Statistical Dating of Finno-Mordvinic Languages through Comparative Linguistics and Sound Laws

Peter Sauli Piispanen

ABSTRACT

Through comparison of Swadesh-200 word list cognates and the employment of lexicostatistics, accompanied by detailed cognate sound changes, the branching of some Uralic languages have been statistically determined. Assuming linear branching from a line originating in Proto-Uralic and leading to modern Finnish, Moksha (Mordvinic) originated from 3423 BP, Northern Saami (Finno-Saamic) from 3038 BP and Estonian (Balto-Finnic) from 1058 BP. The resulting cognacy rates with Finnish (35.6 %, 40.0 % and 72.7 % respectively) and acquired dates are well in accordance with previous estimates as acquired by other methods.

Keywords: Dating, Estonian, Moksha, Northern Saami, sound change, Swadesh, Comparative Linguistics

Running short title: Dating Finno-Mordvinic by statistics & sound laws

1. Introduction

1.1. The linguistic analysis of lexical items is often employed to find and trace the development of genetically related languages. That two languages actually have cognates, i.e. words originating from the same etymon in their common proto-language, can be statistically proven by finding the same cognates in multiple languages. While a majority of scholars have criticized lexical mass comparison methods (cf. Campbell, L. 1986:488, 2001:45 & 2004:348), others, like for example Bengtson and Ruhlen (1994) have presented some very interesting and thought-provoking statistical arguments in favor of multi-language cognates actually being genetically affiliated cognates and not only separate inventions or borrowings.[1] Traditional lexical comparisons of this kind are done only through classification, i.e. multilateral lexical comparison. On a deeper second level, only some sound laws have been proposed for some cognates. Clearly, the addition of sound laws to the studies would deepen them and connect them to the linguistic mainstream. Thus, in this report great care has been taken to find acceptable sound laws for cognates.

1.2. No one has ever directly attempted the dating of the Uralic languages through lexicostatistics while also taking into account sound laws to the best of my knowledge. This study therefore aims to fill in some dating gaps by proposing statistical results from comparative linguistics and also from a mainstream linguistic standpoint by employing sound laws. This study also serves as a test of the lexicostatistical method *per se* as employed on the Uralic languages.[2] Interestingly, Sammallahti studied the relationships of the Saamic languages to each other by employing the Swadesh word list (Sammallahti, P. 1998), while Janhunen addressed the problems of analyzing languages and time depth, including for the

[1] For example, one particularly important argument is that, statistically speaking, basic words, of similar phonologic form and often identical semantics, can so often be identified with such a high number of common lexical roots between some of the prospective genetically related languages that they simply cannot be mere look-alikes or accidents; instead they must represent either a valid distant genetic language relationship or, at the very least, have extensively borrowed lexicon from the same sources.

[2] Merlijn De Smit, Jenny Larsson and two anonymous reviewers are gratefully acknowledged for their valuable input on the manuscript during preparation. This paper is a much updated and detailed version of old work performed during the author's Master's thesis in Finnish.

Uralic languages, elsewhere (Janhunen, J. 2008:223-239).

1.3. The lexicostatistical method was first described by Morris Swadesh (cf. Swadesh, M. 1950, 1952 & 1955), and later significantly improved, among others, by Starostin, S. (2000). The lexicostatistical method remains controversial and discredited in some circles (see for example: Dixon, R. 1997 & Renfrew, C. et al. 2000). However, improvements (cf. Starostin, S. 2000) and quite promising results (for example: Indo-European: Gray, R.D. & Atkinson, Q.D. 2003, Hamito-Semitic: Fleming, H.C. 1973, Semitic & Afro-Asiatic: Militarev, A., Austronesian: Sirk, Y. and also within Chinese languages and Native American languages) – which are well in accordance with the results from both archeology and genetics – may have increased the viability and accuracy of the method to a level where it is now ready as a proper tool for dating.[3] This should be particularly true if the method were also accompanied by the study of sound laws of the compared languages.[4] As a matter of fact, a lot of the criticism towards lexicostatistical methods is rooted in its usage as a replacement for normal comparative work, and, in particular, its application to possible language families where no comparative work has yet been done (for example Pama-Nyungan, as studied by Dixon). Deployment of the methods used herein into Uralic and Indo-European studies can be argued to be "control" experiments. The question should be: do we get sensible results and what does this imply for the validity of lexicostatistical results in general?

1.4. More recently, different statistical tools and methods have been employed and presented for dating the separation of languages. A very recent example pertaining to the Uralic languages includes Syrjänen, K. *et al*. 2013 & Honkola, T., *et al*. 2013 (using Bayesian phylogenetic analysis). An example from the dating of Indo-European languages is found in Bouckaert, *et al*. 2012, while the Uralic and Indo-European languages were statistically compared to each other in terms of the Swadesh word list in Kassian, A., *et al*. 2015 (also using Bayesian phylogenetic analysis). In particular, borrowings – and how they are handled in the analysis – often cause huge discrepancies in the results. Such results, while often lauded by news media, often have problems of their own, one of which is the arrival at very much older datings for various proto-language stages than what is usually accepted by mainstream linguistics. In the dating of Indo-European languages, for example, different methods and results are often presented, agreeing either with the Anatolian homeland hypothesis or the Kurgan hypothesis, while for example the work of Syrjänen's team arrived at higher time depths for various proto-languages than what is currently accepted by most Uralists. With such techniques, systematic sound laws are rarely employed for the determination of true cognates, thus differentiating them from

[3] In all fairness it has to be mentioned that while the method appears to work with "average" languages, certain languages are uniquely conservative and cannot at this point be dated using traditional lexicostatistics. There are known issues when comparing the time-depth of Indo-European languages (IE) acquired through lexico-statistical methods to the results from mainstream comparative IE studies; some accept the Kurgan hypothesis while others instead accept the Anatolian hypothesis, and the debate is still ongoing. A noteworthy – but far from uncontroversial – example of ultra-conservatism is Kusunda, a language in Nepal, which appears to have a rather clear genetic language relationship, to Indo-Pacific languages (cf. Whitehouse, P. , Usher, T., Ruhlen, M. and Wang, W.S-Y. 2004) that would go back perhaps 40-50 000 years (!)(cf. Bowler, J.M., Johnston, H., Olley, J.M., Prescott, J.R., Roberts, R.G., Shawcross, W. & Spooner, N.A. 2003) – according to lexicostatistical methods all lexical traces should have completely vanished at this point, and yet the relationship seems to remain strangely and unexpectedly clear. I postulate that other exceptions, of "too rapid" change, may be expected from languages that have experienced extremely frequent language contacts, social upheaval, (artificial) scholastic reforms and spread to groups of speakers of completely different original languages, and which may include English, Bulgarian, Albanian, and Norwegian Bokmål.

[4] This second stage, that is the changes by sound laws, consists of the traditional comparative historical method (and could here also be called Modified Etymostatistics or, perhaps, Phono-Glottochronology).

borrowings, the difference obviously being very important for the outcome of the results. To this author, the determination of actual cognates seemingly often boils down to too much guesswork in the absence of clearly defined sound correspondences, with the resulting lack of proof leading to unacceptably high margins of error in many studies. While sound laws are indeed an after-construction derived from lexical comparisons[5] such laws should regularly be employed in finding and comparing additional cognates. I note that a specific problem with this kind of analysis is the presence of synonyms, one of which may be vernacular; e.g. Fin. *missä* and *kussa* (see Appendix). In order to solve this and eliminate internal diversity as much as possible, I have explicitly based the comparison on literary, standard Finnish.

1.5. Hence, the study performed in this paper takes great care, as far as is reasonable for this large a data set, in formulating all cognates in terms only of what is provable both by regular and irregular, but often observable, sound changes. The modern starting point in the dating is set to the year 2010 when most of the lexical data was collected. Any lexical, orthographical or semantic errors are mine since I am unfortunately only a fluent speaker of Finnish and not of Estonian, North Saami or Moksha Mordvin. Indeed, some lexical items are very difficult to analyze regarding cognacy according to the given lexicostatistical rules and principles. Some new possible cognate suggestions between the Uralic languages will also be given in the appendix and some possibly hitherto non-discussed sound correspondences are given in the tabulated sound changes of Moksha.

1.6. In this study, Finnish was chosen as the token language to which to compare the others. The compared languages were chosen specifically to have separated in a tree-like model at different time depths – Estonian, Northern Saami and Moksha – and these were thus limited to before the Finno-Volgaic period.[6] Sound law analyses complemented the lexical comparison of the Swadesh-200 list for these languages in order to find the exact semantically unchanged cognates between them. This places the results within a more broadly accepted linguistic methodology (as has often been suggested for lexicostatistics but has actually not been carried out in detail as far as this author can ascertain). Naturally, the sound changes are more numerous the higher the time depth. The obtained *root glottochronologic* results were then, if possible, compared to the dating results obtained through paleontology, archaeology and genetics.

