**Supplements to:**

# Drastic demographic events triggered the Uralic spread

Riho Grünthal,[1] Volker Heyd,[1] Sampsa Holopainen,[1, 2] Juha A. Janhunen,[1] Olesya Khanina,[1,3] Matti Miestamo,[1] Johanna Nichols,[1,4,5] Janne Saarikivi,[1] Kaius Sinnemäki[1]

[1]University of Helsinki
[2]University of Vienna
[3]Institute of Linguistics, Russian Academy of Sciences
[4]University of California, Berkeley
[5]HSE University

***Diachronica*** **2022** (open access)

Link to the article: https://doi.org/10.1075/dia.20038.gru
Links to Appendixes: Proto-Uralic cognates https://doi.org/10.5281/zenodo.6035956
Indo-Iranian loans https://doi.org/10.5281/zenodo.6035900
Ugric vocabulary https://doi.org/10.5281/zenodo.6035881
Link to all supplements: https://zenodo.org/deposit/6345559

**Contents of these supplements:**



R script for S4.2
Data spreadsheet for R script

How to cite this Supplements file:

Grünthal, Riho, Volker Heyd, Sampsa Holopainen, Juha A. Janhunen, Olesya Khanina, Matti Miestamo, Johanna Nichols, Janne Saarikivi, Kaius Sinnemäki. 2022. Supplements to 'Drastic demographic events triggered the Uralic spread' (*Diachronica,* c. 2022).

## S1. The Uralic family tree

The Uralic daughter languages and their branches are shown in Table S1.

**Table S1.** The Uralic daughter branches and languages, in approximately east to west order of branch homelands.[1] In each branch the first line is languages used for comparison here; the second line is the remaining daughters. In each line languages are listed from south to north. For a full listing with coordinates and some bibliography see Hammarström et al. 2019 > Families > Uralic. Ugric (and within it, Ob-Ugric = Khanty and Mansi) is at least an areal grouping and possibly genealogical (areal: Salminen 2001; Helimski 1982, 2003:161; J. Häkkinen 2009; genealogical: Honti 1998). Western Uralic is more clearly genealogical, but the internal structure is debated. Finno-Ugric is a term of convenience for the non-Samoyedic branches but not a firmly demonstrated branch (see discussion below).

| | | | |
|---|---|---|---|
| Samoyedic | †Kamass, Selkup, Tundra Nenets, Nganasan<br>†Mator, Forest Nenets, Forest Enets, Tundra Enets | | |
| Finno-Ugric: | | | |
| Hungarian | Standard Hungarian | | Ugric |
| Mansi | North Mansi<br>South Mansi, East Mansi | Ob-Ugric | |
| Khanty | North Khanty, East Khanty<br>South Khanty, West Khanty | | |
| Permic | Udmurt, Komi (Zyrian)<br>Komi (Permiak) | | |
| Mari | Meadow Mari<br>Hill Mari | | |
| Mordvin | Moksha, Erzya | | Western Uralic |
| Finnic | South Estonian, Livonian, Estonian, Votic, Veps, Karelian, Finnish | | |
| Saamic | South, North, Inari, Skolt<br>Ume, Pite, Akkala, Kildin, Ter | | |

The family is now increasingly being dated at about 4,500 years old based on the following evidence: (1) The Indo-Iranian contact episode (§2.2), firmly dated at about 4,000 BP, affected an already incipiently divergent set of early Uralic varieties (branch ancestors and probably others now extinct); (2) approximately 500 years is the time frame generally taken as

[1] An ordering based on possible shared sound changes and order of separation begins Samoyedic, Mansi, Hungarian... (see Supplement S3).

sufficient to produce distinct daughter languages from one ancestor. By this reckoning the internal divergence of Proto-Uralic began no later than 4,500 BP.[2] The traditional and still widely held view is that the family is older, perhaps 6,000 years or more, based on the low rates of PU lexical retention and I-I borrowing in the Samoyedic branch, which have been explained as due to longer time since separation. Favoring the date closer to 4,500 years are the many similarities between Samoyedic and other Uralic inflectional morphology and the un-Uralic phonotactics and stem canon among the large portion of the Samoyedic vocabulary that lacks Uralic cognates, suggesting that the dearth of cognates in Samoyedic is due not to gradual loss but to an intense contact episode that produced sweeping replacement of native vocabulary by substratal or borrowed vocabulary.

