Murray Gell-Mann
(Santa Fe Institute)

Ilia Peiros
(Santa Fe Institute)

George Starostin
(Russian State University for the Humanities)

Distant Language Relationship: The Current Perspective

The idea that it may be possible to use data from modern languages in order to trace their origins back ten thousand years or even further, let alone even partially reconstruct the shape of their remote ancestors, remains a subject of discussion. Our article, partially based on results of research conducted by a joint team of Russian and American scholars within the *Evolution of Human Languages* project, discusses the methodological basis of this approach and the various ways of overcoming the most common obstacles associated with it. In the second part of the article we summarize the main results of the project and suggest further specific directions for exploring the question of deep-level linguistic relationships and the prehistory of human languages.

The study of genetic relationships among the world's languages has, for the last two hundred years, steadily remained one of the most fascinating and productive areas of research in linguistics. Yet, despite great advances achieved in the field during that time, even more remains to be done, due primarily to the immense quantity of the material, widely varying degrees of our knowledge about different linguistic areas of the world, and a lack of proper coordination of efforts between specialists.

This is why, in an attempt to overcome at least some of these problems, *Evolution of Human Languages* (*EHL*), an international program of the Santa Fe Institute, was launched in 2001. The primary goal of the present paper, then, is to outline what we consider the major achievements of the *EHL* program, along with a brief description of its methodological basis and various problems encountered along the way.

1. *General background*.

One of the most important concerns of *EHL* is that all the work conducted within the project should rest on a properly defined and well-grounded foundation, one that would acknowledge and respect all the achievements of traditional comparative historical linguistics, yet at the same time give clear instructions as to how these achievements should be used in further research on deeper level relationships between languages — research that is, for the most part, still in its early stages and the standards for which have not yet been carved in stone.

This means that it is perhaps reasonable to begin our article with a very brief summary of some of the axioms of comparative linguistics (even though, no doubt, all of them will be familiar to the majority of our readers), after which we shall add to those some of the (usually) less emphasized methodological assumptions as they are shared by *EHL* participants.

It is commonly recognized today that all languages are subject to historical change; regardless of the environmental or social context, no human language can be completely immune to this process. One inevitable type of change occurs when languages are passed down from parents to children: this includes modifications within the phonological system, shifts of meanings for various morphemes, loss of certain morphological and lexical oppositions and emergence of new ones, &c. This process of "vertical" transmission is in stark contrast with the "horizontal" process of borrowing features from other languages. Since both types of changes can occur independently in different speech communities using a particular language, that language may eventually split into several different ones, each being a direct descendant of the original one (the "proto-language").

Given that the splitting process frequently leads to the appearance of an entire family of related languages in the place of just one descendant, this process is best illustrated by a tree-type representation[1].

[1] Two major alternative models should be mentioned: the "wave theory" and "language convergence". The "wave theory" talks mainly about possible causes of language changes (wave-like spread of innovations) and thus

Each family tree consists of a number of nodes connected by arrows. Nodes represent (proto-)languages, while arrows are used to represent relations between the nodes. Therefore, while studying any particular language family, we need to address two different questions: "which nodes will we identify? (grouping)" and "how are they linked to each other in the tree? (branching)".

With the exception of a tiny handful of ancient languages attested in writing, modern languages are our only source of information on previous stages of language development. Reconstruction of these stages based on a scientific study of modern languages is the main task of comparative linguistics.

The major principles of comparative linguistics, after half a century of preliminary research, were explicitly and systematically formulated by the Neogrammarian school in the second half of the XIXth century on the factual basis of the Indo-European language family. Today, despite occasional criticism from the theoretical side, they continue to remain as the backbone of this field of study. For the most part, these principles concern reconstruction of proto-languages through a comparison of their descendants that necessarily involves discovery and description of systematic patterns of language change. Primary among these is the presence of regular (recurrent) phonetic correspondences between compared lexical items. Over the last century or so, the comparative method has been successfully applied to a whole set of families whose languages vary widely in phonological and grammatical structure, area of dispersal, lifestyle of the speaker community, and cultural background. We therefore accept the universal nature of the comparative method[2].

2. *Language families*: *the three types*.

Expanding beyond general textbook information on basic premises of comparative linguistics, we may classify all well-established and hypothetical linguistic families[3] into three types, depending on the level of transparency of relationships exhibited by their daughter-languages. This classification has both a practical / intuitive origin and a more formalized one, and we believe it to be of serious use for macrocomparative purposes.

In practical terms, type A would be associated with compact, well-defined families (which we may simply call "groups") like Germanic, Polynesian, or Turkic. As a rule, it is intuitively evident for specialists and native speakers alike that such families have arisen not very long ago, and it is usually not very difficult to obtain a proto-language reconstruction for each of them in strict accordance with the comparative method. These reconstructions normally constitute the most reliable results in historical linguistics.

Type B families (below called "stocks") consist of languages whose relationship is not intuitively evident for speakers, so that special research is needed to demonstrate it. Nevertheless, a strong case often can be made for genetic relationship even without resorting to strict phonological reconstruction, since the languages still preserve a large number of the common forms and features of their proto-language. The best-known case of such a stock (usually called "family") is Indo-European; other examples would include, for instance, Turkic, Dravidian, or Uto-Aztecan. The number of reliable reconstructions for this type of family is significantly smaller than for those of type A.

As a rule, the best results at this level are achieved not through direct comparison of modern languages, but in two steps: (i) reconstructing proto-languages for all the type A groups within the larger family and (ii) using these reconstructed "intermediate" proto-languages as the base material for reconstructing the proto-language of the larger family, thus increasing the level of transparency[4].

does not contradict the "tree model". The language convergence model claims that new languages can arise through convergence of formerly unrelated idioms. In most known cases, "convergence" can be simply interpreted as massive-scale borrowing from language *A* to language *B*, without affecting its genetic affiliation. For a good discussion of the subject see, e. g., [Dimmendaal 1995].

[2] For a more detailed explanation of the comparative method and the way it can be applied to different language families any number of general handbooks on historical linguistics may be consulted. Some of the best include, in English, [Hock & Joseph 1996]; [Campbell 2004], in Russian, [Burlak & Starostin 2005].

[3] The term 'family' is extended here to signify any type of genetic group of languages, families of any age and internal diversity, superfamilies, and even super-super-families. For a detailed discussion, see [Peiros 1997].

[4] One notable exception from this procedure is Indo-European. Although the family (= stock) unquestionably belongs to Type B, its reconstruction was effectuated in the 19th century not on the basis of intermediate reconstructions of Proto-Slavonic, Proto-Germanic, Proto-Indo-Aryan, &c. but on the basis of historically attested ancient languages (Old Indian, Ancient Greek, Classical Latin, &c.); these were used as entities close to proto-language status for their respective branches.

To use an example from Dravidian languages, a successful direct comparison between, for instance, Kui *siṅg* and Toda *üȥ* ʻfive’ would hardly be imaginable; however, if the former is first traced back to Proto-Gondwan **say-* and the latter to Proto-South Dravidian **ay-* (with both *-ṅg* and *-ȥ* elements going back to old numeric suffixes), the comparison becomes easier to interpret; all that is needed is to verify the regularity of the correspondence “Gondwan **s-* : Proto-South Dravidian *Ø-*”, which is, indeed, regular [Burrow & Emeneau 1984: 245].

In addition, relying on intermediate proto-languages usually leads to a drastic increase in the number of suggested etymologies: there will always be more matches between Proto-Germanic and Proto-Slavic, for instance, than there will be for modern German and Russian, due to the inevitability of lexical loss in individual daughter languages. A relevant form may be lost in Russian, but retained in Polish and, say, Bulgarian and thus reconstructed for the proto-language.

Finally, Type C, for which the name “superfamily” or “macrofamily” is often reserved, represents a situation in which comparison between attested (usually modern) languages may yield so little evidence that any hypothesis of their being related risks falling within the limits of chance coincidence. With thousands of years of phonetic change obscuring potential regularity of correspondences, and, even more importantly, thousands of years of gradual semantic and lexical replacement drastically reducing the numbers of true cognates and exact semantic matches, credible reconstructions of proto-languages for superfamilies cannot usually be significantly advanced by comparison of modern languages. However, one may always try to increase the level of transparency by employing reconstructions of intermediate stages (A > B > C).

Thus, in a situation similar to that of the previous example, hardly anyone would propose to compare directly Hungarian *víz* ʻwater’ with Greek *hydōr* id., due to lack of obvious phonetic similarity. In addition, even if we made a wild guess, proposing that the correspondence between Hungarian *v-* and Greek *h-* is regular, we would not be able to support our proposition with a significant number of equally strong lexical matches between Hungarian and Greek.

