George Starostin
HSE University, Russia, Moscow / Santa Fe Institute, New Mexico; gstarst1@gmail.com

How we could show that Hadza is Afroasiatic: a response to Militarev's "Hadza as Afrasian?"[1]

In this brief response to Alexander Militarev's paper on the Afroasiatic (Afrasian) affiliation of Hadza, I attempt to identify the major theoretical issues with his lexicostatistical analysis of the similarities between Hadza and the various branches of the Afroasiatic family, explaining why this analysis cannot be accepted as conclusive, and suggesting some steps that could be taken in order to weed out insignificant evidence (e.g. more attention toward meticulous step-by-step reconstruction of the proto-wordlists potentially involved in such a comparison).

Keywords: language isolates; Hadza language; Khoisan languages; Afroasiatic languages; long-distance relationship; lexicostatistics.

Dr. Alexander Militarev's paper, in which he presents a (seemingly) impressive amount of linguistic evidence for the Afrasian (Afroasiatic) affiliation of Hadza, one of the most famous and intriguing isolates on the African continent, is of immense interest to myself — not only because I, too, have been seriously involved for more than a decade in figuring out the genetic and areal connections of Hadza on a lexical basis (Starostin 2008, 2013, etc.), but also because the arguments laid out in Militarev's paper have very wide-reaching historical implications. Essentially, the paper could be interpreted as a specific case study in trying to answer a complicated general question — *is it possible at all, and if yes, how is it possible to convincingly demonstrate the genetic affinity between a modern day linguistic isolate and an entire macro-family (= super-family, phylum, etc.) of languages going back to the very limits (some might argue, even beyond the limits) of the classic comparative method*?

In fact, this particular case study is as perfect as it could ever be. On one side of the equation, we have Hadza, a language that clearly has no living relatives that would not be separated from it by thousands of years (how many thousands — remains to be seen); has been first attested no more than a hundred or so years ago (if we take something like Obst 1912 as the starting point); and, because of the small number of speakers, shows very little, if any, dialectal variety, making internal reconstruction based on dialectal comparison impossible. (Careful and detailed study of the language itself allows for a little bit of internal reconstruction based on morphemic analysis of its lexemes, as is shown in Sands et al. 2023 in this volume, but one should never overestimate the potential of such a reconstruction). In short, we know almost nothing of the linguistic history of Hadza, as compared to, say, an average Indo-European or an average Bantu language, the chronological distance between which and its hypothetical distant relatives (e.g. from other branches of "Nostratic" or "Niger-Congo") can be easily and significantly shortened through proper historical reconstruction based on comparison with its close relatives.

On the other side of the equation, we have Afroasiatic — one of the few "macro-family" level linguistic taxa that continues to enjoy widespread mainstream acceptance, despite many significant issues, such as the internal constituency of the taxon (e.g., with serious doubts cast upon the inclusion of the Omotic branch, see Theil 2006), its internal classification, and multiple disagreements on the right way to reconstruct Proto-Afroasiatic phonology and lexicon. None of these problems, however, have managed to shatter the historical validity of Afroasiatic as a genetic unit; and no matter how old it is, how many Afroasiatic etymologies can be considered established beyond reasonable doubt, or how many languages it actually contains, there is — at least in theory — absolutely no reason to consider its current borders permanently closed to the acceptance of new branches.

Furthermore, *pace* the concerns expressed in the final section of the response offered by Sands et al. 2023, I cannot see any waterproof historical reasons to regard the Hadza-Afroasiatic scenario as an *a priori* highly unlikely one. Militarev's explanation proposes the situation of a language shift, in which the Hadza people would adopt the (possibly more "prestigious", under historically obscure circumstances) language of their Afroasiatic neighbors; such an idea not only explains why Hadza, as opposed to early Afroasiatic speakers, remain foragers rather than pastoralists or agriculturalists, but is also relatively consistent with the African situation as a whole, where numerous cases of similar language shifts, from the «Pygmies» to the more geographically close ethnicities linguistically belonging to the South Nilotic group (Rottland 1982), have been detected. While I agree with Sands et al. that a more concise scenario, grounded in ethnographic and archeological reality, would be welcome in this situation, I also concur with Militarev that it makes little sense to spend time and effort on building up such a scenario *before* the Hadza-Afroasiatic connection has been demonstrated on solid linguistic ground. The question, then, is whether the connection has really been demonstrated, or, as Sands et al. claim, the presented evidence is thoroughly insufficient for such a purpose.

