Jaakko Häkkinen

After the protolanguage: Invisible convergence, false divergence and boundary shift

I discuss the processes involved in the birth of a language family: what kind of processes can happen or may have happened between the common protolanguage and the present-day languages. I do not consider the subject at a purely theoretical level, but rather through examples drawn from the Uralic studies. I name certain processes which have not (to my knowledge) previously been explicitly analyzed. I also argue that the taxonomic structure of a language family cannot be reliably reconstructed on the basis of the lexical level, and even less so if based on lexical retentions, which has been the common practice in lexicostatistic studies.[1]

1. Divergence and convergence

The primary process after the protolanguage is divergence: without divergence there is no language family. The process of divergence is normally represented by the figure of the family tree: the protolanguage splits up to daughter languages. Divergence affects all levels of language – phonology, morphology, syntax and lexicon – although not all of these have been made use of to the same extent.

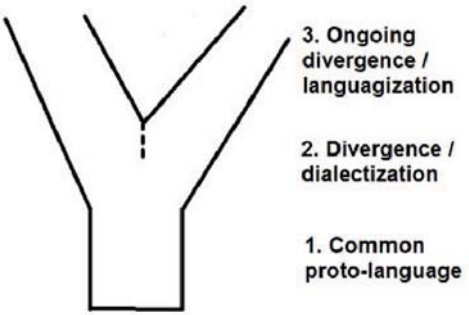


Figure 1: family tree (phonological)

In certain situations the separate languages may (more or less) unite again; this process is known as "convergence". (In this article I use the term only in reference to a contact-induced process; other denotations are irrelevant in this context).This occurs primarily at the lexical level. To illustrate convergence, another kind of family tree is needed:

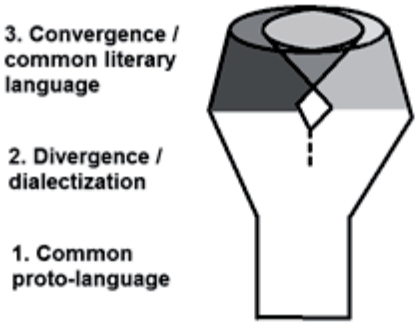


Figure 2: family tree (lexical)

If we compare Finnish to Hungarian, the first family tree is accurate enough: we see the divergence, and there is no convergence (since the languages have never been spoken in nearby areas). In illustrating the language families, the divergence tree is most often used, but under certain conditions adding the convergence level to the tree is informative – for example in comparing East Finnish to West Finnish, or North Estonian to South Estonian:

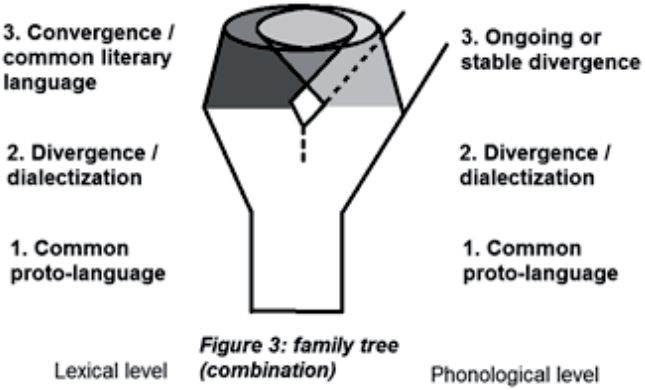


Figure 3: family tree (combination)

In both the abovementioned cases the earlier divergence can still be seen at the phonological level, while the later convergence can be seen primarily at the lexical level. This factor is connected to the separation of languages (language split): if the vernaculars cease to be mutually intelligible, a split occurs. Intelligibility, then, is dependent on the lexical level: separating sound changes are irrelevant, as long as there are shared words with shared meanings. Thus lexical convergence is the main factor behind the reunion of two dialectally split vernaculars.

It is very important to understand the coexistence of divergence and convergence: they are not mutually exclusive but may occur simultaneously. Normally it is quite easy to tell which features are due to divergence and which to convergence. However, there are certain processes which can lead to a false result, if we are not aware of them: *invisible convergence* and *false divergence*.

2. Invisible convergence

An example of invisible convergence is found in the case of the common Finno-Saamic vocabulary. By the calculations of Peter A. Michalove, Finnic and Saami share more inherited Proto-Uralic words than any other pair of branches (73 of 123 = 59,3%). (Michalove 2002.)

Over the last decade further evidence has emerged in favour of younger datings for different Uralic protolanguages. While the results are somewhat more vague for the earliest levels (Proto-Uralic), they are quite indisputable for the later levels (Late Proto-Finnic, Late Proto-Saami), which can be verified by the earliest written examples from the Germanic languages. (Kallio 2006.)

Finnic and Saami are the westernmost branches of the Uralic language family, and mutual contacts have occurred since the "beginning". This is also reflected in the external loanword layers: Finnic and Saami have received common loanwords from different stages of the Germanic lineage. We find Northwest-Indo-European, Pre-Germanic, Palaeo-Germanic, Early Proto-Germanic, Late Proto-Germanic and Northwest-Germanic loanwords, spanning a continuous period of about two millennia (Koivulehto 2002; Aikio 2006; Kallio 2009; Häkkinen 2010). There are also younger Scandinavian loanwords, but these are rarely shared by the two branches and thus are not equally diagnostic.

Elsewhere I have argued for the view, based on phonological evidence, according to which Proto-Uralic first split into two dialects, East-Uralic (> Hungarian, Mansi, Khanty and Samoyed) and West-Central-Uralic (> Finnic, Saami, Mordvin, Mari and Permic); the latter soon splitting further into West-Uralic (> Finnic, Saami and Mordvin), and possibly Central-Uralic (> Mari and Permic) or directly into Mari and Permic (Häkkinen 2007). West-Uralic is about the same level as Early Proto-Finnic and Early Proto-Saami.