However, if, through systematic reconstruction, the Hungarian form is first traced back to Proto-Uralic (Type B family) **wete*, and the Greek form to Proto-Indo-European (also type B) **wedōr*, the reconstructed forms, being much closer to each other phonetically, provide a strong incentive for further comparison. In addition, operating with Proto-Indo-European and Proto-Uralic rather than Hungarian and Greek allows us to demonstrate the regularity of this correspondence in a much more satisfactory manner; cf. such additional comparisons as PIE **u̯egh-* ʻto carry’ : PU **wiɣe* id., PIE **u̯isu̯-* ʻall’ : PU **weńće* id., PIE **u̯onk̂-* ʻcrooked’ : PU **waŋka* id., &c. (in these and other examples the Greek and/or Hungarian equivalents are either missing or show different reflexes of the original reconstructed phoneme **w-*).

The basic principle underlying this and many other such examples is as follows: if the overall number of cognates, their degree of phonetic similarity, and the degree of transparency of regular phonetic correspondences between proto-languages X and Y all show a significant increase compared to the same features of any given daughter language of X and any given daughter language of Y, this is a strong signal in favour of genetic relationship between X and Y.

This semi-intuitive understanding of different types of families is conveniently formalized on the basis of such methods as *lexicostatistics* and *glottochronology*. The former measures degrees of relationship between languages based on the relative number of cognates in their basic lexicon (most conveniently represented by the so-called “Swadesh wordlist” of 100 items); the latter carries this procedure even further by trying to calculate approximately the absolute age of separation between related languages[5]. Thus,

[5] A detailed discussion of glottochronology (frequently defined as the method of dating the age of language separation based on the assumption of a constant rate of change within the basic lexicon) lies beyond the scope of this paper, and acceptance or rejection of this method is generally irrelevant to the main points that the current article is trying to emphasize. It should, nevertheless, be stated that the actual formula that is being consistently applied within *EHL* is not the original one, first suggested by Morris Swadesh in the early 1950s and heavily criticized by many specialists in historical linguistics. Instead, we use a significantly revised version of it put forward by S. Starostin. The revised variant, in contrast with the Swadesh formula, works well on all historically attested cases (this has been demonstrated in [Starostin 2000]) and has so far yielded credible results for all analyzed families (for some of the figures, see, e. g., [Starostin 2003]). Further methodological improvements, such as introducing individual replacement rates for each word on the 100-wordlist ([Starostin 2007]), were gradually introduced by Starostin until his demise in 2005; today, the work is carried on by some of his disciples, e. g. [Vasilyev & Militarev 2008].

Several serious objections against the methods of lexicostatistics and glottochronology are worth a brief mention. One is based on misleading percentages of cognates obtained while comparing Icelandic with other Scandinavian laneach of the three family types can be roughly associated with (a) certain ranges of percentages of lexical matches between related (modern) languages and (b) a certain range of ages.

Type	Percentage of matches on the Swadesh 100-wordlist	Approximate age[6]
A	≈ 45%–95%	≈ 2,000–4,000 KYA[7]
B	≈ 10%–45%	≈ 4,000–7,000 KYA
C	less than 10%	older than 7,000 KYA

In reconstructing a proto-language we first start with attested languages, thus moving in the opposite direction from actual historic development. However, it is only in this way that we can preserve and account for all the relevant information maintained in recorded languages. Relatively simple reconstructions of type A families are used to advance to more ancient stages (type B); with a sufficient number of type B reconstructions at our disposal, we have the right to begin the study of an even deeper, Type C, relationship, where "B-reconstructions" perform the same function as modern languages for the first step of the research.

It is necessary to specify here what we mean by "reconstruction", since the term can be used differently in various kinds of historical research. An actual reconstruction consists of several steps:

(i) formulation of the initial relationship hypothesis. This is usually done on an intuitive basis through observing certain kinds of similarities between a number of languages — primarily lexical and morphological (if the languages in question *do* have morphology);

(ij) accumulating a representative list of similar-looking morphemes in the languages under investigation ("similar-looking" first and foremost presupposes straightforward phonetic similarity between the morphemes, but may also signify a preliminary, sketchy notion of patterns of correspondence). At this stage such a collection may include all sorts of comparisons: borrowings, chance resemblances, forms of common origin, even mistakenly transcribed data (particularly for languages where such data are the only available kind);

(iij) performing a reconstruction of the phonological system of the protolanguage by tracing recurrent patterns in the accumulated material; these might suggest rejecting some look-alikes or adding previously neglected comparisons;

(iv) reconstructing the morphemic inventory of the proto-language by sifting through the list of accumulated lookalikes and modifying it according to one's current understanding of the languages' historical phonologies;

(v) reconstructing (wherever possible) the grammar of the proto-language.

It must be stressed that the order of these stages cannot be changed: phonological correspondences can only be established on the basis of morphemic look-alikes, while reconstructing morphological structure is impossible without a list of reconstructed grammatical morphemes. The minimum amount of information about any given proto-language must necessarily include a reconstructed phonological system accompanied by a representative list of reconstructed morphemes (both lexical and grammatical). In other words, a list of *segmental units* presented for the proto-language is always more substantial evidence for its existence than any proposals on its typological structure, means of word formation, &c., which represent the final stage of research.

3. *Problems of linguistic grouping.*

Of the several hundred language families mentioned in general linguistic literature, some are well-studied, with credible proto-language reconstructions, while others may be rejected as errone-

guages [Bergsland & Vogt 1962]; this issue has been eliminated by S. Starostin [2000], who showed that such errors are usually generated by large amounts of borrowings in the wordlist, and that elimination of borrowings from the calculations leads to much better results. The idea of diversity of retention rates put forward by R. Blust on the basis of Austronesian, (see, for example, [Blust 2000]), has been demonstrated as unnecessary by one of the authors of the present article [Peiros 2000]. Also, Blust's observation that existing lexicostatistical classifications of Austronesian, such as [Dyen 1965], do not confirm the mainstream views on this family, is not particularly relevant, since so far, neither the lexicostatistical, nor the generally accepted classification of this family have yet received proper justification.

[6] We crudely estimate the margin of error of our datings at an order of about 10%.

[7] Thousands of years ago.

ous or unconvincing. A list of more or less accepted language families is given in *Ethnologue* [Gordon 2005], which can be used here as a temporary starting point for further discussion.

The number of families identified within Eurasia is not very large[8]; it includes mainly B-type families like Indo-European, Uralic, Dravidian, Sino-Tibetan, Tai-Kadai, &c. There are also a few smaller A-type families that require further comparative study. Some languages of Eurasia are treated as isolates, without any known genetic connection[9]. No C-type families in Eurasia are accepted by *Ethnologue*.

When it comes to classification of African languages, the same source gives us quite a different picture. Here it lists but four "families", *all* of them C-type: Khoisan, Afroasiatic, Niger-Congo, and Nilo-Saharan. The Khoisan family is the smallest and, through recent research, probably, slightly better understood than the others[10] (although its external connections are hard to define because of unique phonetic features such as the so-called 'click' sounds, making it hard to compare even the intermediate reconstructions with the other African taxa). There is also a large amount of information about the Afroasiatic superfamily, including lists of etymologies and major phonological correspondences, but much further work is needed to raise it to expected standards, especially within its Chadic, Cushitic, and Omotic branches. Both the Niger-Congo and Nilo-Saharan superfamilies were identified by Joseph Greenberg ([1966]) based on his "mass comparison" method (see below). Unlike Greenberg's hypotheses on "Amerind" (see below), his identification of African superfamilies has in large part been accepted, despite the lack of conclusive comparative results.

The Indo-Pacific region remains linguistically the worst known part of the world. Most of the languages spoken here (both Austronesian and non-Austronesian = Papuan) are not well-documented. For the Papuan languages it has so far only been possible to group them into a dozen families, ranging from small A-type ones to major C-type ones like the Trans-New Guinea superfamily, plus a significant number of isolates. None of these families has been extensively studied based on the comparative method. All the languages of Australia are grouped by *Ethnologue* in a single C-type super-family[11], even though this plausible suggestion cannot yet be firmly backed up by the comparative method either.

The most diverse situation, according to *Ethnologue*, is found in the Americas. Here it lists 79 groups[12], most of them formed by a small number of languages. Only 18 families (all of them A or B-types) have more than 15 languages each. For some of these larger families there are reconstructions (including etymological dictionaries).

Table 1. A summary of *Ethnologue* groupings

Regions	Number of languages in families				126
	large (40+)	medium (15–39)	small (2–15)	isolates	total
Eurasia	8 (B-type)	2 (B-type)	7 (A/B-types)	5	22
Africa	3 (C-type)	1 (C-type)	—	—	4
Indo-Pacific	3 (B/C types)	3 (B/C types)	8 (A-type)	6	20
Australia	1 (C-type)	—	—	—	1
Americas	8 (B-type)	10 (A/B type)	36 (A-type)	25	79

These groupings are based on a number of reasons, ranging from solid comparative evidence to affiliations based on little more than 'general' agreement rather than facts. This brings up the question: is there actually a way to evaluate various grouping proposals according to the same standard?

One common reproach directed towards traditional comparative linguistics is that it lacks an objective methodology of language classification[13]. With more and more linguistic stocks added to the

[8] Very similar classifications are found in *Encyclopedia Britannica* and other respected sources.