A counterargument to the younger age is that the comparison of phonotactics pits the young non-Uralic vocabulary (reconstructable only to Proto-Samoyedic, c. 2,000-2,500 years ago) against PU forms of native vocabulary. If the Samoyedic phonotactic canon evolved gradually to its present state, the foreign vocabulary could reflect non-drastic borrowing over a long period of time rather than a single canon-changing influx. Potentially decisive as to the time depth is the question to what extent it is whole inflectional paradigms (or coherent whole subparadigms) vs. individual endings that can be reconstructed to PU on the Samoyedic evidence: whole (sub)paradigms are powerful evidence while individual endings are no better than individual lexemes (and in fact often weaker, as affixes are usually monosyllabic and monoconsonantal while PU lexemes are usually disyllabic and contain two or three consonants). Another is whether the low cognacy rates of Samoyedic are statistically significantly lower than those of the other branches, but to determine that we need a larger cognate base than the maximally ~250 items found in Samoyedic (see main text).

The nine basic branches shown in Table S1 and Table 1 of the main text are uncontroversial as building blocks of the Uralic family tree, but many questions remain about the higher-level branching structure. There have been two main proposals:

First, and more traditional, is a left-branching (west-branching) tree, with the earliest branches in the east, beginning with a split of Samoyedic vs. Finno-Ugric, and Finno-Ugric then splitting into successively lower branches as the family spread west (Figure S1.1). All splits in this tree are binary.[3] This structure was assumed in early work, based on informally observed lexical affinities, but never demonstrated with shared lexical innovations or shared sound changes (Salminen 2002; K. Häkkinen 2001). For the dashed lines see below in this supplement.

[2] Kallio (2006) is an early proponent of this chronology.

[3] This is consistent with biological phylogenetics, where trees are assumed to be binary (Warnow 2018:29; Nichols and Warnow 2008).

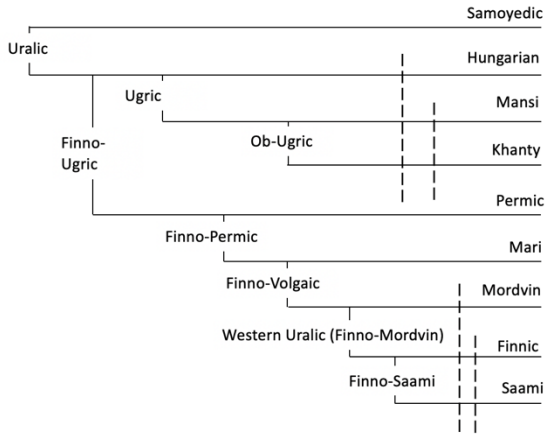



Figure S1.1. Traditional west-branching family tree (after Syrjänen et al. 2013:327; Korhonen 1981:27). East above, west below. For dashed lines see text below.

A variant of this tree has a first split between Samoyedic-Ugric and the rest (J. Häkkinen 2007, 2009). See Supplement S8 for other evidence of an eastern division.

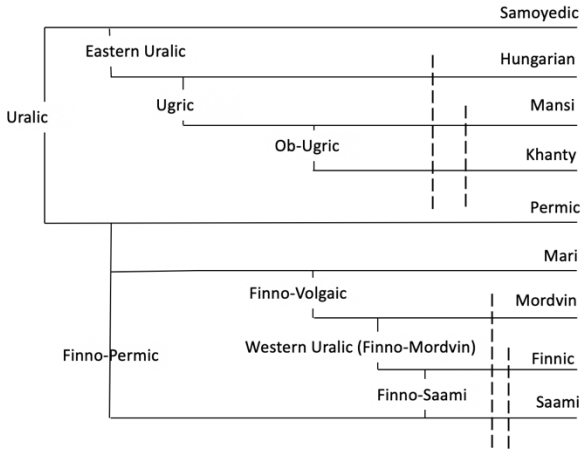



Figure S1.2. Same traditional tree with upper branching following J. Häkkinen (2007, 2009). Dashed lines as for Figure S1.1.