Phonologically Finnic and Saami began to differentiate only during the Late Proto-Germanic layer (Häkkinen 2010), and it is only from this point onward that we can identify mutual loanwords from Finnic to Saami or from Saami to Finnic. The first word is a Germanic loanword, but the cognates in Finnic and Saami do not differ from the old words inherited from Proto-Uralic, while in the second word they do differ.

Old cognates: Finnic **h* ~ Saami **s*

Finnish *rauha(nen)* 'gland' < Middle Proto-Finnic **ravša* ← Proto-Germanic **hrauza-*
→ Early Proto-Saami **rawša* >Middle Proto-Saami **ravsa* > Late Proto-Saami **ruovsē* > North-Saami *ruoksa* 'udder' (Aikio 2006: 11)

Mutual loanwords: Finnic **h* ~ Saami **š*

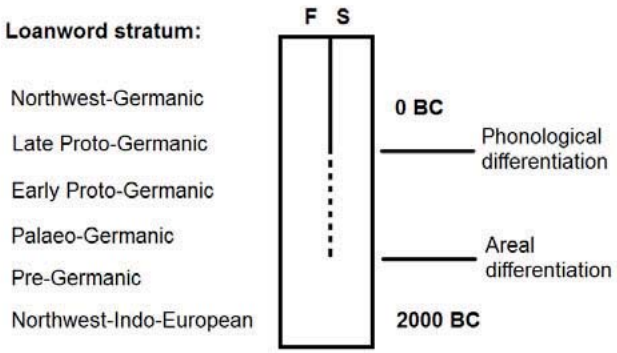
Finnish *paha* 'bad, evil' < Middle Proto-Finnic **paša* (← Northwest-Germanic **bāga-* < Proto-Germanic **bēga-*)
→ Middle Proto-Saami **paša* > Late Proto-Saami **puošē* > North-Saami *buošši* 'bad-tempered woman' (Koivulehto 1999: 202; Aikio 2006: 41)
(This word was borrowed into Saami after the old **š* had changed to **s*, and new **š* had occurred in the phoneme system; for the phonological development of Saami, see Korhonen 1981; Sammallahti 1998.)

Although Finnic and Saami are phonologically distinguished only at the end of the first millennium BCE, areally they separated much earlier, as can be seen in the substitution patterns of shared loanwords: even some Palaeo-Germanic loanwords have different derivatives in Finnic and Saami:

Finnish *kavio* 'hoof of a horse' < Late Proto-Finnic **kapja* < Early Proto-Finnic **kapa-ja*
~ South-Saami *guehpere* 'hoof, claw' < Late Proto-Saami **kuopērē* < Early Proto-Saami **kapa-ra*
← Palaeo-Germanic **kāpa-s* 'hoof' > Late Proto-Germanic **χōfa-z* > English *hoof* (Kallio, forthcoming)
(The Saami word was later borrowed to Finnish: *kopara* 'hoof of a reindeer'.)

With the help of the Germanic loanword strata, we observe that after the first signs of areal separation it took several centuries, perhaps half a millennium or more, before Proto-Finnic and Proto-Saami were phonologically distinguishable from each other. Consequently, it is very likely that words which are restricted to these two branches only, and which look like old cognates, are actually mutual or parallel (external) loanwords, borrowed during this long period of vicinity when the languages were still phonologically identical.

It is not always easy to distinguish between the phonetic and phonological levels in the reconstruction, but insofar as we are dealing with reconstructed languages mainly without any allophonic variation, we can speak of the phonological level. Losing allophonic variation, however, seems to be an innate feature of the comparative method: we can only reconstruct the stable, invariable situation underlying the variation, not *vice versa* (Korhonen 1974).



The names of different reconstruction levels in the Finnic and Saami lineage, together with concrete examples from different loanword layers, have been presented elsewhere (Häkkinen 2010). Only under very fortunate conditions can this kind of invisible convergence be linguistically "triangulated": in this case we were fortunate to find three branches (Finnic, Saami and Germanic) which share multiple layers of mutual loanwords. In this case the invisible convergence gives an erroneous result concerning the number of true inherited lexical cognates: the relationship between the Finnic and Saami branches seems to be closer and the split between them more recent than it actually is.

Another distorting effect is caused by the fact that Finnic seems to be the branch which has preserved the greatest number of inherited Uralic words (Michalove 2002): the lowest percentage of common words is shared with Hungarian (48 of 123 = 39%), and Finnic even shares more such words with Mansi and Khanty than Hungarian does (see below).

3. False divergence

A strong foreign influence may cause a false divergence. In an article about inherited Proto-Uralic words, Peter A. Michalove (2002) finds that Hungarian shares the smallest number of common Uralic words with the other Finno-Ugric branches: the greatest proportion is shared with Finnic (48 of 123 = 39%). Even Finnic shares more common words with Mansi and Khanty than Hungarian does:

(Uralic words)	**Finnic**	**Mansi**	**Khanty**	**Hungarian**
Finnic	—	**56**	**56**	48
Mansi	56	—	62	44
Khanty	56	62	—	45
Hungarian	48	**44**	**45**	—

If we also take into consideration the younger words of the Finno-Ugric layer, and the words shared by the Ugric branches (Mansi, Khanty and Hungarian) only, as in László Honti (1998), we get quite a different picture:

(All words)	**Finnic**	**Mansi**	**Khanty**	**Hungarian**
Finnic	—	**317**	**334**	287
Mansi	317	—	547	404
Khanty	334	547	—	381
Hungarian	287	**404**	**381**	—