[9] The list of isolates will be longer if we add extinct languages such as Sumerian, Etruscan, and others.

[10] See Sands [1998], where the idea of all Khoisan languages with the exception of Hadza belonging to a single family is deemed realistic, and G. Starostin [2003] and [2008], offering lexical and lexicostatistical evidence for such a relationship.

[11] The extinct languages of Tasmania, however, must be classified differently ([Crowley & Dixon 1981]).

[12] Campbell [1997] gives even more groups and strongly rejects most attempts to search for deeper connections.

[13] The most commonly used method is that of 'shared innovations', but it has not been formalized enough to resolve numerous complicated cases of relationship, nor is it possible to use 'shared innovations' as a universal classificatory standard for all of the world's languages. A painfully obvious example of the limitations of this

list of families subject to comparative research, the problem of finding strict criteria for classification becomes even more immediate, as answers are required for such questions as 'which languages can be put in one family?' (absolute classification) and 'which groups form the family and how they are related to each other?' (relative classification).

One potential answer is to rely on a properly formalized and standardized application of lexicostatistics, already mentioned above. This solution is frequently chosen by specialists in African languages and, to a lesser extent, those in the Pacific region. It is also embraced by all members of the *EHL* project. However, application of lexicostatistics to a given language family cannot be fully successful without preliminary identification of cognates — meaning that lexicostatistical classifications can be of two types: (a) preliminary, based on limited sets of historical data and used only in the exploratory stages of research and (b) advanced, based on a long process of etymological research on compared languages.

4. *Etymological dictionaries*.

We have already pointed out that related languages always possess morphemes of common origin, inherited from their common ancestor. If no such morphemes are found, this may indicate that: (i) the languages studied are related on such a remote level that one cannot find those morphemes, or (ii) they are not related at all (even though there is no certain method to prove lack of relationship). The first criterion, then, is to use morphemes of potentially common origin as a tool to evaluate various grouping proposals. The logic is straightforward: if a grouping proposal is supported by a sufficient etymological list (EL) of morphemes sharing a common origin, we accept it; otherwise, we remain unconvinced.

One commonly asked but rarely answered question is how many actual etymologies we need to constitute a satisfactory EL. To show that the morphemes in question may truly share a common origin and not be just look-alikes, we need to establish regular phonetic correspondences between the languages. At least 300–400 proto-morphemes are needed, on average, to support and illustrate any given phonological reconstruction. This approximate number — reflecting the idea that each phoneme of our reconstructed proto-language has to be illustrated by a statistically significant number of examples — can be seen as the lower limit for our EL.

Ideally, however, for each proposed family we need to have a full etymological dictionary, containing all the morphemes that can be reconstructed for its proto-language. Etymological dictionaries of well-studied A-type families usually have about 2–3 thousand proto-morphemes, with reflexes found in at least two daughter languages. The size of etymological dictionaries for B and C-type families depends mainly on the level of transparency of the relations among their daughter families and the number of proto-morphemes known for each of them. Dictionaries of Indo-European, Altaic, Uralic, and several other well-studied B-type families have more than 2,000 entries; by induction we can assume that this number should be expected for any proposed family regardless of its age (although it depends on such a technical factor as the number of languages / branches used for reconstruction).

Apart from quantity, reliable ELs also have to meet certain qualitative requirements. These are based on the well-known (although approximate) division of the lexicon into 'basic' and 'cultural', as well as the empirically and theoretically backed assumption that 'basic lexicon' is generally less subject to processes of horizontal transmission than 'cultural lexicon'. Therefore, if two or more languages are related, they will always have etymologies that belong to the sphere of basic lexicon.

The process of phonological reconstruction is so well defined in comparative linguistics that if two linguists have the same systems of phonological correspondences and lexical data, they are usually expected to get the same proto-form (or, at least, be able to agree upon one). Semantic reconstruction is different: we still lack a formal procedure to evaluate proposed semantic connections. Historical dictionaries of an individual language, be it English, French or a less familiar language, contain many non-trivial semantic changes (e. g. English *clean* = German *klein* 'small') some of which would never be accepted by linguists unless recorded in literature; likewise, etymological dic-

method is the still unresolved issue of the tree of Indo-European languages. While there is little question about the consistency of its 'small', basic branches — Celtic, Germanic, Slavic, etc. — 'shared innovations' do not seem to work where larger sub-groupings of Indo-European are concerned (such as, e. g., the very hypothetical Italo-Celtic). Likewise, 'shared innovations' have not managed to resolve numerous problems with the internal classification of such families as Uralic, Dravidian, Austronesian, and not a few others.

tionaries of language families frequently feature unusual semantic changes that some scholars may consider controversial (cf., for example, Nenets *sārmik* 'wolf' = Selkup *suurem* 'bird', out of a Proto-Uralic root with a presumed original meaning 'wild animal' [Rédei 1988: 490]). The balance between 'straightforward' and 'loose' semantic reconstruction is a good indication of the quality of any given EL: a reliable one always has a large proportion of straightforward semantic comparisons.

Finally, if a certain proto-stem is attested in all the daughter languages, it is usually attributed to the proto-language level; if not, we need to decide which chronological level it represents. If very few of the etymologies are represented widely enough in the daughter languages, this increases the probability that we may actually be dealing with language contact or even chance resemblances.

Etymologies that meet these conditions can be called 'core', loosely defined as etymologies that

(a) belong to the basic lexicon;

(b) satisfy the proposed system of phonological correspondences, either fully or partially (in the latter case, accompanied by a reasonable explanation of the irregularly corresponding parts of the forms; this remark is introduced specially to account for the fact that even in well-established families of A or B types occasional irregularities are practically inevitable — e. g., Latin *quattuor* and Proto-Germanic **fiðwor* 'four' do not correspond to each other ideally, but can still be counted as a 'core' etymology since the latter form quite likely is the result of a phonetic dissimilation from **hwiðwor*);

(c) constitute either a perfect semantic match (i. e. the meanings of the compared words are the same) or a near-perfect one (the meanings are not the same, but the semantic development is trivial and frequently attested in languages, e. g. 'water' → 'rain', 'belly' → 'stomach', 'eye' → 'see', &c.);

(d) undoubtedly belong to the proto-language level, i. e. are represented in most or at least non-contacting daughter languages / branches.

The more core etymologies we find in an EL, the more reliable it becomes; and with this growing reliability we can allow ourselves to include more numerous 'peripheral' etymologies, ones that are based on the same system of correspondences as the core ones but either do not belong to the basic lexicon or violate the 'near-perfect semantics' rule. The ultimate proof of relationship, however, always lies with the 'core' part.

The 'core etymology' criterion is not always precisely formulated in various handbooks on historical linguistics, but probably no one would argue that there is even a single well-studied language family for which it does not work. And when properly applied to such linguistic situations as we find between English and French, Tamil and Old Indian, Chinese and Japanese, Turkic and Mongolic, it rarely, if ever, fails to indicate the difference between traces of old contacts and genetic inheritance.

However, this criterion has so far remained unapplied to many lesser known families that do not have a long tradition of scientific research. The *Ethnologue* classification recognizes many of these not because their proto-languages have been reconstructed and core etymologies established, but because scholars have used various similarities (lexical, grammatical, even typological) as indications of possible groupings. Such an approach is typical, for example, of the language families of New Guinea, where preliminary identification was not accompanied by proper comparative work[14]. As a result, further study suggests that some of these groupings have been mapped out incorrectly, e. g. the Sepik-Ramu family [Foley 2005], whereas others are better supported by data.

The immense Austronesian family with its twelve hundred languages represents an intermediate situation: some of its members are tied together on solid comparative grounds, while others (mainly in West Oceania) are added on the basis of unsystematic similarities. Such a basis may be useful at a very early stage of research, but the problem is that unsystematic similarities can be traced between almost any languages, and only a transition to systematic phonetic correspondences and 'core' etymologies guarantees a convincing relationship/classification proposal.

5. *Language families and linguistic tradition*.

From a purely practical point of view, it helps to distinguish between what we have tentatively named 'classic' and 'non-traditional' language families. The first type includes families with a long his-

[14] At this stage of research it is practically impossible to find hard evidence in favor of Torricelli, Left May, and some other groupings given in *Ethnologue*.

tory of comparative studies, like Germanic, Semitic, or Turkic. Current specialists in Germanic linguistics, for example, have access to a vast amount of information, not just literature on Proto-Germanic, but also etymological and dialectal dictionaries of modern languages, detailed descriptions of histories of individual languages, and, in many cases, native informants. Only a meticulous study of this literature and the major languages of the family enables scholars to find a new topic for research and investigate it.

A 'non-traditional' family never enjoys the same level of attention. Some or even most of its languages are poorly known, and historical research here is usually carried out by one or two scholars (who often have to collect their own data in the field before using them for their reconstructions). Each of these scholars is doing research equivalent to that of dozens of specialists in 'classic' families; yet at the same time they are expected to produce results of the same quality, which is often not possible. This is not to suggest that much looser requirements should be applied to the investigation of 'non-traditional' families. But one should always remember that their study is conducted in a less favorable environment and must be treated accordingly.

