulary, which has been extensively reconstructed by Usher, and Peiros' (at that time highly tentative) reconstructions of Mon Khmer looked more promising than between Great Andaman and subfamilies of Trans New Guinea. This would accord with recent findings in population genetics which show Andamanese Y chromosome lineages to group with those of of some Tibetan populations and Ainu (DM174), rather than with those of New Guinea and Australia. Blevins (2007) connects Little Andaman directly with Austronesian; this is certainly wrong, but may be worth revisiting at the Proto-Andaman and Proto-Austric level. Unfortunately, no one has been in a position to evaluate this, due to a lack of clarity in the very extensive colonial-era Andamanese data (Portman 1887, 1898, Man 1923) and inattention to Andamanese by students of Southeast Asian families. As Usher is currently focused on the bulk of mainland families which likely constitute their own group(s), it is recommended that the Andamanese reconstruction be considered by EHL scholars who are evaluating relationships in Southeast Asia and between Eurasian families.

Australia: It has long been generally, though not universally, accepted that all the languages of Australia excepting Tasmanian form a single family (Capell 1956). While the settlement of modern humans is ancient (40k+ b.p.), the languages are far more similar than such a date would suggest. Various explanations have been proposed for this, including a lower rate of change than found in other regions of the world (Dixon 1997). The more straightforward explanation is a replacement event. Preliminary lexicostatistics by Ilia Peiros indicate a distance between the Gunywinygan family of the Top End and several subfamilies of Pama-Nyungan of 7,500 b.p.

The larger question is, does Australian have a special relationship to any New Guinean families? The probable answer is yes, but this is very distant. Very preliminary lexicostatistics conducted by Timothy Usher and the late Sergei Starostin suggested a distance of c. 20k b.p. between Australian subfamilies and several New Guinean families. While hardly impressive, this contrasted with results approaching zero between these and the West Papuan family of North Halmahera.

E. «The Great Borean Bottleneck»?

EHL research is very explicitly NOT targeted at trying to reconstruct or even prove the historic reality of «the language of Adam and Eve», i. e. a single modern-type language that would be the common ancestor of all of the world's linguistic diversity. This refusal is based on two considerations:

> (a) there is no reason why «X = the single common ancestor of all known world languages» should necessarily be the equivalent of «Y = the first modern-type language in the world», since it is perfectly plausible that Y, whenever it was actually spoken, had other descendants, in addition to X, all of which had disappeared without a trace;
>
> (b) even if X = Y, the chronological distance between the world's deepest linguistic groupings can easily be too deep for historical linguistics to be able to do anything with it. For instance: *supposing* that [1] most of the world's languages are traced back to a «Globalese», spoken 20,000 years ago, [2] except for «Khoisan» languages, traced back to a «Proto-Khoisan», spoken 15,000 years ago, [3] «Adamese», the common ancestor of «Globalese» and «Proto-Khoisan» was spoken 40,000 years ago — all of which is quite possible in theory — there is most likely no way one could arrive at any definitive conclusions about «Adamese» whatsoever: the data would be too scarce and the chronological distance too great to distinguish any useful information from pure noise.

Consequently, the main priority of EHL is *to explore the limits of genetic classification*, which means arriving at a finite number of linguistic taxa (of widely varying sizes — most likely, including huge «macro-macro-families» along with several smaller ones and linguistic isolates) and showing that this number is truly «final» in that genetic relationship between these taxa is impossible to demonstrate by any known means.

On an essential level, this priority is not at all different from the usual goals of all comparative-historical linguistics: the difference between «bold» projects such as EHL and «conservative» research is in that EHL advocates a «relaxed» methodology that allows to formulate realistic, historically plausible, data-supported hypotheses of deep level relationship that go much further than the somewhat artificial limit of anything from 6,000 to 10,000 years, somewhat arbitrarily imposed by «conservative» schools of research.

