Preliminary lexicostatistics as a basis for language classification: A new approach

The article discusses the basic methodology that underlies the construction of a global lexicostatistical database for all of the world's languages, currently one of the main tasks of the Evolution of Human Languages project at the Santa Fe Institute. The author presents several important modifications of the traditional lexicostatistical procedure, such as: replacing the traditional 100-item wordlist with a more compact list of 50 “ultra-stable” items; use of low-level protolanguage reconstructions as primary construction nodes; a combination of the comparative-historical method and principles of phonetic similarity as the basis for the cognate scoring procedure; and, most importantly, a heavy emphasis on semantic precision and severe restrictions on the use of synonyms.

Keywords: lexicostatistics, taxonomy, comparative method, language relationship, semantic reconstruction, Swadesh wordlist.

1. The issue: how to set up the proper criteria for judging language relationship

For over a decade now, the author of this paper has been involved in the long-term scientific project of establishing an up-to-date classification of the world's languages and understanding how far back in linguistic prehistory it is possible to penetrate by using the comparative method — first, within the framework of the Moscow-based “Tower of Babel” project, later, within the broader “Evolution of Human Language” project, centered around the Santa Fe Institute; the major results and conclusions of EHL have been recently summarized in [Gell-Mann, Peiros, Starostin 2009].

At the moment, these results remain largely unendorsed by what may be tentatively called “Western mainstream linguistics” (tentatively, since the very notion of “mainstream lin-
guistics” eludes any precise definition), mainly due to the current trend in thought that tends to emphasize the importance of language contact and areal convergence over that of genetic relationship (for a solid overview of the interaction between the two in different regions of the world, see, e.g., [Aikhenvald & Dixon 2001]). It may, in fact, be noted that the old distinction between the so-called “lumpers” (i.e., those who believe in the historical reality and demonstrability of linguistic macrofamilies) and “splitters” (those in firm opposition to at least the idea of demonstrability of such macrofamilies) can, today, be all but reinterpreted as a distinction between “heritagists” and “arealists”. Macrofamily hypotheses such as Altaic, Nostratic, Austric, Amerind, Khoisan, etc., are nowadays most commonly declined by their opponents not so much because the similarities between their members are perceived as random (this factor is still frequently wielded as a counterargument; however, the more rigorous work is being done on these hypotheses, the more it recedes into the background), but primarily because their proponents — so we are told — lack the proper means of separating true traces of common genealogical descent from the effects of “horizontal transmission”.

This problem — the difficulty of differentiating between cognate and contact — is, of course, not restricted to hypotheses on long-range comparison; it regularly manifests itself in just about every branch of historical linguistics, which has so far been unable to offer it a uniform, objective solution or set of solutions — or, at least, to set up a certain number of strict “rules of conduct” that all historical linguists would agree to obey when dealing with the issue.

Thorough analysis of available data (first and foremost, Indo-European, later augmented by data from other well-studied families) has shown that, in any comparison of two or more related languages, the best way to distinguish between inherited and borrowed lexical strata is to set up two subsystems of phonetic correspondences — one, reflecting the older inherited layer, will inevitably be more complex and difficult to establish, the other one, representing borrowed items, will be more immediately obvious and consist of generally simpler rules. In this way, it has become possible, for instance, to distinguish between the old layer, inherited from Proto-Indo-European, and the new layer, borrowed from Iranian, in the Armenian language [Hübschmann 1875]; in the same way we distinguish between the “colloquial” — inherited — and “literary” — borrowed — readings of Chinese characters in Sinitic languages (see, e.g., [Starostin 1989: 61–65] for the description of such a differentiation within the Min dialect group).

This criterion, however, is unusable in many types of situations — for instance, when the historical phonetic distance between the languages in question is too small to allow us to distinguish between phonetic laws responsible for vertical transmission and those governing horizontal one; such is the case with, e.g., certain non-literary Dravidian languages (such as Kolami or Gondi), where it is frequently impossible to determine whether a certain item has been retained from the Proto-Dravidian state or borrowed from Telugu. An even more typical situation concerns language families that have not been studied well enough for scholars to arrive at a definitive list of phonetic correspondences, so that distinguishing between any possible layers of lexical interrelation is out of the question. A good example of this is the Khoisan language grouping, where linguistic standards for identifying borrowings are generally substituted for sociological ones [Sands 2001; Gältemann 2006] — i.e., similarity between non-closely related languages is a priori attributed to areal diffusion and contact because (a) areal diffusion as such is known to occur in that region and (b) specialists (for now, at least) are unable to explain it properly in accordance with the canon of comparative-historical linguistics.

Even when rules are set up to differentiate between “old” and “new” correspondences, this does not serve as a guarantee that the “old” layer will be recognized as representing verti-
Preliminary lexicostatistics as a basis for language classification

For instance, the necessity of disentangling the layers of cross-borrowings within Altaic languages has always been recognized by Altaicists as a sine qua non of their field of study, and it is hardly a coincidence that the most recent serious compendium of Altaic etymology [EDAL] attempts to deal with this problem in the very first chapter of its lengthy introduction (pp. 13–21), even before going into the general description of the phonological system of Proto-Altaic itself. In this chapter, the authors take on the issue of two of the most troublesome types of convergence between the descendants of Proto-Altaic (Turco-Mongolic and Mongolic-Tungusic contacts) and, in accordance with the above-mentioned principle, point out the differences between phonetic correspondences that reflect relatively recent contact and those that should rather be interpreted as reflecting relationship; e.g., Middle Mongolian *aǯu ‘fang’, corresponding to Proto-Turkic *ańg, is a borrowing from some form of Old Turkic (aźiǰ), whereas Mongolian ara id. is a genetically related form, reflecting the regular correspondence “Turkic *ŕ : Mongolian r” [EDAL: 16].

This argument per se, however, does not appear sufficiently convincing to many specialists, who put forward the alternate hypothesis — namely, that this different set of correspondences merely reflects areal contacts that belong to an earlier layer; a particularly appealing theory here is that of a series of “Mongolo-Bulgar” relations during which many Turkic words in a specifically “Bulgar-like” shape must have penetrated the direct ancestor of all medieval and modern Mongolic languages [Georg 1999/2000]. Although non-linguistic evidence to support such a claim seems to be lacking, and a systematic linguistic scenario is hard to construct, theoretically, no matter how many different layers of phonetic correspondences we succeed in establishing, nothing prevents us from simply assigning each of them to a different layer of contact relationships, going back as deep in time as it suits our imagination. (The idea that it must take exactly the same amount of “rigorous proof” to justify a situation of historical contact as it takes to justify a theory of genetic relationship, for some reason, is usually missing in works critical of long-range relationship hypotheses — as if there were something wrong with the idea!)

It seems, therefore, reasonable to assert that, in differentiating between inherited and borrowed lexical layers in the language, we cannot rely on “mechanistic” phonetic criteria alone; each situation of alleged “contact” must also be subject to additional scrutiny, conducted from a statistical (“how much has been borrowed?”), sociolinguistic (“what exactly has been borrowed and why?”), and typological (“how often does this kind of borrowings happen?”) points of view. Yet it is precisely these points, particularly the last one, that still remain rather obscure in today’s work on language contacts.

The situation has, perhaps, been best summarized in a frequently quoted passage from a paper by Werner Winter: “the inspection of a wide array of observations... leads to the conclusion that in this field nearly everything can be shown to be possible, but... not much progress has been made toward determining what is probable” [Winter 1973: 135]. The quotation is now more than thirty years old, yet, despite the huge rise of interest in contact linguistics, its intonations still ring true; every now and then, we learn something new about the possibilities of borrowing, but we still have no idea of how to estimate the probability of borrowing on a reconstructed, pre-historic level, because there exists nothing like a general typological framework of contact situations to help us with this task.

Should this, however, mean that, simply because we do not have a fully operational model, the linguist should be prohibited from a genetic interpretation of the facts as the likeliest one, and should such a “ban” be equated with scientific caution and healthy skepticism, or would it rather represent an unnecessary hyper-reaction, inhibiting real progress in historical linguistics? I would say that it depends significantly on the situation, and that it
is our duty to learn to distinguish, as objectively as possible, between different types of situations.

A crucial component of the language on which it is reasonable to base our decisions is, of course, the basic lexicon, and more or less every serious linguist recognizes that the best place to look for non-contact-induced, non-chance similarities is somewhere in and around the Swadesh wordlist. On practice, however, the “skeptics” never fail to remind that the basic lexicon is only more rarely borrowed than the cultural one, and that it is fallacious to automatically count every non-chance similarity on the Swadesh list as reflecting genetic relationship; the very fact that we know for certain that English mountain < French montagne or that Japanese niku < Middle Chinese njuk should be enough to keep us wary whenever we spot any similarities in the basic lexicon. It is, however, never stated precisely just how wary one should be, and what is the “breaking point” at which these similarities should become universally convincing as indications of relationship. Judging by such recent publications as [Yeon-Ju & Sagart 2008], in which it is argued that the Bai language in Yunnan has borrowed as much as 47% of the lexicon from Han-era Chinese (unconvincingly, in this author’s opinion), one should be wary just about always, but surely this is a rather unsatisfying conclusion, were it to be judged as final.

Another equally unsettling problem, but this time coming from the other side — long-rangers’ own elaboration of their hypotheses — is the issue of evaluating competing hypotheses and determining degrees of relationship rather than the simple fact of relationship. Certain evidence exists, as stated in one of our previous publications on the subject ([Gell-Mann, Peiros, Starostin 2009]; the evidence in question is available at http://starling.rinet.ru, the “Global etymologies” database), that suggests deep-reaching genetic relationship between all major macrofamilies of at least Eurasia, and possibly much of Africa and America as well. Within that scenario, supposing it were true (whether it is true does not matter for now), how do we find the means to set up internal subclassification? And how do we choose between mutually contradicting hypotheses, such as, e. g., Starostin’s Sino-Caucasian [Starostin 1984] and Sagart’s Sino-Austronesian [Sagart 2005], or multiple different models of Nostratic/Eurasiatic?

These and certain other issues can all, in fact, be reduced to a single one — the quest for the Holy Grail of historical linguistics: a set of stable, rock-solid “genetic markers”, ones that would be generally stable and guaranteed against the pressures of both internal (ultra-slow rate) and external (resistance to borrowing) change. Since such a set would only make sense if all, or most, of its elements were applicable to all of the world’s languages, it is clear that morphological markers and paradigms, one of the most popular types of data in establishing genetic relationship, cannot be part of it.

The typological approach, such as, for instance, is advocated for in [Nichols 1992] and is currently gaining more popularity in diachronic typology, certainly has this advantage of universal application: languages around the world may lack synthetic morphological markers, but no language is known to lack grammatical meaning as such. Nevertheless, it will probably take a lot more time before historical linguists learn to properly rely on typological data as serious argumentation supporting genetic relationship. For now, we have literally heaps of evidence from all the levels of language — phonology, morphology, syntax, semantics — showing how quickly a genetically non-related language can shift its typology once locked in a Sprachbund with languages from other families.

To quote but one example, it is rather hard to locate a significant number of typological features that would easily separate Modern Chinese in its Beijing form from the Thai language; the reconstructed Proto-Sino-Tibetan, from which Modern Chinese is unquestionably descended, however, looks seriously different from Proto-Zhuang-Tai in many more respects.
Perhaps some time in the future our understanding of linguistic typology and the mechanisms of its evolution will reach such heights that the “inherent” Sino-Tibetan traits of Modern Chinese will become easily detachable from its areal innovations, but for now it is safe to say that not only do we lack a strict set of rules to separate the wheat of genetically significant typological isoglosses from the chaff of typological diffusion, we do not even know where to begin in order to establish them.

2. Some basic thoughts on lexicostatistics

Coming around full circle, it can be seen that, for the moment at least, we still do not have any serious alternative to basic lexicon when it comes to issues of external relationship and internal classification that involve significant time depths. Discarding lexically based classification as such simply because it runs into certain problems will leave us with either classification methods that are even more questionable, or with no classification methods at all. A far more productive approach would be to tackle these problems head-on in an attempt to minimize their negative effects.

The main goal of this paper is to advocate, once more, the use of the lexicostatistical method in both testing hypotheses of relationship and establishing the internal classification of well-demonstrated taxa. In general, I propose nothing new: ever since the popularization of lexicostatistics by Morris Swadesh in the 1950s, it has been used for these purposes over and over again, in many different ways and with widely varying results. The Moscow school of comparative linguistics, in particular, has embraced it as the primary tool due to the works and influence of S. A. Starostin [Starostin 2000, 2007a, etc.], and, in recent years, Vaclav Blažek, working in close association with the Moscow school, has initiated a continuing series of papers [2006, 2008a, 2008b and others] that consistently apply Starostin’s modified formula of “glottochronological decay” to various language families of Eurasia and Africa, with generally credible results.

(It should be quite specifically stressed at this point that I see it fit to distinguish between lexicostatistics, as a procedure that builds genealogical trees based on percentages of cognates on the Swadesh wordlist, and glottochronology, as an “add-on” to lexicostatistics that assigns absolute dates to nodes of separation. I am sympathetic to and, with some technical reservations, generally endorse glottochronology, but my primary concern in this paper and the intended follow-ups is with relative, rather than absolute, chronology, and the use of cognate matching in assessing the chances of genetic relationship. Glottochronological dates will be given from time to time merely for the sake of convenience; they are of no crucial importance for the method I am describing.)

Alternate methods and models of classification using the basic lexicon have recently been suggested by non-linguists based partially on their experience in other branches of science, such as Russell Gray [Gray & Atkinson 2003] and Mark Pagel [Pagel et al. 2008]. All of this means that lexicostatistics is still an active field of study, maybe even more active today than during the “lull” period in the 1970s and 1980s, and that the testing of its scope and general capacities is far from over.

---

2 Prominent representatives of this school who have, over the last twenty years, offered lexicostatistical classifications for various families, include A. Dybo (Altaic), A. Militarev (Semitic, Afro-Asiatic in general), O. Mudrak (Altaic, Chukchee-Kamchatkan, Eskimo), I. Peiros (Austro-Asiatic, Kra-Dai, Sino-Tibetan), E. Helimski (Uralic) and others; unfortunately, only parts of this data have been published officially.
It must be stressed, however, that, as of now, the word ‘lexicostatistics’ itself is frequently applied to two significantly different procedures, causing deep confusion among proponents as well as opponents of the method. This confusion is perhaps best exemplified by the following quotation:

“…glottochronology cannot find or demonstrate remote relationships; rather, in the application of the method, forms which are phonetically similar in the languages being compared are checked/ticked as possible cognates and then, based on the number counted, a date is calculated for when the languages split up. That is, the method does not find or test distant genetic relationships, but rather just assumes relationship and proceeds to attach a date. This is illegitimate for research on possible remote linguistic relationships” [Campbell 1998: 185–186].