Today, Indo-European comparative linguistics is still — for good reason — seen as the best standard to which one can hold comparative research done on the basis of other language families, 'classic' and 'non-traditional' alike. Yet there are also certain types of research for which it makes little sense to hold them up to such high standards.

Imagine, for instance, a situation when someone asked to evaluate a 17th century map of North America responds in the following manner: 'Since this map is plagued with mistakes and there are much more precise recent maps available, this one is only valuable as an antique.' This answer is at best incomplete, since it forgets to mention that

- at the time of its charting, this map was the best of its kind;
- many of the locations are mapped quite precisely;
- it may have been used as a source for further geographical discoveries, &c.

The same historical approach is normally applied to an 'obsolete' piece of linguistic research. We evaluate it according to its impact on the respective field and not just according to the current state of linguistic knowledge. It would certainly look odd if, for example, a modern reference to R. Caldwell's trail-blazing *Comparative Grammar of Dravidian Languages* ([1856]) were to simply list all of its (quite numerous) errors, but failed to mention that it was actually the first major publication on Dravidian historical linguistics, which opened this field of research to subsequent generations of linguists who were later able to improve on its achievements.

The same logic must be applied to evaluating historical research done for 'non-traditional' language families. It would not be consistent, for example, to compare reconstructions of Proto-Arawakan by E. Matteson or D. Payne ([Matteson 1972], [Payne 1991]) based on modern languages fresh from the field with, e. g., J. Pokorny's dictionary of Indo-European ([1958]), based on at least a century of linguistic and philological studies. Yet even with that impressive background, Pokorny's dictionary is still known to contain factual mistakes and dubious etymologies, with scores of linguists working to improve it; but are there currently even two linguists in the world working on Proto-Arawakan?

Finally, the same reservations that should be made for poorly studied and 'traditionless' groups apply equally well to research on Type C families. Proto-languages for these families are derived from reconstructions of Type B and Type A families which themselves may be incomplete and partially incorrect (just as work on recently transcribed 'minor' languages can be), and research on them is usually carried out by just a tiny handful of specialists (as is the case with Arawakan). From a technical standpoint, this inevitably decreases the general quality of the output; yet there is still no denying its significance for further generations of researchers.

Within the linguistic community, there are two approaches to evaluating research on 'suspicious' families of these kinds. We may provisionally call those approaches 'hypercritical' and 'bona fide'. The former is perhaps best illustrated by a series of publications criticizing the Altaic theory (the hypothesis of a genetic relationship among Turkic, Mongolian, Manchu-Tungus, Korean, and Japanese languages) by authors such as G. Clauson, G. Dörfer, and, more recently, J. Janhunen, S. Georg, and others. This kind of criticism usually aims at weaker, less easily defensible parts of the theory and, upon discarding them, uses induction to carry the skepticism over to its stronger sides. It is possible, for instance, to concentrate one's attention on several etymologies containing mistakes on the part of the authors (usually mixing together *factual mistakes* and *alternate hypotheses*, so that it is hard for the general reader to tell one from the other) and omit the better etymologies from the discussion altogether[15].

What is forgotten in the process of such criticism is that the same procedure can easily be used to discredit even commonly accepted, 'traditional' theories of genetic relationship, essentially bringing comparative research to a standstill altogether. The first effects of this may already be observed in such works as [Marcantonio 2002], in which the author presumes to discredit the well-proved theory of Uralic (Fenno-Ugric and Samoyed) relationship. Fortunately, the work met more criticism than appraisal from specialists in the field, but it should be noted that the author is a professional linguist, and that the book was issued by no less than the Philological Society series at Cambridge. It is alarming to imagine that this may, indeed, be but the first sign of things to come; surely there are plenty of things that can be criticized about Indo-European as well.

Evaluations and reviews that we call 'bona fide', on the other hand, are based on assuming that:

- the publication may contain important new data;
- its author(s) is/are not intentionally trying to deceive the linguistic community;
- new approaches, especially in the domains not fully explored by traditional comparative linguistics, should not be prohibited unless they lead to conclusions that directly contradict the basic methodology of the science;
- factual mistakes may undermine, but not necessarily invalidate the major conclusions;
- if these conclusions go against the views of the 'mainstream', this does not justify their immediate rejection.

This by no means equates a 'bona fide' review with one that willingly endorses the proposed genetic relationship; the only serious difference is that 'bona fide' reviewers may be willing to weigh the positive evidence and explicitly evaluate it along with the negative. Yet such a position is not frequently found in the literature.

As a typical example we might mention the treatment of Joseph Greenberg's Amerind hypothesis [Greenberg 1987]. Several books and dozens of articles were then written in response, contradicting the proposal that all native languages of the Americas that are not Eskimo-Aleut or Na-Dene belong to just one Amerind super-family. The authors gave theoretical reasons why they think Greenberg's method is wrong[16]; listed his numerous typos and mistakes; argued — in many cases, quite convincingly — that certain languages should actually be assigned to different low-level families, &c. (a representative example is [Campbell 1997]). Yet at the same time it still remains unexplained why there are so many similar-looking words found, for instance, in Greenberg's Equatorial family. Does this mean that all of his Equatorial comparisons have no value whatsoever? Is the same also true for other major families within "Amerind"? How can we provide a reasonable explanation for lexical similarities among the major families listed by Greenberg? Shouldn't we carefully investigate each proposed connection in more depth instead of bluntly ascribing all of them to 'chance' and 'contact' factors? No 'bona fide' review of this pioneer work has ever been published, thus leaving the Amerind problem unresolved and the evidence in favour of either "Amerind" as a whole or subdivisions of "Amerind" as separate families still waiting for a proper treatment.

6. *Evolution of Human Languages Project.*

In the light of this state of affairs, the main goal of *Evolution of Human Languages* (*EHL*), has been stated as a thorough scientific investigation of degrees of genetic relationship among the world's languages. Its principal unique feature is the creation of a comprehensive network of etymological data based on the following general principles:

[15] For a detailed discussion of such methods, see [Dybo & Starostin 2007].

[16] The Amerind hypothesis does not follow the methodology advocated within this publication; instead, Greenberg applies a method that eschews explicit intermediate reconstructions in favour of direct language comparison, even within Type C families. We consider this method useful at times for forming preliminary hypotheses on genetic relationship, which then have to be strengthened through a careful application of the comparative method as described above. This, however, does not mean that objective evaluation of Greenberg's method and results should focus strictly on their negative aspects.

(i) we strongly believe that, in the highly computerized world of today, etymological data must be stored in electronic databases and not only in the old fashioned published format;
(ij) etymological information should be easily and instantly available to every scholar, meaning that the databases should have public access;
(iij) the best outcome is usually achieved when at least two scholars are working on the same topic, and active interaction between researchers specializing in different language families is of crucial importance as well;
(iv) even the 'raw' and preliminary results of etymological research should be open for public inspection and suggestions for improvement (according to the 'bona fide' strategy), although creation and editing of the databases themselves should, of course, be restricted to specialists.

The general database network can be described as follows[17]:

1. It is a hierarchical structure, reflecting the step-by-step method of comparative reconstruction. Starting at the lowest levels (Germanic, Turkic, South Dravidian, &c.), the databases are then systematically linked to higher level ones, gradually progressing from A-type families to B- and C-type taxa.

2. All the databases follow one standard unified format (proto-form with approximate meaning; reflexes and meanings in daughter languages; commentaries and bibliography), although minor variations are possible.

3. The hyperlink system works in both directions, so it is possible to search for related forms in various distantly related modern languages, e. g. find potential cognates between Chinese and Chechen, or Finnish and Japanese (see below).

4. Efforts are undertaken to make each database as comprehensive as possible by incorporating all the lexical (and sometimes grammatical) material from the compared languages that is historically relevant. Naturally, the size and reliability of each database are directly related to the size and reliability of available lexical corpora, but the general principle of *EHL* is to preserve and make use of every bit of information. Most of the databases are available for public viewing (some even for downloading) at http://starling.rinet.ru. At the present time they mainly focus on linguistic families of Eurasia and North Africa, where four 'superfamilies' can be identified with varying degrees of reliability: Sino-Caucasian, Eurasiatic, Afroasiatic, and Austric. A brief perspective on each of them is given below.

Sino-Caucasian

The existence of this 'superfamily' was first scientifically motivated by S. Starostin [Starostin 1984]. His evidence was based on three lower level language families: North Caucasian, Sino-Tibetan, and Yeniseian. Later two additional languages — Basque and Burushaski, usually treated as isolates, were added to the list.