Now Hungarian shares more words with Mansi and Khanty than Finnic does. The Finnic values are still remarkably close, even though Finnic is at the other end of the language family from the three Ugric languages. Finnic has apparently preserved a large proportion of the common Uralic and Finno-Ugric vocabulary. When we go back to the calculations of Michalove (2002) and include the Samoyedic branch from the eastern end of the Uralic language family, the picture changes again:

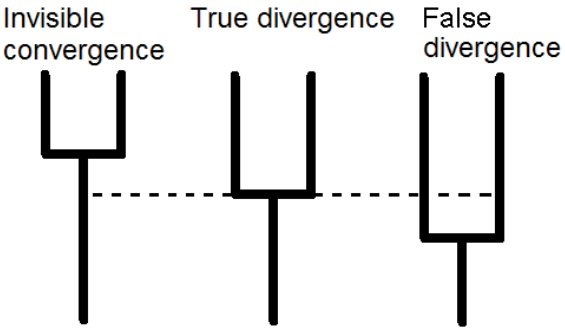
(Uralic words)	**Finnic**	**Mansi**	**Khanty**	**Hungarian**	**Samoyedic**
Finnic	—	**56**	**56**	48	96
Mansi	56	—	62	44	67
Khanty	56	62	—	45	73
Hungarian	48	**44**	**45**	—	61
Samoyedic	**96**	**67**	**73**	**61**	—

Samoyedic shares the greatest number of words with all the other branches, but this is due to the very nature of the classification: to be counted as Proto-Uralic, a word must have a cognate in Samoyedic. Thus Samoyedic has preserved 100% of the Proto-Uralic vocabulary, and all the Uralic words preserved in all the other branches are automatically shared with Samoyedic. Thus the figure in the "Samoyedic" column simultaneously shows the number of preserved Proto-Uralic words: Finnic 96 etc.

Consequently, it would indeed be difficult to draw a family tree based on Michalove's lexical data: Hungarian would be the farthest branch from every other branch (indicating early separation), while Samoyedic would be the closest branch to every other branch (indicating a very late separation). Therefore Michalove correctly omits Samoyed. He then recognizes the (lexically) isolated status of Hungarian and derives it directly from

Proto-Uralic, paralleled by the Finno-Permic, Ob-Ugric and Samoyedic branches (Michalove 2002). However, Michalove fails to see that the case of Hungarian is similar to that of Samoyedic; see below.

The above tables are presented to illustrate how difficult it is to estimate the taxonomic structure of a language family by counting words alone. A strong foreign influence can lead to a false divergence: the relationship between Hungarian and the other branches seems to be more distant and the split between them older than it actually was. Thus the false divergence distorts the actual relationship (true divergence) in the opposite direction compared to the invisible convergence.



4. Importance of the phonological level

In both cases, invisible convergence and false divergence, lexical evidence alone leads to an uncertain result. We cannot know for sure whether a common Finno-Saamic word is inherited or an early mutual or external loanword; similarly, we cannot know for sure whether the small number of Uralic words in Hungarian is due to the massive loss of inherited words (caused by intense foreign influence), or whether Hungarian was the first branch to separate from the Uralic unity. Only the phonological level can tell us which option is more credible: when different sound changes have the same distribution, it is highly probable that they indicate an ancient dialectal boundary. It would be against all odds to expect **1)** that numerous sounds were borrowed from the neighbours, **2)** that they all showed an identical distribution, and **3)** that they all replaced precisely the same sound the etymological cognate of which was borrowed. (Häkkinen 2007.)

In the case of Hungarian, we find that it shares many early sound changes with other Ugric and even Samoyedic languages; I have named this intermediate stage the East-Uralic dialect (Häkkinen 2007: 71–76):

From Proto-Uralic to East-Uralic:

1. **s* > **š* (coalescence with original **s*)
2. **š* > **L* (both original **š* and **s* change to voiceless fricolateral)
3. **ś* > **s* (secondary **s* occurs)
4. **ę* > **į* ~ **ę* (sporadic split; conditions not known)
5. **k, *w* > **γ* (coalescence with original **γ* < **x* between vowels)
6. **Sk* > **γS* (sibilant metathesis in some obstruent clusters)

There are old Indo-European loanwords, demonstrating that the East-Uralic sibilants are indeed innovations, while West-Uralic represents the original sibilant (Häkkinen 2009: 21):

Evidence from Proto-Aryan loanwords:

Hungarian *száz* ~ Mansi KM *sęęt* ~ Khanty V *sàt* '100'
< East-Uralic **sęta*
< Proto-Uralic **śęta* (> Mordvin *śado*)
←Proto-Aryan **ćata-*/ Proto-Indo-Aryan **śata-* '100'

Mansi K M *uutər* 'lord, prince; hero'
< East-Uralic **aLəra*
< Proto-Uralic **asira* (> Mordvin *azoro*)
← Proto-Aryan **asura*> Iranian *ahura* 'lord'

Thus Hungarian seems to be descended from the East-Uralic (i.e. Ugro-Samoyedic) dialect, and cannot have split off first right after the Proto-Uralic stage. The phonological level confirms that the lexical level leads to an erroneous result.

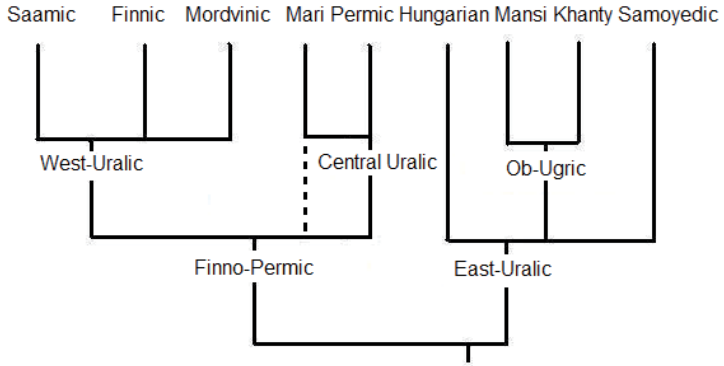
During its separate development and its spread from the Ural area to Central Europe, Hungarian borrowed for example many layers of Iranian loanwords (including some unidentified Old and Middle Iranian languages, Alan and Persian), and likewise many layers of Turkic loanwords (including an unidentified Old Turkic language, Old Bolgharian, Khazar, Cuman,

Pecheneg and Osman [Turkish]). There are also large numbers of Slavic and German loanwords in Hungarian. (Kulonen 1993.)