Lyle Campbell’s unwillingness to distinguish (at least, on a practical level) between “lexicostatistics” and “glottochronology” is of no great concern in this context, but his use of the expression “phonetically similar” may be so. The original application of lexicostatistics, as demonstrated in the earliest works of Morris Swadesh on the subject [Swadesh 1952, 1955], was essentially limited to languages whose relationship had already been demonstrated through more “conventional” means — such as systematic morphological evidence or the use of the basic comparative method, either thorough (in the case of Indo-European test languages) or partial, but effective (in the case of Eskimo-Aleut). This means that, for Swadesh and everybody else, it is not the forms that are “phonetically similar” which hold the most relevance, but the forms that correspond to each other historically, regardless of whether they remain “similar” or not. Were it otherwise, we would hardly expect words like English eye and German Auge, quite dissimilar phonetically, to be checked as cognates on the list given in [Swadesh 1955].

This original application of the method should, perhaps, be called classic lexicostatistics (CL), and it is strange that, in his rejection of the lexicostatistical procedure as such, Campbell does not even refer the reader to its existence. In the general framework of comparative-historical research, CL constitutes merely the final phase of the lengthy process of suggesting and testing language relationship through other means such as the ones listed above. Once the process is finished, or, at least, has reached a “respectable” stage at which the relationship is no longer doubted, CL is applied to certify the internal classification of the taxon. CL is, therefore, applicable to language families like Indo-European, Uralic, Eskimo-Aleut, or Mayan, for which we know (or mostly know) the phonetic correspondences, but — at this stage — unapplicable to (in comparison) poorly studied families like Pama-Nyungan, Kwa, or Ñe, for which we do not have reliable proto-language reconstructions, even if there is little general doubt of their existence. Even less possible is the application of CL to hypothetical macrofamilies like Austric or Nilo-Saharan, whose very reality is questioned by numerous specialists in the field(s).

The other way of using lexicostatistics — namely, applying it to assembled wordlists before the proper historical research has been performed on them — may be called preliminary lexicostatistics (PL). It is true that Swadesh rarely, if ever, explicitly stated the difference between CL and PL, and if his earliest works, meant to present and explicate the method, did not stray away from well-studied language families, some of his later theories, such as the “Dene-Finnish” relationship [Swadesh 1965], were based on a very crude and superficial application of PL, lacking any conclusiveness whatsoever. This, unfortunately, is one possible reason for the fact that the two procedures have also been mixed in works like [Campbell 1998] and others. Below I summarize the crucial differences between both methods:
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Classic lexicostatistics</th>
<th>Preliminary lexicostatistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object of analysis</td>
<td>Basic lexicon wordlists for 2 or more languages known to be related</td>
<td>Basic lexicon wordlists for 2 or more languages suspected of being related</td>
</tr>
<tr>
<td>Previous research on object</td>
<td>Relationship demonstrated; phonetic correspondences worked out; protolanguage reconstruction performed</td>
<td>None necessary</td>
</tr>
<tr>
<td>Main point of analysis</td>
<td>Cognates scored based on the established system of correspondences</td>
<td>Cognates scored based on phonetic similarity (along with some knowledge of the general typology of phonetic change, if and where possible)</td>
</tr>
<tr>
<td>Main result of analysis</td>
<td>Establishing the internal classification of the family</td>
<td>Confirming relationship (and only then establishing internal classification), or rejecting relationship</td>
</tr>
</tbody>
</table>

Contrary to Campbell’s generalization of PL as the most common understanding of lexicostatistics in general, examples of its application in scholarly literature are quite scarce compared to examples of CL. PL does serve as a major source of classificatory explorations in surveys carried out by members of the Summer Institute of Linguistics (for understandable reasons, given that, for the most part, SIL members work with very poorly studied languages), but very little of their data is actually published in any printed or Internet sources, and, besides, even in their work PL is mostly applied to closely related languages rather than any complicated cases.

I do not, therefore, feel any need to justify the existence and usefulness of lexicostatistics as such; in its CL form the method, applied many times over to relatively well-studied families all over Eurasia and the other continents, has yielded results that are perfectly well compatible with uncontroversial results obtained by other methodologies of classification (such as the “shared innovations” approach), especially with the addition of Sergei Starostin’s correction that loanwords detected on the 100-wordlist must be excluded from calculation ([Starostin 1989]; unfortunately, this correction still remains largely unnoticed by critics of the idea of a constant rate of retention, even though it by and large eliminates the issues raised in [Bergsland & Vogt 1962] that once threatened to bury the idea, but, eventually, only helped to reinforce it). Situations in which CL results enter into direct and sharp contradictions with classifications obtained by different means are, by comparison, rare and indecisive, such as the Austronesian case (see, e.g., [Blust 2000], and the counter-argumentation in [Peiros 2000]); their existence no more discredits lexicostatistics than the existence of alternate Indo-European classifications, all of them supposedly based on the same foundation of “shared innovations”, discredits the very concept of “shared innovations”.

It is also not easy to understand Campbell’s argument that, since lexicostatistics/glotto-chronology simply “assumes relationship and proceeds to attach a date”, “this is illegitimate for research on possible remote linguistic relationships”. The argument is obviously wrong in the case of CL, but even in the case of PL, where its observation on “assuming relationship” is correct, the conclusion remains obscure. Surely every demonstration of relationship, regardless
of the kind of evidence it is based upon, “assumes relationship” and then proceeds to prove it with this evidence. Knowledge or suspicion of language relationship does not fall on us from the sky; we arrive at it through various ways of analyzing data, and one such way can be PL, just as another way could be, for instance, analysis of morphological connections between languages. Perhaps “assumes relationship” is supposed to mean “assumes the relationship as having already been demonstrated beyond doubt by other methods, even though it has not”? But this would be untrue for any application of PL.

The crucial difference between CL and PL — the one that is responsible for widespread application of the former and only marginal and highly controversial application of the latter — is that the former rests on far more rigid standards: reliance on phonetic correspondences rather than phonetic compatibility, working as a solid and, in many ways, objective anchor for the cognate scoring procedure.

Of course, practical application of both procedures shows that, in quite a few cases, the distinction between the two is somewhat blurred, because even for well-studied families like Indo-European, there is always a “fringe” area where uncontroversial etymological decisions are impossible — for instance, do we judge Latin canis ‘dog’ to be cognate with Old Indian çvan-, Greek κυών, etc., despite the blatant discrepancy in vocalism, or do we consider it to be a different root altogether (or, perhaps, a contamination of the old root with some other lexeme, leading to the vocalic irregularity?). Another troubling issue is that, according to the procedure as modified by S. Starostin, we are required to filter out borrowings, and it is not always easy to understand if a particular form that has replaced the old root represents an old “native” morpheme in the language or represents a borrowing.

Nevertheless, it goes without saying that, on the average, the better we understand the history of a given language family, the better we can rely on the CL procedure to provide us with a fairly secure genealogical model for it. Complex cases like the one described above can be dealt with on a semi-formal basis, and it is reasonably safe to assume that they will not distort the picture to the point of rendering it useless, especially when the comparison is conducted not on a binary, but on a multi-lateral basis.

Much more troubling is the realization that, for an absolute majority of the world’s languages, we simply lack the means to conduct CL in any way, because no proper work has been done on establishing a well-defined system of correspondences between them. This does not merely include such “infamous” potential stocks (“pseudo-stocks” from the “mainstream” point of view, which is, technically, not a good term because it intentionally discourages further work on these promising hypotheses) as Indo-Pacific or Amerind, large chunks of which have not even begun to be subject to the appropriate comparative-historical treatment; similar problems crop up with families that are generally thought of as much better understood — e. g. Sino-Tibetan, where the understanding of comparative phonology seriously differs from linguist to linguist (cf., for instance, the many disagreements between models offered in [Peiros & Starostin 1996] and [Matisoff 2003]), or Afro-Asiatic, where some general agreement on the basic correspondences does exist, but the issue of proper matching of cognates still stands tall for each

---

3 I will be using the term phonetic compatibility to refer to situations when two or more words can be judged as cognates either due to their phonetic similarity or because their phonetic shapes, although dissimilar, can nevertheless be reasonably connected due to either our general knowledge of the typology of phonetic change or supporting data from other languages. E. g., to quote an example from the Bongo-Bagirmi group of languages, Bagiro fàɗù ‘fire’ would be phonetically similar to Kenja pòòɗò (the consonantal matches $f : p$ and $d : d$ are quite straightforward) and phonetically dissimilar, but compatible with Mbay hɔr id. (phonetic developments $p > f > h$ and $d (d) > r$ are well-known in the world’s languages; also, cf. such related “intermediate” forms as Ngambay păr and Deme hàɗé id.).
second, if not first, etymology (cf. the numerous discrepancies between, e.g., [Orel & Stolbova 1995] and the more recent and advanced, but still constantly changing, “Database of Afro-Asiatic etymology” by A. Militarev and O. Stolbova, available online at http://starling.rinet.ru).

It may be argued that, since CL is impossible to apply to such families and PL rests on shaky methodology and overestimated intuition, lexicostatistics as such should be ruled out in trying to determine both their internal classification and external relations. But, if so, then what other criterion should not be ruled out? Morphological isoglosses between languages are not a universal means of classification, and, besides, they are only as good as the phonetic correspondences they are based upon — which brings us back full circle: no genealogical classification of any family will be resting upon a rock-solid foundation unless a proper amount of historical research has been previously done on it. On the other hand, researching the history of a language family can hardly be done without at least some idea of the internal structure of this family, leading to a vicious circle of sorts.

Still, there can hardly be anything wrong in submitting compiled lexical data to a PL investigation as long as we do not forget to state that the resulting classification is not “final” or “proven”, but merely a working model — a phylogeny that has to be validated further through much more detailed comparative research. By its very nature, PL will inevitably share some of the flaws of J. Greenberg’s “mass comparison” method — although, as will be shown below, many of them will be greatly reduced or completely eliminated — but an a priori admittance of its relative non-robustness should save us the trouble of engaging in the same kind of spirited debates that have always accompanied “mass comparison”. The statement I want to make is not that “PL is sufficient to establish, beyond reasonable doubt, a general classification of the world’s languages”, but only that “PL is sufficient to establish a general working model of the classification of the world’s languages, prone to refining or refuting, in part or even in toto, through ensuing research founded strictly on the comparative method in its Neogrammari an application”.

Use of PL as a valid technique to form hypotheses on language relationship and classification is not at all new; it has been employed, in various shapes, by many members of EHL as well as other linguists outside the project. The primary goal of this paper is, therefore, not to introduce and promote it as some radically different technique guaranteed to yield quick and ready solutions, but rather to define, as precisely as possible, the exact conditions under which PL, the way I see it, can and should be used to arrive at a preliminary picture of the world’s linguistic situation. First and foremost, this involves answering the following set of questions:

a) What should be the object of PL? How much, and what kind of, data, should the compared wordlists include?

b) What should be the basic principle of cognate scoring? Should it be “phonetic similarity”, “phonetic compatibility”, or something else, and how should we avoid subjectivity in this matter?

c) What is the solution offered for the “common plague” of lexicostatistics — the synonymity issue? Should synonyms be allowed on the list?

4 In particular, the author of the present paper has himself tested one variant of PL on the Elamite language, leading him to reject the dubious theory of Dravidian-Elamite relationship [G. Starostin 2002], and on the hypothetical Khoisan macrofamily, resulting in a preliminary classification of Khoisan as well as the elimination (for now) of Hadza from the phylum [G. Starostin 2003]. The EHL team also possesses numerous 100-wordlists on Papuan, Australian, Siberian, and Native American families that have been subjected to PL treatment (by O. Mudrak, S. Nikolaev, I. Peiros, and T. Usher), although the results are still being refined and not yet ready for publication. Finally, some PL on the “macro-macro-family” level has been performed by S. Starostin [Starostin 2003], although he usually preferred relying on lexicostatistics exclusively in its “classic” form.
d) In the particular situation when the PL procedure is testing potential long-range relationship, should there be any “special” rules for cognate scoring (distinct from the basic rules for testing relationship between chronologically more shallow units)?

e) Is there any particular safeguard about mistaking old contacts for cognates, and what kinds of PL lists would decrease the risk of this happening?

Below I will try to answer, one by one, all of these questions, based on both theoretical considerations and practical results already obtained by myself and my colleagues in the process of applying PL to a wide range of families across the world.

3. Selection and compilation of wordlists for preliminary lexicostatistics (PL)

The first issue to be settled within the general task of applying the common PL procedure to all of the world’s major and minor linguistic families is the degree of shortcutting that will be permissible and reasonable in this procedure. To compile Swadesh 100-wordlists — better still, 200-wordlists; better still, 500-wordlists, etc. — for all languages all over the world is a grand endeavor indeed, but, unfortunately, one that is completely out of the question for now due to serious lack of manpower, working hours, and, above all, reliable linguistic data, or, in fact, any kind of data on at least half of these languages.

Fortunately, such an endeavor is also quite obviously excessive if our main goal lies not in the establishment of a fine-grained internal classification of small, chronologically shallow groups, but rather in the creation of a general framework, within which it will later be possible to ascertain individual relationships with increased precision. To be more exact, we need not be significantly concerned with the inner structure of compact groupings that descend from proto-languages whose age is commonly estimated not to exceed 2,000 — 2,500 years, such as, e.g., Germanic, South Dravidian, Mongolic, Athapaskan, Daju, North Khoisan, etc. The very existence of such groupings is generally undisputed (and, more often than not, intuitively evident even to native speakers), and, for our purposes, it would be more productive to have each such “primary grouping” represent one node on our future “global” tree than to insist upon “maximum splitting”.