1.7 It must be noted that with this methodology and these results I do not claim that all languages change at a constant rate; rather, logically, there are periods of faster change and of slower change, which eventually do result in an average speed of change, and which is what the proposed retention factor of lexical change should represent. I imagine that the socio-linguistic conditions for language change are very different for hunter-gatherers and farmers/city dwellers. R.M.W. Dixon's punctuated equilibrium hypothesis – referring to his large body of work, particularly 1997,

[5] That is to say they are in the heuristic phase of the work: we find "similar" words and construct sound changes, but these sound changes are then used "predictably" to find cognates not at all similar, e.g. Finnish *hiiri* 'mouse' and Hungarian *egér* 'mouse'.

[6] I generally regard Proto-Finno-Volgaic merely as a lexical layer. This limited time depth was chosen to avoid any potential problems with the proposed alternative bush (cf. Häkkinen, K. 1984), linear comb and rake models (cf. Salminen, T. 2002) of the Uralic language family group. It also eliminates any possible problems that might arise from a case where there never has been a Proto-Uralic language, but merely a large *Lingua Franca* (cf. Wiik, K. 2002). On the other hand, others have instead divided Proto-Uralic into Western and Eastern Uralic (cf. Itkonen, E. 1966; Häkkinen, J. 2009), an agreeable take.

2002 – may support this hypothesis as it suggests that stationary hunter-gatherer groups will experience very slow grammatical convergence often with lexical retention (e.g. Australia, New Guinea) whereas expansive farmer/pastoralist groups will show a typical “language tree” with substratum effects from displaced languages.[7] The patterns and speed of language may no doubt change depending on social structure as well as geographic surroundings (for example, compare the Eurasian plainlands vs. the New Guinean jungle and mountains; these factors has been discussed by Nichols, J. 1992, *passim*). These factors probably mean different rates of language change for the language groups studied here with the exception of the nomadic Saami population; the languages studied in this work should mostly relate to settled populations. As such, the majority of results by the methodology used here may represent an approximate statistical dating result of when the studied proto-language was spoken somewhere.

2. Traditional dating of the Uralic language tree

2.1. Throughout the years a number of datings and places of origin of the Uralic languages employing several different methods have been presented. A select few of these are presented in the table below (summarized from: Anttila, R. 1989:301, Kallio, P. 2006, Janhunen, J. 2009 & Honkola, T. *et al*. 2013:1248 – here adding further materials to the table presented by Kallio).

Ref.	PU	PFU	PFP	PFV	EPF	LPF
Kettunen & Vaula 1938	4000 BCE	2500 BCE	1000 BCE	1000 BCE		
Toivonen 1953	3500 BCE	2500 BCE	1500 BCE	1000 BCE	500 BCE	
Décsy 1965	4000 BCE	2500 BCE	1500 BCE	400 BCE	1 CE	1000 CE
Hajdú 1975	4000 BCE	2000 BCE	1500 BCE	500 BCE		1 CE
Korhonen 1981	4000 BCE	3000 BCE	2000 BCE	1500 BCE	1000 BCE	1 CE
Anttila 1989	7000 BCE	5000 BCE	3000 BCE	1500 BCE	1250 BCE	
Taageperä 1994	4000 BCE	2100 BCE	1500 BCE	1400 BCE	1000 BCE	
Kallio 2006[8]	3650 BCE				1000 BCE	1 CE
Janhunen 2009[9]	3000 BCE	2500 BCE	1500 BCE	1000 BCE	500 BCE	1 CE
Honkola et al. 2013[10]	5300 BCE	3900 BCE	3700 BCE?	3200 BCE	1200 BCE?	800 BCE?
Abbreviations: PU (Proto-Uralic), PFU (Proto-Finno-Ugric), PFP (Proto-Finno-Permic), PFV (Proto-Finno-Volgaic), EPF (Earlier Proto-Finnic or Proto-Finno-Saamic), LPF (Later Proto-Finnic), BCE (Before Common Era), CE (Common Era)						

The table well shows the contrasts between various estimates and methods.[11] Clearly, the dating of the various proto-languages is difficult for several reasons. The perhaps currently accepted dating of the Proto-Uralic language places it at around 3000 - 4000 BCE.

2.2. A modified version of Janhunen’s (2009) previously published Uralic time depth table is presented below. It gives the representative languages with the number of speakers

[7] However, it must be noted that this hypothesis has been heavily criticized from various angles since its conception in 1997, see, for example: Crowley, T. 1999:109-115; Watkins, C. 2001:44-63; Janda, R. & Joseph, B. 2003: 3-180; Koch, H. 2004: 17-60; Campbell, L. & Poser, W.J. 2008:318-329.

[8] The datings of Kallio are mainly based on archeological/paleolinguistic findings.

[9] Janhunen notes that the south-to-north dimension of the Uralic language belt has a chronological depth of less than two millennia, but the geographic length of the east-to-west chain and the Mesolithic cultural level reflected by Proto-Uralic suggest a very early dating for the language family as a whole. The first split must have taken place before any contact with the Indo-Europeans due to the lack of loanwords in the earliest branches. Proto-Uralic itself is likely at a level earlier than the earliest stages of Proto-Indo-European (namely Indo-Hittite). It is noted that Janhunen’s dating estimates appear to be based under the assumption that a fully formed protolanguage forms every five hundred years.

[10] The datings of Honkola *et al*. are based on the principles of *Bayesian phylogenetics*. It also very tentatively estimates the breakup of Hungarian and the Ob-Ugric languages from Proto-Uralic to 3300 BCE, while the Ob-Ugric languages separated in 1900 BCE. The study notes that Mari separated from Proto-Finno-Volgaic in 3200 BCE, while Erzya broke up in 2900 BCE. The great difference between these dating estimates and those of mainstream linguists is noteworthy and seemingly suggest a scenario something akin to the suggested breakup of Proto-Uralic first into West and East Uralic as has been suggested in recent years (Häkkinen, J. 2009).

[11] Furthermore, one could divide up the estimates into different schools of thoughts, for example: M. Korhonen (traditional), K. Wiik (long time-depth), J. Janhunen (somewhat long time-depth), J. Häkkinen (shallow time-depth), and so on.

(mostly as given at the World Ethnologue website) and previously estimated time of separation from the Proto-language. The languages compared in this study have been underlined in bold for clarity, and represent the time depths in the rightmost columns of the table above.

0. **Finnish + Para-Finnish**
Modern time – 2010 CE
Finnish (~5.7 millions) (+ Meänkieli and Kven)

1. **Finnic**
(LPF) – 1 CE –1000 CE
Estonian (1.05 million)
Võro (70 000 spk.)
Ingrian (360 spk.)
Karelian Proper (45 000 spk.)
Ludic (3 000 spk.)
Olonetsian (19 270 spk.)
Livonian (2 spk.)
Veps (5 750 spk.)
Votic (15 spk.)

2. **Finno-Saamic**
(EPF) –1250 BCE – 1 CE
Western Saamic:
Southern Saami (600 spk.)
Ume Saami (20 spk.)
Pite Saami (20 spk.)
Lule Saami (2 000 spk.)
Northern Saami (20 700 spk.)
Eastern Saamic:
Kainu Saami (extinct)
Kemi Saami (extinct in the 19th century)
Inari Saami (300 spk.)
Akkala Saami (extinct in the 21st century)
Kildin Saami (500 spk.)
Skolt Saami (400 spk.)
Ter Saami (2-10 spk.)

3. **Finno-Mordvinic**
Erzya (696 630 spk.)
Moksha (614 000 spk.)

4. **Finno-Volgaic**
PFV – 1500 BCE – 400 BCE
Hill Mari (30 000 spk.)
Meadow Mari (460 090 spk.)
Merja (extinct in the 17th century)
Muroma (extinct in the 16th century)
Meshcherian (extinct in the 16th century)

5. **Finno-Permic**
PFP – 3000 BCE – 1000 BCE
Komi-Zyrian (217 000 spk.)
Komi-Permyak (94 300 spk.)
Udmurt (479 800 spk.)

6. **Finno-Ugric**
PFUg – 5000 BCE – 2100 BCE
Hungarian (12.5 millions)
6b. Ob-Ugric
Khanty (13 600 spk.)
Mansi (2 750 spk.)
6c. Northern Samoyed
Forest Enets (20 spk.)
Tundra Enets (10 spk.)
Nenets (31 300 spk.)
Nganasan (500 spk.)
Yurats (extinct)
6d. Southern Samoyed
Kamassian (extinct in the 20th century)
Mator (extinct in the 19th century)
Selkup (1 640 spk.)

