Comment on:

[Author *et al*]. 2010. [...] *Indo-European language evolution*. [xxxx]

Arnaud Fournet

The original paper purports to be a survey of a number of Indo-European languages. It resorts to the so-called "basic vocabulary", a concept dating back to the 1950s. The paper's dataset and first part appear to follow quite closely existing works by Dyen *et al*. The 84 supposedly "Indo-European" languages surveyed in the original sources include Creoles like Surinam's Takitaki, a situation overlooked by the authors. In addition the paper resorts to an inadequate approach to the concepts of cognacy and borrowing. The claim that a huge number of words (61%) in Indo-European languages would be borrowings must be rejected.

Keywords: Indo-European languages; phylogenetics; historical linguistics

---

Note: As the reader will discover this *Comment* is extremely critical of the article. It has been refused by the Journal. The names of the authors and the Journal are withheld.

The original "article" by Authors *et al.* (2010) purports to be a lexical survey of a number of Indo-European languages. It resorts to the so-called *basic vocabulary*: words that have been hypothesized by Morris Swadesh to be universally present in all languages and that were once considered to be best suitable for comparative purposes. Some linguists tend to consider that the so-called basic vocabulary is in addition more stable through time than other lexical items but this view has never gained widespread acceptance. On the whole the concept of *basic vocabulary* plays very little role in present-day mainstream linguistics. Originally the list included 200 words in 1952, but in the following decade Swadesh proposed an emended list with only 100 words. Later on another Russian linguist, the now deceased Sergei Yahontov, proposed an even shorter list of 35 items. Naturally each language has its own equivalents for these words. On account of the items shared or not shared by the subset of languages it takes into account Authors *et al*. (2010) generates a network of lexical connections between them and determines an internal branching which is supposed to have a phylogenetic relevance. It is claimed that it can detect borrowings and that an extremely high number of words (61%) should be considered borrowings.

To put it rather short and simple from the start, Authors et al. (2010) presents major problems of all kinds: in the data, in the perceived competence of the authors, in the framework, wordings and reasonings, not to speak about the absurdity of the "conclusion" reached at the end. Nearly nothing makes much sense from the point of view of a linguist with minimal competence and interest in comparative linguistics and in the Indo-European family and issues.

A first problem with Authors et al. (2010) is the data: its origin, its coherence and its relevance. Two sets of data are mentioned: (1) The first one is available online (Dyen *et al.* 1997). This source provides the 200 word list used as one primary corpus in the survey. It would appear that most of the paper is based on Dyen *et al.* (1992 and 1997). Authors et al. (2010) only cites Dyen *et al.* (1997) and does not mention the monograph by Dyen *et al.* (1992) which is nearly entirely readable on Google-

books. It can also be noted that this dataset is actually much older than 1997 as indicated in the web reference itself: "This file contains data that were placed on punched cards in the 1960s [sic], and transferred to disc circa 1990." (2) The other "reference" mentioned reads: [note] "30 Starostin, G[eorgij]. 2008 *Tower of Babel: an etymological database project*. See http://starling.rinet.ru". It includes word forms for 110 basic vocabulary items for a total of 98 languages" (Authors et al. 2010). In fact the address given by Authors *et al*. (2010) is the home page of a website, where a number of Russian macro-comparatists belonging to the Moscow school publish data and works. This "reference" just does not exist: neither the name of Georgij Starostin nor the dating exist. I have not been able to find any word list on the site contrary to what Authors et al. (2010) state: "The second dataset is based on etymological dictionaries and Swadesh lists [sic] published by the ToB project [30]." The website indeed has a wealth of very useful etymological databases but it does not seem to include any Swadesh-Yahontov word-list. The authors have added on their own website a number of supplementary "resources". None of these files can be directly traced back to the database files available on the Russian website. These files obviously contain an inverifiable, unexplained and untraceable number of modifications and operations. In all cases the second "dataset" cannot be referred to as being "Starostin, G[eorgij]. 2008." It is not even based on any works from that person. As the Russian website explains: "the Indo-European database [has been] compiled on the basis of Walde-Pokorny's dictionary by S. L. Nikolayev." The second "reference" is just completely impossible and unacceptable.