One way of achieving that would be to have each such grouping be represented on our tree by just one “diagnostic” member — e.g., have German (or Dutch, or English, or Swedish) represent Germanic, Tamil (or Kannada, or Kota) represent South Dravidian, Khalkha Mongolian represent Mongolic, etc. However, such an approach would be painfully anti-historical to the point of irrationality. Thus, for language groups whose history is relatively well understood, we would frequently find ourselves forced to throw away important data. Limiting ourselves to German as our “Germanic representative”, we would have to note that the word for ‘bone’ is Knochen, and intentionally ignore that it has nothing to do with the common Germanic word *bair-an for this item [Orel 2003: 32]. Limiting ourselves to Tamil, we would have to acknowledge (and, in accordance with the procedure, discard) the obvious Sanskrit borrowing nakam for ‘fingernail’, instead of the perfectly legitimate Common South Dravidian *ugur(u) [DEDR: 55], etc.

Things would work even worse in the case of poorly studied or described language families, where individual languages almost always are less reliable than comparative data. Thus, were we to take Mursi as our representative for the Surmic subgroup of Eastern Sudanic, we would be stuck with the word hoho for ‘heart’, even though the other languages mostly agree in having an entirely different root: Tennet zinzet, Baale simi, Chai hini, Koegu šen, Mešen šini,
Didinga dhinit, etc. Here, not only would we have to discard more important evidence, but we would also have problems with certifying the status of Mursi hoho — is this a native Surmic word or a borrowing from some extraneous source?

For these and other reasons, it seems preferable to have the primary nodes represented not by any “diagnostic” forms from particular languages, but rather by the likeliest common invariant — in historical terms, the protoform for each of the primary groupings.

Usage of reconstructed rather than attested forms in lexicostatistical lists is a slightly controversial, but, perhaps, inevitable application of the method. Its most ardent supporter used to be the late S. Starostin, who was particularly adamant about using reconstructed forms to test hypotheses of long-range relationship [Starostin 2003], an issue which we shall consider in more details below. Most Western linguists have generally refrained from following his example, but this mostly has to do with the fact that, for their particular purposes — usually having to do with building an internal classification for just one family — this was simply unnecessary. Even if we want to build a grand lexicostatistical tree for such a huge family as, say, Austronesian ([Dyen 1965], [Blust 2000], [Greenhill et al. 2008]), we do not require the use of reconstructions: most of the attested Austronesian languages have preserved sufficient quantities of “Proto-Austronesian lexical stock” for us to be able to measure and grade these quantities. But if our aim is to cover the entire globe, this is a different matter; it requires “shortcuts”, and reconstructions are both the most logical and the most honest ones.

There are, however, two obvious questions that arise from using low-level reconstructions. These are: (a) how can we be certain of the validity of the reconstructions, especially for families that have not been well studied in the historical perspective?; and (b) in the case of several alternatives, how do we select the one root to represent the entire family?

The first question requires a special answer, and we will tackle it in the corresponding section; for the moment, let us assume that in general, low-level reconstructions for our list can be obtained relatively easily and with plenty of confidence. As for the second question, it is tightly connected to the issue of dealing with synonymy on the wordlist, and will also be discussed specially. For now, I will simply say that both issues are problematic, but that there also are ways to minimize these problems or, at least, to deal with them on a formal basis.

Now that we have chosen low-level reconstructions\(^5\) as our main object of study, the next obvious issue is quantitative: how many items do we need for our list? The initial consideration would, quite naturally, be to simply use the “classic” 100-item list as originally selected by Morris Swadesh, especially since for many languages, ready-made 100-wordlists are already available.

However, given our stated purpose, it can be argued that use of the entire list will be excessive, both for technical and substantial reasons. From a practical viewpoint, requiring that all the positions on the list be filled in would inevitably hinder the inclusion of quite a few low-level language groups in Africa, America and the Pacific region, where for many languages we only have very short — but, nevertheless, still informative — wordlists collected under specific “rapid survey” conditions. While these wordlists may, and should, be used as valuable data for genetic classification, demands for more data would force us to reject them as evidence, which would hardly be reasonable.

\(^5\) For our purposes, here and below “low-level reconstructions” will be understood as “most probable lexemes with a particular meaning that can be reconstructed for the immediate ancestor of a group of languages that is uncontroversially understood to be related and whose members share, on the average, no less than 50% of cognates on the regular Swadesh 100-wordlist.” It should be noted, of course, that language isolates, having no close relatives, will, in any case, have to be represented by modern attested forms on our list.
Another, more serious, consideration is that for our purposes 100 items may simply be excessive. It has always been clear, both to opponents and proponents of lexicostatistics alike, that some words on the Swadesh wordlist are generally more stable than others (e.g. the words for ‘eye’ or ‘two’ are empirically known to be replaced far less frequently than the words for ‘round’ or ‘yellow’), and this, in turn, led to suggestions about replacing the original Swadesh “stability quotient” of 0.14 (or the “improved” Starostin quotient of 0.05) with individual stability quotients for each item on the list.

An attempt at empirically calculating the individual “stability level” for all 100 items was actually carried out by S. A. Starostin [Starostin 2007a], based on a simple procedure of calculating a “stability index” for the items within a particular family (the general criterion here is the number of different roots that are used within the family to denote the item) and then averaging the indexes across the world (calculations were performed for wordlists of the following families: Afro-Asiatic, Altaic, Australian, Austro-Asiatic, Austronesian, Daic, Dravidian, Indo-European, Kartvelian, Khoisan, North Caucasian, Sino-Tibetan, Uralic, Yeniseian). Since the results have not been published in English, it makes sense to reproduce the resulting list here, ranged from the most stable items to the least stable ones (I omit the 10 “additional” elements to the 100-wordlist that are sometimes used in calculations):

<table>
<thead>
<tr>
<th>1. we</th>
<th>21. one</th>
<th>41. stand</th>
<th>61. meat</th>
<th>81. night</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. two</td>
<td>22. tooth</td>
<td>42. tree</td>
<td>62. road</td>
<td>82. see</td>
</tr>
<tr>
<td>3. I</td>
<td>23. new</td>
<td>43. ashes</td>
<td>63. know</td>
<td>83. walk (go)</td>
</tr>
<tr>
<td>4. eye</td>
<td>24. dry</td>
<td>44. give</td>
<td>64. say</td>
<td>84. warm</td>
</tr>
<tr>
<td>5. thou</td>
<td>25. liver</td>
<td>45. rain</td>
<td>65. egg</td>
<td>85. red</td>
</tr>
<tr>
<td>6. who</td>
<td>26. eat</td>
<td>46. star</td>
<td>66. seed</td>
<td>86. cold</td>
</tr>
<tr>
<td>7. fire</td>
<td>27. tail</td>
<td>47. fish</td>
<td>67. knee</td>
<td>87. woman</td>
</tr>
<tr>
<td>8. tongue</td>
<td>28. this</td>
<td>48. neck</td>
<td>68. black</td>
<td>88. round</td>
</tr>
<tr>
<td>9. stone</td>
<td>29. hair</td>
<td>49. breast</td>
<td>69. head</td>
<td>89. yellow</td>
</tr>
<tr>
<td>10. name</td>
<td>30. water</td>
<td>50. leaf</td>
<td>70. sleep</td>
<td>90. lie</td>
</tr>
<tr>
<td>11. hand</td>
<td>31. nose</td>
<td>51. come</td>
<td>71. burn</td>
<td>91. green</td>
</tr>
<tr>
<td>12. what</td>
<td>32. not</td>
<td>52. kill</td>
<td>72. earth</td>
<td>92. cloud</td>
</tr>
<tr>
<td>13. die</td>
<td>33. mouth</td>
<td>53. foot</td>
<td>73. feather</td>
<td>93. big</td>
</tr>
<tr>
<td>14. heart</td>
<td>34. full</td>
<td>54. sit</td>
<td>74. swim</td>
<td>94. bark (of tree)</td>
</tr>
<tr>
<td>15. drink</td>
<td>35. ear</td>
<td>55. root</td>
<td>75. white</td>
<td>95. sand</td>
</tr>
<tr>
<td>16. dog</td>
<td>36. that</td>
<td>56. horn</td>
<td>76. bite</td>
<td>96. good</td>
</tr>
<tr>
<td>17. louse</td>
<td>37. bird</td>
<td>57. fly</td>
<td>77. fat</td>
<td>97. many</td>
</tr>
<tr>
<td>18. moon</td>
<td>38. bone</td>
<td>58. hear</td>
<td>78. man</td>
<td>98. mountain</td>
</tr>
<tr>
<td>19. fingernail</td>
<td>39. sun</td>
<td>59. skin</td>
<td>79. person</td>
<td>99. belly</td>
</tr>
<tr>
<td>20. blood</td>
<td>40. smoke</td>
<td>60. long</td>
<td>80. all</td>
<td>100. small</td>
</tr>
</tbody>
</table>

Prior to the compilation of this index, Starostin and other EHL/Moscow school members would occasionally rely, instead of or in addition to the standard Swadesh wordlist, on a

---

*See, e.g., [Merwe 1966]. In [Starostin 1989], the idea was reflected indirectly by introducing a special parameter — deceleration of the rate of change of the original wordlist depending on the amount of unreplaced items remaining on the list at any given time — but later on, the method of using individual quotients instead of a fixed one was successfully incorporated by him into STARLING computer software, and is now tested by EHL members and their colleagues (as the “experimental method”) along with calculations based on a fixed quotient (called the “standard method”). In most cases, “experimental” and “standard” calculations yield surprisingly similar results, although the “experimental” method tends to yield slightly earlier glottochronological dates.
shortened 35-item version of it, compiled by S. Jaxontov (the list originally appeared not in any
of Jaxontov’s own publications, but in [Starostin 1991: 59–60]). This 35-item list, in Jaxontov’s
opinion, constituted the generally more stable part of the Swadesh list, and the theoretical idea
behind it was that any two or more related languages always had to show a larger percent of
matches within this section than within the remaining 65-item section, the reverse situation
indicating language contact rather than language relationship. This idea was heartily em-
braced by Starostin in much of his work (in particular, to validate the Altaic theory); more im-
portantly, the 35-wordlist has been used by him as a possible “shortcut” to arrive at a prelimi-

The major problem with Jaxontov’s list, however, has always been that the exact consider-
ations underlying the selection of 35 items out of a total of 100 have never been stated ex-
pressly; it seems that, for the most part, the words had been chosen simply based on his own
linguistic experience, gained from working on the history of language families in one particu-
lar area — Southeast Asia. However, the list from [Starostin 2007a], compiled on the basis of a
somewhat more formal and objective principle, shows that Jaxontov’s intuition has misled him
‘mouth’, ‘dry’, ‘hair’, ‘drink’).7

Now that we stand on somewhat firmer ground in determining which items are more sta-
ble and which ones are not, it is only natural that, for the purpose of establishing a general
classification scheme even for one macrofamily, we do not really need all one hundred items.
To take but one example: S. Starostin quotes 26 cognate matchings between Indo-European
and Uralic on the list [Starostin 2003: 482], but if we split the list into two equal parts — the
generally more stable items 1–50 and generally less stable items 51–100 — the first part, pre-
dictably and in accordance with “Jaxontov’s law”, will yield more matches (17) than the sec-
ond part (9); in addition, these 17 matches are generally less questionable from a phonetic, se-
mantic, and distributional point of view than the other 9. The situation does not change much
if we look at more shallow time depths: out of the 42 direct matches between Finnish and
Saami, 28 belong to the “stable” half of the list, and only 14 — to the “non-stable” part of it.

constitute part of Swadesh’s original 100-wordlist (taken from the second half of the 200-wordlist instead).

A radically different approach has recently been advocated by Mark Pagel and others [Pagel et al. 2007],
who propose to predict “stability” of particular items based on their relative frequency in the language (more fre-
cquently used items tend to be more stable), illustrating this on the example of large lexical corpora drawn from
four Indo-European languages. While it would be rash to claim that frequency of usage has nothing whatsoever
do with “stability”, it is also safe to assume that it is but one of the supposedly many factors influencing “stabil-
ity”. Pagel and his co-authors do not give individual statistics for each word, but it is very hard to believe that, for
instance, the word for ‘fingernail’ in Indo-European (very high stability rate of 0.92, according to Starostin) is used
more frequently by active language speakers than the word for ‘blood’ (very low stability rate of 0.18), or that the
word for ‘new’ (0.90) is used more frequently than the word for ‘many’ (0.19). In addition, what works for Indo-
European will not necessarily work for other language families. Thus, numerals ‘one’ and ‘two’ are almost never
replaced in Indo-European, which may be accounted for by the extremely high frequency of both words; outside
Indo-European, however, we constantly find that the word for ‘two’ has a much slower rate of replacement than
the word for ‘one’ (cf. in Uralic: 1.0 vs. 0.65, in Daic: 0.79 vs. 0.55, in Kartvelian: 0.86 vs. 0.57, in Sino-Tibetan: 0.92
vs. 0.37), even though there is little reason to think that speakers of these languages resort to saying ‘one’ far less
often than they say ‘two’. As attractive as Pagel’s model is on the surface, at this point it cannot be used for any
practical purpose.
To cut a long story short, it is not very likely, given their observedly poor “performance” on shallow chronological levels, that words like ‘road’, ‘swim’, ‘cloud’, or ‘yellow’, to name but a few, will persevere over several millennia\(^9\), yielding precious lexicostatistical information about long-distance relationship. Since there is no general, exceptionless “law of retention” for each individual word, occasional exceptions must and will occur, but their efficiency will be quite low compared to the troubles of compiling full-fledged 100-item wordlists and, more importantly, the troubles of cognate matching between poorly studied families, which will increase significantly for unstable words (any historical linguist who has seriously studied existing reconstructions, or contributed to any of them him/herself, knows how much more difficult it generally is to reconstruct the protoform for ‘big’ or ‘warm’ or ‘root’ than it is for ‘ear’ or ‘eye’ or ‘die’).

It may be argued, in fact, that testing relationship hypotheses on different chronological levels requires wordlists of different sizes. Obviously, if we want to measure the lexicostatistical distance between closely related languages or dialects, such as East Slavic, Scandinavian, Oghuz, or North Khoisan, limiting ourselves to the “stable” half of the Swadesh wordlist will almost certainly result in an incorrect classification: most, if not all, of the words will simply match, and we will get, at worst, a zero degree of separation, at best, minimal degrees that will all lie within the margin of error and tell us virtually nothing. For such purposes, we would definitely need all 100 words, or perhaps, better still, the full original 200-word list. But already for Indo-European, utilizing only the “stable half” seems to yield results that are not too far removed from results of the regular classification based on all 100 items — at the very least, all the subgroupings are “recognized” properly.