Second, and more recent, is a flat structure (i.e., rake, star phylogeny), recognizing only the nine basic branches and regarding that division as the initial one (Salminen 2001) (Figure S1.3). The argument is based on the near-total absence of shared phonological innovations between branches.

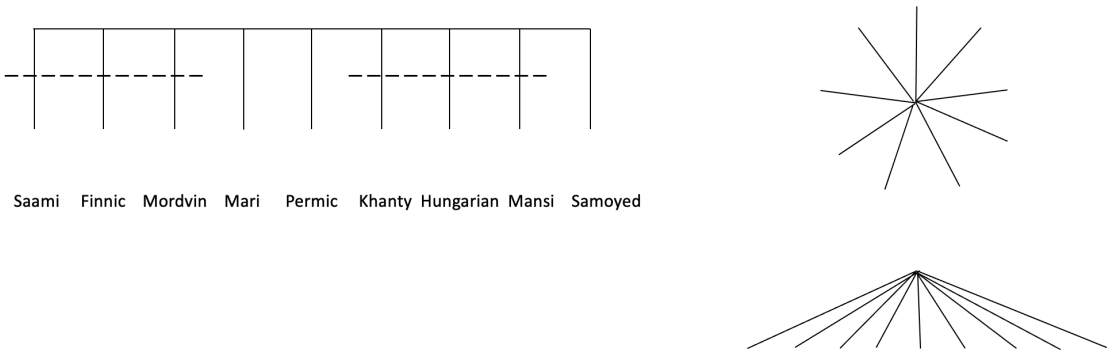



Figure S1.3. Rake tree (left) and star phylogenies (right), showing simultaneous separation of all branches. The three diagrams are phylogenetically equivalent. Dashed lines as for Figure S1.1; see text.

On either proposal, the status of Saamic-Finnic-Mordvin (Western Uralic) and Khanty-Hungarian-Mansi (Ugric) is not fully resolved in the literature. Western Uralic has shared innovations in lexicon, morphology, and phonology but there is discussion as to whether the Finnic-Saamic resemblances are inherited or borrowed. For Ugric, there are many sharings but it is debated whether they are inherited or contact-based. The orthogonal dashed lines in Figures S1.1-3 capture these uncertainties (this convention is patterned after Ross 1988, where a similar representation is used for dialect chains and other non-treelike groupings within the Oceanic language family; see below).

Recent work applying computational phylogeny reaches a variety of solutions, always identifying the nine basic branches and generally finding Finnic-Saamic and Samoyedic to be sharply distinct and maximally far from each other, e.g., Lehtinen et al. 2014, Syrjänen et al. 2013, and Maurits 2018, which use cognate lexical material; Dellert 2019 using very large numbers of resemblant but not necessarily cognate lexemes; Nichols et al. (in prep.) using typological characters, with or without cognates. Higher-level structure along fairly traditional lines can also be inferred from Dellert 2019 and perhaps Nichols et al. (in prep.), but the largest and most visible splits in both of those run between Finnic plus Saamic and the rest.

There are in fact grammatical innovations supporting nearly every split and node in the traditional binary tree (Janhunen 2009 and unpublished material). Only Mari is difficult to fit into a binary tree.

Samoyedic vs. Finno-Ugric: The unity of Samoyedic is uncontroversial, supported by many branch-internal innovations not reported here (see e.g., Janhunen 1998) and about 900 Proto-Samoyedic etyma (Saarikivi in press). The unity of Finno-Ugric is supported by a number of pan-Finno-Ugric innovations in word formation such as suffixation in Finno-Ugric *ńoma-la 'hare' ((Hungarian nyúl, North Saami njoammil) (Janhunen 2009:67). Samoyedic has only unsuffixed Proto-Samoyedic *ńama. Also, Finno-Ugric shares the numerals 'three', 'four', and 'six', absent from Samoyedic.