When we postulate a linguistic scenario for the affinity between a linguistic isolate and a macro-family, from a purely phylogenetic point of view there are two logical possibilities. One is that the isolate may be a top level division of the macrofamily, i.e. historically the first branch to split from the common protolanguage, or one of the first several branches in the potential case of an original multiforcation. In the case of Afroasiatic, which Militarev currently dates to (approximately) the 11th millennium BC (Militarev & Nikolaev 2020: 200), this would mean approximately 12–13 thousand years of independent development (including phonetic change and gradual cognate loss) for Hadza. Lexicostatistical calculations, be they carried out according to the Swadesh formula ($c = e^{-0.14t}$, where c = percentage of retentions, t = time elapsed) or the revised Starostin formula ($c = e^{-0.05ct^2}$), would in this case both agree that Hadza could theoretically retain about 20% of its basic vocabulary (= 100-item Swadesh wordlist) in this time interval, although, of course, (at least if we adopt Starostin's method) the actual number of items would be much smaller than 20 because of all the extra borrowings received along the way (and also because at least a few of these items could be retained in Hadza, but not in any other branch of Afroasiatic and therefore become technically unrecoverable as genuine Afroasiatic legacy).

While 20% (give or take a few) could look like an impressive figure by itself, lexicostatistics and glottochronology typically deal with pairwise comparisons; and the same formulae, when applied to pairs of languages, naturally yield much smaller figures. Again, both in Swadesh's and Starostin's formulae a period of divergence between two languages that is equal to approximately 10,000 years translates into a figure of approximately 5–6% of matches between the two ($c = e^{-0.28t}$ for Swadesh, $c = e^{-0.1\sqrt{ct^2}}$ for Starostin); for 12,000–13,000 years of separate development we should expect an even smaller number of matches. Glancing at Mili-

tarev's figures, one does indeed often find these kinds of values (around 5–6% of matches with many random Omotic, Cushitic, Chadic, etc. languages), although there are occasional unexpected surges (on which I shall comment later) and equally unexpected drops (e.g. only 3% with South Omotic and a measly 2% with Qwadza).

The problem is — how would a 5-6% amount of matches (remember that these are *percentages*; elimination of borrowings from the 100-item wordlist means that the actual number of matches is even lower) be distinguishable from chance resemblance? Recent experiments conducted, e.g., on the basis of data included into the Global Lexicostatistical Database, show that 5-6% of accidental similarities between two unrelated languages is quite a realistic figure. The only way to increase the significance of such pairwise comparisons, then, would be to demonstrate that they are not pairwise, but N-wise, i.e. that Hadza regularly yields matches with not just *one*, but *several* branches of Afroasiatic at the same time — and that in all such cases, we are genuinely dealing with a solid lexical candidate for the Proto-Afroasiatic level.

Unfortunately, Militarev's paper does not summarize specific numbers of three-way, four-way, etc. matches between Hadza and the different branches or subbranches of Afroasiatic; these have to be calculated by the reader based on the data he provides. However, while this is not an impossible task, it is not quite clear whether it is a useful one, because it is not enough to merely find look-alikes — it is just as important to demonstrate their *reconstructibility*. That is, if a certain item in Hadza is *lexicostatistically* compared to a certain item in one or two Chadic languages and a certain item in one or two Cushitic languages, it is of vital importance to the comparison to be able to show that both in Chadic and Cushitic, the item in question has a solid chance of reflecting the respective Proto-Chadic and Proto-Cushitic item.