The choice of 50 as the “magic number” is somewhat arbitrary, but not entirely so: a 50% discrepancy between the wordlists of two different languages (corresponding, according to the glottochronological formula of S. Starostin, to approximately 3,000 years of divergence time) is generally the threshold beyond which relationship ceases to be “intuitively obvious” and requires resorting to more sophisticated methodology in order to become transparent, and we may reasonably expect that the non-stable elements will, overall, be the first to go, or, at least, will fade away about twice as fast as the stable ones. On the other hand, the 35-item list, previously employed in some long-range calculations, will not be convenient for us if we want to utilize the material of units like Proto-Germanic and Proto-Slavic (on Indo-European territory), or Proto-Ethiosemitic and Common Arabic (on Semitic territory) — there will be way too few differences to be of any statistic relevance. At the moment, 50 items looks like the most promising alternative, by way of compromise between the different extremes.

On the other hand, mechanistically selecting the first half of the list (stopping at the word ‘leaf’) will inevitably lead to certain practical difficulties and imbalances. Up until the number 24, I have no problems with it, but beyond that number I propose nine replacements of “more stable” items by “less stable” ones in order to facilitate the work on both the compilation of the wordlists and the scoring. The following items are to be discarded:

a) ‘this’, ‘that’: first, the wordlist is already heavily biased towards pronouns (‘I’, ‘thou’, ‘we’, ‘what’, and ‘who’ are all included), second, stems like a ‘that’, i ‘this’, etc., are nearly universal, rendering them of little use for global classification purposes, and third and most important, many languages around the world show far more than these two basic degrees of

---

\(^9\) It should perhaps be strongly emphasized that, in the strict lexicostatistical spirit, I am talking about words, i.e. “form : meaning” pairs, not etymological roots, prone to meaning shifts. A root with an original meaning like ‘swim’, ‘yellow’, etc., obviously has a better chance of being preserved over lengthy time periods than the original bundling of its meaning with its form.
deixis (e.g., triple systems like ‘this near’ — ‘this/that neither near nor far’ — ‘that over there’, etc.), complicating the already pressing synonymity issue;

b) ‘liver’: this word, despite its relative stability, is very frequently not included in short wordlists collected on data survey trips, and would have to go missing in quite a few cases anyway;

c) ‘fish’: this item is frequently lacking in desert communities (e.g., it is not attested at all, or represents an obvious recent borrowing, in quite a few Khoisan languages), for the languages of which it will be of no use whatsoever;

d) ‘neck’, ‘breast’: these words are not only at the very bottom of the “stable” list, but they also frequently tend to be sound-symbolic (‘neck’ frequently is the same as or stems from ‘throat’, where onomatopoeic forms like *kur*/*qur* are of little diagnostic value, and ‘breast’ is frequently the same as ‘mother’, representing nursery forms; also, confusion frequently arises as to whether the intended meaning is ‘male chest’ or ‘female breast(s)’);

e) ‘full’, ‘stand’, ‘give’: the semi-abstract semantics of these verbal/adjectival roots has been frequently found a big “nuisance” (they tend to have multiple synonyms where it is frequently impossible to tell the difference), and, in general, it is advisable to have as few verbal roots on the list as possible.

For these reasons, it looks justified to remove these nine items and replace them, respectively, with nine other ones that may not be as stable, yet, on the average, turn out to be less of a bother on practice: ‘kill’, ‘foot’, ‘horn’, ‘hear’, ‘meat’, ‘egg’, ‘black’, ‘head’, ‘night’. I shall not give out detailed reasons for these particular choices; let us simply assume that the swap will hardly make any profound substantial difference, but will inevitably facilitate the overall work process.

We will designate this array of 50 lexical “genetic markers” as the main wordlist (MW), opposed to the original wordlist (OW) that contains all 100 items. The presumption is that the slots on the MW are occupied by low-level reconstructions; these low-level reconstructions, in turn, are generally based on OWs (and, where possible, on even more detailed etymological databases) for the respective low-level families — data that actually allows us to produce a low-level reconstruction, as well as establish the internal classification of the low-level family.

E.g., the MW for “Slavic” looks like [1] *pepel-* ‘ashes’, [2] *pъt-a ‘bird’, [3] *čьrn-ъ ‘black’, etc.; the reconstructions are validated by OWs for several Slavic languages, which not only confirm these reconstructions, but also contain etymological information on other words like ‘all’, ‘bark’, ‘belly’, ‘big’, etc., to ensure more accurate internal classification of Slavic languages.

4. Cognate scoring: a compromise between the comparative method and “phonetic compatibility”

Now that we have established the basic constituency of the MW and the type of information in it (low-level reconstructions), the most important question is setting up the rules for scoring potential cognates. This is tricky, since any such procedure, unless operating on a fully automatic, machine-conducted basis, could easily lead one into the trap of subjectivity. Even well-established families frequently show irregularities that allow for different interpretations  

---

10 M. Robbeets [2005: 50], on the contrary, advocates for an increased use of verbal roots to demonstrate relationship, claiming that verbs tend to be borrowed far less frequently than nouns. Her observation is quite correct, but this advantage is completely annulled by the tendency of verbal roots to be generally less stable within the basic lexicon than nominal ones — a tendency that is fully confirmed by the adduced stability index, where we find only 5 verbs (‘die’, ‘drink’, ‘eat’, ‘stand’, ‘give’) in the upper half and 14 in the lower half, and the ratio is even worse for adjectives (which are frequently undistinguishable from verbs in languages around the world) — 3 vs. 13!
George Starostin

(a typical example would be Latin canis ‘dog’, whose correspondence to Proto-Indo-European *kwon-~*kun- is obviously irregular, but no consensus has been reached on whether the form itself represents an entirely different root or a regional ‘permutation’ of the original entity, and the situation becomes much worse when we start dealing with medium-level (or even low-level) families that have not been subject to a great deal of historical research, not to mention any possible long-range connections.

On the other hand, use of a fully automated procedure, completely wiping out subjective approaches to etymology, would deprive us of the same factor of historicity that we tried to bring in by choosing low-level reconstructions as the main point of entry. Such procedures chiefly operate on the principle of “phonetic similarity” — matching phonemes (usually consonants) across compared languages according to their belonging to one of several distinct phonetic classes — and, in the end, this is exactly what is actually being measured: the degree of phonetic similarity, meaning that, for instance, languages that are in reality more distantly related to each other but more archaic in their phonetic systems may end up as more closely related than languages with innovative phonetic structures.

The major weaknesses of getting history out of the picture are, perhaps, most clearly illustrated by the recent results of the international ASJP (Automated Similarity Judgment Program) project hosted by the Max Planck Institute, whose major aim is presented as “achieving a computerized lexicostatistical analysis of ideally all the world’s languages” (http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm). The selected method — a moderately sophisticated procedure of estimating “degrees of phonetic similarity” between pairs of words — results in the construction of a phylogenetic tree [ASJP 2009] where historically correct nodes are hopelessly mixed with nodes that reflect either areal convergence (e.g. the closest branch to Sinitic turns out to be Hmong-Mien instead of Tibeto-Burmese), differences in the rate of phonetic evolution as mentioned above (e.g. Kota is not recognized as a South Dravidian language, although it most certainly is), or straightforward absurdities (e.g. the closest neighbour of Khoisan languages turns out to be... Kartvelian!)

Participants of ASJP obviously understand these limitations of the method and are able to correctly identify most of the underlying causes [Wichmann et al. 2009]. This understanding, however, does not really answer the inevitable question — of what particular use is the produced tree? The importance of assessing an average degree of “lexical similarity” between the world’s languages without distinguishing between various factors that cause this similarity is quite dubious, since such information cannot be reliably used for any further scientific purposes. And if our specific purpose is to arrive at the likeliest — in the light of available data — genealogical tree for the world’s languages, then the importance of the ASJP assessment drops to zero, as it is quite liable to rewarding us with large quantities of false positives and equally false negatives.

Less “global” applications of various statistical procedures measuring and analyzing degrees of phonetic similarity have yielded interesting, but inconclusive results. Thus, Baxter & Manaster-Ramer [2000] have, based on the comparison of only one phonetic segment (the initial consonant), shown that the number of phonetic resemblances on Jaxontov’s 35-wordlist between English and Hindi exceeds chance expectations and serves, therefore, as proof of relationship (presumably, contact is all but excluded in this particular situation), disproving the popular myth that it is impossible to demonstrate the existence of Proto-Indoeuropean without having access to ancient language data. However, there is no guarantee that the same procedure would work equally well on any pair of languages known to be related11.

11 Baxter & Manaster-Ramer’s method established nine potential cognates between English and Hindi, only five of which were true from a historical point of view. The method determined that number to be sufficient; how-
Recently Turchin, Peiros & Gell-Mann [2010] have tested a similar, but slightly more sophisticated method, with extra safeguarding against the effects of language contact, that seems to yield true positives for the case of Altaic relation. Their case, however, is not one of simply measuring “pure” phonetic similarity between attested languages: the procedure is tested on reconstructions of Proto-Turkic, Proto-Mongolic, etc., meaning that they are not unwilling to take historical information into consideration. Tests that have tried to verify hypotheses of long-range relationship based exclusively on data from modern or historically attested languages — e. g., [Ringe 1992], [Kessler 2001] — have almost invariably failed to come up with any positives (but it must be noted that Ringe does report “weak positive” results for Indo-European and Uralic; somehow, though, even this has not brought mainstream linguistics any closer to a common acceptance of “Indo-Uralic” as a historically valid taxon).

Of course, automatic procedures need not necessarily be as simple as that. In addition to estimating degrees of phonetic similarity between compared words (either absolute or relative), such a procedure can attempt to establish patterns of potential correspondences — essentially, doing much the same things that a real comparative linguist, equipped with knowledge of Neogrammarian methodology, would try to do with a bunch of unfamiliar material. This implies that the algorithm will try to match not merely similar, but, in fact, any consonantal classes, and try to determine those matches that are statistically significant. One such procedure, designed by the author of this paper with the help of programmer Phil Krylov (see [G. Starostin 2008]), does show far more promising results for relatively closely related languages; results report, among other things, a total of 64 out of 77 cognate forms between modern English and modern High German recognized — a number which is further increased to 72 out of 77 when the comparison procedure is extended from binary to multilateral (including lexicostatistical data from other Germanic languages). The algorithm even seems robust enough to recognize some of the “controversial” intermediate level groupings, such as Altaic or North Caucasian (relationship between Nakh-Daghestanian and Abkhaz-Adyghe languages).

On the other hand, the capacity of this procedure is, even at this point, insufficient to match quite a few of the obviously correct etymological decisions that comparative linguists have “manually” established over the years. The main reason is clearly the insufficiency of data present on the 100-wordlist. For instance, the algorithm was incapable of understanding the cognacy of English *mouth* and German *Mund*, because the regular correspondence “English zero : German n” (more precisely, of course, “English th : German nd”) could not have been substantiated by any other examples. Stepping outside the wordlist, it is easy to ascertain that the correspondence is indeed regular even without resorting to the more archaic stages of both languages (cf. such examples as other : ander, youth : Jugend, lithe : linde, un-couth : kunde, etc.), but this would require having the algorithm run through the entire compared vocabularies and, in addition to valuable information, picking up a huge lot of “noise” (false cognates, shared borrowings from third sources, etc.) that could seriously distort the desired results.

The conclusion is that “rough” automatic data handling is, at present, unable to arrive at the same level of precision in its results that can be provided through manual handling of the same data; the obvious benefit of “weeding out subjectivity” does not fully compensate for the
lack of fine-graining analysis techniques — techniques which, more often than not, are a very serious influence on classification schemes. This does not mean that automatic procedures should be abandoned; on the contrary, one of our major goals should be to refine and readjust them in accordance with the basic principles of historical linguistics. In the meantime, however, we can only place more trust in manual procedures, all the while attempting to enforce maximally formal criteria. In other words, it may be too early to teach the machine to behave like a human, but it is, in some respects, easier to make the human behave like a machine.

Therefore, for our classification based on 50-item wordlists we will ultimately be relying on manual rather than automatic cognate scoring. This gives us the important bonus of being able to use all kinds of historical information and reliable historical conclusions accumulated over two hundred years of incessant work by specialists in language comparison. The two basic principles of scoring will be defined in the following way:

1. For language groups already studied by the comparative method, judgements about the cognacy of particular items will be made on the grounds of recognized **regular phonetic correspondences** between said groups.

2. For language groups that lack serious comparative study, judgements on cognacy will be made on the grounds of (a) **phonetic similarity** of the items concerned, or (b) **phonetic compatibility** of the items, provided it is possible to base the judgement on **traces of regularity**.

Both points require more precise comments. First of all, it must be made clear that in a lot of situations it is hard to make a clear distinction between the two types of scoring. “Historically studied” is not an absolute definition: no two language groups in the world have received a completely equal amount of study, and our knowledge of the regularity of correspondences is always relative rather than absolute. Even Indo-European is prone to cases where it may be reasonable to sacrifice regularity and resort to scoring on the grounds of phonetic similarity instead.

Case in point: do we judge Old Indian **ḥṛḍ** ‘heart’ as cognate to Germanic **xirt-**, Slavic **sǫrdь-ce**, Greek κηρ, etc. ← IE **ǵhrd-**, or do we score it differently, since it violates the regularity principle (the Old Indian form should reflect IE **ǵhrd-**)? In Pokorny’s dictionary, an authoritative but by no means dictatorial source, the Indo-Iranian root is judged to represent a separate “Reimwort” [Pokorny 1958: 580], not to be related to **ǵhrd-**. Intuitively, however, it is extremely hard to think of the two variants as having nothing to do with each other — apart from complete regularity in every other respect, there is also the important issue of **representativity**: the two variants are in complementary distribution throughout Indo-European, and no non-conjectural evidence can be found as to their co-existence in at least one branch of the family. Hence, probably, the “compromise” solution of **ǵhrd-** as a “rhyme word”, adopted by Pokorny — a solution that achieves nothing, since nothing is explained about the mysterious
origins of this “rhyme word” (did it exist in Proto-IE? was it an original concoction on Indo-Iranian grounds? how did it originate? are its origins related to the existence of *krd- or is it just a fortunate coincidence? etc.), but at least spares the author from the painful Neogrammarian duty of declaring the phonetic similarity between the two variants as the result of pure coincidence.