Ugric vs. Finno-Permic: The numeral 'two', PU *kekta, is one of only two numerals reconstructable for PU. In Finno-Permic, 'two' was reshaped to *kakta.

Within Ugric, in the word for 'three', Mansi and Hungarian (Ugric) retain PU *r while Finno-Permic languages innovate *l, doubtless sequential analogy with medial *l in 'four'. Thus Hungarian *három* vs. Finnish *kolme*. This suggests an original split of Mansi-Hungarian from Finno-Khanty, followed by a split of all of Ugric (plus Samoyedic) vs. Finno-Permic.

These are not just isolated individual words; they are part of the sequence of numerals 1-10, of which only two are attested in Samoyedic, more in other parts of the

family, and the full sequence in Finnic and Mordvin. 'Seven' is a loan in every Finno-Ugric branch (from Iranian in Ugric), and there is a Finno-Ugric cognate set for '20'. The numeral system is thus a microcosm of the evolution of the Uralic family tree (Janhunen 2000:60-61, 2009:67). The position of an individual item in the reconstructable sequence and the position of the branch in the lexical evolution of the family gives each numeral much more evidentiary value than a random word drawn from the lexicon.

Saamic-Finnic-Mordvin-Mari (i.e., Western Uralic-Mari) vs. Permic: Saamic et al. innovate an internal local series of case forms *-s-na inessive 'in', *-s-ta elative 'out of'; Permic lacks this series.[4]

Saamic-Finnic-Mordvin (Western Uralic) vs. Mari and/or others: Merger of PU *ï and *a. In these two sets Finnish, North Saami, and Erzya Mordvin have back vowels corresponding to Mari front vowels:

| PU | Finnish | North Saami | Erzya | Mari |
|---|---|---|---|---|
| *ńïli 'arrow' | nuoli | -- | *nali > nal | nöl(ə) |
| *ïpti 'hair' | hapsi | vuik'ta | -- | üp |

Compare, with back vowels in all branches:

| | | | | |
|---|---|---|---|---|
| *kali- 'die' | kuole- | -- | kuli- | kole- |

The details of the merger of originally two distinct stem types *ï-a and *a-a in Western Uralic are discussed in Aikio (2015).

Finnic-Mordvin vs. Saamic: Finnic and Mordvin have the full set of numerals 1-10 including '10' (Finnish *kymmenen*, Erzya *kemeń*) which is lacking in Saamic (where *logi* '10' is a more archaic word, shared with Mari *lu* and Mansi *low*, based on the verb 'count'). An alternative interpretation is that *logi,* etc. reflect an original Finno-Ugric numeral. One or the other set is due to semantic innovation: for Finnic and Mordvin cf. Finnish *kämmen* 'palm (of hand)', and for Permic, Mari, and Saamic cf. Finnish *luetella* 'count', *lukea* 'read'. In Permic (Udmurt *das*) and Hungarian *tíz* the word for '10' was replaced by an Iranian loanword.

Also, there are many more unique Finnic-Mordvin cognates than unique Saamic-Mordvin ones (Itkonen 1997). However, it has been suggested that Saamic and Mordvin share a set of common unique phonological innovations with a common relative chronology (Zhivlov 2014: 115-117).

Saamic-Finnic: 220 unique cognates with regular correspondences (Aikio 2012).

These are strong pieces of evidence, involving numerous cognates and/or items fitting into paradigms or ordered sequences. However, it is not certain that sheer preponderance of cognates, even strong cognates, suffices to establish a subgroup. Until the whole family and its component wordstocks can be surveyed for this purpose, and until word formation has been better studied across the family, we do not know whether the frequency of any of the pieces is enough to exceed chance. The number of cognates required to exceed chance depends on the extent of formal and semantic latitude and the size of the lexical database searched, and that increases with the number of languages and extent of documentation. Saamic and Finnic have more daughter languages than other branches including some of the best-researched ones (notably Hungarian), so their high number of unique cognates may not suffice to

[4] Unless the elative in the possessive declension preserves a lone trace of Permic *-s-ta (IN-ELA): Udmurt *busy* 'field' (elative suffix underscored): *busy-os-ys' (*field-PL-ELA) but *busy-os-ys't-ym'* (field-PL-ELA-1SG), etc. (Bartens 2000). For the history of the *-s- case series and its impact on taxonomy see Ylikoski (2016).

show that they form a single higher branch. Until these statistical issues are resolved, all of the above evidence must be regarded as suggestive rather than diagnostic.