7. **Uralic**
PU – 7000 BCE – 3500 BCE

8. **Pre-Uralic**
Pre-PU
Tundra Yukaghir (150 spk.) – Para-Uralic
Kolyma Yukaghir (50 spk.) – Para-Uralic

Additionally, prospective Para-Uralic entities have been found in the Yukaghiric languages (cf. Collinder, B. 1940 & 1957, Fortescue, M. 1998 (also additionally arguing for Eskaleut and Chukchi-Kamchatkan as Para-Uralic entities), Rédei, K. 1999, Wurm. S.A. 2001:27, Piispanen, P.S. 2013, 2015 & 2016) and perhaps also in the Chukchi-Kamchatkan languages (cf. Blažek, V. 2006).

3. The lexicostatistical method

3.1. The basis of the method used in this paper is according to the postulates of Starostin's *etymological statistics* (Starostin, S. 2000) given below verbatim:

> 1. In every language there are some roots that are original, i.e. not borrowed during the period of separate existence of this language. According to preliminary estimates, there are not much more than two or three hundred roots of this type in any modern language.[12]
> 2. These roots have different frequencies of occurrence, in other words they have different probabilities of being found in a chosen text.[13]
> 3. The frequency of occurrence (as just defined) of a given root in some language at a fixed period of time, t, is stable, and does not depend (or hardly depends) on the type of text.
> 4. All roots can 'age' – their frequency of occurrence gradually approaches zero, after which the root is considered to have disappeared from the language. At the same time, however, the rate of loss of different roots is not identical: roots, like words, may be divided into stable and less stable.[14]
> 5. The loss of roots from a language proceeds at a steady rate – that is, from some set of roots, characterized by a fixed frequency, over a given period, Δt, a fixed number of roots will be lost.

[12] Here it must be assumed that Starostin referred to very old protolanguages such as Proto-Uralic and Proto-Indo-European or perhaps languages from even higher time depths.

[13] i.e. there are differences between, for example, the rate of borrowing for cultural, technical, social and basic lexical items.

[14] The Swadesh-200 word list is considered to be a list of more stable word items.

Also, according to Starostin, the shorter Swadesh-100 word list will have a cognacy rate of 90 % or more for dialects, 70-80 % for closely related languages (such as Slavic, Romance, Germanic and Turkic), 25-30 % for Indo-European, 10-20 % for Uralic and 5-9 % for macro-families such as Nostratic. The rate of word loss from the Swadesh list has been calculated to 14 of 100 items per 1000 years (giving a retention constant of 0.86). The formula to calculate the point of divergence for two genetically related languages is thus:

$\Delta t = \log c / 2 \log r$

; where Δt = point of language divergence before present (in years), c =current rate of cognacy between two languages (0-1.00) and r is the so-called *glottochronological constant* of 0.86. In this study the larger Swadesh-200 word list has been employed for increased granularity and accuracy.[15]

3.2 It must be noted that while the Swadesh-200 list is perhaps not completely culturally or faunistically neutral it does seem to work well to describe Uralic cultural vocabulary.[16] As a non-speaker of Estonian, North Saami and Moksha it has been difficult to determine the most neutral lexical item appropriate for the comparisons; however, in order to not miss cognates between any of the languages, synonymic items, if they are deemed semantically identical, have been included in the analysis. Likewise, dialectal items are sometimes presented. In other cases dictionaries have been able to provide original cognates, but which are now clearly semantically different from each other, and hence considered different words in the study.

3.3 Further clarifications of the method of cognate selection are required. Acceptable cognates in the study are the words that clearly share a common origin in an earlier proto-language, are traceable through sound changes and retain the same semantic meaning still today in the two languages, i.e. semantically unchanged cognates as reasonably proven by sound laws. As such, a word can have been subject even to unexplainable or non-categorized phonemic changes, since not all sound changes in the studied languages are yet known (categories 9, 34 and 36 in the respective tables for Estonian, North Saami and Moksha sound changes), but still be considered a cognate. In contrast, words that are clearly borrowed only into that particular and narrow branch of languages, invented into that language or a recent proto-language or semantically changed words from what appears to have been the original meaning in the older proto-language (categories 10, 35 and 38 in the respective tables for Estonian, North Saami and Moksha sound changes) are disqualified as cognates in the study since they have been subject to the mechanisms of change. Finally, lexical items that are cognates with the common Finnish form are marked in blue color. Assuming uniformitarianism, items with a group of words in the entry of both Finnish and the compared language are particularly revealing in their comparison; if at least half of such word groups are judged cognates that word entry gets checked as a cognate set, indicating close genetic relationship. Relevant proto-items for the comparisons are given (usually as UEW refs.). Relevant, proposed sound changes for each language pertain only to the upmost lexical item if several are presented.

[15] Clearly the word choices consist of excellent ’eternal’ words that well fulfill the postulates of lexicostatistics. Should the list ever be extended, good candidates to add could be ‘to go’ (Fin. *mennä*), ‘home dwelling/nest/hole’ (Fin. *pesä*), ‘a lot’ (Fin. *paljon*), ‘to swallow’ (Fin. *niellä*) and ‘ear’ (Fin. *korva*).

[16] For example, since there are no snakes in New Zealand, Iceland, Greenland, Antarctica and Ireland a word for *snake* (found in the Swadesh-200) would not be a good basic word to include in the studies of languages in those areas, in contrast to all “Uralic” lands where snakes are plentiful.

3.4. In addition to the references given in the appendix of each respective language comparison, the following basic etymological dictionaries have also been employed for further tracing the Finnish language: Collinder, B. 1955, Häkkinen, K 2011, SSA 1992-2000, SKES 1955-1981 and UEW 1988-1991.

4. Estonian – Finnish

4.1. Estonian is known to be one of the closest relatives of Finnish. Finnish has at least 205 000 words, and if all adverbials and archaic words are included the total is 300 000 words (cf. Turunen, A. *in* Sinor, D. 1988:79). The majority of the modern language (80 %) was derived from the Earlier Proto-Finnic language with the remaining being borrowings and inventions.[17] Estonian too has the same amount of original words, borrowings and inventions. Most borrowings in Finnish are from Swedish, while in Estonian they are mostly from High and Low German (between the 13th century to the 1940s), but also from Finnish (since the 1870s), Old and Modern Russian and Latvian.[18] Still, the majority of the languages' words originate from Proto-Uralic, and the number of similar lexical items and the cognacy rate can be expected to be relatively high. The distribution and diachronic properties of the word borrowings suggest a geographic movement of the ancestral forms of Finnic (and Estonian and Saami) across the forest belt between the Urals and the Baltic Sea (cf. Janhunen, J. 2009). Estonian and Finnish likely separated sometime after Later Proto-Finnic, whose homeland may have included southern Finland, Estonia and Karelia[19] (cf. Kallio, P. 2009), and which occurred, according to the various estimates in the previous table, around 1 A.D.-1000 A.D.

4.2. While the languages may be similar from a syntactic standpoint, they differ somewhat from a phonologic standpoint. The known sound changes from Later Proto-Finnic into modern Estonian[20] are summarized below:

Sound Change	
1	Exchanging second or last syllable voiceless plosives into voiced plosives Example: Fi. *jalka* – Est. *jalg* 'foot'
2	The loss of vowel harmony Example: Fi. *suolet* – Est. *söakus* 'guts'
3	Assimilations (usually progressive) of the types: Example: -tk- -> -kk- ; -lt- -> -ll- ; -ns- -> -s(s)-
4	Only verb roots are compared, completely ignoring the infinitive endings Example: Fi. *-tV(k)* – Est. *-mV*
5	Other sound changes, such as elimination through gradation Example: Fi. *sidon* – Est. *seon* 'I bind'
6	The loss of certain diphthongs[21] Example: Fi. *punainen* – Est. *punane* 'red'
7	Elimination of end consonant
8	Elimination of end vowel
9	Unexplainable sound changes and non-categorized sound changes. Example: iśä -> *isa* 'father'
10	Borrowed, invented or semantically changed word
Notes: The sound laws are summarized from Turunen, A. *in* Sinor, D. 1988:65-69.	

4.3. By comparing Finnish and Estonian cognates on the Swadesh-200 list (Appendix A)[22] while taking into account the sound laws, the cognacy rate has been determined. Proposed sound changes are given in the last column for each word.

[17] Although there are only about 5 500-6 000 undivided stems in the Balto-Finnic languages (cf. Turunen, A. *in* Sinor. D. 1988:81) the rest being changes acquired through local changing circumstances of life. Of course Finnish is not a discrete Proto-Finnic branch, but rather a merger of at least three or so branches, with SW Finnish dialects closer to Estonian/Livonian and East Finnish dialects closer to Karelian (Sammallahti, P. 1977).

