Review of Pagel et al. 2013: "Ultraconserved words point to deep language ancestry across Eurasia"

Jaakko Häkkinen, 14th May 2013

The aim of the authors is to reach beyond the conventional language families, all the way to the Ice Age superfamily from which many of the present Eurasian language families would have evolved. However, their statistical method can never reliably define any single concrete word to be inherited from this supposed "Proto-Eurasiatic" parent language.

Words change by form and meaning

It is well known that words change by their form. Within the language families there are many regular cognates, which are beyond recognition without historical linguistic training: the Latin word *rex* 'king' is a regular cognate of the Hindi *rājā* 'prince', and the Hungarian word *egér* 'mouse' is a regular cognate of the Finnish *hiiri* 'mouse'. It is rather difficult to find words which would be very similar by form in the remote ends of a language family.

Not only form, but also meaning (the semantic property of a word) can change. Etymological cognates give a very different picture from that of "normal" Swadesh list cognates based on meanings. A few observations between Finnish and Hungarian:

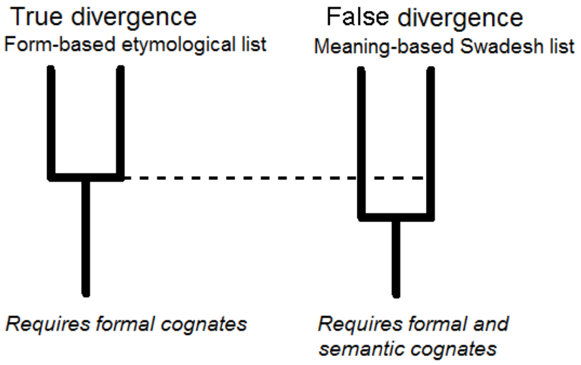
– For the meaning 'snake' the words are not cognates: Hungarian has *kígyó*, Finnish has *käärme*. The Finnish word is a Baltic loanword (cf. Lithuanian *kirmis* 'snake'), but Finnish also has a cognate for the Hungarian word: *kyy* 'adder' < **küji(wa)* 'snake' > *kígyó* 'snake'. The Finnish word has gone through a specializing semantic shift.

– For the meaning 'bird' the words are not cognates: Hungarian has *madár*, Finnish has *lintu*. But Hungarian also has a cognate for the Finnish word: *lúd* 'goose' < **lônta* 'bird' > *lintu* 'bird'. The Hungarian word has gone through a specializing semantic shift.

There are also other words in the Swadesh lists, which have cognates both in Hungarian and Finnish but which have undergone a semantic shift or become rarefied by newer words in one of the languages: for example Proto-Uralic **mälki* 'breast', **näki-* 'to see', **kuji* 'fat', **piŋi* 'tooth', **kêri* 'bark'. Such cognates are not seen in the Swadesh lists, because in the modern languages a more common word is used instead of the old cognate: in the respective order Finn. *rinta* 'breast' instead of *mälvi*, Hung. *lát-* 'to see' instead of *néz-*, Finn. *rasva* 'fat' instead of *kuu*, Finn. *hammas* 'tooth' instead of *pii*, Finn. *kaarna* 'bark' instead of *keri*.

These examples suffice to show that the meaning-based Swadesh list gives as a result clearly fewer cognates than the form-based etymological list, because the latter requires only *formal* cognates, while the former requires both *formal and semantic* cognates.

Taxonomically this means, that the meaning-based Swadesh list gives much more distant relationship between Hungarian and Finnish than the form-based etymological list. A possible source of error arises, if the two Uralic branches in question happen to have above the average amount of semantic shifts (changes in the meaning of the word). Such an occasion could result as an erroneous taxonomic illustration (family tree) of the language family: some other language with less semantic shifts or rarefications caused by newer words might seem to be closer than a language with more semantic shifts or rarefications.


True divergence
Form-based etymological list
False divergence
Meaning-based Swadesh list
Requires formal cognates
Requires formal and semantic cognates

Temporally the situation means, that the common protolanguage behind Hungarian and Finnish is erroneously assessed much older than it actually is. The meaning-based Swadeshian lexicostatistics simply ignores many cognates existing in the modern languages, even though all the cognates should definitely be taken into account when assessing the age of the split between the related languages. **Therefore the best basis for any lexicostatistic or glottochronological study is the form-based etymological list.**

There are also other problems in lexicostatistics and glottochronology, concerning the many distorting processes affecting the lexical level. Because of these processes the lexical level can never be reliable at its own, but the phonological level should always be applied, as it is the most reliable level for finding the true taxonomic structure of a language family. More about this subject can be found in my recent article "After the protolanguage: Invisible convergence, false divergence and boundary shift" in *Finnisch-Ugrische Forschungen* 61 (Häkkinen 2012b). I also mention the main points in my draft "Problems in the method and interpretations of the computational phylogenetics based on linguistic data – An example of wishful thinking: Bouckaert et al. 2012" http://www.elisanet.fi/alkupera/Problems_of_phylogenetics.pdf (Häkkinen 2012a).