The case of Samoyedic is quite similar to that of Hungarian, although the earliest Palaeo-Siberian contact languages have been lost. There were contacts at least with Tocharian (Kallio 2004), Yukaghir (Rédei 1999) and Turkic (Janhunen 1998). Samoyedic also:

a) has moved far from the related languages and has been exposed to strong foreign influence
b) shares a small number of common words with other branches (from Sammallahti 1988: only 123 "Uralic" words, versus 390 "Uralic"+"Finno-Ugric" words found in other branches than Samoyedic = 31,5%)
c) derives phonologically from the East-Uralic dialect.

The phonological level is taxonomically more reliable, since it lacks the distortion caused by invisible convergence and false divergence at the lexical level. Thus we can conclude that the traditional taxonomic model, according to which Samoyedic was the first branch to split off from the Proto-Uralic unity, is just as incorrect as the view that Hungarian was the first branch to split off. The strong foreign influence and other processes, which reduced the number of inherited Uralic words, were mistaken for a sign of early divergence. Thus the most credible (or the least uncertain) family tree is that based on phonology, although the Central-Uralic node is still uncertain:



Uralic family tree according to Jaakko Häkkinen (2007)

5. Morphological level

In this article I have not focused on the morphological level, but tentatively it can be treated similarly to the lexical level: morphological convergence may likewise be invisible (if the already split languages/branches are still phonologically similar), and foreign influence may also cause morphological false divergence. However, the wearing of old case suffixes may give rise to a need for new, secondary case suffixes, as in the Permian languages and Hungarian. This might be seen as an *internally motivated* subtype of false divergence, while foreign influence represents an *externally motivated* type. The phenomenon can be included in the category of false divergence, since the more worn languages (such as Hungarian and the Permian languages; Kulonen 1993; Bartens 2000) will diverge morphologically from each other more than the less worn languages or branches (such as Saami and Samoyedic; Sammallahti 1998; Janhunen 1982).

It is also noteworthy that morphological categories are often compared loosely, for instance in terms of the case system. As there are fewer grammatical cases than words, and as cases often develop in clusters (e.g. three directional/local cases), two languages may seem to differ far more at the morphological than at the lexical level. The percentage of common morphological items may quickly be reduced to a much lower level than the percentage of common lexical items.

Due to these properties and the dual false divergence (external and internal), the morphological level may be even more vulnerable to distorting processes than the lexical one. Thus morphology is no more reliable or suitable for the taxonomic analysis of a language family than lexis.

6. Boundary shift (unstable split)

Stable split is the normal case: two vernaculars are first areally differentiated pre-dialects, then linguistically differentiated dialects, then finally they become different languages. In the context of prehistoric languages we need not be concerned with sociolects or with shifting political borders.

Unstable split is the case when the earlier dialectal split occurs in a different spot than the later language split. Here I present a possible case within the Uralic language family, which is ancient enough to represent the natural type of boundary shift. A similar unstable splitting process has also been suggested for Indo-European languages: Andrew Garrett argues

that the Graeco-Mycenean protolanguage was still almost identical with archaic Indo-European, and that Proto-Greek, Proto-Italic and Proto-Celtic arose only later, due to areal convergence in Greece, Italy and Central Europe respectively. There are some ancient features shared by the East-Greek and Anatolian branches, and others shared by the West-Greek and Italian branches, pointing to the possibility that the original dialect boundary was located differently from the later language (branch) boundary. (Garrett 1999; 2006.)

Mordvin consists of two languages, Erzya and Moksha, both having split into several dialects. The disintegration of (Late) Proto-Mordvin is dated to ca. 1000 CE (Bartens 1999: 15–16). The boundary between the two languages lies at the western (original) end of the Mordvin language area, along the lower Moksha River (right-bank tributary of the Oka). At this boundary zone we find a group of mixed dialects, counted as belonging to the Erzya language, called the Shoksha dialects. In certain villages belonging to this group there are interesting deviations from the common Mordvin forms of certain ancient words found in the rest of Erzya and Moksha dialects:

E:Kažl *viśkä*	< **viśkə*	~	Erzya, Moksha < **uśkə* 'metal chain' (~ Fi. *vaski*)
E:Kažl, E:Kal *viźir*	< **viśər*	~	Erzya, Moksha < **uźər* 'axe' (~ Fi. *vasara*)

This is not a regular sound change; there are words beginning with *u-* and *vi-* which have the same form in every dialect. These varying words are also quite old, or at least lacking in any recent loan etymology. I have searched for **e*-dialectal (Late Proto-Mordvin **i* goes back to earlier **e*) words only among words beginning with *u-* and *vi-/və-* in Volume 4 of the *Mordwinisches Wörterbuch* (Paasonen 1996), but even this narrow sample (ten words) gives us the maximal dialectal distribution of such words. It is interesting that the **e*-dialects do not follow the boundary between present-day Mordvin languages or even dialect groups. Using the abbreviations from Volume 1 of *Mordwinisches Wörterbuch* (Paasonen 1990), the **e*-dialectal words can be found only in the following areas:

Erzya:

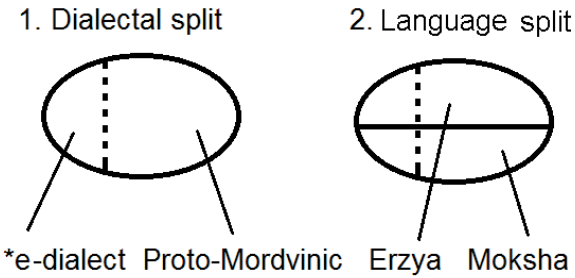
1. Shoksha dialects: Kad, Kal, Kažl, Šir
2. Samara area: Af, Ba, Nsurk, Večk

Moksha:

1. Pensa/Insar area: Al, P, Pš (SO), Sučk (M), Vert (Z)
2. Kazan area: Jurtk (M), Ur (M)
3. Saratov area: Sučk (M)

(M = mixed; SO = southeastern; Z = central)

In both languages the first group represents the older (more western) Mordvin area, while the other groups represent later migrations. All the villages in the older Mordvin area, which have preserved some **e*-dialectal word forms, are located along the Moksha River and its upper tributaries – from downstream to upstream they are near the towns of Tengushevo, Temnikov, Spassk, Narovchat, Insar, and lastly Penza by the Sura River. It is noteworthy that the **e*-dialects included in the Moksha language are divided into three different dialect groups: mixed, southeastern and central. This kind of independent distribution can be seen as supporting the idea that the **e*-dialects truly are the last remnants of an ancient **e*-dialect, which spread upstream along the Moksha. Later the language boundary within the Mordvin branch (between Erzya and Moksha) split the former **e*-dialectal area in two.



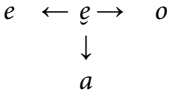
We can even follow the history of **e*-dialect further back in time. This dialectal split seems to have occurred even before the Late Proto-Mordvin stage (before the vowel changes **o* > **u* and **e* > **i*, and before the loss of **v* before a round vowel), and at the Early Proto-Mordvin stage we have the forms like:

**e*-dialect **veśka* ~ Early Proto-Mordvin **vośka* 'metal chain'
**e*-dialect **veśara* ~ Early Proto-Mordvin **vośara* 'axe'

Interestingly, these two words are included in the group of loanwords in West-Uralic, which show an irregular cognate set between Saami, Finnic and Mordvin:

Saami *e ~ Finnic *a ~ Mordvin *e / *o			
*veaćērē	~ *vasara	~ *viźər / *uźər	← A *vaśara* 'club, thunderbolt' (Joki 1973: 339)
*veaškē	~ *vaski	~ *viśkə / *uśkə	← Pre-Permic *węśka* 'copper' < U *wäśka* (Häkkinen, forthcoming)
*keaćē	~ *kasa	~ —	← IE *h_2aḱyā* 'edge (of axe)' (Koivulehto 2001: 241)
*leakšę	~ *la(a)kso	~ —	← Unknown

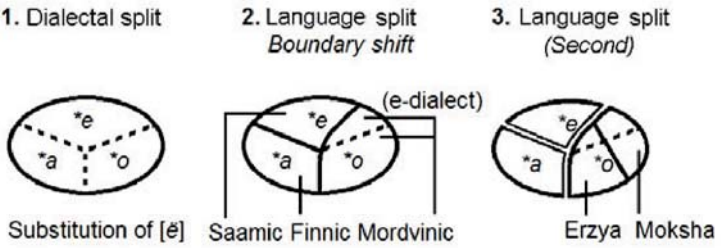
Elsewhere (Häkkinen, forthcoming) I have suggested that these words may reflect different substitutions of [ę] in the source language:



For the first and third word, phonologically there is an **a* in the source languages (Late Proto-Aryan and Northwest-Indo-European); these languages, however, had only three and five different vowel qualities respectively, versus ten vowel qualities in Proto-Uralic, and the phonetic value for **a* in the source languages was thus probably closer to the [ę] (IPA [ë]) of Uralic speakers, due to the larger realization spaces of the Indo-European vowels. There are examples of this same phonetic adaptation already in the Aryan loanwords of Proto-Uralic: U **śęta* '100' ← A **ćata-* | U **sęrńa* 'gold' ← Iranian, cf. Avestan *zaranya* (Häkkinen 2009: 21, 23). Pre-Permic is a label for an indefinite level before (Late) Proto-Permic but after the Permic branch had separated at least areally from other Uralic branches. One of the reflexes of Proto-Uralic **ä* in Permic is **ę* (Sammallahti 1988: 531), and the presence of large copper deposits in the Permic area would help to explain the spread of a Pre-Permic word into the West-Uralic pre-dialects.

While Proto-Uralic had a phoneme **ę* (Janhunen 1981 and Sammallahti 1988: **i*; Häkkinen 2007: **ę*), which in certain words was therefore substituted for an Indo-European/Aryan **a*, there was no longer such a phoneme in West-Uralic: it had coalesced with **a*. Speakers of the West-Uralic pre-dialects heard the reduced quality of Indo-European/Aryan **a*, but they replaced it differently by the closest phonemes: some with **a*, some with **e* and some with **o*.

In Mordvin, the usual reflex in these words is **u* (< Early Proto-Mordvin **o*), but in the **e*-dialects it is **i* (<**e*) – the latter representing identical substitution with Saami, where [ę] → **e*. Thus it is possible that here we see a true *boundary shift*: the language boundary emerges in a different place than the earlier dialectal boundary. It should be noted that the Finnic, Saami and Mordvin branches are based on many changes at every level of language, while the hypothetical connection between the Mordvin **e*-dialects and Saami is merely based on this one substitution pattern.