[18] Lexical borrowings: Low German 770-850 words, High German 490-540 words, Balto-German 60 words, Swedish 100-150 words, Russian 300-350 words, Latvian 30-45 words and Finnish 100 words (cf. Sinor, D. 1988). All in all Finnic has about 200 loanwords of apparently Baltic origin (cf. Kallio, P. 2008).

[19] As is rather strongly suggested by the various strata of loanwords in the different languages: Finnish, Estonian and Northern Saami. A list of loanwords from various eras appearing in Finnish (and closely related languages) can be found in Itkonen, E. (1966).

[20] i.e. after the assibilation took place, ex: *tulka > sulka (feather), *käti > käsi (hand).

[21] It should be noted that while Estonian has 25 diphthongs and Finnish only 16, these are much more commonly used in Finnish.

[22] Lexicon in Estonian-Finnish cognate appendix is collected from Greenberg, J.H. (2002) common Estonian dictionaries.

The cognacy rate, i.e. words derived from the same original word, is 149 words out of 205, i.e. 0.7268. The so-called *glottochronologic formula* gives:

Log 0.7268 / 2 log 0.86 = 1.058 = 1058 BP

This places the split between Estonian and Finnish at around 952 A.D. As such, the result is reasonable and fits rather well with previous dating results (i.e. LPF: 1 AD – 1000 AD).[23] The cognacy rate between Estonian and Finnish is higher than expected perhaps suggesting a few hard-to-determine borrowings already before the 18th century.

5. Northern Saami – Finnish

5.1. The Saami languages are spoken in a very large area ranging down from south of Idre in Dalarna (Dalecarlia in English) in central Sweden to the tip of the Kola Peninsula in Russia along the coast in an area 150-300 km wide. The representative among the Saamic languages was chosen to be Northern Saami which has the most speakers.[24] While three different hypotheses have been presented on the origin of the Saamic languages[25] the currently believed theory is that Proto-Saamic developed from Earlier Proto-Finnic (cf. Korhonen, M. 1981:23).[26] As such, Saamic apparently branched off much earlier than Estonian (as Pre-Proto-Saamic), which probably happened during the period of 1000 BCE – 700 CE (see Table above). More precisely, Proto-Saamic is believed to have disintegrated into a very diverse dialect by the middle of the first millennium A.D. while less than perhaps a millennium earlier (Pre-)Proto-Saamic had been a dialect of Proto-Finno-Saamic (cf. Kallio, P. 2009:38).

Interestingly, Proto-Finno-Saamic seems to have taken Pre- and Paleo-Germanic loanwords that spread from about 1700 BCE and onwards (cf. Kallio, P. 2009). Importantly, Aikio's research includes a language exchange hypothesis showing substratum vocabulary apparently, based on phonology, as being contemporaneous with Scandinavian loanwords, which must have consequences for the dating of the break-up of (Pre-)Proto-Saamic (Aikio, A. 2004, 2006, 2007).

Further, some Proto-Indo-European,[27] Proto-Indo-Iranian or Proto-Balto-Slavic loanwords[28] are found in the Saamic languages (cf. Kallio, P. 2009), but which are missing from Finnish, and thus suggest old origins, branching and northwards coastal language contacts for Pre-Proto-Saamic speakers,[29] likely in southern Finland during the Iron Age. While researchers such as Aikio and Junttila clearly contend that all Baltic items in the Saamic stem from indirect borrowings through Finnic, the viewpoint is probably correctly countered, for example, by the research of L-G. Larsson (2001:237-254).

Common Proto-Balto-Slavic loanwords that exist in Finnish (cf. Sammallahti, P. 1990), but only partly in Saamic are dated to ca. 1000 BCE (cf. Kallio, P. 2008), suggesting Finno-

[23] It is perhaps noteworthy that this places the split at perhaps approximately the same time, BP, as the tentative split between Catalonian and Spanish (*strict cognacy rate* 0.72; Harris, M. 1997).

[24] More specifically, Northern Saami belongs to the Western Saamic languages and is one of ten current Saami languages.

[25] Namely these: (a) Proto-Saamic developed when Samoyed people exchanged their original language with Proto-Finnic at some stage (*language exchange hypothesis*), (b) The speakers of Proto-Saami had originally spoken an unknown language, but started to heavily borrow from Finnish lexicon and morphology (*contact borrowing hypothesis*) and (c) Proto-Saamic developed from Earlier Proto-Finnic, which is the main, current theory for which evidence can be readily found.

[26] Namely: Earlier Proto-Finnic -> Middle Proto-Finnic + Proto-Saamic. Proto-Saamic then produced Western Saamic and Eastern Saamic.

[27] Example: Proto-Indo-European *ḱṷōn > Proto-Balto-Slavic *ś(u/v)ōn 'dog' -> Early Proto-Saamic *śa/ōvonji > Proto-Saamic *śuovunjë > Northern Saami *šūvon* 'well-trained dog'.

[28] Of the about 40 Baltic loanwords from this era, 30 also have cognates in Finnic. Finnic, however, has at least five times more Baltic borrowings than Saamic.

[29] Which is also implied by recent genetic studies of Saami populations in comparison to other European and Asian subsets (for example in: Tambets, K. et al. (2004)).

Saamic linguistic uniformity, but areal divergence at that time. Other interesting borrowing and language contact phenomena in Northern Saami are also indicated (Piispanen, P.S. 2012).

While the modern words originate from the same roots, numerous language innovations and sound changes render cognate recognition and understanding difficult. Still, the languages no doubt have a relatively close genetic relationship. Additionally, semantic changes sometimes make understanding more difficult.[30] The Saamic languages have around 550 words completely lack etymology (cf. Sammallahti, P., 1998:125)[31] – as such, these words have no cognates in the other Uralic languages or in any other, known language.[32] It can be assumed that the ancestral speakers of Saami have been into language contact with some other, now likely extinct language (cf. Aikio, A. 2006). At this time depth intelligibility between Finnish and Saamic can no longer be expected, and the cognacy rate will be relatively low.

5.2. The known sound changes from Earlier Proto-Finnic into various stages of Saamic are summarized below:

	Earlier	Western		Earlier	Northern
	Proto-Finnic	**Saamic**		**Proto-Finnic**	**Saami**
1	*-š(t,k,n)-	-*jh(t,k,n)-	16	*ä	á, i(e), ea
			17	*a-,*-a-	á, uo
	Earlier Proto-Finnic	**Proto-Saamic**	18	*o	uo, o, oa
2	*-k(l,ń,j)-	*-v(l,ń,j)-	19	*ō	uo
3	*-kj-	*-kš'-	20	*e	a, ie, ea
4	*-mp-	*-mb-	21	*ē	ie
5	*-nt-	*-nd-	22	*u	o
6	*-ŋk-	*-ŋg-	23	*ū	u
7	*ś- / -ś-	*ć- / -č-	24	*ü	a
8	*-š-	*-s-	25	*i	a
9	*-x-	*-k-	26	*ī	i
10	*-č-	*-c-	27	*-VKV-	-VKKV-
11	*-t(n,v)-	*-r(n,v)	28	*p-,t-,k-	*b-,d-,g-
12	*-tj-	*-rš-	29	*u(o)-a	o-i
13	*-p(δ,l,r)-	*-b(δ,l,r)-			
14	*-kη	*-gη-			
15	*-č'm-	*-ʒˇ'm-			
30	Assimilations (usually regressive) of the type: Example: *-nt-* -> *-nd-* -> *-dd-*				
31	Only verb roots are compared, completely ignoring the endings Example: Fi. *-tV(k)* – Saam. *-Vt*				
32	Other sound changes, such as insertions Example: *-au-* -> *-av-* & *-ok-* -> *-ohk-*				
33	Elimination of end consonant				
34	Unexplainable, non-categorized sound changes				
35	Borrowed, invented or semantically changed word				
Notes: V = vowel. The Western Saamic languages are Southern, Ume, Pite, Lule and Northern Saami. The sound laws are summarized from Sammallahti, P. 1998 & Korhonen, M. 1981.					

5.3. Comparing Finnish and Northern Saami cognates on the Swadesh-200 list (Appendix B),[33] while taking into account both known word etymology (mainly Sammallahti, P. 1998) and, importantly, the Àlgu tietokanta and the sound laws (as outlined above and presented in Appendix B), yields a cognacy rate of 82 words from 205 available words, i.e. 0.40 or 40.0 %.[34] The proposed sound changes are

[30] An example would be Fin. *pakkanen* 'frost, cold' <> N. Saami *báhkas* 'warm' (< Proto-Saamic *pakka- 'hot, cold').