The etymological procedure – or lack of it

Pagel et al. state:

> *"We have recently extended this result to include speakers from the Uralic, Sino-Tibetan, Niger-Congo, Altaic, and Austronesian families, in addition to Indo-European, plus the isolate Basque and the Creole Tok Pisin. Even in languages as widely divergent as these, we found that a measure of the average frequency of use predicted rates of lexical replacement as estimated in the Indo-European languages."*

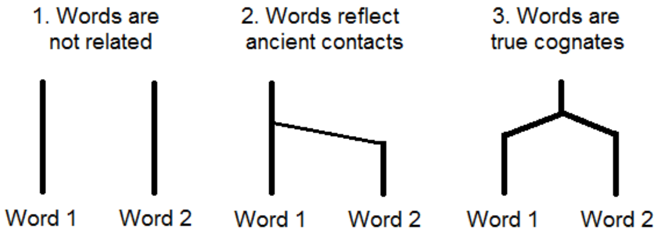
This is a very interesting and promising result. Yet it is important to understand what can and what cannot be deduced from this. The main problem with their method is the lack of normal etymological procedure, which is so elementary in historical linguistics. There is a certain procedure how they should proceed with their method, when trying to prove the words to be cognates:

1. Search for the formal cognates with similar enough meaning.
2. Try to distinguish the regular inherited cognates from the mutual borrowings by defining sound correspondences.
3. See if the words fulfill the criterion for being frequent in use.
4. Propose that a word in the present language families is inherited from the common Eurasiatic superprotolanguage.

The authors of the article do not follow this order – they totally skip the second item and subcontract the first item to their source database. But if they cannot prove the true cognateness of the words, their claims are baseless: **there is absolutely no way how high frequency and thus high age of the words could testify the words to be of the same origin!**

Without the correct procedure there are three equally possible options:

1. Words are not related (they resemble each other by chance).
2. Words reflect ancient contacts (borrowed from one language to another).
3. Words are true cognates (inherited from the common protolanguage).



All these kind of words can be old and in frequent use, so the method of Pagel et al. (2013) cannot distinguish between these options. How much they try, they cannot avoid the problem of the lack of proper etymological study of the words concerned. The authors rely on the etymological database of the Eurasiatic languages, but a glimpse is enough to reveal that this database does not contain reliable etymological data, either:

- It applies rather vague semantic criteria (like so many long-range comparisons), making it all the more probable that there are many chance resemblances included.
- It does not systematically present regular sound correspondences and does not exclude words if they do not fit into the regular pattern. Also the many perils of *omnicomparativism* should be pointed out, but this is not the place for that.
- It contains many loanwords (mainly due to the previous point), like Uralic **teki-* 'to do' from Indo-European **dheh-*, presented erroneously as cognates. It is highly improbable, that the words which are as similar between Proto-Uralic and Proto-Indo-European as the current "Indo-Uralic" words, could be inherited from their common protolanguage. It is impossible to explain how the languages could have developed so different from each other and at the same time not have gone

> through many major sound changes – especially when we see that all Indo-European and Uralic daughter languages/branches have gone through many differentiating sound changes.

The database merely collects similar-looking words with similar-looking meanings together, but this is not a scientific result yet – it is only a maximal corpus of data, which still requires a lot of etymological work with the help of historical phonology. **Before this work is done, we cannot tell true cognates, loanwords and chance resemblances apart from each other.**

So, neither the authors nor their source database possess any reliable etymological corpus. Therefore their method even theoretically could not prove any distant affinity between the Eurasian protolanguages. Considering the three possibilities above, their method has merely a chance of guessing to get it right: there is a probability of 33,3 % with every single word to be inherited, borrowed or chance resemblance.

the Eurasiatic words found with the method of Pagel et al. are kind of "Schrödinger's words": they might be or might not be true cognates, and there may not be any reliable method to ever define the state of any single word for certain.

Confusing stability with constancy

The authors write:

"Both objections can be overcome if it can be shown that: (i) a class of words exists whose members' sound-meaning correspondences are expected to last long enough to retain traces of their ancestry between language families separated by thousands of years; and (ii) these ultraconserved words can be predicted a priori and independently of their sound correspondences to other words."