One factor obscuring phylogenetic relations is post-dispersal intra-family contact effects. On the one hand, close contact between sister languages can result in inter-branch loanwords and grammatical patterns borrowed outright or calqued. Loanwords borrowed early enough in the languages' histories, before criterial branch-identifying sound changes have taken place, are generally indistinguishable from inherited vocabulary and inflate the number of cognates between those languages. The presence of a word in one contacting dialect or language can favor preservation of the cognate in the other, when the item might otherwise have been lost from one of the languages. This last point is *contact-induced retention* or *contact-induced inheritance* (Seržant in press). These processes are well attested cross-linguistically and plausible for Uralic, but in fact Uralic has few cases other than sporadic loans. Metsäranta (2020) gives a thorough survey of the Proto-Permic vocabulary and finds almost no examples of inter-branch loans into or from Proto-Permic, where there was good reason to expect them to exist.

On the other hand, bilingualism in sister languages can also result in lexical divergence happening at higher rates than expected, conditioned by speakers' conscious efforts to strengthen their linguistic identities through lexical differentiation (see Ellison and Miceli 2017 for an overview of the reported cases and a possible psycholinguistic explanation). This is also plausible for Uralic, but so far there is no particular evidence for such lexical differentiation being a major problem for reconstructing inter-branch relationships.[5]

There is no single standard way to reflect intra-family contact episodes in family trees. Above we have adapted the convention proposed in Ross 1988 for representing dialect/language chains and dialect/language networks using an orthogonal dashed line for the chain or network. Chains and networks are both sets of languages in which each shares grammatical traits and/or lexical items with adjacent languages but none are shared across the entire set. A chain is a linear, i.e., one-dimensional, set; a network is two-dimensional. Ross deals with chains and networks descending from a single ancestor, while Uralic has groupings that may be secondary associations of initially separate branches, which we call clusters. Ugric and Western Uralic are regarded as two such clusters by some. Clusters themselves can overlap differently in time as contact relations change due to changes in language range such as migration, as appears to have happened in both Ugric and Western Uralic.

[5] Within Samoyedic, Forest Enets and Tundra Enets might be an example of lexical differentiation with their identical morphosyntax and several dozens of non-cognate lexemes belonging to the most frequently used part of the lexicon (see Khanina et al. 2018 for a history of contacts between Forest Enets and Tundra Enets).

## S2. Linguistic paleontology

This supplement addresses linguistic paleontology based on species terms (§2.1) and metal names. Outside of technical usage, simplex plant and animal terms usually refer to genera, with modifiers added to create binomial terms for species (as with *black spruce, blue spruce, red spruce,* etc.). Therefore, it is genus terms and not species terms that are reconstructed to protolanguages, unless a particular species is distinctive or economically important. Since genera have wider ranges than species, reconstructed terms usually cannot identify a homeland with any precision.

Reconstructed PU plant names from Appendix 1 are shown in Table S2.1.