To take one specific example of how this principle is undermined, let us look at the word 'eat' (Hadza *seme ~ simi*), which is compared by Militarev to two counterparts in Egyptian and Chadic, both of which are marked with = (symbolizing exact lexical match). Upon first glance, the parallel with Egyptian *sꜥm* 'eat' and Proto-Chadic **sVm-* 'eat' may seem like an impressive three-way consonantal match. However:

(a) the Egyptian word is by no means the original Egyptian basic word with this meaning. It is not attested at all in Old Egyptian, and its first and most common meaning is rather 'to swallow' ("schlucken lassen, hinunterspülen lassen" in Hannig 2006: 2107), from which figurative extensions to both 'drink' and 'eat' are occasionally encountered. Meanwhile, the basic and most common equivalent for 'to eat' throughout the entire history of Egyptian and Coptic is the verb *wnm*, which by no means matches Hadza (or Chadic, for that matter);

(b) the Chadic equivalent is put together from the data of two Western languages (Angas and Sura) and a few more Central languages, clearly insufficient to ensure the Proto-Chadic status of this item, much less in the basic meaning 'eat'; in terms of semantics and distribution, the ideal candidate for the status of Proto-Chadic 'eat' is the lexical root **ti/aw/ʔ-* (Stolbova 2016: 317), well represented in all the three major branches of that family in precisely the requested meaning. Again, no match with Hadza.

From this point of view, Hadza *seme ~ simi* would be much better compared with, for instance, Austroasiatic, where Ilia Peiros in his database reconstructs a hypothetical **sVm* 'eat' based on Proto-Palaungic **som* and Proto-Munda **ʒVm* (although the root is not attested in any other branches of the family, it is at least quite safely reconstructible in this basic meaning to the top level of Proto-Palaungic and Proto-Munda; cf. also **cuum ~ *cəm* 'to eat' in Shorto 2006: 364).

Just a few paragraphs away, we find Hadza *furu-ne* 'to be many, plenty, full' (not even the most default equivalent for the meaning 'full' in Hadza itself) compared with Zenaga *tu-fur-t* 'full' (at least this item has a very slim chance of going back to Proto-Berber, given the isolated status of Zenaga within this family); a single dialectal Hausa form (!) representing the entirety

of Chadic; and a late Egyptian form ʕpr that is clearly not the most basic term for this notion in Egyptian (it is mḥ, well attested at all stages from the Old Kingdom to Coptic). Below that is the comparison of Hadza *kwe-* 'to give' with a small handful of Berber and Chadic forms whose Proto-Berber and Proto-Chadic statuses are not assured in the least, let alone their Proto-Afro-Asiatic antiquity. Meanwhile, from Peiros' Austroasiatic database, compare *b(ʔ)ɨːr* 'full', with reflexes in Khmer and Vietnamese, and *ʔVk* 'to give', with the same monoconsonantal match as in the presented hypothetical etymology (the Austroasiatic comparison is, of course, not to be taken seriously, merely to underscore how generally easy it is to find potential cognates in a significantly large pool of comparative data extracted from one family).

Admittedly, the author himself understands the issue at stake, adding in a footnote that "Hadza-AA matches representing a common AA... or at least a common AA branch root... are, of course, of much better *quality* than Hadza matches with a few isolate and disperse AA terms". The ensuing appeal to relative statistics, however, does not seem like a direct way to solve the outlined problem. Even if it *is* true, as claimed by Militarev, that Hadza consistently yields higher numbers of phonetic and semantic matches with various branches of Afro-Asiatic than Nubian or Kuliak, genetic relationship is not the only possible explanation in this case; at least some such surges may be caused by areal contact, and this is why it is particularly important to analyze them closely in order to understand whether genetic inheritance from a common ancestor is a more logical and economical reason for the matching.

This is where the second phylogenetic possibility comes into focus: namely, a specific surge in matches between Hadza and one specific branch of Afro-Asiatic could hint at Hadza being not one of the top-level branches, but rather a historical offshoot of some subdivision of this macrofamily, such as Chadic or Cushitic. Indeed, were this actually the case, it might have been easier to demonstrate the Afro-Asiatic affinity of the language isolate in question — simply because the absolute chronology of a "Hadza-Chadic" or "Hadza-Cushitic" taxon would be smaller, and thus, we could expect a relatively larger number of cognates with better identifiable phonetic correspondences.

Unfortunately, this possibility manifests itself in arguably the least useful way of all: according to Militarev, Hadza does in fact show a slightly closer affinity with both the Omotic and the Cushitic families of the macrofamily, but since Militarev's own glottochronology has Cushitic and Omotic splitting around the 10th millennium B.C. — barely a thousand years after the primary split of Proto-Afro-Asiatic itself — this has hardly any phylogenetic or etymological significance. Moreover, a slight increase in the number of matches between Hadza and "Cush-Omotic" (a highly dubious taxon in itself, according to my own lexicostatistical calculations) could be easily explained — at least in theory — by areal contact between Hadza and these branches, both of which (especially Cushitic) are Hadza's closest neighbors of all the Afro-Asiatic stock.