The scenario presented here, whereby the Mordvin **e*-dialects were originally part of the dialect which later became the Saami branch (and possibly also other, now lost branches), is also geographically plausible. The West-Uralic dialect is a phonologically relevant stage, and at least the following changes are common to all surviving West-Uralic branches (Saami, Finnic and Mordvin; Häkkinen 2007: 71–76):

1. **ę* > **a*
2. **δ'* > **δ* (between vowels)
3. **w* > ∅ (word-initially before a round vowel; occurred in Mari, too)

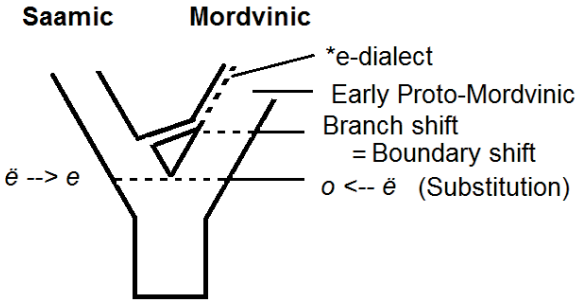
Moreover, the tripartite set of local cases with the co-affix **-s* (inessive **-sna*, elative **-sta*, illative **-sin*), as well as the development of the Uralic ablative **-ta* into the object-marking case, are common innovations exclusively in all three surviving West-Uralic branches (Korhonen 1981: 210–222; Grünthal 2007). Even the translative **-ksi* could be added to the list (Korhonen 1981: 229–230), unless it has a cognate in Samoyedic. West-Uralic seems to be an even clearer unit (genetic or areal) morphologically than phonologically. Mari is again similar to some extent: the Mari inessive *-štə, -štə̑* may also go back to **-sna*, and the Mari lative *-(e)š* may be compared to

Mordvin illative *-s* and the Finnic unproductive lative *-s*, but judging by the quality Mari is not a core member of West-Uralic.

On the lexical level, it suffices to note that Terho Itkonen has listed words shared (in different combinations) between Finnic, Saami, Mordvin and Mari; Mordvin is clearly closer to Finnic and Saami than Mari (Itkonen 1997). However, I have previously stressed that the lexical and morphological levels are far more unreliable indicators for protolanguage than the phonological level (Häkkinen 2007: 63–68), and in the present article I have further argued that distorting processes, such as invisible convergence and false divergence, affect only the morphological and lexical level (see the section "Morphological level" above).

When the West-Uralic dialect began to disperse and differentiate, this most probably happened near the mouth of river Oka. Saami has preserved the old Uralic numeral **luka* '10' (North-Saami *logi* ~ Mari **lŭw* ~ Mansi **låw*), while Finnic and Mordvin share the common innovation (loanword?) **kümmen* '10', as well as the words for 'oak' and 'maple', which also show irregular correspondences (Häkkinen 2009: 37–40).

As the ancient Pre-Proto-Mordvin area was on the right bank of the Oka and Pre-Proto-Saami remained north of the Upper Volga (concluding from the lack of external influence shared between Finnic and Mordvin), we can locate Pre-Proto-Finnic somewhere on the left bank of the Oka. Thus the *origo* of the figure above can be located approximately near the mouth of the Oka. As the distance between the mouth of the Oka and that of the Moksha is only about 200 km on the map, it is not too daring to assume that a group of speakers of ancient **e*-dialect (spoken north of the Upper Volga) moved southward to the mouth of the Moksha, where they were assimilated by the future Mordvin speakers of the **o*-dialect. This process can also be represented by a family tree figure:



It is of course also possible that the Mordvin **e*-dialects could be explained in some other way, but so far it seems economical to explain the situation by assuming a boundary shift: the **e*-dialects represent the descendants of the West-Uralic dialect, which replaced foreign [ę] by **e* and which today is mainly represented by the Saami branch. It would indeed be less economical to suppose that the Mordvin **e*-dialects independently changed these words – and only these words, not all possible **o*-words – into the **e*-pattern, which only accidentally resembles the Saami **e*-substitution; the more so as the **e*-dialects represent different dialect groups of both the Erzya and the Moksha languages.

It should nevertheless be noted that the figure above does not necessary represent the geographic situation: the gap between the branches is drawn only for the sake of clarity.

7. Effects on family tree

A branch in a family tree must be based on innovations, not retentions, since retentions do not distinguish the branch (and the corresponding intermediate protolanguage) from the ultimate protolanguage. On this basis we have to reject those family trees which are based on the amount of vocabulary inherited from Proto-Uralic – only the younger vocabulary, shared by the languages of a single branch, can be used as the basis for intermediate protolanguages in the family tree. (Salminen 2002.)

Even when we take into account only lexical innovations (shared intra-branch vocabulary), there are possible distorting processes, such as invisible convergence and false divergence, which cause the result to be highly uncertain. The phonological level should thus always be applied to verify the result; if there is a contradiction, the result based on the phonological level should be considered more certain and more credible.

Even if phonological convergence does occur, as in the case of the high reduced vowels between Proto-Permic and Volga Bolghar (Bartens 2000: 60), this does not blur the chronological relations of the phonological changes, since a sound change is visible in all words where a certain sound is present. Unlike a single word, a sound change does not simply disappear. By the means of internal reconstruction and external comparison, phonological changes can be placed in chronological order.

Earlier in this article I have argued that the most uncertain and thus most irrelevant level for the taxonomic point of view is the lexicon inherited from

the common protolanguage. However, this is precisely the level which is mostly applied in lexicostatistics. In the Swadesh lists, for example, the method is to calculate the number of common inherited words and construct family trees on this basis. This method can be considered trustworthy only under certain laboratory-like conditions; as in Oceania, where the various islands are located far from each other, effectively preventing invisible convergence, and were uninhabited before the spread of the Austronesian languages, totally preventing externally motivated false divergence. (Gray & Jordan 2000.)

We need to be aware that such favourable circumstances are very rare, and that distorting processes affect the dispersal of every continental language family – and most insular ones as well. We must not take the exception as a rule.