**Table S2.1.** PU etyma for plant names. Those in bold have diagnostic value for the PU homeland.

| PU term | Gloss | Name | Range | Branches | Comment |
|---|---|---|---|---|---|
| *jVxi | tree; pine | *Pinus* spp.? | | Samoyedic, Ugric | Any homeland |
| *pVwi | tree | | | All but Khanty, Mordvin, Saamic | Any homeland |
| *ńulka | fir | *Abies* spp. | Boreal forest | Selkup, Ob-Ugric, Permic, Mari | Any homeland |
| *kVwsi | spruce | *Picea* spp. | Boreal forest | All but Hungarian | Any homeland[6] |
| ***si̯ksa** | Siberian pine (Siberian stone pine, cedar pine) | *Pinus sibirica (Pinus cembra sibirica)* | Western Siberia | Samoyedic, Ugric, Permic | Food plant. **East of Urals** |
| *pVjV | willow | *Salix* spp. | Widespread | All but Mansi, Hungarian, Mari | Any homeland |
| *kVji | birch | *Betula* spp. | Widespread | All but Ugric, Permic | Any homeland |
| ***d'i̯mi** | bird cherry (chokecherry, hackberry) | *Prunus padus* | North of steppe; to 60° in Siberia, to tundra in Scandinavia | All but N Samoyedic | Any homeland; **food plant east of Urals.** |
| *mura | cloudberry | *Rubus chamaemorus* | North of ~55°; in Urals to ~52°. | Samoyedic, Ugric, Finnic | Any homeland |

Most of these terms denote genera and are of little value for identifying a homeland. The Siberian pine, a term at the species level, is a good diagnostic: it is found only east of the Urals, and its referent is a valued food plant (it yields nuts very similar to those of the Italian stone pine). The nuts are a market commodity across Russia and China today, so conceivably they might have been traded outside of their range in ancient times, but we have seen no evidence of this. In any case the term is found only in the three branches found in the tree's range: Samoyedic, Ugric, and Permic (the range extends to the western Ural slope where it barely overlaps Permic). This branch distribution is diagnostic of PU status on the traditional definition, and with attestation in three separate branches it qualifies on other definitions.

The bird cherry (hackberry, chokecherry) is a food plant in Siberia; in Europe, where the berries are small and astringent, it is not a food plant but has other uses (e.g., making liqueur and dye).[7] Its wide range gives it little diagnostic value, but its use as a food plant in the east must account for some of its lexical stability and can be considered weak supporting evidence.

[6] Depending on the exact chronology for Uralic, the form for 'spruce' may point to an eastern homeland. It originated east of the Urals but spread westward, reaching the Baltic area by the mid-Holocene, some two millennia before the Uralic languages (Friedrich 1970:36-37, map p. 40).

[7] http://www.luontoportti.com/suomi/en/puut/bird-cherry

Cloudberries are a phenotypically distinctive northern plant and a valued food plant. The southern edge of their range touches the northern edges of the branch homelands, accounting for the survival of the term at the far eastern and western Uralic range but giving it little diagnostic value.

PU animal terms are all glossed at a very general level ('squirrel', 'bird').

A word for 'honeybee' (*Apis mellifera*),[8] *mekši, is an I-I loan found in all the branches in the natural range of honeybees: Hungarian, Permic, Mordvin, Finnic. (Another I-I loan, *meti 'honey', has a nearly identical distribution. See Appendix 2; both words are in the sample there.) It has been used to argue for a Volga homeland, when the I-I loans were described as borrowed into PU (as they have been in previous work); but as the loans came into CU they tell us nothing about the PU homeland. They are, however, revealing as evidence for the CU distribution: the branch ancestors of the languages now in the honeybee range were in that range 4,000 years ago as well, as the survival of the terms proves. Importantly, the term argues for an early location of Hungarian near the southern Ural range: the natural range of the honeybee reaches the southernmost Urals and then trends SSE to the western Altai. If the I-I loan tells us nothing about the PU homeland, the absence of a native Uralic term in the honeybee range may be diagnostic. Beekeeping developed early in the middle Volga area (Carpelan and Parpola 2001: 115-120)[9] and in any case a term for a prized food and source of alcohol and wax would surely have existed in PU if the homeland had been in or near the honeybee range, and it could have survived in three western branches (Saamic-Finnic-Mordvin, Mari, Permic) to count as PU on post-traditional definitions.