By 2003, S. Starostin had already compiled a comparative database on the basis of protoforms, reconstructed for four hypothetical proto-languages of four macro-families: Proto-Nostratic, Proto-Sino-Caucasian, Proto-Afro-Asiatic, and Proto-Austric. The database, labeled «Long-range etymologies», implied that there is some evidence for an ultimate genetic relationship between all four of these macrofamilies, covering most of the territory of Eurasia and a large part of Northern Africa. The evidence was mostly circumstantial, based on phonetic similarities (rather than correspondences) between items reconstructed with widely varying levels of certainty, but the sheer amount of these similarities — more than a 1000 — suggested that the hypothesis deserved further exploration.

Informally, the hypothetical «proto-proto-language» was named «*Borean*» (a term borrowed from an earlier, only tangentially related, proposal by H. Fleming), and a tentative dating for this linguistic entity was set at around 16,000 BP (*terminus post quem*, since the «oldest daughters» of «Borean» are themselves dated to around 14,000 BP) to 20,000 BP (*terminus ante quem*, since the amount of phonetic-semantic similarities observed between descendants of Borean could hardly correlate with more than five or six thousand years of independent development of its four major branches).

Later on, S. Starostin compared the «Borean» data with various «Amerind» data (taken from J. Greenberg's and M. Ruhlen's comparisons), and still later, G. Starostin added in parallels from Niger-Congo (particularly Bantu), which led to a significant revision of the «Borean» hypothesis. The primary questions concerning this term — understood as a blanket expression for «a circa 20,000 year-old macro-macrofamily» — are now as follows:

— are the 1,000 series of comparanda, collected in S. Starostin's original database, truly indicative of a single nearest common ancestor for the four major macrofamilies of Eurasia? is the evidence that ties together these four macrofamilies *notably stronger* than the evidence that *also* ties in various forms of «Amerind», Niger-Congo, and Nilo-Saharan?

— is it true that some linguistic groupings in New Guinea and Australia show more similarities to «Borean» in their basic lexicon than others? if so, could this indicate that these groupings should also be included in «Borean»?

— is it possible to speak of a particular «Borean bottleneck», i. e. a relatively brief historical period during which speakers of «Proto-Borean» and / or its immediate linguistic descendants rapidly spread across a large part of the world, with «Borean» languages wiping out previously existing linguistic diversity in a process, analoguous to well-known «bottlenecks» in later history, such as the Indo-European expansion in Eurasia or the Bantu expansion in Central and Southern Africa? If so, what could be the reason for such a «bottleneck», and with what sort of genetic and archaeological evidence could it be correlated?

It is important to keep in mind that, although much of the evidence for «Borean» is flimsy and circumstantial, and many of the ideas are purely speculative, the «quest for Borean» is still a far better grounded, more plausible, and more modest enterprise than a search for «the root of human language», «Ursprache», «global etymologies», etc. The primary difference is that EHL does not operate with *random*, *isolated* words culled from different languages — it strictly observes the «bottoms up» principle of comparison that helps filter out linguistic innovations and promote linguistic archaisms. This makes the comparative data historically plausible on the whole, even if each individual comparison may still be criticized for specific reasons.

V. A concise list of actual problems and tasks for future resolution

[a] Methodology

In general, EHL is looking for ways to improve the reliability of comparative linguistic data, so that it would be easier and safer to work out hypotheses of deep level relationship between languages, which could then be proven beyond reasonable doubt. This implies research on the following issues:

[i] *Localizing the most «genetically stable» layers of language on the whole and of specific linguistic areas in particular*. In particular, this means research on the so-called «basic lexicon», with empirically based selection of meanings that are typically more stable and, therefore, more easily reconstructible.

[ii] *Introducing rigorous control factors in linguistic comparison*. This implies research on the typology of phonetic and semantic change, with empirically-based databases to be created that document both «typical» and «rare» types of such change — data that could later be used in order to assess the probability of particular reconstructions.