Is there any way to avoid or nullify the effect of the distorting processes on the lexical level? Some kind of calibration is clearly needed; however, accurate measurement of the distortion is thus far not possible. We can only get some hints by examining the circumstances: if two branches have been located in adjacent areas for a very long time, like Finnic and Saami, there is a real risk of invisible convergence. Prolonged adjacency is detectable by constructing the mutual external loanword layers (between Finnic and Saami) or shared ones (between the two branches and Germanic, for example). Phonological analysis of the loanwords may allow the identification of invisible convergence.

On the other hand, if a branch has continuously been located far from the related branches and shares a suspiciously low number of common words with all of them (as in the case of Hungarian or Samoyed), there is a real risk of false divergence. Phonological data can again confirm whether this is a case of true or false divergence.

8. Summary

There are two critical points to note:

1. A family tree must be based on *innovations*, not retentions. Finnic and Samoyed have some common features, but these are retentions inherited from Proto-Uralic and thus irrelevant. Similarly, the common Uralic lexicon represents retentions and is thus less important, while the lexicon common to some branches only can be seen as testifying more reliably to an intermediate protolanguage.

2. The lexical level is more uncertain than the phonological one, because of the bias due **a)** to *invisible convergence* (= common loanwords [mutual or external] increase the common vocabulary), **b)** to *false divergence*, caused by strong external (foreign) influence (= loanwords replace inherited words, reducing the common vocabulary) or by internal processes (wearing of morphological endings).

There are also different points in the process of branching in the family tree; it is, however, only on rare, fortunate occasions (as in the case of Finnic and Saami) that there are enough datable external loanword layers to make these all identifiable:

1. Common protolanguage
   West-Uralic dialect (~ Finno-Saamic)
2. Areal differentiation (seen at the morphological/derivational level?)
   Between Early Proto-Finnic and Early Proto-Saami
3. Dialectal differentiation (seen at the phonological level)
   Between Middle Proto-Finnic and Middle Proto-Saami
4. Language differentiation (seen at the lexical level)
   Between Late Proto-Finnic and Late Proto-Saami

It may not always be necessary – or even possible – to take all these phases into account. It is nevertheless important to be aware of the different phases and levels, in order to understand what the particular material can actually tell us, and what kind of taxonomic interpretation is even possible.

The following processes can be seen to affect the lexical level:

Process	**Effect**	**Example**
Invisible convergence	Separation of two branches seems shallower than it actually is.	Finnic vs. Saami
Lexical conservativeness	Portion of vocabulary shared with all other branches is suspiciously high.	Finnic
False divergence	Separation of two branches seems deeper than it actually is.	Hungarian vs. Ob-Ugric
Lexical innovativeness	Portion of vocabulary shared with all other branches is suspiciously low.	Hungarian, Samoyedic

In this article I have not been concerned with lexical conservativeness or innovativeness in any depth, but these factors are nevertheless important to take into account to avoid misleading conclusions. Fortunately they are quite easy to identify by comparing the shared vocabulary between the branches of the language family.

Jaakko Häkkinen
Department of Finnish, Finno-Ugrian
and Scandinavian Studies
University of Helsinki
jaakko.hakkinen@helsinki.fi

Note

1. I thank Juho Pystynen, Janne Saarikivi and Jussi Ylikoski, as well as two anonymous referees, for their valuable comments.

Literature

Aikio, Ante 2006: On Germanic-Saami contacts and Saami prehistory. – *Journal de la Société Finno-Ougrienne* 91. 9–55. Helsinki: Finno-Ugrian Society. http://www.sgr.fi/susa/91/aikio.pdf

Bartens, Raija 1999: *Mordvalaiskielten rakenne ja kehitys*. Mémoires de la Société Finno-Ougrienne 232. Helsinki: Finno-Ugrian Society.

— 2000: *Permiläisten kielten rakenne ja kehitys*. Mémoires de la Société Finno-Ougrienne 238. Helsinki: Finno-Ugrian Society.

Garrett, Andrew 1999: A new model of Indo-European subgrouping and dispersal. – *Proceedings of the Twenty-Fifth Annual Meeting of the Berkeley Linguistics Society, February 12–15, 1999*. 146–156. Ed. by Steve S. Chang, Lily Liaw, and Josef Ruppenhofer. Berkeley: Berkeley Linguistics Society. http://linguistics.berkeley.edu/~garrett/BLS1999.pdf

— 2006: Convergence in the formation of Indo-European subgroups: Phylogeny and chronology. – *Phylogenetic methods and the prehistory of languages*. 139–151. Ed. by Peter Forster and Colin Renfrew. Cambridge: McDonald Institute for Archaeological Research. http://linguistics.berkeley.edu/~garrett/IEConvergence.pdf

Gray, Russell D. & Jordan, Fiona M. 2000: Language trees support the express-train sequence of Austronesian expansion. – *Nature* vol. 405 / 29 June.

Grünthal, Riho 2007: The Mordvinic languages between bush and tree. – *Sámit, sánit, sátnehámit. Riepmočála Pekka Sammallahtii miessemánu 21. beaivve 2007*. 115–137. Ed. by Jussi Ylikoski & Ante Aikio. Mémoires de la Société Finno-Ougrienne 253. Helsinki: Finno-Ugrian Society. http://www.sgr.fi/sust/sust253/sust253_grunthal.pdf

Honti, László 1998: Ugrilainen kantakieli – erheellinen vai reaalinen hypoteesi? – *Oekeeta asijoo. Commentationes Fenno-Ugricae in honorem Seppo Suhonen sexagenarii*. 176–187. Mémoires de la Société Finno-Ougrienne 228. Helsinki: Finno-Ugrian Society.