The first comprehensive etymological dictionary of North Caucasian (B-type family) was published in 1994 [Nikolayev & Starostin 1994]. Descriptive sources for all languages and major dialects of the family were used in a step-by-step reconstruction of the proto-language that generally meets the criteria listed above. The reconstruction, with 2759 entries, is supported by 8 subordinate databases representing A-type daughter families (Nakh, Lezghian, Abkhaz-Adyghe, &c.). The main point of the work, tying together Nakh-Daghestanian (East Caucasian) and Abkhaz-Adyghe (West Caucasian) languages, is still far from being universally accepted in Western linguistics, mainly due to insufficient scientific debate (apart from one or two very brief critical reviews, not a single detailed discussion of the dictionary has appeared in Western press up to now), but enjoys much support among the majority of Russian experts on historical Caucasian linguistics.

Due to a lack of descriptive data for most languages, the only existing etymological dictionary of Sino-Tibetan languages [Peiros & Starostin 1996] was compiled through systematic comparison of only five available languages: Old Chinese, Tibetan, Burmese, Kachin (Jingpaw), and Lushai; however, even on the basis of these five it is possible to establish detailed phonological correspondences and arrive at a reconstruction. After publication the work was continued by adding new etymologies from published sources and improving the correspondences. An updated electronic version of the dictionary (2823 entries) also contains two subordinate databases: comparative Kiranti and a substantial historical database of Chinese characters (more than 4,000 entries), both prepared by S. Starostin.

[17] The network is supported by the StarLing software package developed by S. Starostin.

The Yeniseian etymological dictionary was published in [Starostin 1995], following the phonological reconstruction in [Starostin 1982]. An updated version, which includes additional information from [Werner 2002], contains 1059 entries.

There are several publications on Sino-Caucasian, most of them belonging to S. Starostin. The family, like most C-type ones, has not yet found significant recognition. In this particular case, besides the usual reasons, there is also the complexity of correspondences (North Caucasian languages, in particular, are among the most phonetically complex in the world) and lack of full-length publications. Just before his premature passing away in September 2005, however, S. Starostin finished a book-long manuscript on Proto-Sino-Caucasian, including detailed discussion of its comparative phonology and lexicon, which will be published in the near future (and is already available online). So far, the etymological database contains 1361 entries.

Glottochronology suggests the following rough dates of splitting for Sino-Caucasian and its descendants:

Sino-Caucasian	10 KYA	Sino-Tibetan	6 KYA	Basque	modern
North Caucasian	6 KYA	Yeniseian	2 KYA	Burushaski	modern

It is also highly likely that the Na-Dene family of North America belongs to the Sino-Caucasian superfamily. This has been postulated with the help of modern comparative data by S. Nikolayev [Nikolayev 1991], and supported on the basis of "mass comparison" by J. Bengtson and M. Ruhlen; today, work in this direction is mainly advanced by J. Bengtson. However, no EL is currently available for Na-Dene, and further research is needed to verify this hypothesis. Occasional attempts to broaden the scope of the family by including extinct languages of Eurasia such as Sumerian or Etruscan, as well as other language families in the Americas (e. g. Salishan), have not been successful so far.

Eurasiatic

In the 1960s V. M. Illich-Svitych presented a set of arguments in favour of a genetic relationship among several families of Eurasia and North Africa; the proposed 'superfamily' was called Nostratic ([Illich-Svitych 1971–1984]). Evidence in favour of Nostratic includes regular correspondences between daughter proto-languages, large etymological lists that rely heavily on the basic lexicon, and a small inventory of reconstructed grammatical morphemes. Several decades later S. Starostin has proposed that it is better to talk about two coordinate superfamilies: Eurasiatic and Afroasiatic (for the former the old name 'Nostratic' is sometimes retained, to avoid confusion with Greenberg's 'Eurasiatic' — a similar classification proposal, although based mainly on the "mass comparison" method, that excludes Dravidian and Kartvelian). The *EHL* etymological network includes several databases of Eurasiatic languages:

- the Altaic comparative dictionary, published in 2003; its electronic version contains the proto-language database (2805 entries) supported by five subordinate reconstructions: Turkic, Mongolian, Tungus-Manchu, Korean, and Japanese;
- the Eskimo database (1774 entries); according to O. Mudrak [Mudrak 1984], Eskimo forms part of Eurasiatic[18] and within it is particularly close to Altaic. It should be mentioned, however, that no database on Aleut is available so far;
- a detailed database on comparative Dravidian, prepared by G. Starostin (2211 entries), based primarily on etymological data available in [Burrow & Emeneau 1984], but incorporating a modified variant of the traditional reconstruction;
- the Indo-European database by S. Nikolayev is arguably the largest collection of proto-morphemes of this well-studied family (3178 entries). Two subordinate databases — Germanic and Baltic — were also compiled by the same scholar;
- the Uralic database (1898 entries) is loosely based on K. Rédei's etymological dictionary [Rédei 1988] with significant additions from a team of Uralic linguistics scholars working in Moscow. No subordinate databases are available so far, but the daughter families are undergoing intensive study;
- the Kartvelian database (1310 entries) is based on G. Klimov's etymological dictionary ([Klimov 1998]) with some additions by S. Starostin that reflect his study of the external contacts of this family.

[18] The Eurasiatic origin of the Eskaleut languages was also proposed by J. Greenberg.

No other B- or A-type families can be demonstrated to belong to Eurasiatic (one possible hypothesis suggests that the Chukchi-Kamchatkan languages also form parts of it, but it is also possible that in reality they belong to a different superfamily; the whole issue needs very serious work). The combined Eurasiatic database brings together reliable etymologies both from already published sources and those newly discovered by members of the *EHL* team (altogether 2077 entries of varying quality, but including 'core' ones).

The age of Eurasiatic and its major descendants may be crudely estimated as:

Eurasiatic	12 KYA				
Altaic	8 KYA	Dravidian	5 KYA	Uralic	6 KYA
Eskimo	2 KYA	Indo-European	7 KYA	Kartvelian	4–3 KYA

Afroasiatic

Afroasiatic is arguably the only 'superfamily' that is generally recognized by mainstream linguistics, despite (or, perhaps, due to) the fact that its languages are actually far less studied than those of Eurasiatic. It is formed by five daughter families (Semitic, Berber, Chadic, Cushitic, and Omotic) and one isolate (Egyptian). None of these families except for Semitic can boast a comprehensive etymological dictionary, and even for Semitic with its lengthy tradition of study, existing ELs are either outdated [Cohen 1974] or, as of now, still incomplete [Militarev & Kogan 2000–2003]. The situation with other branches of Afroasiatic is even less fortunate.

As a result, daughter families are represented differently in the Afroasiatic network of *EHL*. The Semitic database (compiled by A. Militarev) contains 2852 etymologies, each found in at least two languages of the family. Next to it is the Egyptian one (also compiled by Militarev) with 1495 entries. Other databases have about several hundred etymologies each.

The main Afroasiatic database consists of 3212 etymologies of varying quality. Its compilers, A. Militarev and O. Stolbova, brought together acceptable published etymologies (from dictionaries such as [Orel & Stolbova 1995]) and then enriched it with hundreds of newly discovered ones. Further research on the family in general is tightly connected with improving existing reconstructions for Cushitic, Omotic, and Chadic.

The ages of Afroasiatic and its descendants are roughly estimated as:

Afroasiatic	12 KYA	Cushitic	9 KYA	Semitic	7 KYA
Omotic	7 KYA	Chadic	7 KYA	Berber	3 KYA

Austric

The existence of the Austric superfamily is less clearly shown than that of the others. In contrast to Eurasiatic, Sino-Caucasian, and Afroasiatic, no detailed proto-Austric glossaries or equally detailed tables of correspondences between the various daughter branches of Austric have been produced. The number of these proposed daughter branches is four, and they are grouped in pairs: Austronesian / Tai-Kadai and Austroasiatic / Miao-Yao[19]. Until recently none of these daughter families was represented by a comprehensive etymological dictionary. In terms of known etymologies Austronesian fares significantly better than the others, reflecting a fairly long tradition of scholarship. The work, however, is mostly on languages of Western Indonesia, the Philippines, and parts of Eastern Oceania. Other territories are underrepresented, which often makes it difficult to attribute an etymology to the appropriate chronological level. An *EHL* Austronesian database is still under construction.

The situation with Tai-Kadai is slightly different. A representative collection of etymologies for one of its principal branches, Zhuang-Tai, has been published [Li 1977] and is available (with a few modifications) on the *EHL* network (1329 entries). A limited number of common Tai-Kadai forms (taken mainly from [Peiros 1998]) are also available electronically.

The Austroasiatic family is represented much better. A preliminary version of its comparative dictionary by I. Peiros contains 2457 entries, identified mainly through comparison of ten lower level families that include about all of Austroasiatic[20].

[19] See [Peiros 1998].

[20] Etymologies from *Austroasiatic Etymological Dictionary* by H. Shorto have not yet been included.

For Miao-Yao no etymological dictionaries or convincing phonological reconstructions are available; as a result, no corresponding database is found in the *EHL* network either. All the Miao-Yao proto-forms used in Austric comparisons are taken from [Peiros 1998].