In summary, all evidence from plant names point to a homeland in or near the northern forest, and among them, plant names diagnostic of a more specific location point to an origin east of the Urals. Where the term for 'honeybee' has been used to argue for a western homeland, the argument mistakenly identifies CU as PU, which obscures the importance of the term for identifying the early Hungarian homeland. All previous literature argues for a homeland in or near the boreal forest (which, NB, extends to the southernmost Urals), and this is the unanimous consensus. Important literature on plant names includes Hajdú 1969, 1975ab. K. Häkkinen (1996: 108-111) discusses reliability issues. Kallio 2015:84-85, supporting a western Oka-Volga homeland, argues that absence of evidence for PU hardwood tree names does not constitute evidence of absence.

Terms for metals and metal artifacts can be critical in determining language family ages and/or ranges relative to technological states such as the Copper Age (Eneolithic) or Bronze Age. In the case of Uralic, a single term for a metal, *wäśkä, can be posited for PU (Table S2.2), and that term has some phonological irregularities suggesting a post-PU areal Wanderwort (Aikio in press:§1.6.2, 2015:42-43).[10]

The absence of a developed terminology for metals and metal artifacts has been used to argue against the claim that the PU spread can be equated to the ST phenomenon (e.g., Kallio 2006; Parpola 2012; but now see Kallio 2015): traders in bronze should have had a larger and more specific terminology for the metal and probably some artifacts. Two important distinctions need to be made. One is that, while only one *root* is reconstructed as a PU metal term, it could well have produced binomial terms for different metals when accompanied by a lexicalized modifier, as in Khanty, with (in the modern terminology, reflecting Iron Age developments) *wăX* 'iron', 'red'+*wăX* 'copper', 'blade'+*wăX* 'steel', 'white'+*wăX* 'silver', etc. (Viitso 2012:188, 190). Table S2.2 illustrates the semantic range and gives examples of such compounds.

The other distinction concerns the earliest stages of Uralic. The lack of extended metal terminology in PU has long been an obstacle to seeing PU as connected with the Seima-Turbino phenomenon. The sole term *wäśkä is reconstructible with the ambiguous meaning 'copper, metal' as

---

[8] Another Eurasian honey-making species, *Apis cerana,* has a south and southeast Asian range well south of Siberia.

[9] Carpelan and Parpola do not specify the time frame of the earliest Volga beekeeping (nor does their source Hajdú 1975: 33). They emphasize that beekeeping on an industrial scale was needed for the industrial-scale bronze casting of the Bronze Age, for which lost-wax casting is essential. But for the PU homeland we need pre-Bronze Age support.

[10] This word is not in Appendix 1 because the irregularities hinder straightforward reconstruction of a protoform. However, it can confidently be attributed to PU, albeit with some ambiguity as to the exact reconstruction.

might be expected for a culture that made little or no use of metals and no use that required a specific metal. However, it was not PU but CU that was spoken c. 4,200 BP after ST had begun to link all of western Siberia and at least the northern periphery of the Caspian steppe in a single trade system. CU would have been different from PU due not only to a few centuries of evolution but also to the lexical consequences of this interregionalization. By 4,200 BP or later, CU should have acquired terminology for the key materials and artifacts of the metal trade. See Map 1: ST sites cluster densely enough along the major waterways and the Altai-Sayan foothills to indicate that probably most languages of the western Siberian forest zone had some words for the important ST materials and artifacts and aspects of the trade traffic. As is noted in Chernykh 2008 (one of the sources of our Map 1), the western Siberian S-T sites are somewhat older than the Volga ones.

In what follows we review the Uralic forms and reconstructions for the word and solutions offered in the three branches with irregular correspondences. A term of approximately a PU-like shape was a Wanderwort in early inner Eurasia, complicating the analysis of the Uralic word. By now, despite the inherent complexity of tracing Wanderwörter, the facts can mostly be resolved into a PU form, its largely regular developments, and non-Uralic protoforms with regular reflexes in the adjacent language families. We then summarize the arguments for the CU range and the PU homeland.