[31] Examples: *atnit* (to use), *bivvat* (keep warm), *coagis* (low, shallow), *čáhppat* (black), *čiekčat* (to kick), *heavdni* (spider), *jalŋŋis* (stump), *jorrát* (to purr), *láhppit* (to loose, to spend), *nagir* (to sleep), *njivli* (mucus), *ohca* (lap), *oakti* (downpour), *ravgat* (to fall), *sarrit* (blackberry), *šiehttat* (to agree) and *uhcci* (small). A few of these words are present in the Swadesh-200 word list.

[32] Proto-Saamic speakers may have assimilated several layers of earlier languages in Europe from the early hunter-gatherers they encountered (cf. Aikio, A. 2006). However, it has also been theorized that when speakers of Uralic languages arrived in the eastern Baltic Sea region (*Textile Ware*, ca. 1900 BCE) they may have met people already there and in Finland who spoke versions of the Pre-Uralic languages they had brought with them during the Stone-Age waves from the Volga-Oka region (*Sperring Ware*, ca. 5100 BCE and/or *Combed Ware Style 2*, ca. 3900 BCE). These factors, of course, would make the identification of non-Uralic substrate items in Finnic more difficult to identify (cf. Kallio, P. 2009).

[33] Lexicon in Northern Saami-Finnish cognate appendix is collected from: Sammallahti, P. (1990), Sammallahti, P. (1998), Korhonen, M. (1981), Greenberg, J.H. (2002), and various dictionaries and from the Álgu tietokanta: http://kaino.kotus.fi/algu/index.php?t=etusivu . Further, Professors Erling Wande and Mikael Svonni are gratefully acknowledged for help with Northern Saami glossary.

[34] It has been suggested that there are a number of Finnic loanwords in Northern Saami, i.e. not inherited words, some of which are found on the Swadesh-200 list (cf. Aikio, A. 2007 and Sammallahti, P. 1990). The proposed loans are items 74, 146, 161, 186 & 204. This presents a possible conundrum where a language may have borrowed words from another branch of a

given in the last column for each word. The so-called *glottochronologic formula* gives:

Log 0.400 / 2 log 0.86 = 3.038 = 3038 BP

This places the split between Northern Saami and Finnish at around 1028 BCE The result fits rather well with earlier estimates (i.e. EPF: 1250 BC – 1 AD).[35] This split concretely represents Pre-Proto-Saamic, which much later became Proto-Saamic, the ancestor of all Saamic languages.

6. Moksha – Finnish

6.1. Moksha (and Erzya) are Mordvinic languages that originated from Proto-Finno-Mordvinic, perhaps as early as 1500 BCE. A third of the current Mordvinic speakers live in the Republic of Mordovia, a federal subject of Russia. The languages have fairly certainly been spoken in the area already since at least 1 CE, and the ancestors of the Gorodets culture of 500 CE are believed to have been speakers of the Mordvinic languages. There are currently 614 000 Mokshan and 696 630 Erzyan speakers, most living in the old RSFSR area.[36] The two languages are assumed to have separated around 1000 CE and are no longer mutually intelligible.

Mordvinic vocabulary consists of around 30 % each of inherited, invented and loanwords, while 10 % has no known etymology. Inherited vocabulary, of course, is inherited from Proto-Finno-Ugric. Among the invented vocabulary onomatopoetic words can naturally be found; all languages have them, but sometimes, such as in Estonian, onomatopoetic words are created consciously, like was done by Aavik who removed awkward compounds in favor of new Finnish borrowings and artificial new sounds to revitalize standard Estonian (Aavik, J. 1919). Loanwords in the language[37] can be dated to Finno-Volgaic,[38] Finno-Permic,[39] Finno-Ugric and even Uralic times.[40] In general, most loanwords are derived from Germanic, Baltic,[41] Indo-Iranian,[42] Turkic,[43] Tatar[44] and Russian words. Further, Mordvinic has been heavily influenced by Russian and it can be concluded that earlier speakers have been in intense language contact with speakers of Uralic and other languages (cf.

genetically related language, or perhaps even adjusted its own form of the word with the form of the related language. Borrowing itself is usually one of the mechanisms of language change that would make its inclusion as a cognate in lexicostatistics non-valid. However, there may be no way of knowing what the original Saamic word would have been before borrowing, or if Saami-speakers have merely corrected their speech to more closely resemble that of Finns with whom they were in close contact. If the proposed items are indeed true loanwords into North Saami at a stage later than Proto-Saamic and their inclusion as cognates were to be non-allowed, the cognacy rate would instead change to 77 words out of 205 (37.56 %), yielding a dating of 3246 BP.

[35] It is interesting to note that the Saamic languages between them have cognacy rates of 80-90 % on the Swadesh-100 word list (cf. Sammallahti, P. 1988:37). For this reason the Finnish-Northern Saami cognate comparison is quite representative as a whole for the Saamic languages. The loanwords refer to various proto-linguistic stages, eg. Indo-Iranian loanwords may go back to PFU times, etc. Further, curiously, the UEW reports several North Saami words that seem to have cognates only in the Samoyed languages, but nowhere else, for example North Saami *čävddë* (skin, bark) & Nenets *śāpt* (bark), North Saami *čâllât* (rub the antlers against something to get the skin off) & Nenets *śelā-* (id.) and North Saami *čoaw'je* (stomach, belly) & Kamassian *šɯjə* (id.).

[36] The speakers of Moksha live in Western Mordva, and its literary language is based on the dialects of Krasnoslobodskin and Temnikov. Erzyan speakers live in Northern and Eastern Mordva, and its literary language, written with the Cyrillic alphabet, is based on the dialect of Kozlovska. Previously there were also speakers of the extinct Muroman and Meshcherian.

[37] Examples: *meš* (bee) & *mije-* (to sell), both possibly Indo-European items.

[38] Example: *ĺišmä* (horse, from *lešmä, cf. Finnish *lehmä* (cow)).

[39] Example: *kunda* (cover, lid).

[40] For exhaustive lists of loanwords arrived in various eras and protolanguages, see Itkonen, E. (1966) and Häkkinen, J. (2009), and for the layers and sources of loanwords in Uralic, see Comrie, B. (2008:482) and Dahl, Ö. et al (2001).

[41] Example: *karda*, *kardo* (sheep; Lithuanian *gardas*).

[42] There are 18 such, known words. Mokshan examples: *pavas* (God, luck), *vaŕgas* (wolf), *śada* (a hundred) and *śura* (horn), as given in Bartens, R. (1999). A curiosity is that there are only 6 such items in Erzya, but 30 in Hungarian. Additionally, there is also at least one apparently pre-Ossetic loanword, namely *lomań* (person, Modern Ossetian *lymæn*/*limæn*)(found in the dialects, i.e. Mordvinisches Wörterbuch by H. Paasonen, 1990-1999, presented in Lexica Societatis Fenno-Ugricae XXIII). A useful source for Moksha Mordvin has been Herrala, E. & Feoktistov, A. 1998.

[43] There are perhaps around 190 such loanwords in Mordvinic from either Bolgar-Turkic or Tatar, as given in Paasonen, H. (1897).

[44] Example: *baška* (except, Tatar *baška*, id.). Many more examples can be found in Hasselblatt, C. *et al*. (2011).

Abondolo, D. 1998:211-217). For these reasons, and due to the high time depth involved, a low cognacy rate between Finnish and Moksha can be expected.

6.2. The known sound changes from Proto-Uralic and Proto-Finno-Ugric to the Moksha language are summarized below:

	PU or PFU	Moksha		PU or PFU	Moksha
1	*-i-	-e-	16	*-kk-	-k-
2	*-ü-	-e-	17	*t-	d-
3	*-ä-	-e-	18	*-t-	-d-
4	*-u-	-o-	19	*-tt-	-t-
5	*(-)e-ä	(-)i-e	20	*-p-	-v-
6	*-ī-	-i-	21	*-kp-	-kb-
7	*-ē-	-e-	22	*-pp-	-p-
8	*-ō-	-a-	23	* ń-	n-
9	*-ū-	-u-	24	*-ŋ-	-v- or –ø-
10	*-a	-ä	25	*č-	š-
11	*-ä	-ə	26	*sä-, se-, si-	śV-, śV-, śe-
12	*-e	-ä or -ə	27	*-es-	-iz-
13	*ü-	v-	28	*-ś-	sometimes -ź-
14	*-VkV-	sometimes -VvV-	29	*w-,-w, -w-	sometimes v-, v-, -v-
15	*-eke-	-ijä	30	*- δ-	-d-, -t-
31	Elimination of the first or last syllable Examples: *šukšna –> *šna* & *śüdämi –> śeďi				
32	Synkope, i.e. elimination of end vowel				
33	Only verb roots are compared, completely ignoring the endings Example: Fi. -tV(k) – Moks. -Vms				
34	Other changes, such as the voicing of consonant clusters Example: *tulka –> *tolga*				
35	Palatalizations, such as t- into t’-. Examples: *tä- –> t’ä & *näke- –> ńäjəms				
36	Uncategorized sound changes such as insertion, suffixation or irregular vowel changes. Words for which most - but not all - sound changes are known are also placed into this category.				
37	Further sound correspondences between Finnish and Moksha (see the text and footnote 41 for details)				
38	Borrowed, invented or semantically changed word				
39	Observed common change counter to the established sound laws (i.e rule 4): *-o- –> -u-				
Notes: V = vowel. The sound laws are summarized from the works of Gábor Bereczki *in* Sinor, D., 1988:316–331. For example, *t- means that the sound change pertains to the first syllable of the word, and *-ä pertains to the last syllable.					