A systematic search for Austric etymologies has not yet been conducted. In 2004–2005 S. Starostin and I. Peiros collected some lexical similarities in support of Austric, included in the highly provisional 'Austric' database. The 900 proposed etymologies represent mainly lexical similarities between Austroasiatic and Austronesian, although they are actually supported by basic (relatively simple) phonological correspondences. Thus, we can say that our knowledge of Austric is much more limited than that of other superfamilies; however, the discovery of a number of comparanda that fit the 'core etymology' requirements makes the Austric hypothesis quite plausible.

The approximate age of Austric and its descendants may be estimated thus:

Austric	10 KYA	Austronesian	5 KYA[21]	Austroasiatic	7 KYA
		Tai-Kadai	5 KYA	Miao-Yao	4 KYA

“Borean”?

The *EHL* network of databases presents what we consider extremely strong evidence in support of the four superfamilies discussed above. Two of them — Sino-Caucasian and Eurasiatic — are based on reconstructions performed in full accordance with the comparative method, with regular phonetic correspondences established between reconstructed intermediate proto-languages. Knowledge of the other two — Afroasiatic and Austric — has not so far reached the same level of confidence; however, since it is possible to discuss the two theories in terms of regular phonetic correspondences rather than mere similarities, both can be considered scientific, and further work on them is promising.

Since Afroasiatic and Austric are still to be reconstructed and the other two superfamilies are still in need of serious improvements, it is not yet possible to apply strict comparative methods of investigation to even deeper chronological levels. However, the obstacles here are technical rather than theoretical. The widespread idea that comparative research has an impassable threshold of about 10 KYA (such a period of time is claimed to be enough to make related languages lose all relevant similarity) does not take into account the fact that the main objects of research in this case are not modern languages, but reconstructed proto-languages which turn out to be more similar to one another than their modern day descendants. Thus, solid reconstructions for C-type families like Eurasiatic, Sino-Caucasian, and others should eventually help[22].

At the present time it is possible to discuss such 'ultra-deep' relationships only on a very speculative level. Numerous morphemic similarities between various language families of Eurasia have already been spotted in the past as potential indication of such a relationship; many, if not most, of these similar forms (traced back to high level reconstructions) were compiled by S. Starostin into a special database and later supplemented by some of his own findings. Since such morphemic comparisons are rather numerous (several hundred at least), chance resemblances are not very probable, and a 'Borean' super-superfamily hypothesis, open to 'bona fide' discussion, has been formulated[23]. Statistical analysis of attested similarities shows that if such a taxon really exists, its initial division was as follows:

(i) Eurasiatic and Afroasiatic (= Illich-Svitych's 'Nostratic');
(ij) Sino-Caucasian
(iij) Austric.

The estimated age of 'Borean' would be around 15–17 KYA.

[21] Our calculations exclude Austronesian languages of New Guinea and some surrounding islands.

[22] It should be noted that the amount of information recoverable for protolanguage states always remains in direct proportion to the number of languages used for comparison. Since lexical loss and replacement, in most cases, occurs independently in daughter languages, the probability of any given morpheme disappearing without a trace in three of them is less than in only two, and so on. This ensures that, given a sufficient number of languages or language branches, the morphemic inventory of the reconstructed ancestral language will be just as large (sometimes even larger!) than that of any single one of its descendants. Such is the case for commonly accepted families like Indo-European, Turkic, Semitic, etc., and there is little reason to doubt that a different situation has to be proposed for Type C families.

[23] The term was originally used by H. Fleming for a somewhat different linguistic entity [Fleming 1991].

Other families

The 'Borean' hypothesis currently links together the 4 superfamilies described above. At the same time, the real scope of 'Borean' remains unknown, since we still lack deep level reconstructions of many families in the American, African, and Indo-Pacific, and Australia. This means that we have no certain means of verifying whether some of these families can also form part of the hypothetical 'Borean'.

So far, preliminary research has been carried out for Khoisan in South Africa, resulting in a classification and a set of provisional reconstructions by G. Starostin [2003, 2008], whose Khoisan databases are already incorporated into the *EHL* etymological network. Comparison with the 'Borean' data has not produced any conclusive results, suggesting that Khoisan, at least, cannot be included in 'Borean', although genetic connections on an even deeper level might be possible.

The preliminary lexicostatistical study of Sub-Saharan African languages identifies at least 4 other superfamiles: Niger-Congo (not quite identical to the Niger-Congo super-family proposed by J. Greenberg), East Sudanic, Central Sudanic, and Kordofanian, plus a number of smaller branches whose position is so far very unclear (such as Songhay or Atlantic languages). We still need to investigate how these superfamilies are connected to each other and to Borean. This can be done, however, only through an extensive etymological study of available data.

The situation with native languages of the Americas is different, and still far from being resolved. According to J. Greenberg, they can all be classified into two small stocks (Eskimo-Aleut, Na-Dene) and one huge super-family — Amerind. We have significant evidence that the Eskimo-Aleut family is part of Eurasiatic, while Na-Dene seems to be related to Sino-Caucasian languages (although in the latter case a completely convincing demonstration is still lacking).

The status of the Amerind proposal remains unclear. The main source for lexical lookalikes between these languages remains J. Greenberg's monograph. We have already mentioned that it was and still is heavily criticized, sometimes for good reasons; this, however, does not eliminate the problem itself — why exactly are there so many lexical similarities found not just between some of the proposed daughter families, but frequently between all the families of the hypothetical Amerind superfamily? Even if we rule out everything that is 'non-core' (i. e. forms with scarce distribution, far-flung semantics, &c.), explaining all the rest away as chance resemblances would simply be closing our eyes on the problem. This leaves us with two options: either the similarities are of common genetic origin, or they result from intensive language contact, which is a less probable option given the peculiarities of their dispersal.

A very preliminary proposal based on *EHL* exploratory studies suggests that in the Americas one can find at least three types of grouping:

(i) The Almosan superfamily (Algic, Salishan, Wakashan, and some other languages) might be related to Chukchee and Nivkh languages of North Asia, forming a so-called "Beringian" superfamily; connections with "Borean" have been noticed as well[24].

(ij) A number of families (Penutian, Hokan, Mayan, Mixe-Zoque, Maipuran, Pano-Tukanoan, &c.) presumably form a different superfamily, also with resemblances to "Borean"[25].

(iij) At the same time, we were not able to detect any external relations for such well-established families as Siouan, Gulf, or Otomanguean.

Distant relationships among the Papuan (= Non-Austronesian) languages of New Guinea and aboriginal languages of Australia remain to be investigated. It is possible that in that region we could distinguish up to 4–6 superfamilies (Trans-New-Guinea, Australian, East Papuan, &c.), none being properly reconstructed. Some lexical similarities have also been spotted between Trans-New-Guinea morphemes and some of the alleged 'Borean' roots, but these remain too scarce to establish a firm connection.

7. *The 'bottleneck' scenario.*

Anatomically modern human beings seem to have evolved in Africa around two hundred thousand years ago ([Alemseged, Coppens, & Geraads 2002]). It is not known for certain when they acquired language of the modern type. The second wave of migration of anatomically modern human beings out of Africa, according to genetic and archaeological data, seems to have taken place around 60 to 50 KYA. Compared with the previous known wave (which occurred of the order of a hundred and thirty KYA and is known to have reached as far as Palestine) it was definitely more successful, populating Eurasia and the Indo-Pacific region, including Australia. In Western Europe it gave rise to the Aurignacian culture,

[24] Proposed by S. Nikolaev, who is now working on a detailed justification.

[25] Proposed by S. Nikolaev and I. Peiros in their survey of existing linguistic classifications.

with its remarkable paintings, engravings, and sculpture in addition to the widespread Upper Paleolithic tools, suggesting that the human beings of the second wave were behaviorally as well as anatomically modern. That makes it likely that they already possessed a language or languages of the modern type.

For tens of thousands of years afterwards, the usual kinds of linguistic transformation presumably took place, producing daughter languages, which themselves gave rise to daughter languages, and so forth. In that way a considerable degree of linguistic diversity would have been achieved. However, that amount of diversity need not necessarily be reflected in the diversity of attested languages. Instead, we may be dealing here with a "bottleneck" effect, in which a great many languages (but not necessary all of them) descend from a single ancestor.

The climatic changes near the height of the last Ice Age some twenty thousand years ago shrank drastically the territories suitable for human habitation, with ice caps and deserts occupying a large fraction of the land mass. We may picture the human beings of that time confined to refugia often separated by hostile areas. Under those conditions linguistic diversity could have been greatly reduced and it may therefore be the case that all or most of the languages of subsequent times are descended from a single ancestor, the tongue of a particular refugium. If the similarities of attested languages are found to suggest a common origin for all or most of them, that origin could well be a speech that survived the height of the Ice Age when most others did not. With the improvement of climatic conditions, humans began to move out of their refugia, colonizing territories previously unsuitable for permanent occupation. This led to growth and subsequent division of their communities, resulting in the development of new languages.