An additional observation about "the high percentage of coincidences with individual Chadic languages (Tera, Mubi, etc.), which is not easy to explain" puts the entire comparison in even higher jeopardy. Clearly, if Hadza is genetically related to Chadic, its lexicostatistical matches must be with Proto-Chadic, not with individual Chadic languages. If there are more such matches between Hadza and Tera or Mubi than there are between Hadza and Proto-Chadic, such a circumstance may be explained in three ways:

(a) the respective Tera or Mubi equivalents (for instance, Mubi *mḏé* 'good' = Hadza *muta-*; Tera *kiya* 'who?' = Hadza *ʔakʷ-*) do indeed go all the way back to Proto-Chadic roots with the same basic meanings. In this case, credible etymological scenarios must be proposed to demonstrate this, and explain why it is those roots and not the ones with wider distribution across the Chadic continuum that should be projected back to Proto-Chadic. Until this has been done to general satisfaction, such an explanation must be rejected;

(b) the extra links with Tera, Mubi, etc. represent areal contacts between speakers of Hadza and those of various subbranches of Chadic already after the split of Common Chadic into several distinct lineages. This is quite implausible on geographical and historical grounds, and would be nothing short of a miracle were it convincingly demonstrated;

(c) the extra links with Tera, Mubi, etc., are accidental resemblances. Given that, in absolute numbers rather than percentages, what we are talking about here is, at most, 2-3 cases out of 50, this is much less incredible than it might seem to the author of the hypothesis[2].

Returning to the extra links with Cushitic and Omotic, Militarev's occasional reasoning against explaining these as results of borrowing feels equally unsatisfactory. Thus, when commenting upon the striking similarity between Hadza *mitl'a* 'bone' and Dahalo *miĉĉ-o* id., he writes: "...lack of other known Dahalo loanwords in Hadza speaks against the idea of 'bone' (which belongs to the most stable part of the core wordlist and is borrowed extremely rarely) to be the only word borrowed into Hadza from Dahalo". However, if the Dahalo word is indeed traceable back to Proto-Cushitic **miĉ-*, this means that it is not necessarily Dahalo that might have served as the source of the borrowing, but any other Cushitic branch or language that was still preserving the lateral articulation of the affricate at the time of contact. As for the argument about rare borrowing, consider the situation of the nearby Ethiopian isolate Shabo (whose speakers are, in many ways, sharing the same conditions today as the Hadza) — its basic lexicon is, to a large degree, autochthonous, but the small bunch (about 10% out of the Swadesh 100-item wordlist) of recent borrowings from nearby Ethiosemitic, Omotic, and Surmic languages does include *ema-ka* 'bone' ← Majang (North Surmic) sg. *eme-nan*, pl. *eme-k* 'bone', implying that such a borrowing is not at all implausible in that region (Starostin 2017: 715).

In the end, arguably the only genuinely impressive piece of evidence that could tie Hadza to Afro-Asiatic as a whole is probably the Hadza paradigm of 1st/2nd person pronouns, which seems to be patterned along the same "N/T" principle as the majority of Afro-Asiatic: the opposition of *ono* 'I' vs. *tʰe* 'thou' is comparable to Proto-Afro-Asiatic **ʔanV* 'I' : **(ʔan-)tV* 'thou' — moreover, this seems to be a more or less exclusive isogloss between Hadza and AA, as the other pronominal systems on the African continent all seem to follow significantly different patterns. Considering that such pronominal paradigms are indeed among the most stable and long-lasting «building blocks» of language, this, in itself, would be a fairly strong argument in favor of the Afro-Asiatic roots of Hadza. Unfortunately, accidental paradigmatic matchings of this kind do occasionally happen, and without any additional corroboration this single argument will probably be not enough to validate the hypothesis.