Concerning the first point, they have only shown that the *meaning* of certain words is expected to last long without changing, and a word is expected to last without being *replaced* by newer words. They have not shown that this permanent word with a permanent meaning would not change radically *by form* (go through sound changes). Indeed, even though there may exist many words inherited from a protolanguage which are still used in modern languages with the original meaning, there are practically *no words* which would not have gone through sound changes in *all the languages*.

The authors seem to confuse *stability* (no formal shift) with *constancy* (no formal change).

1. Semantic shift (new meaning replaces the old meaning)
2. Formal shift (new word replaces the old word)
3. Formal change (the old word goes through sound changes)

The authors can thus exclude the first two processes but not the third one. Concerning their second point, there is no way to find out whether the "ultraconserved" words (conserved only in the sense of semantic shift and formal shift) represent chance resemblance, old borrowings or true cognates (see afore). The authors write:

"We use this framework to predict words likely to be shared among the Altaic, Chukchi-Kamchatkan - -, Dravidian, Eskimo (hereafter referred to as Inuit-Yupik), Indo-European, Kartvelian, and Uralic language families. These seven language families are hypothesized to form an ancient Eurasiatic superfamily that may have arisen from a common ancestor over 15 kya, and whose languages are now spoken over all of Eurasia."

These words are actually not likely to be shared between the mentioned language families, but they only form a basic corpus of the most promising words for further research. There may still equally well be both chance resemblances and ancient loanwords in addition to true cognates (if any). A thorough research with the methods of historical linguistics (etymology, sound history, loanword studies) are needed **before any single concrete word can be claimed to be inherited from the common Eurasiatic protolanguage.**

Computational phylogenetic programs cannot do the etymological research on behalf of the human researchers. A program can only calculate the words which have been defined as cognates by the researchers. No phylogenetic tree can ever make any assumption of a superfamily more probable – the program only gives out what is put in.

Problems concerning the dating

The first problem was already presented afore: the meaning-based Swadesh lists do not reach all the preserved cognates inherited from the common protolanguage, which makes the split between the languages erroneously look more ancient than it actually is. This innate dating bias concerns all the glottochronological studies and lexicostatistic datings which apply meaning-based lists for basic words.

Another serious problem is, that the authors rely on the results of Gray and Atkinson (2003). According to them Proto-Indo-European is ca. 8 700 years old. As I have presented elsewhere, the lexical level is prone to many processes distorting the dating, and there is no basis whatsoever to believe the method to be reliable. On the other hand, the traditional historical linguistic dating methods (including linguistic paleontology) are still credible – and so far the only credible method for dating a protolanguage.

When we have regularly inherited Proto-Indo-European words clearly refuting any pre-Copper Age dating versus computational phylogenetic result pointing to as early as the Neolithic dating, there is no question that the former one is the only scientifically valid result. (Häkkinen 2012a.)

An erroneous calibration point (Gray dating for Proto-Indo-European) of course affects the dating of Proto-Eurasiatic parent language, too.

Finally

It seems that the Gray School of computational phylogenetics, which includes Pagel et al., has taken sort of a rebel role: they try to shake the very basics of historical linguistics. The aim itself is praiseworthy, but the prime requisite for any success is the knowledge and thorough comprehension about the methods and results of historical linguistics. They seem to believe blindly in their own methods, which still seem to be full of problems and uncertainties, and they have arbitrarily decided to overrule all the results gained through the procedure of historical linguistics.

In order to become a credible agent in the field of historical linguistics they should communicate more with the methods and results of the field, not just ignore all the evidence and results opposing their view and swarm in their own views, which so often fail to see any problems and weaknesses in their own method.

Computational phylogenetics without a doubt may have something to give for historical linguistics, but it is not a miraculous device without shortcomings, which could totally replace all the traditional methods of historical linguistics.

Literature

Gray RD, Atkinson QD (2003): Language-tree divergence times support the Anatolian theory of Indo-European origin. – *Nature* 426 (6965): 435–439. http://www.nature.com/nature/journal/v426/n6965/abs/nature02029.html

Häkkinen, Jaakko 2012a: Problems in the method and interpretations of the computational phylogenetics based on linguistic data – An example of wishful thinking: Bouckaert et al. 2012. (Draft) http://www.elisanet.fi/alkupera/Problems_of_phylogenetics.pdf

Häkkinen, Jaakko 2012b: After the protolanguage: Invisible convergence, false divergence and boundary shift. – *Finnisch-Ugrische Forschungen* 62: 8–28.

Pagel et al. 2013: Ultraconserved words point to deep language ancestry across Eurasia. – *PNAS* May 6th 2013. http://www.pnas.org/content/early/2013/05/01/1218726110.abstract