The representativity criterion — which, in this case, merely represents a particular application of Occam’s razor — would strongly speak in favor of judging the Old Indian form as cognate with the rest of Indo-European. The exact reason that underlies the irregularity remains unknown, with several ad hoc explanations possible (idiosyncratic development of some old non-trivial cluster, perhaps with a laryngeal; assimilatory influence of two ensuing voiced segments; analogy/contamination with some other word; taboo, etc.) but none of them supported by strong independent arguments. But the assumption of a lexical replacement in this case would reduce the Neogrammarian model to absurdity, and, more importantly, leave us with a far larger number of unanswered questions (see above) than the assumption of an unexplainable irregularity.

Therefore, in making cognation decisions even for families with a generally elaborated historical phonetics and a large etymological corpus, it is reasonable to allow for occasional irregularities in the forms, especially when the two irreconcilable variants appear to be in complementary distribution and there is no easy way to “explain away” one of the variants as having an entirely different origin. A demand for utmost mechanistic rigor will inevitably result in our throwing away true historical cognates and coming up with unnecessarily distorted classification schemes. We may formulate the main rule of exception as follows:

1a. Phonetic irregularities between potential cognates within groupings for which a system of phonetic correspondences has been established may be ignored if [a] they concern not more than one consonantal segment of the root (out of two or more), [b] the phonetic distance between the two segments does not make them phonetically incompatible, [c] the two variants — “regular” and “irregular” are in complementary distribution across languages and cannot be clearly shown to fall under two different etymologies14.

Concerning the second type of situations — those for which comparative studies are in their initial phases, or non-existent — it is also easier to illustrate the exposed methodology with real examples, this time taken from the African area. Let us consider the following forms from various “branches” of the hypothetical Nilo-Saharan macrofamily, all of them with the meaning ‘to drink’15:

a) East Nilotic: Teso akr-mát-à, Turkana akr-mat, Nyangatom xe-met-, Karimojong aki-mát, Maasai, Sampur a-mát, Ongamo-mát-à, Lotuko a-mádh-à, Oxorok mat-à, Lopit mád-à, Dongotono a-mát, Lokoya a-mát-á. East Nilotic is a relatively compact and well-recognized language family, with a preliminary reconstruction published by R. Vossen, who reasonably reconstructs this particular root as PEN *mat- [Vossen 1982: 356], and there are no grounds to doubt that it functioned as the main root for ‘drink’ in that proto-language. (It is unclear if the Bari subgroup form *mō-ʒu is also related — probably not, but in any case it will not affect our selection of *mat-, since it is overall better represented in the family).

14 A counter-example to ‘heart’ would be the case of Slavic *kostь ‘bone’ vs. IE *(H)ost- id. → Hittite hastai, Old Indian asthi-, Latin os, etc. Not only is the correspondence “Slavic *k- : IE zero” completely irregular and phonetically incomprehensible, but, more importantly, IE *(H)ost- is easier relatable to Slavic *ostь ‘sharp edge, awn’, while Slavic *kostь is better etymologized together with Latin costa ‘rib’. There may have been semantic contamination between the two words in Proto-Slavic, but there is little reason to doubt the presence of two roots on the IE level, and the Proto-Slavic item on the list should be scored differently from the rest.

15 Since this is merely a methodological example, I do not quote all the data sources for particular forms so as not to inflate the list of references too much. Only the sources for protoform reconstructions are quoted.
b) West Nilotic: Nuer, Shilluk *math, Anywa *màath, Luo *màd-, Pàri *maath, Lango *mato, Mabaan *màča, Jumjum *maŋ-ŋà. All the forms are clearly related (with nasal assimilation in Jumjum), and, although no special published reconstruction of West Nilotic is available, we may safely follow G. Dimmendaal in setting up the proto-form *maŋ [Dimmendaal 1988: 38].

c) Surmic: Chai *mat, Koegu *amátiya, Me’en *mad-. These three forms are phonetically similar and most likely related, even though we have so far had no attempts at a Proto-Surmic reconstruction. We may provisionally set up a reconstruction *maT-, indicating lack of knowledge about the exact manner of articulation of the intervocalic coronal stop.

d) East Jebel: Aka *mèet, Molo *moot, Kelo *mad-æ, Beni Sheko *mådi, Gaam *maδ-. This is also a well recognized language group, and we feel justified following M. Lionel Bender’s preliminary reconstruction *mVt- [Bender 1998: 56].

e) Berta: *meera. Berta is an isolated cluster of several extremely similar dialects, with no uncontroversial “relatives” to speak of (C. Ehret thinks of it as the closest relative of East Jebel, but this classification is highly disputed).

f) Central Sudanic: Moru *mɛvu, Avokaya, Mådi *mɛvu, Logo, Keliko, Lugbara *mɛvù, Lulubo mbù, Lenu *mbu, Ngiti *mɛvù, Mangbetu *mbuo, Kresh *mɛj, Aja *amú. This is one of the primary roots for ‘drink’ in this large language family, and its proto-invariant should be approximately (for lack of an overall credible Central Sudanic reconstruction) reflected as *mɛvu (Ehret [2001: 275] reconstructs East Central Sudanic *mbu, but the root has a wider distribution, since Kresh and Aja are not ECS).

All of these six branches are included by J. Greenberg within his “Nilo-Saharan”, and this decision is upheld by such prominent Africanists as M. Lionel Bender, C. Ehret, and others. However, at the moment, only the relationship between (a) and (b) happens to be completely uncontroversial. The grouping of Surmic and East Jebel languages together with the large Nilotic family as separate units of “Eastern Sudanic” is generally questionable; the grouping of Berta within the same family even more so; and the relations of the whole ensemble, on a seriously “macro”-level, with Central Sudanic, are a problem of about the same scope as Nostratic or Austric relationship, if not more so.

In the light of this, we approach all of these groups as potentially related, but consider this relationship, for the moment, insufficiently substantiated through the comparative method, meaning that the situation here clearly falls under type (2). The scoring will, therefore, be conducted as follows:

— West Nilotic *maŋ and East Nilotic *-mat- are scored as cognates, based on phonetic similarity as well as preliminary correspondences, established in [Dimmendaal 1988] and elsewhere;

— Surmic *maT- and East Jebel *mVt- are also scored as cognates both between themselves and with Nilotic, based on phonetic similarity;

— Berta *meera, theoretically, could be scored as cognate to all four. However, there is a serious problem with the second consonantal segment: it belongs to a somewhat, if not crucially, different consonantal class16, and, in order to be more secure about the cognacy, we need to support it by finding traces of regularity, i.e. at least one or two more exact or near-exact se-

16 On the basic principles of classifying consonants into non-intersecting “classes” based on similarity of articulation, see [Baxter & Manaster Ramer 2000; Dolgopolsky 1986; G. Starostin 2008]; proposed models frequently differ as to the degree of detailization (e.g., do we place such front consonants as t, s, r in the same class or in three different ones?) — I would opt for a more detailed classification, so that such forms as [pata] and [para] be judged phonetically compatible rather than phonetically similar, and require the presence of additional “traces of regularity” to be scored as cognates.
mantic matches — not necessarily within the Swadesh wordlist — that would support the correlation. So far, I have been unable to do that, and this means that Berta *meera*, for now, has to be judged as a different root\(^7\);

— the Central Sudanic forms certainly share the initial consonant with the rest, but there is no evidence (for now) that the protoform at one time suffered the loss of the root-final coronal consonant, or, vice versa, that the Eastern Sudanic form had, at one point, become expanded through the addition of some sort of coronal suffix. There is also no question here that these forms should be scored differently from Nilotic/Surmic/Jebel, on one hand, and Berta, on the other.

Let us now take a different example, one that illustrates how “traces of regularity” can influence the scoring. In Khoisan languages, the word for ‘star’ is represented in the Northern (Ju) and Southern (!Wi-Taa) families by two roots that are significantly different as to their segmental structure:

— North Khoisan: Ju’hoan *tũ*, Ekoka !Xù !lũ, etc. ← Proto-NK *ṭũ (* = alveolar click);
— South Khoisan: !Xóõ ||ona, N-Jul ||q=e-si, etc. ← Proto-SK *||?lo- with different suffixes (actually, not fully clear if !Xóõ and N-Jul forms themselves are related, but our main concern here is !Xóõ; || = lateral click).

The biggest obstacle that prevents us from scoring NK and SK as cognate forms is the difference in click articulation, which cannot be overlooked, since clicks are as different from each other as “regular” consonants with different manners of articulation. Cf., however, the following additional comparisons, relatively easy to come by: Ju’hoan *ṭu* ‘cold’ : !Xóõ ||aũ id., Ju’hoan *q’e* ‘young man’ : !Xóõ ||qũ ‘new, young’, Ju’hoan *ṭh* ‘old (of things)’ : !Xóõ ||hɑ ‘old, mature’. These (and other) examples — impeccable from the semantic side and quite convincing phonetically as well — show that, despite the dissimilarity, there is reason to consider this set as displaying traces of regularity. The obstacle is, therefore, overcome, and we can safely score the forms for ‘star’ as cognate.

It is important to stress that the requirement of *traces of regularity* is more lax than that of a complete system of regular correspondences, but should not be underestimated. The principal difference is that finding traces of regularity does not require us to thoroughly explore all the lexical evidence of the compared idioms and present a detailed reconstruction. But it does require us to demonstrate that our comparison is not completely ad hoc. It is not enough to take Proto-Japanese *pa* ‘tooth’ and compare it with Proto-Dravidian *pal id., saying “final -l probably got lost in Proto-Japanese”; at the very least, it is necessary to find and quote several other transparent examples in which Japanese loses its final or intervocalic *-l- compared to the rest of Altaic, such as Japanese *á- ‘receive’ = Tungus-Manchu *al- id., *k̡- ‘to come’ = Turkic *gel-id., *kái ‘hair’ = Turkic *Kɪl etc. (examples quoted from [EDAL]).

Obviously, scoring two or more forms as ‘cognate’ based on PL-related considerations of similarity or compatibility is not the same as demonstrating “beyond reasonable doubt” that said forms are cognate. Nevertheless, if this procedure is relatively strictly adhered to, it is to be expected that mistakes in scoring will be reduced to a minimum, and, furthermore, their negative effect will decrease in direct proportion to the number of language families enlisted in the scoring, since a global perspective will tend to “even out” individual distortions.

---

\(^7\) Ehret [2001: 282] finds the correspondence between Berta *meera* and the East Jebel forms (but not the Nilotic ones!) to be regular, reflecting Proto-Nilo-Saharan *ŋ* (the entire root is reconstructed as *mę*: ‘to lick’). However, I have been unable to find any other satisfactory examples for this correspondence, and have every reason to doubt its regularity (unfortunately, similar situations arise with a great many more examples of particular correspondences given in this work, which cannot be said to give a reliable account of Proto-Nilo-Saharan historical phonology).
5. The issue of synonymity on micro- and macro-levels

One major problem that has pursued lexicostatistics and glottochronology from the very beginning is that of choosing, for a particular language, the correct equivalent for each item on the Swadesh list — and sometimes realizing that a single choice is all but impossible to come by, since “for many items on the list, languages often have more than one neutral equivalent” [Campbell 1998: 181].

This problem is very frequently exposed in works that are critical of lexicostatistics, sometimes in a very grave tone, as if its very existence automatically rendered the whole method useless. In reality, there are multiple reasonable ways to overcome it. For instance, S. Starostin, in all of his writings and calculations, advocated to disregard the issue as such and simply include both (or even more than both) synonyms in the calculations; e. g., if, for a particular item, language 1 yields synonymous lexemes A and B, and language 2 yields B and C, the situation should be qualified as “lack of replacement”, since at least one out of two different synonyms is the same in both languages.

This solution is highly practical, but may create an uncomfortable illusion of “lack of rigor”. Alternatively, one can simply tighten the demands by more precisely specifying the semantics of the “Swadesh notions”, whose principal flaw arguably lies in their having been originally rendered in standard English, thus reflecting all the ambiguities of that language. E. g., a word like ‘hair’ is quite problematic, since it can be understood in at least three different ways: (1) ‘hair’ as material, i. e. ‘wool, body hair’; (2) ‘hair’ as collective ‘head hair’; (3) ‘hair’ as a singulative noun, ‘one hair’. Quite a few languages have a different root for each of the three meanings, and entering them all as synonyms would clearly be excessive. The “default” (i. e. most frequent) usage would probably be (2), and this is the more precise meaning that I would advocate for the word — but it would be hard to get linguists all over the world readily agree upon one universally approved semantic standard 18.

Nevertheless, for the purposes of our global PL enterprise, conducted in accordance with a single standard, all of these technicalities are easily overcome, so that the issue of making the right choice with historically attested languages will depend exclusively upon the quality of known lexical descriptions for these languages.

In our situation, however, there exists a much more serious and important problem that also has to do with synonymity: selection of the appropriate synonym for the protolanguage form, both on low levels that serve as the starting nodes in our tree and on higher ones. The seriousness of this problem, in fact, goes way beyond the needs of lexicostatistics, as it is directly tied in with the whole issue of semantic reconstruction in historical linguistics — a sphere that, even today, is still barely tapped, despite certain theoretical breakthroughs, achieved above all in the works of J. Trier [Trier 1981] and in A. Dybo’s monograph on semantic networks [Dybo 1996].

Even limiting ourselves to low-level reconstructions and a total of 50 most stable items, we will frequently fall upon cases where it is difficult, or even impossible, to ascertain one particular choice over the other (or, perhaps, even more than the other ones). Only in two types of situations do we find ourselves in a relatively secure position; these types have been explicitly formulated in [Kogan 2006], an article specifically dedicated to the issue of reconstructing a reliable wordlist for Proto-Semitic, but whose methodology is equally applicable to any other language family:

18 Several recent sessions of the Nostratic seminar were dedicated to this particular issue, and a paper suggesting a set of more precise specifications for meanings on the Swadesh list — based on setting these meanings within particular sentential contexts — is under preparation by A. Kassian.
“If a PS (Proto-Semitic — G. S.) root functions with the same basic meaning in all Semitic languages, there is hardly any reason to doubt that it did so also in the proto-language... the same conclusion can be safely achieved if the root in question lost its basic function in a limited number of languages or minor subdivisions... finally, if a term is lost in some languages of a minor subdivision but persists in others, its archaic status is strengthened” (p. 465);

“if a PS root functions as the main term for the respective basic notion in several geographically distant languages without special genealogical proximity, it is likely that this meaning goes back to the proto-level. In this case, too, it is usually preserved as peripheral in other languages and, importantly, no alternative basic term suggests itself” (p. 474).