Note also that the pronominal argument, *pace* Militarev, does not extend to the 1st pl. pronoun: Hadza inclusive 'we' *uni-bii*, compared by Militarev with Afro-Asiatic **nV(ħ)-* 'we', is to be morphologically analyzed as *u-ni-bii*, where *-ni-* is not a pronominal morpheme but rather a suffixed marker of inclusivity — cf. *u-bii* 'we (excl.)', as well as the corresponding female forms *o-bee* (excl.), *o-ne-bee* (incl.), clearly showing that the pronominal root here is **u ~ *o*. Accepting Militarev's comparison is only possible if we decide that Proto-Afro-Asiatic **nV(ħ)-* is originally a reinterpreted marker of inclusivity — an extremely implausible decision based on distributional grounds, and one that finds no support on properly Afro-Asiatic grounds. Consequently, this "match" has to be rejected, like so many others.

---

[2] To illustrate the possibility of chance resemblances, consider the following semantically exact and phonetically plausible "matches" between Ari (South Omotic) and Modern English: Ari *deʔ-* = 'die'; Ari *ʔaːni* = 'hand'; Ari *ʔi* = 'I'; Ari *naːmi* = 'name'; Ari *kay-* = 'go'; Ari *woʰ* = 'we'. Note that all of these words, with the exception of 'go', belong to the "ultra-stable" 50-item part of the Swadesh wordlist; and that a few more cases could be easily added by setting up some simple phonetic rules (e.g. Ari *seːn-* = 'stone' if one sets up a perfectly plausible rule of initial cluster simplification in Ari).

Of the non-pronominal comparisons, probably the only impressive match between a fairly widely distributed and reliably reconstructible (in the required basic meaning) Afro-Asiatic root and Hadza is the word for 'blood': Hadza *átʰaʔmá-* = AA *(ʔa-)dam-*, which is also likely to be the optimal candidate for 'blood' on the Narrow Afro-Asiatic (Semitic + Berber + Chadic) level. The biconsonantal match is difficult to brush away as a mere accident (though, of course, accidental biconsonantal matches are quite well-known in comparative linguistics), and impossible to explain as the result of contact with Cushitic or Omotic (since it is not attested in precisely those branches). Even so, a hypothesis of Hadza-Afro-Asiatic relationship whose only strong points — by "strong" I mean "etymological arguments resting on significant phonetic, semantic, *and* distributional evidence" — are confined to the 1st-2nd pronominal paradigm and the word for 'blood' would be considered a shaky hypothesis indeed.

My own general methodological stance, on which I have written in detail in many previous publications, is that we do not necessarily require some sort of "bulletproof", 100% rigorous evidence of genetic relationship in order to label some particular hypothesis of the latter as "promising" or "deserving of further investigation"; all hypotheses of genetic relationship can — at least in theory — be ranged along a probability scale. Thus, even the staunchest opponent of the Altaic hypothesis would probably have to admit that it is more *likely* for Turkic languages to be genetically related to Mongolic or Tungusic than to, say, Niger-Congo — and not just for reasons of geographic proximity, but for actual linguistic evidence as well (e.g. the remarkable similarity in pronominal systems). For this reason, I do not rule out the Hadza-Afro-Asiatic connection as impossible or implausible: a small part of the evidence collected by Militarev allows for a genetic explanation. However, this is such a small part that even accepting the Hadza-Afro-Asiatic connection as a "working" hypothesis can hardly be done before it is clearly and explicitly shown that this evidence is statistically more significant than, for instance, the evidence that links Hadza to the various "Khoisan" families, or, for that matter, any other genetic lineage on the African continent.

In conclusion, I have to stress once again — as I have already done previously in another critical assessment of the same author's conception of Afro-Asiatic (Starostin 2021) — that our current state of awareness on lexical data around the world, their internal connections and their degrees of (both accidental or non-random) similarity, strictly prohibits to treat macrofamilies like Afro-Asiatic the same way we would treat more shallow families like Indo-European. A lexicostatistical comparison between a set of living Afro-Asiatic languages and Hadza is about as useful as a comparison between a set of living Indo-European languages, Finnish, and Turkish to confirm or refute the Nostratic hypothesis (to the best of my knowledge, not even Sergei Starostin, to whose authority the author of the discussed paper constantly refers, ever engaged in such comparisons, sticking instead to calculating matches between more or less reliably reconstructible Proto-Indo-European, Proto-Uralic, Proto-Turkic, etc. etyma). The only truly meaningful procedure in this case requires, first and foremost, a diligent and accurate preparation of lexicostatistical lists for all the reliably reconstructible stages of the branches of Afro-Asiatic, removing all or most of the transparent innovations from comparison as obvious sources of "noise" for any further external comparison — a procedure that, as some of the examples discussed above demonstrate, would certainly take a lot of the Hadza-Afro-Asiatic comparanda proposed by Militarev out of the equation, but at the same time could perhaps strengthen the validity of some of the others. Until such lists (accompanied with all the necessary etymological explanations) have been made available, lexicostatistical demonstration of the Afro-Asiatic affinity of Hadza (or, in fact, any other potentially Afro-Asiatic language) is, almost by definition, impossible.