It is further noted that, in certain cases, words - particularly often verbs - seem to have changed contrary to the established sound laws.[45] In addition to the sound changes presented in the table, several other not completely regular consonant cluster sound correspondences between Finnish and Moksha cognates could be observed[46] (such cognates are marked with 37 among the suggested sound changes).

6.3. Comparing the Swadesh-200 word list for the two languages,[47] while taking into account the outlined sound laws, gives 73 cognates out of 205 words, i.e. a cognacy rate of 0.356 or 35.6 %. The so-called *glottochronologic formula* gives:

Log 0.356 / 2 log 0.86 = 3.423 = 3423 BP

This would place the Finno-Mordvinic split at a rather high time depth of around 1413 BCE and chronologically shortly after the Finno-Volgaic split (a lexical layer estimated by most at around 1500 BCE). It is an interesting note that the cognates in Moksha Mordvin are more often found among the nouns than among the verbs. Indeed, the sound changes of verbs appear to have been complicated. Moksha Mordvin (MM) words are often very long, which may perhaps be due to more recent grammaticalizations (keeping the cognates; example: PFU *jäŋe ‘ice’ (UEW 93) > Fin. *jäätyä* ‘to freeze’, MM *äjənda(kšńə)ms* ‘to freeze’),[48] radical semantic changes or

[45] For example: *o- ->u- (noted as rule 39) is observed numerous times. Likewise, the sound *k* behaves oddly and occasionally changes into unexpected sounds (i.e. other than *j*, *v* or *g*) or completely disappears. Also, the change *ś- -> s-, and other palatalizations, seem to be rather common (noted as rule 35), as is *-e(-) -> -a(-) and perhaps not fully unexpected. While uncategorized sound changes (marked as rule 36) cannot be fully explained, such words can still be cognates.

[46] Such sound correspondences would tentatively include (Finnish <-> Moksha) and word item number given in brackets: (-)t- <-> (-)t’- (9, 149), s- <-> ś/sť- (68, 86, 125), (-)l(C)- <-> (-)lj(C)- (74, 96, 160, 180, 181), k <-> kj (78, 83), -si <-> -dj (83, 150, 183) and (-)h- <-> (-)sh- (35, 157). Further, I note that Moskhan *-ə-* appears in prosodically predictable positions.

[47] Lexicon in Moksha-Finnish cognate appendix is collected from Abondolo, D. (1998), Greenberg, J.H. (2002) and Moksha and Erzya glossaries found on the Internet.

[48] As can be inferred from Bartens, R. (1999:122-161).

borrowings from both Uralic and non-Uralic sources (exchanging the cognates; example: PU *ićä 'father, big' (UEW 78) > Fin. *isä* 'father', but MM *oćä* 'brother of father' – instead MM *al'ä* 'father' is used). Clearly, the time depth here is great enough that the similarities between the languages are quickly disappearing, rendering the recognition of true cognates a daunting task. All of these factors lead to a higher margin of error in the analysis and dating of Moksha in this study. Likewise, the known, proposed sound changes are not complete and in some cases not even exhaustive.

7. Khanty – Mansi - Hungarian

7.1. Here it is also of interest to discuss the question of Khanty (Ostyak in older literature), Mansi (Vogul in older literature) and Hungarian in greater detail. These Ugric languages originate from the Proto-Ugric (PUg) branch of Proto-Finno-Ugric, and later from Proto-Ob-Ugric (POUg). The cognacy rates on the Swadesh-100 word list between these languages have been reported as a mere 45 % for Khanty and Mansi, 34 % for Mansi and Hungarian and only 28 % for Hungarian and Khanty (cf. Sammallahti, P. *in* Sinor, D. 1988:499). This suggests that Hungarian split off from Proto-Ugric[49] (and what would become Proto-Ob-Ugric) 4220 BP, i.e. in 2210 BCE – a figure that actually fits quite well with later dating estimates of the somewhat earlier Proto-Finno-Ugric of 2000 BCE-3000 BCE Much later the Proto-Ob-Ugric entity broke up into Khantic and Mansic 2647 BP, i.e. in 637 BCE.[50] However, these tentative dates for PUg and POUg, respectively, would need to be verified by word comparisons on the larger Swadesh-200 word list while employing sound law changes, as well as complemented by a new dating of Proto-Finno-Ugric itself by the methods outlined in this paper.

8. Results and Conclusions

Compared Languages	Cognacy rate	Point of divergence	Proto-language
Finnish-Estonian	72.7 %	1058 BP	LPF ~ 952 CE
Finnish-Northern Saami	40.0 %	3038 BP[51]	EPF ~ 1028 BCE[52]
Finnish-Moksha	35.6 %	3423 BP	PFM~ PFV ~ 1413 BCE
Khanty-Mansi	45 %	2647 BP	POUg ~ 637 BCE
Hungarian-Khanty	28 %	4220 BP	PUg ~ 2210 BCE

8.1. The here obtained results are in rather good accordance with previously established datings done by linguists, suggesting that earlier linguists were on the right track from the start. All obtained points of divergence are, of course, only approximate years, leaving room for a certain margin of error.[53]

More noteworthy, the very brief time lapse between the apparent break-up of Finnic-Saamic-Moksha unity and Finnic-Saamic unity may actually cast doubt about the existence of a discrete Proto-Finno-Saamic language. Such a scenario would instead support the idea presented by Itkonen (Itkonen, T. 1997) that there was no such proto-language. At the

[49] And later turning into Proto-Finno-Permic, although Hungarian may have prior to this developed from a branch tentatively called Proto-Finno-Hungarian (cf. Sammallahti, P. *in* Sinor, D. 1988:491) - as suggested by etymological difficulties when comparing Hungarian to Proto-Ob-Ugric and Proto-Finno-Permic - which somewhat complicates the picture.

[50] It has been noted by Sammallahti that, while some have dated the disintegration of the Ob-Ugric unity to about the year 500 A.D., the relatively small number of common etymologies do suggest a considerably earlier disintegration (cf. Sammallahti, P. in Sinor. D. 1988:502). The result above does agree with this hypothesis.

[51] Alternatively, due to a possible loanword conundrum (see notes in the Northern Saami section), the cognacy rate may be only 37.6 %, which would yield a somewhat earlier dating of 3246 BP, and EPF ~1236 BCE.

[52] This would surprisingly suggest the existence of a Finnic dialect continuum starting with Early Proto-Finnic to change over two thousand years through a tentative Middle Proto-Finnic into Late Proto-Finnic, from which sprung a multitude of languages closely related to modern Finnish.

[53] Since most of the used lexical data was collected in 2010, this is the year used with BP to calculate the year of respective proto-language breakup.

same time, however, the relatively late break-up of Finnish and Estonian is in accordance with recent estimates (eg. Kallio) of a relatively “late” Proto-Finnic language.

The results suggest that the lexicostatistical method may work with the Uralic languages back to at least 4000 BP. (North) Saami and Moksha Mordvin appear to have separated from the proto-language that led to Finnish at only a few centuries in between.

8.2 The accuracy of some lexicostatistical methods has been estimated to have an error margin of up to 10 percent (Gell-Mann, M., Peiros, I. and Starostin, G. 2009:16). I suggest that the proper inclusion of sound laws for cognate recognition reduces this margin of error – as it eliminates human guesswork of look-alikes – possibly at best cutting it down by half or even two thirds. Thus, the error margin in this work is estimated to be approximately 3-5 %, which can be set as the +/- value to any of the above acquired BP-values. In conclusion, basing comparison on a steadfast phonetic ground probably instead makes the semantic classification the largest problem and source of error of the method.

8.3 Lexicostatistics is often used for languages on which virtually no historical-comparative work has been done. The herein used combination of traditional lexicostatistic methods with detailed traceable, sound change correspondences of cognates should be considered a new and quite innovative methodology, which seemingly allows for more sensible dating result to be obtained. This implicitly also partially confirms the reliability of lexicostatistical methods.

8.4 These results do encourage a future direction of research: the dating of higher and/or lower time depths by select, representative Uralic languages.