This is a different story from that of "monogenesis", according to which the latest common ancestor of all or most attested languages would be the earliest human language of the modern type. According to the bottleneck scheme, by contrast, all or most of the diversity of attested languages would have developed over some twenty thousand instead of fifty thousand years, making it somewhat more plausible that one might discover evidence of common descent if such common descent is actually correct. In addition, the bottleneck idea allows the age of modern language to be pushed back to any time between fifty thousand years ago and two hundred thousand years ago, when anatomically modern humans appeared.

Suppose it is true, as mentioned above, that an enlarged Borean taxon embraces most of the African superfamilies as well as the "Amerind" languages, then the age of such a "supersuperfamily" would agree very roughly with the 20 000 years we attribute to the bottleneck. If some or all of the 'global roots', found across most of the world's super-families[26], are genuine, then they could relate to such a time horizon. Another feature of the bottleneck scheme with a protolanguage less than 20 000 years old is that the migrations of the speakers of the descendant languages (such as Eurasiatic, Afroasiatic, &c.) need have nothing to do with the 'out of Africa' notion which refers to much earlier times.

The phenomenon of linguistic bottleneck is encountered on smaller scales, as in the case of the Australian aboriginal languages. It is believed that humans reached the continent at least 40 KYA. Yet it is generally agreed that all or nearly all of the attested Australian languages form a single superfamily and the similarities among them make it extremely difficult to believe that the proto-language of that superfamily could be older than twelve thousand years or so[27]. It seems that a single language or the descendants of a single language spread over all or nearly all of Australia at that comparatively recent date. It could have been a local language of Australia or of New Guinea; if we had to guess, we would probably choose the latter, on the basis of some lexical similarities. In any case, one can search for genetic or cultural traits that might have been introduced along with the language in question.

A very familiar example of a bottleneck is the domination of Europe by Indo-European languages. Here most of the currently spoken European languages, with but a few exceptions like Basque (whose native speakers constitute a small linguo-geographic "refugium"), are descended from a single language spoken, probably in Southern Russia, some six or seven thousand years ago. Since we are used to that idea, we should be able to entertain the possibility of bottlenecks having occurred on much wider scales.

[26] For more on 'global roots', see e. g. [Ruhlen 1994]. Many of the particular connections proposed by Ruhlen, as well as other researchers, have been heavily — and often justly — criticized (e. g. by D. Ringe, L. Campbell, and others), yet a decisive statistical demonstration that would once and for all reject all of the accumulated evidence as non-evidence is still missing. We apply the same cautionary approach to the issue of 'global roots' as we do towards Greenberg's 'Amerind' and all the other theories arrived at through the 'mass comparison' method: all of the comparative data obtained that way are valid as material for further research, to be gradually accepted or discarded as the families in question are subjected to standard comparative analysis.

[27] I. Peiros, in his unpublished lexicostatistical study of aboriginal Australian languages, estimates the age of Proto-Australian as constituting around 10,000 to 12,000 years.

8. *Conclusion*

The data accumulated within the *EHL* etymological network vary in quality and 'convincing force'. The program, however, manages to collate the results of half a century's research on distant language relationships (so far, mostly within Eurasia) with results recently achieved by the *EHL* team which has the benefits of better data access, methodology that is modernized (but still firmly rooted in tradition), and computer handling of data. Much remains to be done, but even now certain prehistoric scenarios of linguistic development within the last 20 000 years can be drawn up, with the most probable one looking as follows:

(i) At the height of the Ice Age humans were forced to take refuge in one or several zones suitable for survival, causing a decrease of the linguistic diversity that presumably existed before.

(ij) With the improvement of climatic conditions, humans began to move out of their refugia, colonizing territories previously unsuitable for permanent occupation. This led to growth and subsequent division of their communities, resulting in the development of new languages.

(iij) In the process of spreading various linguistic groups suffered different fates; some disappeared with or without any traces, while others expanded, spreading their languages over vast territories or shifting from one language to another[28].

(iv) One of the most succesful survivors of the Ice Age may be a hypothetical 'Borean' super-superfamily, whose age is estimated as 15–17 KYA. Inconclusive, but significant evidence for 'Borean' is provided by preliminary comparison of four super-families, the historical reality of which is no longer questioned by *EHL* members: Eurasiatic, Afroasiatic, Sino-Caucasian and Austric.

(v) Preliminary data also indicate possible connections between Borean and some superfamilies of Africa, America, and the Indo-Pacific region, not included in the four superfamilies mentioned above. Further research into distant relationships of languages is needed to find out whether these additional superfamilies are related to 'Borean' on a higher level or are hitherto unidentified branches of 'Borean'. The question of the original 'Borean' homeland also remains open.

Further research into distant relationships of languages is needed to find out whether there are other Ice Age survivors that are related to 'Borean' on a higher level or turn out to be potentially undiscovered members of the family itself. The question of the original 'Borean' homeland also remains open.

Acknowledgments

We would like to thank all participants of the *EHL* program for providing data and sharing their knowledge with us. We also gratefully acknowledge the generous support for the *EHL* project on the part of Mr. Jerry MURDOCK and Bryan J. and June B. ZWAN Foundation.

References

ALEMSEGED, COPPENS, & GERAADS 2002 — Z. ALEMSEGED, Y. COPPENS, and D. GERAADS. Hominid Cranium from Omo: Description and Taxonomy of Omo-323-1976-896 // *American Journal of Physical Anthropology*. 117 (2).

BERGSLAND & VOGT 1962 — K. BERGSLAND and H. VOGT. On the Validity of Glottochronology // *Current Anthropology*. 3; pp. 115–153.

BLUST 2000 — R. BLUST. Why Lexicostatistics Doesn't Work: the 'Universal Constant' Hypothesis and the Austronesian Languages // *Time Depth in Historical Linguistics*. Ed. by Colin RENFREW, April MCMAHON & Larry TRASK. MCDONALD Institute for Archaeological Research, Cambridge; pp. 311–332.

BURLAK & STAROSTIN 2005 — С. А. БУРЛАК, С. А. СТАРОСТИН. *Сравнительно-историческое языкознание* [*Comparative-Historical Linguistics*]. 2-е изд. М. [Second edition. Moscow]: "Academia".

BURROW & EMENEAU 1984 — T. BURROW and M. B. EMENEAU. *A Dravidian Etymological Dictionary*. Second Edition. Oxford Publishers.

CALDWELL 1856 — R. CALDWELL. *A Comparative Grammar Of The Dravidian or South Indian Family of Languages*, 1st ed. London.

CAMPBELL 1997 — L. CAMPBELL. *American Indian Languages: the Historical Linguistics of Native America*. Oxford University Press, New York.

[28] Technically, language shift is a kind of language loss.

CAMPBELL 2004 — L. CAMPBELL. *Historical Linguistics: an Introduction*. 2nd edition. Edinburgh: Edinburgh University Press, Cambridge MA: MIT Press.

COHEN 1974 — D. COHEN. *Dictionnaire des racines sémitiques ou attestées dans les langues sémitiques*. Paris, MOUTON, La Haye.

CROWLEY & DIXON 1981 — T. CROWLEY and R. M. W. DIXON. Tasmanian // *Handbook of Australian Languages*, Vol. 2. Ed. by R. M. W. DIXON & B. J. BLAKE. Canberra, Australian National University Press.

DIMMENDAAL 1995 — G. J. DIMMENDAAL. Do Some Languages Have a Multi-Genetic or Non-Genetic Origin? An Exercise in Taxonomy // *Cinquième Colloque de Linguistique Nilo-Saharienne*. Éd. par Robert NICOLAÏ et Franz ROTTLAND. Köln: Rüdiger KÖPPE Verlag; pp. 357–372.

DYBO & STAROSTIN 2008 — A. V. DYBO and G. S. STAROSTIN. In Defense of the Comparative Method, or the End of the Vovin Controversy // *Аспекты компаративистики*. [т.] 3 [*Aspects of Comparative Linguistics*. [v.] 3 / *Orientalia et Classica:* Труды Института восточных культур и античности. Вып. XIX. М.: Российский государственный гуманитарный университет [Papers of the Institute of Oriental and Classical Studies. Issue XIX. Moscow: RSUH Publishers]; pp. 119–258.

DYEN 1965 — I. DYEN. A Lexicostatistical Classification of the Austronesian Languages // *International Journal of American Linguistics*. Memoir 19. Bloomington, Indiana.

FLEMING 1991 — H. C. FLEMING. A New Taxonomic Hypothesis: Borean or Boralean // *Mother Tongue. Newsletter*. V. 14.

FOLEY 2005 — W. A. FOLEY. Linguistic Prehistory in the Sepik-Ramu Basin // *Papuan Pasts: Cultural, Linguistic and Biological Histories of Papuan-Speaking Peoples*. Ed. by Andrew PAWLEY, Robert ATTENBOROUGH, Robin HIDE and Jack GOLSON. Canberra: Pacific Linguistics; pp. 109–144.

GORDON 2005 — R. G. Jr. GORDON (ed.). *Ethnologue: Languages of the World*. Fifteenth edition. Dallas, Tex.: SIL International.