References

Hannig, Rainer. 2006. Ägyptisches Wörterbuch II / Hannig-Lexica 5: Mittleres Reich und Zweite Zwischenzeit. Kulturgeschichte der antiken Welt 112. Mainz: Philipp von Zabern.

Militarev, Alexander, Sergei Nikolaev. 2020. Proto–Afrasian names of ungulates in light of the Proto–Afrasian homeland issue. Journal of Language Relationship 18(3): 199–226.

Obst, Erich. 1912. Von Mkalama ins Land der Wakindiga. II. Die Sprache der Wakindiga. Mitteilungen der Geographischen Gesellschaft in Hamburg 26(1): 29–45.

Rottland, Franz. 1982. Die Südnilotischen Sprachen: Beschreibung, Vergleichung und Rekonstruktion. Berlin: Dietrich Reimer Verlag.

Sands, Bonny, Maarten Mous, Mauro Tosco, Andrew Harvey. 2023. Why Hadza is (probably) not Afroasiatic: a discussion of Militarev's "Hadza as Afrasian?" Journal of Language Relationship 21(2): ???.

Shorto, Harry L. 2006. A Mon-Khmer Comparative Dictionary. Canberra: Pacific Linguistics.

Starostin, Georgij. 2008. K voprosu o geneticheskoj prinadlezhnosti jazyka hadza [On the issue of the genetic affinity of Hadza]. In: V. F. Vydrin (ed.). Afrikanskiy sbornik 2007: 262–278. Saint Petersburg: Nauka. 262–278.

Starostin, Georgij. 2013. Jazyki Afriki: opyt postrojenija leksikostatisticheskoj klassifikacii. Tom I: Metodologija. Kojsanskije jazyki. Moscow: Jazyki slav'anskoj kul'tury.

Starostin, Georgij. 2017. Jazyki Afriki: opyt postrojenija leksikostatisticheskoj klassifikacii. Tom III: Nilo-saharskije jazyki. Moscow: Jazyki slav'anskoj kul'tury.

Starostin, George. 2021. The Difficulties of Reconstructing the Cultural Lexicon for a Macrofamily-Level Proto-Language (Based on the Afrasian Example). Etnograficheskoe obozrenie 2021(4): 267–273.

Stolbova, Olga. 2016. Chadic Etymological Dictionary. Moscow: Institute of Oriental Studies, Russian Academy of Sciences.

Theil, Rolf. 2006. Is Omotic Afroasiatic? Ms. Available online at: https://www.uio.no/studier/emner/hf/iln/LING2110/v07/ (accessed 06.10.2023).

Г. С. Старостин. Как можно было бы показать, что хадза — афразийский язык (ответ на статью А. Ю. Милитарева «Хадза — афразийский язык?»)

В настоящей статье, представляющей собой ответ на гипотезу А. Милитарева об афразийской аффилиации языка-изолята хадза, проводится попытка определить главные теоретико-методологические недостатки лексикостатистического анализа А. Милитарева и показать, почему этот анализ нельзя считать доказательным. В качестве конструктивной альтернативы изложено, как метод, используемый А. Милитаревым, может быть усовершенствован для получения более надежных и исторически достоверных результатов; важнейшим условием такого усовершенствования является переход от использования данных живых языков к методу ономасиологической реконструкции 100-словных списков для промежуточных праязыковых состояний внутри афразийской макросемьи.

Ключевые слова: языки-изоляты; хадза язык; койсанские языки; афразийские языки; дальнее родство языков; лексикостатистика.