References

Aavik, Johannes 1919: Uute sõnade sõnastik. Sisaldab üle 2000 uue ja haruldasema sõna ühes tuletuslõppude tabeliga [Glossary of new words. Contains over 2000 new and lesser used words with tables of derivational suffixes]. Tartu: Istandik.

Abondolo, Daniel 1998: The Uralic Languages, London & New York: Routledge.

Aikio, Ante 2004: An essay on substrate studies and the origin of the Saami, Etymologie, Entlehnungen und Entwicklungen: Festschrift für Jorma Koivulehto zum 70. Geburtstag, Hyvärinen, I., Kallio, P. & Korhonen, J. (eds.), Mémoires de la Société Néophilologique de Helsinki, 63, p. 5-34. Helsinki: Société Néophilologique.

Aikio, Ante 2006: On Germanic-Saami contacts and Saami prehistory, *Journal de la société Finno-Ougrienne*, *91*, p. 9-55.

Aikio, Ante 2007: Etymological nativization of loanwords: a case study of Saami and Finnish, in: *Saami Linguistics*, Toivonen, I. & Nelson, D. (eds.), John Benjamins, pp. 17–52.

Anttila, Raimo 1989: *Historical and Comparative Linguistics*, 2nd edition, Amsterdam: Philadelphia: John Benjamins Publishing Company.

Bartens, Raija 1999: *Mordvalaiskielten rakenne ja kehitys*, Helsinki: Finno-Ugrian Society.

Bengtson, John D. & Ruhlen, Merritt 1994: 14. Global Etymologies,in: *On the Origin of Language Studies in Linguistic Taxonomy*, Stanford: Stanford University Press, p.277-336.

Bereczki, Gábor 1988: Geschichte der wolgafinnischen Sprachen, In: *The Uralic*

Languages. Description, history and foreign influences, 1988, Sinor, D. (ed), Leiden, p. 315-350.

Blažek, Václav 2006: Chukcho-Kamchatkan and Uralic: lexical evidence of their genetic relationship, *Orientalia et Classica XI. Aspects of Comparativistics, 2*, Moscow, p.197-212.

Bouckaert, Remco & Lemey, Philippe & Dunn Michael & Greenhill, Simon J., Alekseyenko, Alexander V. & Drummond, Alexei J. & Gray, Russell D. & Suchard, Marc A and, Atkinson, Quentin D. 2012: Mapping the Origins and Expansion of the Indo-European Language Family, *Science, 337*, p. 957–960.

Bowler, James M. & Johnston, Harvey & Olley, Jon M. & Prescott, John R. & Roberts, Richard G. & Shawcross, Wilfred & Spooner, Nigel A. 2003: New ages for human occupation and climatic change at Lake Mungo, Australia, *Nature*, *421*, p.837-840.

Campbell, Lyle 1986: Comment on Greenberg, Turner, and Zegura, *Current Anthropology*, *27*, p. 488. Note: This follows directly after the reviewed article by Greenberg, J. *et al*.

Campbell, Lyle 2001: Beyond the Comparative Method, In: *Historical Linguistics 2001 15th International Conference on Historical Linguistics*, Blake, B.J., Burridge, K. & Taylor, J. (eds), Melbourne, 13–17 August 2001

Campbell, Lyle 2004: *Historical Linguistics: An Introduction*, 2nd Edition, Cambridge, Massachusetts: MIT Press.

Campbell, Lyle & Poser, William J. 2008: *Language Classification*, Cambridge, p. 318-329.

Collinder, Björn 1940: *Jukagirisch und Uralisch*, Uppsala: Almqvist & Wiksell.

Collinder, Björn 1957: *Survey of the Uralic Languages*. Stockholm.

Comrie, Bernard 2009: *The World's major languages*, 2nd Ed, London and New York: Routledge.

Crowley, Terry 1999: *Book reviews – The Rise and Fall of Languages*, Australian Journal of Linguistics, *19*, p. 109-115

Dahl, Östen & Koptjevskaja-Tamm, Maria (eds)(2001) *The Circum-Baltic Languages, volume 1: Past and Present*, Amsterdam & Philadelphia: John Benjamins.

Décsy, Gyula 1965: *Einführung in die finnisch-ugrische Sprachwissenschaft*. Wiesbaden: Harrassowitz.

Dixon, Robert Malcolm Ward 1997: *The Rise and Fall of Language*, Cambridge: Cambridge University Press, UK.

Dixon, Robert Malcolm Ward 2002: Australian Languages: their nature and development (Cambridge Language Surveys), Cambridge: Cambridge University Press, UK, xlii.

Fleming, H.C. 1973: Sub-classification in Hamito-Semitic, In: *Lexicostatistics in genetic linguistics: proceedings of the Yale conference*, Yale University, April 3-4, 1971, Dyen, I. (ed), The Hague & Paris: Mouton & Co., p. 85-88.

Fortescue, Michael D. 1998: *Language Relations Across Bering Strait*, London: Cassell & Co.

Gell-Mann, M., Peiros, I. and Starostin, G. (2009) Distant Language Relationship: The Current Perspective, Journal of Language Relationship, *1*, p. 13-30.

Gray, Russell D. & Atkinson, Quentin D. 2003: Language-tree divergence times support the Anatolian theory of Indo-European origin, *Nature*, *426*, p.435-439.

Greenberg, Joseph H. 2002: Indo-European and Its Closest Relatives, volume 2. Lexicon. Stanford: Stanford University Press, California.

Greenhill, S.J., Atkinson, Q.D., Meade, A. and Gray, R.D. 2010: The shape and tempo of language evolution, Proc. Biol. Sci., Aug. 22, 277(1693), p. 2443-2450.

Hajdú, Peter 1975: Sukulaisuuden kielellistä taustaa, In: *Suomalaisugrilaiset*, Hajdú, P. (ed), Helsinki: Suomalaisen Kirjallisuuden Seura, p. 11-51.

Harris, M. & Vincent N. 1997: *The Romance Languages*. Oxford University Press.

Hasselblatt, Cornelius & Houtzagers, Peter & Pareren, Remco van (eds) 2011: *Language contact in times of globalization*. Amsterdam & New York: Studies in Slavic and General Linguistics.

Herrala, Eeva & Feoktistov, A. 1998: Mokšalais-Suomalainen Sanakirja, Turku.

Honkola, T., Vesakoski, O., Korhonen, K., Lehtinen, J. Syrjänen, K. & Wahlberg, N. 2013: Cultural and climatic changes shape the evolutionary history of the Uralic languages, J. of Evol. Biol., 26, p. 1244-1253.

Häkkinen, Kaisa 1984: Wäre es schon an der Zeit, den Stammbaum zu fallen? Theorien über die gegenseitigen Verwandtschaftsbeziehungen der finnish-ugrischen Sprachen, *Ural-Altaische Jahrbücher, neue folge 4*, *1-4*, Wiesbaden.

Häkkinen, Jaakko 2009: Kantauralin ajoitus ja paikannus: perustelut puntarissa, *Journal de la société Finno-Ougrienne, 92, p. 9-56.*

Itkonen, Erkki 1966: Suomen suvun esihistoria, *Tietolipas*, *20*, p.5-47.

Itkonen, Terho 1997: Reflections on Pre-Uralic and the “Sami-Finnic proto-language”, *Finnisch-Ugrische Forschungen*, *54*, p. 229-266.

Janda, Richard & Joseph, Brian 2003: Handbook of Historical Linguistics, Oxford, p. 3-180.

Koch, Harold, 2004: Australian languages: Classification and the Comparative Method, Bowern, Claire & Koch, Harold (eds.), Amsterdam, p. 17-60.

Janhunen, Juha 2009: Proto-Uralic – what, where, and when? *Suomalais-Ugrilaisen Seuran Toimituksia = Mémoires de la Société Finno-Ougrienne*, *258*, p. 57-78.

Janhunen, Juha 2008: Some Old World experience in linguistic dating, in: *In Hot Pursuit of Language in Prehistory: Essays in the four fields of anthropology*, Bengtson, John D. (ed.), Amsterdam/Philadelphia: John Benjamins Publishing Company, p. 223-239.

Kallio, Petri 2006: Suomen Kantakielen Absoluuttista Kronologiaa, *Virittäjä*, *1*, p.2-25.

Kallio, Petri 2008: On the ”Early Baltic” Loanwords in Common Finnic, In: *Evidence and Counter-Evidence: essays in honour of Frederik Kortlandt*, v1, SSGL 32, Amsterdam, p. 265-277.

Kallio, Petri 2009: *Stratigraphy of Indo-European Loanwords in Saami*. In: Tiina Äikäs (ed.), Máttut - máddagat: The Roots of Saami Ethnicities, Societies and Spaces / Places. Oulu: Publications of the Giellagas Institute 12, p.30-45.