GREENBERG 1966 — J. GREENBERG. *The Languages of Africa*. Indiana University, Bloomington. MOUTON & C°., The Hague, Netherlands.

GREENBERG 1987 — J. GREENBERG. *Language in the Americas*. Stanford University Press, Stanford.

HOCK & JOSEPH 1996 — Hans H. HOCK and Brian D. JOSEPH. *Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics*. Berlin & New York: MOUTON DE GRUYTER.

ILLICH-SVITYCH 1971–1984 — В. М. Иллич-Свитыч. *Опыт сравнения ностратических языков (семитохамитский, картвельский, индоевропейский, уральский, дравидийский, алтайский)*: *Введение. Сравнительный словарь*. М.: «Наука», Главная редакция восточной литературы [*An Attempt at Comparative Dictionary of the Nostratic Languages (Semito-Hamitic, Kartvelian, Indo-European, Uralic, Dravidian, Altaic)*. Moscow: "Nauka" publishers]. V. 1 (*b–K*): 1971; V. 2 (*l–ǯ*): 1976; V. 3 (*p–q*): 1984.

KLIMOV 1998 — G. A. KLIMOV. *Etymological Dictionary of the Kartvelian Languages*. Berlin: MOUTON DE GRUYTER.

LI 1977 — LI FANG-KUEI [李方桂, Lǐ Fāngguì]. *A handbook of Comparative Tai*. Honolulu: University Press of Hawai'i.

MARCANTONIO 2002 — A. MARCANTONIO. *The Uralic Language Family: Facts, Myths and Statistics*. Publications of the Philological Society. Cambridge.

MATTESON 1972 — E. MATTESON. Proto Arawakan // *Comparative Studies in Amerindian Languages*. Ed. by Esther MATTESON. *Janua Linguarum*, series practica, 127. The Hague: MOUTON; pp. 160–242.

MILITAREV & KOGAN 2000–2003 — A. MILITAREV and L. KOGAN. *Semitic Etymological Dictionary*. Münster: *Ugarit* Verlag. Volume 1: 2000. Volume 2: 2003.

MUDRAK 1984 — О. А. МУДРАК. К вопросу о внешних связях эскимосских языков [On the External Relations of Eskimo Languages] // *Лингвистическая реконструкция и древнейшая история Востока*. Сб. тез. М.: ИВ АН СССР [*Linguistic Reconstruction and the Prehistory of the East*. Moscow, Institute of Oriental Studies]; pp. 80–87.

NIKOLAEYV 1991 — S. L. NIKOLAYEV. Sino-Caucasian Languages in America // *Dene-Sino-Caucasian Languages*. Ed. by V. SHEVOROSHKIN. Bochum; pp. 42–66.

NIKOLAYEV & STAROSTIN 1994 — S. L. NIKOLAYEV and S. A. STAROSTIN. *A North Caucasian Etymological Dictionary*. Moscow, "Asterisk" publishers.

OREL & STOLBOVA 1995 — V. OREL and O. STOLBOVA. *Hamito-Semitic Etymological Dictionary: Materials for a Reconstruction*. Leiden & New York.

PAYNE 1991 — D. L. PAYNE. A Classification of Maipuran (Arawakan) Languages Based on Shared Lexical Retentions // *Handbook of Amazonian Languages*, Vol. 3. Ed. by Desmond C. DERBYSHIRE and Geoffrey K. PULLUM. Berlin: MOUTON DE GRUYTER; pp. 355–499.

Peiros 1997 — I. Peiros. Macro Families: Can a Mistake Be Detected? // *Indo-European, Nostratic, and Beyond: Festschrift for Vitalij V. Shevoroshkin* / Ed. by Irén Hegedűs, Peter A. Michalove and Alexis Manaster Ramer. JIES Monograph No. 22. Institute for the Study of Man, Washington D.C.; pp. 265–292.

Peiros 1998 — I. Peiros. Comparative Linguistics in Southeast Asia // *Pacific Linguistics Series* C-142. Canberra, Australian National University.

Peiros 2000 — I. Peiros. Family Diversity and Time Depth // *Time Depth in Historical Linguistics*. Ed. by Colin Renfrew, April McMahon & Larry Trask. McDonald Institute for Archaeological Research, Cambridge; pp. 75–108.

Peiros & Starostin 1996 — I. Peiros and S. Starostin. *A Comparative Vocabulary of Five Sino-Tibetan Languages* (6 vols.). Melbourne.

Pokorny 1958 — J. Pokorny. *Indogermanisches etymologisches Wörterbuch.* 2 Bände. Francke Verlag, Bern und München (Bern–Stutgart 1989, 2. Aufl.; Tübingen–Basel 1994, 3. Aufl.).

Rédei 1988 — K. Rédei. *Uralisches Etymologisches Wörterbuch*. Akadémiai Kiádo, Budapest.

Sands 1998 — B. Sands. Eastern and Southern African Khoisan // *Evaluating Claims in Distant Linguistic Relationships*. Ed. by R. Vossen. *Quellen zur Khoisan-Forschung / Research in Khoisan Studies*, Bd 14. Hamburg: Rüdiger Köppe.

Starostin 1982 — С. А. Старостин. Праенисейская реконструкция и внешние связи енисейских языков [Proto-Yeniseian Reconstruction and the External Relations of Yeniseian Languages] // *Studia Ketica*, vol. 3. Leningrad: “Nauka” publishers; pp. 144–237.

Starostin 1984 — С. А. Старостин. Гипотеза о генетических связях сино-тибетских языков с енисейскими и севернокавказскими языками [A Hypothesis about the Genetic Connections between Sino-Tibetan, Yeniseian, and North Caucasian Languages] // *Лингвистическая реконструкция и древнейшая история Востока*. Сб. тез. М.: ИВ АН СССР [*Linguistic Reconstruction and the Prehistory of the East*. Moscow: Institute of Oriental Studies]; pp. 19–38.

Starostin 1995 — С. А. Старостин. Сравнительный словарь енисейских языков [A Comparative Dictionary of Yeniseian Languages] // *Studia Ketica*, vol. 4. М.: «Языки Русской Культуры» [Moscow: “Languages of Russian Culture”]; pp. 176–315.

Starostin 2000 — S. A. Starostin. Comparative-Historical Linguistics and Lexicostatistics // *Time Depth in Historical Linguistics* / Ed. by Colin Renfrew, April McMahon & Larry Trask. McDonald Institute for Archaeological Research, Cambridge; pp. 223–259.

Starostin 2007 — С. А. Старостин. Определение устойчивости базисной лексики [Definition of the stability of the basic lexicon] // С. А. Старостин. *Труды по языкознанию*. М.: «Языки Славянской Культуры» [*Works on Linguistics*. Moscow: “Languages of Slavic Culture”]; pp. 827–839.

Starostin 2003 — G. S. Starostin. A Lexicostatistical Approach towards Reconstructing Proto-Khoisan // *Mother Tongue*. VIII; pp. 83–128 [“81–126” in the on-cover TOC].

Starostin 2008 — G. S. Starostin. From Modern Khoisan Languages to Proto-Khoisan: the Value of Intermediate Reconstructions // *Аспекты компаративистики*. [т.] 3 [*Aspects of Comparative Linguistics*. [v.] 3] (= *Orientalia et Classica:* Труды Института восточных культур и античности. Вып. XIX. М.: Российский государственный гуманитарный университет [Papers of the Institute of Oriental and Classical Studies. Issue XIX. Moscow: RSUH Publishers]); pp. 337–470.

Vasilyev & Militarev 2008 — М. Е. Васильев, А. Ю. Милитарёв. Глоттохронология в сравнительно-историческом языкознании. Модели дивергенции языков [Glottochronology in Comparative-Historical Linguistics and the Models of Linguistic Divergence] // *Аспекты компаративистики*. [т.] 3 [*Aspects of Comparative Linguistics*. [v.] 3] (= *Orientalia et Classica*: Труды Института восточных культур и античности. Вып. XIX. М.: Российский государственный гуманитарный университет [Papers of the Institute of Oriental and Classical Studies. Issue XIX. Moscow: RSUH Publishers]); pp. 509–536.

Werner 2002 — H. Werner. *Vergleichendes Wörterbuch der Jenissej-Sprachen*. Band 1: *A–K*, Band 2: *L–S,* Band 3: *Onomastik*. Wiesbaden, Germany: Harrassowitz Verlag.

Резюме

Убеждение, что данные современных языков могут быть использованы для реконструкции праязыков с возрастом в 10 000 и более лет, сегодня разделяется далеко не всеми специалистами. Данная статья, частично основанная на результатах совместных исследований российских и американских участников проекта Института Санта Фе «Эволюция языков», посвящена обсуждению методологии макрокомпаративистики и различным способам преодоления связанных с ней теоретических препятствий. Во второй части статьи подводятся основные итоги деятельности проекта, а также предлагаются различные направления дальнейших исследований в области глубинного языкового родства и лингвистической предыстории человечества.