Based on the first criterion, Kogan is able to reliably fill in 39 slots on the 100-wordlist; based on the second, he adds 12 more, bringing the total up to 52. Even without looking, I can reasonably predict that significantly more than half of these words will belong to the 50-item wordlist specified above, and, indeed, 38 of Kogan's semantically reliable Proto-Semitic items coincide with elements on that “ultra-stable” half of the Swadesh wordlist. Since, in general, I agree with both of Kogan's criteria, this means that, for our PL procedure, the problem of choosing the correct entry for (at least) low- and mid-level reconstructions will not be a critical one.

Nevertheless, we still have to find some way to deal with the remaining 12 items, i.e. cases where descendant languages display way too much variability in order to allow for an unambiguous reconstruction. First, it is quite possible to add a few more internal criteria that may raise the chances of a particular choice. These include:

(a) The criterion of internal etymologization: if we have a choice between two items, one of which shows a clearly derived (most likely, recently derived) semantics, while the other one does not, it is the second item that has a better chance of preserving the protolanguage state.

For instance, in trying to establish the proto-root for ‘meat’ in Samoyed languages, we find that the main South Samoyed form (Selkup wêçî, Kamassian uîja ← Proto-Samoyed *âjâ [Janhunen 1977: 17]) differs from the main North Samoyed form (Nganasan ñâmsu, Enets ud’a, Nenet jëmzâ ← Proto-Samoyed *ômsâ [Janhunen 1977: 15]). Without any additional information, selection of the more representative variant is impossible. However, we have every reason to think, following Janhunen, that *ômsâ is, in fact, a nominal derivative from the verbal root *ôm- ‘to eat’ [ibid.]. There is still a chance, of course, that *ômsâ had already been formed and acquired the meaning of ‘meat’ on the Proto-Samoyed level, after which a root *âjâ, of unknown origin, mysteriously replaced it in Proto-South Samoyed; but since we have no clue as to where *âjâ actually came from, yet have every clue for internally etymologizing *ômsâ, it is more reasonable to think of the former as an archaism and of the latter as an innovation 19.

(b) The criterion of polysemy: if one of the roots has several different meanings across languages, while the other one only has the “Swadesh meaning”, this may mean — although it also depends on the representativeness of both forms — that the latter is the more archaic. Case in point: Lettish jaûns means either ‘new’ (of a thing) or ‘young’ (of a person), whereas in Lithuanian jáunas is used exclusively to denote ‘young’ (people), and in the “Swadesh meaning” of ‘new (thing)’ we have the more archaic naũjas.

(c) The criterion of borrowing: if we can reliably show that one of the competing roots is a borrowing from a distantly related or non-related language, this obviously raises the chance of

19 A more detailed analysis shows that both lexemes can actually be traced back to the Proto-Samoyed level, since we also find Selkup apsî (← *ômsâ) in the meaning ‘food; body’, as well as Enets aîja (← *âjâ) ‘flesh’ (not the default Swadesh notion of ‘meat’, for which ud’a is used, as specifically indicated in the Uralic wordlists compiled by E. Helimski). This only confirms the conclusion reached without considering this additional evidence.
George Starostin

the non-borrowed item. Examples are numerous; cf., e.g., the abovementioned case of Tamil
nakam ‘fingernail’ = Malayalam nakham id., both forms replacing the older root ukir = Kannada
ugur, Tulu uguru etc. Since the Tamil and Malayalam forms are transparent borrowings from
Indo-Aryan, this leaves Proto-South Dravidian *ugir as the likeliest candidate for ‘fingernail’ at
that stage.

Nevertheless, all of these criteria have a significant drawback: the reverse situation, in all
three of these cases, is not much less probable. It is not at all excluded that derivation, poly-
semy, or borrowing could have already been present at the proto-level of the families that we
are dealing with, and that new roots were introduced into specific subgroups later, obscuring
the situation. Such solutions are, overall, uneconomical, prompting us to set up extra “dark
horses” that are, in fact, unnecessary (such as, e.g., an obscure “para-Samoyed” substratum
that donated the root *äjä), but they cannot be excluded.

This means that the most important criterion for settling ambiguous cases must be the
external criterion, which we may formulate as follows:

Where two or more equal or near-equal choices are possible for the proto-item, strong priority is
given to one that demonstrates the most reliable external genetic connections.

Let us illustrate this on an example from the Germanic group. Germanic languages have a
wide variety of roots for the notion ‘meat’: Scandinavian *kiut- (→ Icelandic kjöt, etc.), West
Germanic *flaiska- (→ Dutch vlees, German Fleisch, cf. also English flesh, etc.), English meat =
Old Norse mat-r ‘meal’, etc. However, out of all this variety, unquestionably the best candidate
for Proto-Germanic ‘meat’ would be the ancestor of the Gothic form mimz — even though,
apart from Gothic, neither the form itself, nor even any different forms with the same root
have been attested in any other Germanic language.

The reason, of course, lies in the external connections of mimz: it is a perfect phonetic and
semantic match with such forms as Old Indian māṁs(a)-, Armenian mis, Albanian mish, and
Proto-Slavic *męso, all of them related and pointing to Proto-Indo-European *mens- as the
original form. Assuming that *mimz(a)- continued to be used in that function in Proto-
Germanic, we conclude that it was preserved in the Gothic branch of this family (apparently,
until the very end, cf. Crimean Gothic menuš id.), but replaced by different other roots in the
other branches. Assuming the opposite — that it is Gothic mimz that represents a semantic in-
novation — we would have to conclude that Proto-Germanic lost the original semantics of the
Indo-European root, and then restored it in the case of Gothic: a highly unlikely situation, very
rarely (if ever) observed in or surmised for the world’s languages.

There is one obvious and significant problem with this criterion: if it is our aim to use PL
as a means of verifying hypotheses on language relationship and establishing a global classifi-
cation of the world’s languages, how can we allow ourselves to use external data as if we al-
ready knew everything about these relations? Let alone Indo-European, how is this criterion
supposed to work in areas such as America or Papua, where external connections even on
relatively low time depths have been studied so poorly? And is this not, overall, a typical ex-
ample of poorly masked circular logic?

It goes without saying that the external criterion has to be applied very carefully. The best,
and most certain, type of situation in which it may be employed is a sort of “bootstrapping”
mode, in which “proto-list” reconstruction and cognate scoring is achieved in two stages. First,
we only populate those slots on the list for which internal data suggest a non-ambiguous can-
didate, leaving the problematic slots empty. Then we run the first stage of preliminary scoring,
establishing its likeliest external relatives. After this has been achieved, we can now use exter-
nal data to try to solve the internal problems of the low-level family, i. e. populate its “dubi-
ous” slots with those roots that better fit in with the external data.

In the case of Germanic, for instance, we have little methodological reason to worry about
the selection of *mimz(a)- as opposed to, e. g., *flaiska-, simply because the unambiguous en-
tries on the Germanic list — of which there are plenty — clearly demonstrate the Indo-
European character of Germanic. Other situations may not be as immediately transparent, but
careful application of this “two-step” principle is possible practically in all cases.

Of course, it may — and will — frequently happen so that the external criterion is unable
to help us as well, if none of the candidate items have any significant external matches. In the
same Germanic subgroup, for instance, there are at least four or five different roots denoting
‘tail’, but not a single one has any serious ‘tail’-type parallels in other branches of Indo-
European (almost all of which have their own problems with this infamously unstable — in
Indo-European — notion). This means that neither internal nor external data allow us to make
a choice. In this case, for internal needs we should leave the slot open, but for external needs we
may choose any of the forms — it does not make a difference whether it is *swanka- (→ German
Schwanz), or *tagla- (→ English tail), or *xalēn (→ Icelandic hali), because, regardless of our
choice, we will have to count it as a non-match with the rest of the Indo-European subgroups.

We now come to a less obvious, but equally challenging issue that awaits us on levels of
“middle” time depth (such as Indo-European or Semitic), and even more so with macro-family
relationships. Since we are establishing our classification “rung by rung”, it is important to
establish the likeliest candidates for proto-items on every level, i. e. figure out such a candidate
for Indo-European before starting to probe Nostratic, and for Semitic before starting to probe
Afro-Asiatic.

In order to do this, we accept Kogan’s criteria as quoted above, and expand them with
several internal criteria (also quoted above). Note, however, that the second criterion has an
important catch: “...importantly, no alternative basic term suggests itself”. What if, however,
an alternative basic term does suggest itself?

Let us suggest that we have a language family descended from proto-language L, con-
sisting of four branches: A, B, C, D. Out of these four, for a certain Swadesh item N on our list
branches A and B share one cognate (let us call it *X), whereas branches C and D share a dif-
f erent one (let us call it *Y). Let us now suppose that we have already run through the first
stage of scoring for the entire family. If the resulting tree structure looks as follows:

L
  / 
A_X B_X C_Y D_Y

— this is in full agreement with our information on item N. In this case, internal data are con-
sistent, although we will have problems understanding which of the two roots — *X or *Y —
has to be posited at the top node; in order to do this, we will probably have to resort to exter-
nal data. However, it is quite possible that our overall tree structure, based on an overall as-
essment of the lexicostatistical data, will look quite differently, e. g. the following way:

L
  / 
A_X C_Y B_X D_Y
Such a tree would not be in very good agreement with the behaviour of *X and *Y, and would require one of four historical explanations:

(a1) *X and *Y were easily interchangeable synonyms in protolanguage L, as well as the intermediate protolanguages for AC and BD. The situation changed drastically only after the second split, with each of the four new branches “wiping out” one of the synonyms. The four “eliminations” could have been completely and utterly independent, or

(a2) the result of two areal lexical isoglosses that caused the loss of *X in geographical area AB and the loss of *Y in geographical area CD.

(b1) The regular word for notion N in protolanguage L, as well as the intermediate protolanguages for AC and BD, was *X, whereas *Y was semantically close, but not an exact synonym (or vice versa). After the second split, *Y replaced *X in branches C and D, but not in branches A and B. The two replacements could have been completely and utterly independent, or

(b2) the result of an areal semantic isogloss that affected the (supposedly contiguous) geographical area occupied by speakers of C and D, but not of A and B.

Needless to say, explanations (b1–b2) by default look more promising than explanation (a), since they require fewer assumptions (two independent or one common areal replacement vs. four independent or two common areal replacements). Moreover, explanation (a) requires us to set up freely interchangeable synonyms for Swadesh notions, a situation that is typologically rare and should better be avoided in reconstruction. Cases of such “semantic criss-crossing” are not frequent in non-controversial, low- or mid-level families, but they do exist, and it is strange that works on lexicostatistics have so far overlooked the existence of this problem.

A good actual illustration would be the word ‘moon’ in Indo-European languages. The most common and, undoubtedly, archaic root to express this notion is IE *mēns-, yielding Old Indian mās, Iranian *māh-, Baltic *men-, Slavic *měsčŏ, Germanic *mēn- etc. [Pokorny 1958: 731–32]. However, Armenian lusin, Latin lūnā, and certain Slavic forms going back to Common Slavic *lū-nā reflect a different root, usually — and with perfect reason — etymologized as IE *louk-s-nā, derived from the verbal root *leuk- ‘to shine’ and further compared with such forms as Avestan raax-š-na- ‘shining’, etc.

Trying to explain this as a common Armenian-Latin (or Armenian-Latin-Slavic?) isogloss is out of the question; “areal” explanation is excluded[^20], and no other evidence exists to justify the postulation of a special “Armenian-Latin” node within Indo-European. This is, therefore, a typical example of “semantic criss-crossing”, which we can attempt to solve in either of the two ways described above.

First, we can think of *mēns- and *louksnā as two freely interchangeable synonyms already on the Proto-IE level. This is, however, not realistic. Such a situation is not reflected in any of the attested descendant languages, which either only have one of two terms or feature a sharp semantic distinction between the two (as in Latin lūnā ‘moon’ vs. mensis ‘month’, or Russian луна ‘full moon’ vs. месяц ‘crescent moon; month’). Even if we think of a possible stylistic dif-

[^20]: Unless, of course, we declare ourselves adherents of the strongest version of the “wave theory”, according to which “Proto-Indo-European” as such never existed as even a minimally coherent linguistic entity, and that all of its twelve or so main branches have always been, in some ways, distinct from each other, co-existing peacefully on a small piece of territory before dispersing. Such a scenario, rendering useless the very idea of a genetic tree (and replacing it with what would much for just about any “areal” isoglosses between just about any two or more branches of Indo-European, but I regard it as completely absurd and unsubstantiated by hard evidence, more of an artificial “easy way out” of the need to unravel the complex web of genetic and areal isoglosses between different branches of Indo-European than a solid model that makes real historical sense.
ference — e. g., *mēns- as the “neutral” word and *louk-s-nā as a “stylized”, “poetic” moon — this already surmises incomplete synonymy, as it is always stipulated that each slot on the Swadesh wordlist be strictly filled in with the most “neutral” item, and that stylistically embellished quasi-synonyms should be left out.

On the other hand, if we do think of such a difference, or, indeed, consider it in terms of the possible existence of a special compound *mēns louksnā (or, in the feminine, *mēnsā loukānā) ‘shiny moon’, i. e. ‘full moon’ (cf., for instance, Avestan raoxad māyham acc.), it becomes very clear how easily the formerly adjectival form could have independently replaced the former noun *mēns in at least several branches of Indo-European. To this should be added the additional “polysemy pressure” — since *mēns was used both in the meaning of ‘moon’ and ‘month’, its replacement in at least one of these meanings could have been anticipated.

Work on semantic reconstruction for mid-level “non-controversial” families shows that such “criss-crossings” are relatively rare. Generally, if one item is replaced in several branches, it tends to be ushered out by different roots, because for each item on the list at least several different paths of semantic evolution are possible, and the more such paths we know, the less is the probability that the same path will be independently selected by two or more languages.