Kassian, A., Zhivlov, M. & Starostin, G. 2015: Proto-Indo-European-Uralic Comparison from the Probabilistic Point of View, *The Journal of Indo-European Studies*, *43*:3-4, p. 301-392.

Kettunen, Lauri & Vaula, Martti 1938: *Suomen kielioppi sekä tyyli- ja runo-opin alkeet oppikouluille ja seminaareille*. Helsinki: WSOY.

Korhonen, Mikko 1981: *Johdatus lapin kielen historiaan*. Helsinki.

Larsson, Lars-Gunnar 2001: *Baltic influence in the Finnic languages*, in: Circum-Baltic Languages, Volume I: Past and Present, Studies in Language Companion Series, Dahl, Ö. & Koptjevskaja-Tamm, M. (eds.), John Benjamins Publishing Company, Stockholm University.

Nichols, Johanna 1992: *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Paasonen, Heikki 1990-1999: *Mordwinisches Wörterbuch. Mordovskij slovar'*, Lexica Societatis Fenno-Ugricae XXIII:1-6, Heikkilä, Kaino (ed.), Suomalais-Ugrilainen Seura. Also found online at: http://www.sgr.fi/lexica/lexicaxxiii.html

Paasonen, Heikki 1897: Die türkische lehnwörter im mordwinischen, *Journal de la société Finno-Ougrienne*, *15*, 2.

Piispanen, P.S. 2012: Statistical Dating of Uralic Proto-Languages through Comparative Linguistics with added sound change law analyses, *Fenno-Ugrica Suecana Nova Series*, *14*, p. 61 – 74.

Piispanen, Peter 2013: The Uralic-Yukaghiric connection revisited: Sound correspondences of Geminate clusters, SUSA~JSFOu, *94*, p. 165-197.

Piispanen, Peter 2015: Evaluating the Uralic -Yukaghiric word-initial, proto-sibilant correspondence rules, SUSA~JSFOu, 95, p. 237-273.

Piispanen, Peter 2016: A semivowel sound change rule in Yukaghir, Journal of Historical Linguistics, Accepted.

Rédei, Károly 1999: Zu den uralisch-jukagirischen Sprachkontakten, *Finnisch-Ugrische Forschungen*, *55*, p.1-58.

Renfrew, C. & McMahon, A & Trask, L. 2000: Time-Depth in Historical Linguistics, Cambridge: The McDonald Institute for Archaeological Research.

Salminen, Tapani 2002: Problems in the taxonomy of the Uralic languages in the light of modern comparative studies. In: *Лингвистический беспредел: сборник статей к 70-летию А. И. Кузнецовой.* Москва: Издательство Московского университета, 2002. 44–55.

Sammallahti, Pekka 1977: Suomalaisten esihistorian kysymyksiä, *Virittäjä*, *81*, p. 119-136.

Sammallahti, Pekka 1988: Historical Phonology of the Uralic Languages, In: *The Uralic Languages. Description, history and foreign influences*, 1988, Sinor, D. (ed), Leiden, p. 478-554.

Sammallahti, Pekka 1990: The Sami Language: Past and Present, In: *Arctic Languages*, Collis, D. R. F. (ed), Paris: Unesco, p. 437-458.

Sammallahti, Pekka 1998: *The Saami languages – An Introduction*. Kárášjohka/Karasjok.

Sinor, Denis 1988: *The Uralic Languages. Description, history and foreign influences*, Leiden.

Starostin, Sergei 2000: Comparative-Historical Linguistics and Lexicostatistics, *Time Depth in Historical Linguistics, v1*, Cambridge: The McDonald Institute for Archaeological Research, p.223-265.

Swadesh, Morris 1950: Salish internal relationships, *International Journal of American Linguistics*, *16*, p.157-167.

Swadesh, Morris 1952: Lexicostatistic dating of prehistoric ethnic contacts, *Proceedings American Philosophical Society*, *96*, p.452-463.

Swadesh, Morris 1955: Towards greater accuracy in lexicostatistic dating, *International Journal of American Linguistics*, *21*, p.121-137.

Syrjänen, Kaj, Honkola, Terhi, Korhonen, Kalle, Lehtinen, Jyri, Vesakoski, Outi & Wahlberg, Niklas 2013: Shedding more light on language classification using basic vocabularies and phylogenetic methods, *Diachronica*, *30:3*, p. 323-352.

Taageperä, Rein 1994: The linguistic distances between Uralic languages, *Linguistica Uralica*, *30*, p. 161-167.

Tambets, Kristiina & Rootsi, Siiri & Kivisild, Toomas & Help, Hela & Serk, Piia & Loogväli, Eva-Liis & Tolk, Helle-Viivi & Reidla, Maera & Metspalu, Ene & Pliss, Liana & Balanovsky, Oleg & Pshenichnov, Andrey & Balanovska, Elena & Gubina, Marina & Zhadanov, Sergey & Osipova, Ludmila & Damda, Larisa & Voevoda, Mikhail & Kutuev, Ildus & Bermisheva, Marina & Khusnutdinova, Elza & Gusar, Vladislava & Grechanina, Elena & Parik, Jüri & Pennarun, Erwan & Richard, Christelle & Chaventre, Andre & Moisan, Jean-Paul & Barać, Lovorka & Peričić, Marijana & Rudan, Pavao & Terzić, Rifat & Mikerezi, Ilia & Krumina, Astrida & Baumanis, Viesturs & Koziel, Slawomir & Richards, Olga & De Stefano, Gian Franco & Anagnou, Nicholas & Pappa, Kalliopi I. & Michalodimitrakis, Emmanuel & Ferák, Vladimir & Füredi, Sandor & Komel, Radovan & Beckman, Lars & Villems, Richard 2004: The Western and Eastern Roots of the Saami – the story of Genetic “Outliers” Told by Mitochondrial DNA and Y Chromosomes, *Am.J.Hum.Genet.*, *74*, p. 661-682.

Toivonen, Yrjö Henrik 1953: Suomalis-ugrilaisesta alkukodista, Virittäjä, p. 5-35.

Turunen, Aimo 1988: The Balto-Finnic languages, In: *The Uralic Languages. Description, history and foreign influences*, 1988, Sinor, D. (ed), Leiden, p. 58-83.

Watkins, Calvert 2001:Areal Diffusion and Genetic Inheritance, Aikhenvald & Dixon (eds.), Oxford, p. 44-63.

Whitehouse, Paul, Usher, Timothy, Ruhlen, Merritt and William S-.Y. Wang 2004: Kusunda: An Indo-Pacific language in Nepal, PNAS, vol. 111, no.15, April 13.

Wiik, Kalevi 2002: *Suomalaisten juuret*. Atena Kustannus Oy.

World Ethymology website: www.ethnologue.com

Wurm, Stephen Adolphe 2001: *Atlas of the World’s Languages in Danger of Disappearing*, Unesco Publishing.

(Finnish and Saamic) etymological dictionaries:

Àlgu tietokanta: http://kaino.kotus.fi/algu/index.php?t=etusivu&kkieli=fi

Collinder, Björn 1955: *Fenno-Ugric Vocabulary. An Etymological Dictionary of the Uralic Languages*. Uppsala.

Häkkinen, Kaisa 2011: Nykysuomen etymologinen sanakirja, Sanoma Pro.

SSA: *Suomen sanojen alkuperä. Etymologinen sanakirja*. Helsinki: Kotimaisten kielten tutkimuskeskus / Suomalaisen kirjallisuuden seura. 1992-2000.

SKES: *Suomen kielen etymologinen sanakirja*. Lexica Societatis Fenno-Ugricae XII. Helsinki: Suomalais-ugrilainen seura. 1955-1981.

UEW: Rédei, Károly 1988-1991: *Uralisches Etymologisches Wörterbuch*. Budapest: Akadémiai Kiadó.

Abbreviations

PU=Proto-Uralic, PS=Proto-Samoyed, PFU=Proto-Finno-Ugric, PUg = Proto-Ugric, POUg = Proto-Ob-Ugric, PFP=Proto-Finno-Permic, PFV=Proto-Finno-Volgaic, PFM = Proto-Finno-Mordvinic, EPF = Earlier Proto-Finnic (i.e. Proto-Finno-Saamic), MPF = Middle Proto-Finnic and LPF = Later Proto-Finnic, Fin. = Finnish, Kar. = Karelian, Est. = Estonian, N. Saami = North Saami, K. Saami = Kildin Saami, EM = Erzya Mordvin, MM = Moksha Mordvin, KZ = Komi-Zyrian, Udm. = Udmurt, Hung. = Hungarian, Rus. = Russian, Eng. = English, Latv. = Latvian, Lith. = Lithuanian, Swe. = Swedish, PIE = Proto-Indo-European.