Häkkinen, Jaakko 2007: *Kantauralin murteutuminen vokaalivastaavuuksien valossa*. Pro gradu -työ, Helsingin yliopiston Suomalais-ugrilainen laitos. https://oa.doria.fi/handle/10024/7044

— 2009: Kantauralin ajoitus ja paikannus: perustelut puntarissa. – *Journal de la Société Finno-Ougrienne* 92. 9–56. http://www.sgr.fi/susa/92/hakkinen.pdf

— 2010: Jatkuvuusperustelut ja saamelaisen kielen leviäminen (OSA 2). – *Muinaistutkija* 2 / 2010. http://www.mv.helsinki.fi/home/jphakkin/Jatkuvuus2.pdf

— (forthcoming): Kantasuomen keskivokaalit: paluu.

Itkonen, Terho 1997: Reflections on Pre-Uralic and the "Saami-Finnic protolanguage". – *Finnisch-Ugrische Forschungen* 54. Helsinki: Finno-Ugrian Society.

Janhunen, Juha 1981: Uralilaisen kantakielen sanastosta. – *Journal de la Société Finno-Ougrienne 77*. 219–271. Helsinki: Finno-Ugrian Society.

— 1982: On the structure of Proto-Uralic. – *Finnisch-Ugrische Forschungen* 44. 23–42. Helsinki: Finno-Ugrian Society.

— 1998: Samoyedic. – *The Uralic Languages*. 457–479. Ed. by Daniel Abondolo. Routledge, London and New York, 1998.

Joki, Aulis J. 1973: *Uralier und Indogermanen. Die älteren Berührungen zwischen den Uralischen und Indogermanischen Sprachen*. Mémoires de la Société Finno-Ougrienne 151. Helsinki: Finno-Ugrian Society.

Kallio, Petri 2004: Tocharian loanwords in Samoyed? – *Etymologie, Entlehnungen und Entwicklungen. Festschrift für Jorma Koivulehto zum 70. Geburtstag*. 129–137. Herausgegeben von Irma Hyvärinen, Petri Kallio und Jarmo Korhonen. Mémoires de la Société Néophilologique de Helsinki, LXIII. Helsinki 2004.

— 2006: Suomen kantakielten absoluuttista kronologiaa. – *Virittäjä* 110, s. 2–25. http://www.kotikielenseura.fi/virittaja/hakemistot/jutut/2006_2.pdf

— 2009: Stratigraphy of Indo-European loanwords in Saami. – *Máttut – máddagat: The Roots of Saami Ethnicities, Societies and Spaces / Places*, p. 30–45. Ed. by Tiina Äikäs. Publications of Giellagas Institute 12. Oulu: Giellagas Institute. http://www.mv.helsinki.fi/home/petkalli/mattut.pdf

— (forthcoming): The Prehistoric Germanic Loanword Strata in Finnic.

Koivulehto, Jorma 1999: *Verba mutuata. Quae vestigia antiquissimi cum Germanis aliisque Indo-Europaeis contactus in linguis Fennicis reliquerint.* Mémoires de la Société Finno-Ougrienne 237. Helsinki: Finno-Ugrian Society.

— 2001: The Earliest Contacts between Indo-European and Uralic Speakers in the Light of Lexical Loans. – *Early Contacts between Uralic and Indo-European: Linguistic and Archaeological Considerations*. 235–264. Ed. by Christian Carpelan, Asko Parpola and Petteri Koskikallio. Mémoires de la Société Finno-Ougrienne 242. Helsinki: Finno-Ugrian Society.

— 2002: Contact with non-Germanic languages II: Relations to the East. – *The Nordic Languages. An International Handbook of the History of the North Germanic Languages*. 583–594. Ed. by Oskar Bandle et al. Berlin – New York: Walter de Gruyter.

Korhonen, Mikko 1974: Oliko suomalais-ugrilainen kantakieli agglutinoiva? Eli mitä kielihistoriallisista rekonstruktioista voidaan lukea ja mitä ei. – *Virittäjä* 1974.

— 1981: *Johdatus lapin kielen historiaan.* 370. Helsinki: Finnish Literature Society.

Kulonen, Ulla-Maija 1993: *Johdatus unkarin kielen historiaan.* Suomi 170. Helsinki: Finnish Literature Society.

Michalove, Peter A. 2002: The Classification of the Uralic Languages: Lexical Evidence from Finno-Ugric. – *Finnisch-Ugrische Forschungen* 57, p. 58–67. https://oa.doria.fi/handle/10024/20282

Paasonen, Heikki 1990: *Mordwinisches Wörterbuch*. Zusammengestellt von Kaino Heikkilä. Bearbeitet und herausgegeben von Martti Kahla. Lexica Societatis Fenno-Ugricae XXIII: 1 / Kotimaisten kielten tutkimuskeskuksen julkaisuja 59: 1. Helsinki: Finno-Ugrian Society.

— 1996: *Mordwinisches Wörterbuch*. Zusammengestellt von Kaino Heikkilä. Bearbeitet und herausgegeben von Martti Kahla. Lexica Societatis Fenno-Ugricae XXIII: 4 / Publications of the Research Institute for the Languages of Finland 59: 4. Helsinki: Finno-Ugrian Society / Research Institute for the Languages of Finland.

Rédei, Károly 1999: Zu den uralisch-jukagirischen Sprachkontakten. – *Finnisch-Ugrische Forschungen* 55. Helsinki: Finno-Ugrian Society.

Salminen, Tapani 2002: Problems in the taxonomy of the Uralic languages in the light of modern comparative studies. – *Лингвистический беспредел: сборник статей к 70-летию А. И. Кузнецовой*. 44–55. Москва: Издательство Московского университета.
http://www.helsinki.fi/~tasalmin/kuzn.html

Sammallahti, Pekka 1988: Historical phonology of the Uralic Languages. – *The Uralic languages*. 478–554. Ed. by Denis Sinor. Leiden: Brill.

— 1998: *The Saami Languages. An Introduction*. Kárášjohka: Davvi Girji OS.