Nevertheless, semantic typology shows that some paths are more frequent than others, and in such cases, we must be prepared to expect independent developments. For instance, the term for such a body part as ‘ear’ is, every now and then, all over the world, re-formed as a nominal derivative from the verb ‘to hear’ (= ‘hearing-thing’). In Indo-European, there is little doubt as to the original proto-root for ‘ear’ — IE *ous- — but in Tocharian, we find that old root replaced by such a derivative: Tocharian A klots, B klautso ← Proto-Tocharian *kleutsājān-[Adams 1999: 230] ← IE *kley- ‘to hear’. Not surprisingly, we also find a similar (although morphologically slightly different) development in Celtic: cf. Irish, Gaelic cluas, Welsh clust etc. Does this mean that Tocharian and Celtic share a common node on the tree, or, perhaps, this should be considered a special “areal” Tocharian-Celtic isogloss? Hardly likely.

But the one area where the issue of “semantic criss-crossing” hits the hardest is, of course, macro-comparison. Taking advantage of the fact that semantic reconstruction is one of historical linguistics’ weakest spots, macro-comparative lexicostatistics may, in dealing with a particular Swadesh item, take any root which has the appropriate Swadesh meaning in any of mid-level family A’s subbranches (or, in fact, even in any of its individual languages) — and score it as a positive cognate with any root with the appropriate Swadesh meaning in any of mid-level family B’s subbranches (provided, of course, that the scoring is sanctified by phonetic correspondences or phonetic similarity). This approach is more or less explicitly stated by S. Starostin for his lexicostatistical calculations for language of Eurasia: “I have chosen the following principle: a word can be used as representing a particular meaning in the protolanguage if it has exactly this meaning in at least one subbranch of the family” [Starostin 2007b: 807].

Frankly, I have the gravest doubts about the statistical validity of this approach. Suppose that, in a certain language, we have a pair of semantically close roots (e. g. ‘fire’ : ‘light’; ‘star’ : ‘shine’; ‘bird’ : ‘fly’; ‘head’ : ‘top’, etc.), the second of which is easily liable to usurp the functions of the first at some future point in time. How high are the chances of at least two of its future descendants to effectuate that transition independently of each other? Obviously, the primary dependency is on the number of those descendants. In the case of ten — twelve branches of Indo-European, chances for independent unidirectional semantic change will be quite modest (and this is explicitly confirmed by the actual historical analysis of the Swadesh wordlist), but if we multiply that number by a factor of five or six (the number of large families that constitute Nostratic), these chances will increase quite rapidly. (This could relatively easily be illustrated with a probabilistic model).
Not coincidentally, even a brief survey of the comparative tables for lexical matches between nine mid-level families of the Old World (Indo-European, Uralic, Altaic, Dravidian, Kartvelian, representing the Nostratic macrofamily; Semitic, representing the Afro-Asiatic macrofamily; North Caucasian, Sino-Tibetan, Yeniseian, representing the Sino-Caucasian macrofamily), presented in [Starostin 2007b: 807–815], reveals a picture that can only be called “Synonymity on the Rampage”: two, sometimes three roots for each Swadesh item within one mid-level family — and, consequently, three to five roots on average within one macrofamily — are the norm. The word ‘sun’ in Nostratic languages alone, for instance, is illustrated by (a) a match between Indo-European *seHw- and Altaic *sjàgu; (b) a match between Uralic *pųńwV and Altaic *p’iagV; (c) a match between Altaic *nērə and Dravidian *neįir-. Should this be historically interpreted as reflecting three freely interchangeable synonyms for ‘sun’ in Proto-Nostratic (and, further down, in Proto-Altaic)? Apparently not. In order to admit such a possibility, we should either find some typological support for it on less remote time scales — in all likeliness, an impossible task — or suggest that language speakers in pre-Neolithic times had a far more liberal attitude towards synonymity than their descendants, being accustomed to freely sharing two or three words for each meaning. This, however, would simply plunge us into the world of fantasy 21.

Let us look at this situation with ‘sun’ more closely. The three matches, as can clearly be seen, are determined by the three roots in Altaic — itself a “near-macro-family”, still controversial among mainstream linguists. I do not doubt the existence of Altaic — evidence for a special relationship between Turkic, Mongolic, Tungusic, Korean, and Japanese is too overwhelming to make room for skepticism — but I will be the first to admit that this evidence is in dire need of further filtering and refining, and that one of its major problems is the lack of a detailed semantic reconstruction.

The three mentioned Altaic roots for ‘sun’ are not, in fact, “Altaic”: they are rather the main roots to denote this object in separate subdivisions of Altaic. Proto-Altaic *sjàgu (newer reconstruction is actually *sjògu) is reflected as Tungus-Manchu *sīgu-n ‘sun’ and Korean *hǎi id., with a possible further correlate in Japanese *suâ-ra ‘sky’ [EDAL: 1274]. Proto-Altaic *p’iagV is reflected as Japanese *pi ‘sun’, but also Korean *pái ‘dawn’, Tungus-Manchu *piği ‘to warm (smth.), warm oneself’, and Mongolic *heʿe- ‘to heat, be heated’ [EDAL: 1147]. Finally, Proto-Altaic *nērə (*ŋërə in EDAL) is reflected as Mongolic *nara-n ‘sun’, but also Turkic *jær-in ‘morning; tomorrow’, Tungus-Manchu *jēr(i)- ‘light’, Korean *nár ‘day (24 hours); weather’, and Japanese *āri- ‘dawn’ [EDAL: 1028].

Out of these three roots, only *sjàgu has the meaning ‘sun’ in at least two branches of the family, and it is interesting to see that the Japanese parallel shows a suffixal extension, indi-

21 The existence of this problem was well realized by S. Starostin himself, who wrote: “the “protolanguage synonymy” may produce a higher number of coincidences and make the dates of separation somewhat younger” [Starostin 2003: 465]. He, however, believed that the negative effects of this kind of scoring may be counterbalanced and cancelled by a reverse factor: “the impossibility of identifying loanwords may result in an earlier date of divergence (according to the standard procedure adopted by us, a mismatch caused by the borrowing is not taken into consideration; consequently, if loanwords cannot be detected, the percentage of coincidences between the proto-languages becomes lower)” [ibid.].

Perhaps for the full 100-wordlist this may, to a certain degree, be true. But when we pare it down to 50 most stable items, the loanwords issue loses much of its significance, since these items, by default, are expected to contain an absolute minimum of loans (see below). The synonymity issue, on the other hand, is equally disturbing for any version of the list, and I am afraid that, in “macro-calculations”, adoption of a liberal stance on synonymity will inevitably result in an exaggerated number of matches between families and, consequently, younger dates of separation for macro-units like Nostratic or Sino-Caucasian.
cating that the original meaning of *suā-rá may have been something like ‘sunny skies’. In very sharp contrast, the two other roots have only gained the meaning ‘sun’ in one branch each, and show a very different type of semantics elsewhere. In fact, a comparison between *pigí ‘to warm’ and *pí ‘sun’ is hardly imaginable unless the original semantics was that of ‘heat’, because the semantic development ‘sun’ → ‘warm’ is typologically unprecedented (at the very least, I have been unable to encounter any reliable examples in EHL’s huge collection of data). Likewise, *nērā is easily understood as an original ‘day, light time period’, but hardly as an actual designation of the celestial body.

The likeliest candidate for an original Proto-Altaic ‘sun’ is, therefore, only *siągu — for the other two roots, none of the possible scenarios are credible from the point of view of semantic typology. How does this reflect upon the Nostratic comparison? Fairly well: as suggested originally, *siągu is a solid match for Indo-European *seHw-, or, more traditionally, *sāw-el- ~ *sw-en- with fluctuating suffixal extensions [Pokorny 1959: 881–882].

But what of the other two matches, with Uralic and Dravidian respectively? The interesting thing here is that, while Indo-European *sāw-el- ~ *sw-en- is, indeed, unquestionably the primary Indo-European root for ‘sun’, the same cannot be said neither of Uralic *pVjwV nor of Dravidian *ńeįį-. The former, as a polysemous ‘sun; day’, is the main root in Balto-Finnic and Lappic (Finnish päävä, Estonian pääev, Saami ba’eve, etc., see [Rédei 1988: 360]), but not anywhere else. The latter, reconstructable as *ńeįį or *ńeis, is seen only in the South Dravidian subgroup (Tamil nāyiru, nāyiru; Kannada nēśar; Tulu nēṣuru ‘morning’; Toda nēṛ ‘sun (only in songs)’) and, perhaps — although the phonetic correspondences are dubious — in North Dravidian, with different semantics (Malto nīṛu ‘sunshine, heat’); see [DEDR: 252]. It is certainly a far less likely candidate for Proto-Dravidian ‘sun’ than the far better represented *poṛud- [DEDR: 403].

By applying nothing but the basic, simplest principles of semantic reconstruction, we have managed to show that, out of these three instances of ‘sun’ in Nostratic, there is really one strong case — strong on all sides — and two weak ones — weak on all sides. Note that the etymologies as such have not been killed off (at least the Uralic-Altaic connection is still relevant), only their lexicostatistical significance. The evidence in favor of Nostratic has not been weakened; on the contrary, it has only become tighter, as the “evolutionary scenario” for Nostratic ‘sun’ is now more comprehensible and realistic.

There does, however, remain the issue of scoring. We have more or less certified that Proto-Uralic *pVjwV did not necessarily have the meaning ‘sun’, and that Proto-Altaic *pJVjavV almost certainly did not have this meaning. However, our list of proto-languages does not include Altaic and Uralic; the starting nodes are the smaller subgroups that constitute these two large families, and these happen to include Balto-Finnic, where the root for ‘sun’ is *pātova, and Proto-Japanese, where it is *pi. They generally satisfy the requirements for phonetic correspondences in Nostratic languages, and are quite compatible phonetically even without knowing these correspondences — yet they, most likely, do not go back to the respective Proto-Altaic and Proto-Uralic roots for ‘sun’. Should they be scored as cognates or not?

From an etymological point of view, they are cognates — reflecting independent similar semantic development out of an older meaning — and should be scored as matches. However, the epistemological definition of a “match” on the Swadesh list would necessarily surmise the

22 Actually, if the Altaic root *nērā is really to be reconstructed with a temporal meaning (‘bright period of day’), a much better parallel in Dravidian is Tamil nēram ‘time, season, opportunity’, Kodagu nēra ‘time, sun ([f]), Tulu nēr-dē id., possibly (although loss of final -r is irregular) also Brahui dē ‘sun, sunshine, day, time’ [DEDR: 337] — still not the main Proto-Dravidian root for ‘sun’, but a very interesting semantic match all the same.
idea of either *common retention* (the word continues, substantially unchanged, to perform the original function as such in descendant languages) or *common innovation* (the word shifts from its original function in the intermediate language that serves as the specific common ancestor to languages displaying the innovation). In this particular case, as well as plenty of others, there is neither a common retention — chances of this word meaning ‘sun’ in Proto-Nostratic are minimal compared to other candidates — nor a common innovation (Baltic-Finnic and Japanese do not have an immediate common ancestor). Scoring *päivä* and *pi* as a match will, therefore, distort the overall calculation scheme, and, in combination with multiple other distortions of such sort, make the classification results less reliable.

On the other hand, it should not be forgotten that notions such as “Altaic”, “Uralic”, “Nostratic”, etc., already surmise a pre-established idea of branching, and that we run the risk of succumbing to circularity if we modify our scoring results based on preconceived ideas of classification. Moreover, for linguistic areas in which there are no preconceived ideas of classification, or these ideas are at an embryonic stage (= much, if not most of the linguistic world outside Eurasia) such modifications will be impossible in principle. How should we proceed?

I suggest, once again, a return to “bootstrapping” mode. During the first stage of calculations our main goal is to establish the primary “linguistic building blocks” — perform a rough attempt of grouping a large number of families into a smaller number of higher-level units. In the case of Eurasia, this attempt will, without a doubt, let us see all of its principal families — Indo-European, Uralic, Altaic, Dravidian, Sino-Tibetic, Semitic, Austro-Asiatic, etc. — as well as indicate possible higher level connections between them. At this stage, it will be permissible to count *päjvä* and *pi* as (potential) cognates, because we have not yet certified the existence of such “blocks” as Uralic and Altaic.

Once the first stage is completed, we proceed to the second stage: fine-graining the results, using the “block” information we have accumulated as our basis. At this stage, our main task is to wipe out the “false leads”, and this is accomplished through establishing, as precisely as possible, the most likely candidate for the given Swadesh notion at the top of each “block”, i.e. for Proto-Indo-European, Proto-Uralic, Proto-Semitic, etc. By default, only that particular item will be allowed to score as a positive match on the higher level of taxonomy. All other matches will be eliminated, judged as either (a) chance similarities or (b) independent semantic innovations, even if the roots are related etymologically.

Let us demonstrate this on one more example, this time taken from the Sino-Caucasian sphere. In [Starostin 2003: 473], one of the proposed matches is North Caucasian *wěnAx ‘head’ vs. Sino-Tibetic *lŭH id. This comparison satisfies S. Starostin’s own system of phonetic correspondences between the two families (with regular reduction of the initial syllable in Sino-Tibetic) and, at the first stage of comparison, is acceptable. However, since both the “North Caucasian” and “Sino-Tibetic” labels are not quite allowed at this stage, it should rather be noted that the comparison is between (a) Proto-Lezghian *walul, (b) Proto-Dargwa beḳ, (c) Lak (an isolated language) baḳ, (d) Khinalug (another isolate) mīkīr (in other branches of North Caucasian the root is either missing or has such different meanings as ‘beak; mouth; nose’; see [Nikolayev, Starostin 1994: 1041] for details), (e) Old Chinese 首 s-lu, (f) Kuki-Chin *lu (Kuki-Chin is a large, but only one subgroup of Tibeto-Burmese; see [Schuessler 2007: 470] for the etymology). All these forms can be marked as cognates (even such superficially dissimilar forms as Lak baḳ and Kuki-Chin *lu, since we have permission to use our knowledge about the internal and external historical phonology of these languages).

Once the primary stage has been completed, and the North Caucasian and Sino-Tibetic “blocks” established as firm taxonomic units, we run the second stage, checking the validity of *wěnAx* and *lŭH* as the best respective candidates for Proto-NC and Proto-ST ‘head’. First of
all, it should be noted that even the primary stage will clearly indicate a strong binary split in both cases: North Caucasian will be a combination of Northeast (Nakh-Daghestanian) Caucasian and Northwest (Abkhaz-Adyghe) Caucasian, and Sino-Tibetan — a combination of Sinitic (Chinese) and Tibeto-Burmese. Our ideal would be to see *wĕn/V represented in both the Northeast and the Northwest branches, and to see *lŭH in both Chinese and Tibeto-Burmese. The situation is, however, much more complicated.