[iii] *Improving dating procedures*. Most of the absolute datings proposed for language splitting in EHL are based on S. Starostin's modified method of the original «Swadesh glottochronology», assuming a regular rate of lexical change. EHL plans to eventually integrate other methods as well, including character-based rather than distance-based methods that do not need to rely on the assumption of regular rates, as well as further modifications to classic glottochronology.

[b] **«Low-level» issues of classification**

In order to solve the grand problems of linguistic prehistory, one must first take care of small ones. Below we list only a few of the still unresolved issues — the ones that are of particular importance to the successful construction of an organized, unified linguistic classification at all levels of comparison:

Eurasia

— *Eurasiatic*: Finalize the inventory and internal classification of Eurasiatic. Do Eskimo-Aleut and Chukchee-Kamchatkan really form part of this family, or should we rather locate their closest relatives in the New World? How can we strengthen the case for Dravidian and Kartvelian as parts of Eurasiatic?

— *Dene-Caucasian*: Procure a more rigorous system of phonetic correspondences for Dene-Caucasian, particularly in the case of the «Eastern» branch, where neither Proto-Na-Dene nor Proto-Sino-Tibetan have yet been reconstructed in a fully satisfactory manner. Settle the matter with Haida (either exclude it completely, or find better evidence to integrate it somewhere within the macrofamily).

— *Afro-Asiatic*: Provide a new, improved database for Afro-Asiatic etymology with well worked out etymologies for such problematic branches as Chadic, Cushitic, and Omotic. It also remains to be determined if Omotic languages truly belong to Afro-Asiatic.

— *Austric*: Provide reconstructions for «Proto-Austro-Tai» and «Proto-Austro-Miao» in order to finally decide, on the basis of this binary comparison, if «Austric» is really one macro-family or two different ones.

America

— *General*: Finalize the lexicostatistical treatment and analysis of data on Native American languages, including the still poorly elaborated South American part.

— «*Beringian*» and «*West Amerind*»: In order to solidify evidence for these groupings, it is necessary to procure an efficient system of regular correspondences that would work on the basic lexicon of the protolanguages, reconstructed for various branches of these phyla.

Indo-Pacific / Australian

— *General*: Finalize a definitive list of low / mid-level taxa for the languages of New Guinea, with support from regular correspondences, protolanguage reconstructions, and lexicostatistics.

[c] **«High-level» issues of classification**

The most important task in this department is to provide a concise answer to the question, «*What exactly is ‵Borean′?*» In addition to fulfilling most of the tasks listed in section [b], answering this question also involves:

— supplementing S. Starostin's original database for «Borean» with data from non-Eurasian language families, such as Niger-Congo, «Amerind», Australian, etc.;

— thinking of how to implement certain additional standards of phonetic and semantic control in order to minimize the impact of accidental noise on the comparisons;

— conducting a full statistical evaluation of the evidence.

Given the overall rate of EHL progress in recent years, it is difficult to predict the exact moment when this question can be answered. However, the answer as such is not a momentary action; most likely, evidence for «Borean» will be increasing and modifying progressively, along with accumulation of data on everything else.

VI. EHL resources and links

The main on-line resources for EHL-related activity are as follows:

http://ehl.santafe.edu — An introductory page for the project; brief description of the EHL mission, main participants, areas of research, info on meetings, etc. Hosted at the Santa Fe Institute. Cross-linked to EHL linguistic resources.

http://starling.rinet.ru — «The Tower of Babel»: this site hosts all of the etymological databases for various low- and high-level language families, as well as the STARLING software, designed by S. Starostin, that serves as the primary tool for creating and analysing said databases.

http://starling.rinet.ru/new100 — «The Global Lexicostatistical Database»: a sub-site of the «Tower of Babel» that hosts carefully compiled and annotated Swadesh wordlists for languages all across the world, with a small analytical apparatus that allows to calculate percentages of matches, build classificatory trees, and analyze degrees of phonetic similarity between various languages on-line.