A second problem is the languages surveyed in Authors *et al*. (2010). The number and nature of the languages involved in the study are incoherent and absurd. The first source (Dyen *et al.* 1992 and 1997) covers 84 languages and the second source, as represented by the files in the authors' *supplementary resources*, covers 98 languages. Apart from the sheer incoherence of the number of languages the authors never address the issue of the representativity of their two perimeters of languages when it comes to the Indo-European family. In particular the authors appear to be unaware that the 84 languages investigated in the original works of Dyen *et al.* include French Creole and Takitaki, which are egregiously not Indo-European in the usual and established sense of that word. Takitaki is an English-based Creole spoken in Surinam. It is unclear why Creoles were included in the set of languages surveyed in the original work of Dyen *et al.* (1992). The inclusion is not explained in the monograph. In all cases Takitaki is therefore one of the branches in Figure4 in Authors *et al*. (2010) and prides itself on being the closest "relative" of English within Germanic, sharing with English the highest percentage of what Authors *et al*. (2010) mistake as real cognates. This detail about Takitaki being included in the so-called "Indo-European" languages surveyed in Authors et al. (2010) provides conclusive information about the scientific level of the paper. Besides the 84 languages also include two varieties of French Creole, another feature that only makes the situation worse.

At this stage the conclusions from a sheer technical point of view are therefore that Authors *et al*. (2010) is rooted in the intellectual preoccupations of the 1960s. The dataset and the 84 languages are uncritically taken without any modification from Dyen *et al.* (1992 and 1997). The list of languages (or maybe the incoherent two lists of languages) is not updated nor even emended as regards Takitaki or French Creole. Figure2 in Authors *et al*. (2010) is worth comparing with the *tables of lexicostatistical percentages* in Dyen *et al.* (1992, pp. 102-117) and would appear to be significantly (and therefore absurdly) drawn from them.

The next problem is what can be called the perceived competence of the authors. From the very first lines of Authors *et al*. (2010) it becomes glaringly obvious to any linguist that the authors do not have a minimal command of linguistic issues and methods. They confuse a word list with a

lexicostatistical survey. Items in word list are indiscriminately called “cognates”. Their amateurishly approximative and nearly false approach of cognacy, sound correspondences and of the comparative method in the introduction does not bode well for the rest of the paper. Apart from overlooking that Takitaki and French creoles are not Indo-European languages, it would appear that Authors et al. (2010) deals with Irish and Old Irish as being two separate languages and not two historical stages of the same language. It would also appear that the authors are not capable of providing all the names of the languages they try to deal with: for example CRN in a list of Celtic languages is not recognized as being Cornish and left as an acronym.

But the most debatable part comes now. In addition to an unverifiable second dataset, and an uncritical reliance on Dyen *et al.*'s works, which contained languages that should have been eliminated from their survey, Authors *et al*. (2010) make their idiosyncratic claim: “The vertices [the languages] are interconnected either by the branches of the reference tree, representing vertical inheritance, or by lateral edges, representing horizontal transfer [borrowing] (figure 4a).” (Authors *et al*. 2010) This claim which underlies the “conclusion” of massive and so far undetected internal borrowings within the Indo-European family is doubtless false. The difference between the branches of the reference tree and the lateral edges has nothing to do with the status of the words as inherited or borrowed. The branches of each subgroup (Romance, Germanic, Celtic, etc.) correspond to the most typical words that are representative of each subgroup. Nothing proves that the most caracteristic words of each subgroup are really inherited in the first place. A number of them appear to be borrowed from substrates in the course of Indo-European expansions and some are of unknown origin. Such is the case for *hand* or *woman* (< **wīf-man*) for example: words **handu* and **wīf* typical of Germanic languages but of unknown origin. On the contrary words which are shared by separate subgroups are very likely to be cognates and not borrowings, especially when the subgroups are geographically distant, a historically proven situation which is the exact opposite of the idiosyncratic and false claim made above.

From the first paragraph of the paper the authors appear to misunderstand what a cognate is, they misuse the word most of time, being unable to explain what it means and represents in a linguistically acceptable way (Cf. Authors *et al*. 2010). Unsurprisingly they ultimately reach completely unacceptable “conclusions”. Needless to say that the “conclusion” that 61% of the vocabulary of Indo-European languages would be borrowed can be discarded as should have been most of the paper in a reliable peer-review process.

References

Dyen, Isidore, Joseph B. Kruskal, and Paul Black. 1997. *Comparative Indo-European database: file IEdata1*. See http://www.wordgumbo.com/ie/cmp/iedata.txt.

Dyen, Isidore, Joseph B. Kruskal, and Paul Black. 1992. *An Indoeuropean Classification: A Lexicostatistical Experiment*. Transactions of the American Philosophical Society, vol. 82, part 5.

Starostin, Georgij. 2008. *Tower of Babel: an etymological database project*. [This pseudo-reference listed in Authors *et al*. (2010) cannot be verified] See http://starling.rinet.ru.