NC *wĕn/V is not properly NC; it is only encountered as ‘head’ in several Daghestanian branches and is not necessarily even the best candidate for ‘head’ on that level. (In Andian and Tsezian languages the main root for ‘head’ reflects NC *h/qV, and the default West Caucasian root is reconstructed as *Sqla). This is not a death blow, since it merely presumes that we are unable to reach a satisfactory conclusion based on internal evidence alone (see above).

But the situation is worse in the case of Sino-Tibetan. Here, semantic reconstruction strongly indicates that *lŭH may be an independent innovation in Old Chinese and Kuki-Chin — provided the roots are even related in the first place, and do not represent accidental look-aikes. The reason is that the primary root for ‘head’ in Tibeto-Burmese is not *lŭH, but *qh/ohH (reconstruction following [Peiros, Starostin 1996]), reflected in a large number of subgroups: cf. Tibetan m-go, Burmese u-h, Sgaw Karen kho?, Garo s-ko, Pumi khu, Jiarung ko etc. (each language here represents a separate group). The idea that it is *qh/ohH that represents an archaism and not *lŭH is further supported by its very likely cognate in Old Chinese: 后 *gō/glottalstop ‘ruler, sovereign’, suggesting a very usual semantic development from ‘head’. The opposite transition ‘ruler’ → ‘head’ (as body part!) is not at all realistic.

Obviously, we should keep in mind that the general field of Sino-Tibetan etymology at its present state leaves a lot to be desired, and future research may yet show that *lŭH is, in fact, a more firmly grounded reconstruction than *qh/ohH. But the current disposition is hardly in favor of that conclusion, and so, at the second stage of our cognate scoring, we should dispose of this match, since it fails to pass our criteria for choosing the most appropriate synonym.

It is very important to note that there are clear-cut cases when no single item can be unambiguously postulated for the “top of the block” position. The most typical situation here is that of a primary binary split, such as, e. g., Indo-European into Anatolian and “Narrow Indo-European” (or, in other terms, “Indo-Hittite” into Anatolian and Indo-European), Uralic into Feno-Ugric and Samoyed, or North Caucasian into Northeast and Northwest Caucasian. In all such cases, whenever one has to reconstruct different roots for the same notion in each branch, both reconstructions carry the same “weight”, regardless of their size and spread. E. g., “Narrow Indo-European” *onogh- ‘fingernail’ and Hittite sankuwai- id. have an equal chance of reflecting the original root for this notion, despite the fact that *onogh- is seen in at least seven different subgroups of Indo-European.

I predict a certain amount of criticism addressed at this methodology, and understand the main objection: the general inexperience of historical linguistics when it comes to strict semantic reconstruction, the usual uncertainties that we all feel about assigning one particular meaning to a proto-root whenever its descendants show even a slight amount of semantic variety. However, it is exactly this particular objection that makes me insist that the “no synonyms!” principle be applied and tested as rigorously as possible, if only for the reason that we all have to learn to perform strict semantic reconstruction, sooner or later, and that if there is one good place to start with it, it is the Swadesh wordlist. A global lexicostatistical database with an emphasis on semantic change would, in addition to its general goals, serve as an excellent foundation for all sorts of systematic studies on historical semantics.

Finally, a consistent application of the “semantic filter” would, hopefully, help dissipate the major accusation against global-scale lexicostatistics — namely, that the more languages
are added into the pot, the more chances we have of getting accidental look-aliases. Obviously, this accusation is true if we place no limits on "criss-crossing" — score one "Proto-Indo-European" synonym for a given item as a match with Uralic, another one as a match with Dravidian, a third one as a match with Old Chinese, and a fourth one as a match with North Halmaheran. But if it can be shown, for instance, that the best matches between Indo-European and Uralic are truly Proto-Indo-European and Proto-Uralic — most likely candidates for the proto-roots in both families — this leaves no space for such accidence.

6. Contacts, Contradictions, and Conclusions

In the three previous sections, we have attempted to describe the main methodological principles that should, in our opinion, guide the process of constructing a global lexicostatistical database for the world's languages. Their chief differences from previously employed techniques may be briefly summarized as follows: (a) use of a compact, ultra-stable 50-item wordlist with low-level reconstructions serving as the main entries; (b) use of a “mixed” scoring procedure, based on phonetic correspondences where they have been established and “phonetic compatibility with traces of regularity” where they have been not; (c) very strict limits on synonymity both on low, mid and deep chronological levels; (d) a “recursive” approach to scoring, where the first round of calculations is followed by a “fine-graining” round, weeding out false matches with no historical reality behind them.

A careful application of all these conditions, particularly (b) and (c), will minimize the number of accidental similarities in our calculations. But will it be able to neutralize the problem that we described at the very beginning of the paper — the risk of mistaking contact lexicon for genetic cognates? Obviously, words could be borrowed into proto-languages as easily as they can be borrowed into historically attested languages (so strict limitations on synonymity are not necessarily a safeguard), and if the borrowed strata are large enough, they always display “traces of regularity”.

It would be an exaggeration to say that the proposed method is sufficiently robust to let us, in each and every type of imaginable situations, avoid the “contact trap”. Nevertheless, there are two main considerations that make it significantly more waterproof than other methods of classification.

The first one is the choice of the wordlist. None of the 50 items — not even personal pronouns — are 100% immune to borrowing, but, in general, this “core” is much more resilient to being replaced by words of foreign origin than even the remaining half of the Swadesh wordlist. Having analyzed (preliminarily) the 50-item lists for approximately 200 low-level families of Eurasia and Africa, I have been able to detect only three explicit cases in which borrowings amounted to about 1/5 (10–11 items) of the entire list: these were Brahui (one-language group within Dravidian), Albanian (one-language group within Indo-European), and Northern Songhay (a small cluster of closely related dialects with a very heavy Berber influence; Southern Songhay is much more conservative).

Furthermore, Brahui displays a hodge-podge of borrowings from different sources (Indian, Persian, Arabic) that outcancel each other, and some of the alleged “borrowings” from Latin on the Albanian list are etymologically questionable and may actually represent inherited retentions of original Indo-European roots. This leaves the Songhay dialects as just about the only transparent example where one could really make a mistake (provided one had no access to supporting data from Southern Songhay) — and there is no reason whatsoever to think that this ratio of 1 to 200 must have been seriously different ten or more thousand years ago.
The second consideration is one of context. Let us suppose that we are running the first stage of calculations and have no idea of the genetic status of the Brahui language. In this case, we may want to score Brahui *haḍ* ‘bone’ as cognate with Old Indian *asthi*, Brahui *dandān* ‘tooth’ as cognate with *dant-*, and, perhaps, Brahui *draxt* ‘tree’ (although this is a Persian, not an Indian word) as cognate with *daru*. This will give us three false matches that will, nevertheless, be overridden during the tree construction process by the overwhelming number of true matches that Brahui has with the other Dravidian languages. Noticing the sharp increase of Brahui matches with Indo-European, even though the suggested classification clearly puts it with the rest of Dravidian, we will then — at the second, “fine-graining” stage — count the Brahui forms as borrowings (excluding them from calculations), since a true close relationship with Indo-European would require an equally sharp increase in cognation rate between every branch of Dravidian and every branch of Indo-European.

Similar analyses will easily help us weed out false matches between North Songhay and Berber, Fenno-Ugric and Indo-Iranian, Kartvelian and North Caucasian, etc. Counting these pairs of language groups as sharing a close genetic relationship will be out of the question because each of their elements will have a much stronger “attraction” on the part of its true closest relative.

If, on the other hand, potential cognates are found between the respective protolanguages A and B in their “blocks”, and no “stronger” genetic affiliation is suggested between protolanguage A and, for instance, protolanguage C, this should be — by default — considered as indicative of deep-level relationship. “By default” here means that, if we want to interpret such a situation as reflecting contacts, the burden of additional proof here lies on the “arealist”, not on the “heritagist”.

Example: for Indo-European and Uralic, we find such serious matches on the 50-item list as IE *me- : Uralic *mE ‘I’, IE *tu : Uralic *tE ‘thou’, IE *kley- : Uralic *kule ‘to hear’, IE *(H)nom- : Uralic *nime ‘name’, IE *wed-or : Uralic *wete ‘water’, IE *k*-i-s : Uralic *kU ‘who’ (several other, less obvious, cognates will be discussed in further publications on the subject). Similarly strong cognation suggestions also exist between IE, Uralic and some other language families that constitute the traditional “Nostratic”, but none of them override this evidence quantitatively.

Interpretation of these matches in terms of prehistorical contacts is not entirely ruled out, yet, based on our empirical knowledge about contact situations around the world as well as common sense, is significantly less likely than its interpretation in terms of prehistorical genetic relationship. If the “arealist” thinks otherwise, it is up to him/her to provide additional evidence, preferably in the form of at least dozens (if not hundreds) of terms in the cultural lexicon, borrowed from Proto-IE into Proto-Uralic or vice versa — a condition that is, for instance, very easy to satisfy in the cases of Brahui, Albanian, and North Songhay. Until this is done, the default working model will be that of genetic relationship between Indo-European and Uralic.

Before concluding this discussion, three more small, but important technical points should be made on certain procedural aspects of PL:

1. As mentioned above, glottochronological interpretation of the results — with absolute dates of splitting accompanying the classification — is not obligatory, but is nevertheless use-

23 Of course, there always remains the problem of the so-called “mixed languages” (pigins, creoles, etc.), whose existence in prehistoric times can be questioned, but not ruled out. Nevertheless, there are reasons to think that “contextual” considerations such as described above will help us single out and correctly identify such situations as well. For a detailed discussion on the identification of possible “creoles” in lexicostatistical databases, see [Burlak 2006].
ful for those who accept glottochronology as a valid method. However, basing the glottochronological calculations on the old Swadesh quotient of 0.14 or Starostin’s “improved” quotient of 0.05 will be inadmissible, since these rates have been calibrated based on the average stability of the entire 100-wordlist, not its more stable half. We, therefore, either have to recalibrate the quotient — obviously, its value will be somewhat less than 0.05 — or, better still, rely on Starostin’s “experimental” method with individual rates for each item on the list (see fn. 6).

2. It is evident that, no matter how tight we make the rules on scoring, in quite a few cases we will be presented with several alternatives of equal or near-equal probability, sometimes affecting classification results in a serious manner. (Within Indo-European, for instance, Albanian is a particularly difficult case; its position on the tree may depend on as little as one or two questionable etymological decisions). For such cases, it makes sense to consider all the alternate paths of scoring and present all alternate models; additional data will then be necessary to make a more precise choice.

3. Although the principal work should be conducted manually, this does not mean that fully automatic procedures — such as have been described in section 4 — are out of the question; on the contrary, it would make perfect sense to combine manual and automatic handling of the data. Similar results will strengthen the conclusions, while discrepancies may clearly indicate problematic areas in the manual handling as well as help refine the automatic algorithms.

The detailed description of the PL procedure in this paper would, of course, not be possible if the procedure itself still existed only in theory. As it is, 50-item lists have already been compiled by the author of this paper — and are, at the moment, collectively verified and modified at regular sessions of the Nostratic seminar at RSUH’s Center for Comparative Linguistics — for most of the families and sub-families that constitute the traditional “Nostratic”, and are now being compiled for subdivisions of “Afro-Asiatic” and “Sino-Caucasian”.

Sergei Jaxontov, in an overview article on glottochronology, once wrote: “It would be desirable to apply glottochronology among all established and tentative language families. As a result, language groups could be revealed with a maximum divergence of 60–80 (or, probably, 80–100) centuries, as well as language isolates beyond such groups. Also, realistic and comparable classifications could be proposed for each group” [Jaxontov 1999: 59]. With the massive amount of comparative data that members of the EHL project have managed to put together over the past eight years, we now have every possibility of carrying out this work on a more detailed and professional basis than was possible even a decade ago. It is, at present, unclear what the “time ceiling” will be for this kind of approach — whether it will be Jaxontov’s “80–100” centuries or significantly deeper than that — but this really depends on “data behaviour” and can hardly be predicted.

The present paper lays down the basic methodological aspects of PL, yet its real value will only be evident on practice — with the actual discussions of the data for each individual “block” (family) and its comparisons with data from other “blocks”. The paper is, thus, but an introduction to a series of publications (or, perhaps, a collective monograph) that I and other EHL members plan to dedicate to the presentation and analysis of the lexical data relevant for a PL-based global linguistic classification.

Appendix

The proposed 50-item wordlist for the global lexicostatistical database. Items are ranged according to their relative degree of stability. For some of the most ambiguous English lexemes, additional meaning specifications are given in parentheses.
1. we 11. hand 21. one 31. mouth 41. leaf
2. two 12. what 22. tooth 32. ear 42. kill
3. I 24. 13. die 23. new 33. bird 43. foot
4. eye 14. heart 24. dry (e.g. of clothes) 34. bone 44. horn
5. thou 25. 15. drink 25. eat 35. sun 45. hear
6. who 16. dog 26. tail 36. smoke 46. meat (as food)
7. fire 17. louse (head) 27. hair (of head) 37. tree 47. egg
8. tongue 18. moon 28. water 38. ashes 48. black
9. stone 19. fingernail 29. nose 39. rain 49. head
10. name 20. blood 30. not 40. star 50. night

Literature


24 For personal pronouns, as an official exception, synonymity is allowed on the list by taking both the direct and indirect stem of the pronoun into account if they are suppletive (e. g. *I* — *me*).
25 Basic negation, particle or negative verbal stem/suffix.


Hübschmann 1875 — Heinrich HÜBSCHMANN. Über die Stellung des armenischen im Kreise der indogermanischen Sprachen // Zeitschrift für Vergleichende Sprachforschung 23, pp. 5–42.


Статья посвящена методологическим аспектам создания глобальной лексикостатистической базы данных по всем языкам мира — одной из наиболее актуальных задач международного проекта «Эволюция языка» (Институт Санта Фе). Автор предлагает ряд существенных изменений стандартной лексикостатистической процедуры, как-то: замена традиционного стословного списка Сводеша на более компактный список из 50 «сверхустойчивых» лексических единиц; постулирование праязыковых реконструкций «низкого уровня» в качестве отправных узлов общего генеалогического древа; использование как обычного сравнительно-исторического метода, так и представлений о «фонетическом сходстве» для подсчета когната; и, самое главное, упор на максимальную точность семантической реконструкции и на жесткие ограничения синонимии.