**Table S2.2.** PU *wäśkä 'copper, metal' and element in compound metal names. (Viitso 2012:188; Aikio in press; Janhunen 1977, 1981; UEW 560; Appendix 1.) One representative per branch; reflexes are found throughout the branch unless otherwise indicated.

| | Regular | Irregular | Uncertain |
|---|---|---|---|
| Samoyedic | Nganasan basa 'iron'<br>TNenets jeśa | | |
| Khanty | | | wăx 'iron'; ...+wăx in 'copper', 'steel', 'silver' |
| Mansi | | | āt+wəs 'lead' |
| Hungarian | | vas 'iron' | |
| Permic | Udm veś, azveś 'silver' | | |
| Mari | | | |
| Mordvin | E uśke, M uśkä 'wire, chain' | | |
| Finnic | | Finn vaski 'copper, bronze' | |
| Saamic | NSaa veaiki 'copper' | | |

**Table S2.3.** Basic cognates descended from PU *wäśkä 'metal, copper'. Sources: MszFE 1: 169–170, 3: 675–677, SSA 3: 416, SW 175.

| | |
|---|---|
| Saamic | *wɛške ‘copper’ |
| Finnic | *waski : *waske- ‘copper’ |
| Mordvin | *uśke* (~ *viśkɛ*) ‘(metal) chain’ |
| Mari | (-)*waž* ~ (-)*wož* ‘(metal) ore’ |
| Permic | *-veś* ~ *-iś* ‘metal’ |
| Khanty | *wăx* ‘metal, iron’ |
| Mansi | *-weś* ~ *-küš* ‘metal’ |
| Hungarian | *vas* ‘iron’, *ez-üst* ‘silver’ |
| Samoyedic | *wesä ‘iron, metal’ |

***Comments on the forms***

Saamic: The Saamic forms presuppose Proto-Saami *wɛške (> North Saami *veaiki*), which would regularly derive from Pre-Proto-Saamic *weśkä. The expected representation of *wäśkä, containing the vowel combination *ä-ä, would be Proto-Saami *wȧške, cf., e.g., *äjmä 'needle' > Proto-Saami *ȧjme (> North Saami *áibmi*). However, there are other examples of the vowel combination **ä-ä* being reflected as Proto-Saami **ɛ-e*, as, e.g., *päjwä 'day' > *pɛiwe (> North Saami *beaivi*). Some of these apparently irregular examples are loanwords from Finnish, e.g., *nealgi* 'hunger' < *nɛlke ← Finnish *nälkä* (the regular form in Saami would be *ńȧlke, derived from *ńäli- 'to swallow'), but the internal consonantism in *wɛške confirms that this is an old inherited item. Conclusion: the Saamic data can derive from either *weśkä or *wäśkä.

Finnic: The Finnic forms presuppose Proto-Finnic *waski : *waske-. However, in a few items, the vowel combination *a-i can represent an earlier *ä-ä, as also in *talwi : *talwe- < *tälwä (> Saami *tȧlve > modern North Saami *dálvi*) (for more examples see Aikio 2015). The Proto-Uralic consonants **s* and **ś* have regularly merged to **s* in Proto-Finnic. Conclusion: two Pre-Proto-Finnic alternatives are possible for contemporary Finnic, either *waśki or *wäśkä.

Mordvinic: Mordvinic (Erzya) *uśke ~ viśkä* and (Moksha) *uśkä* 'wire, chain' (MdWb 2485) could presuppose an earlier shape of the type *wVśkä, if the Erzya western dialect variant *viśkä* is taken as original. In this case the initial sequence *vi-* would involve a secondary development from *u-, though the vowel combination **u-e* probably nevertheless indicates the presence of an original *w. This development would relate Mordvinic to the rest of the Uralic data by assuming the diachronic sequence *wäśkä ~ *weśkä (> ? *üśkä) > *viśkɛ ~ uśke.*

However, as Aikio 2015: 42 notes, E *viśkä* occurs only in one western dialect (Kazhlodka) where it reflects a local regular sound change (u- > vi- before a palatalized consonant). A further reason why E *uśke*, M *uśkä* cannot descend from PU *wäśkä is that, firstly, in inherited vocabulary a first-syllable *ä should be reflected as *ä* in Moksha, cf. E *pel'e*, M *päl'ä* 'side, half' < *pälä. Secondly, as a rule, word-initial *w- is preserved in Mordvinic, which makes the etymology phonologically problematic. A PU origin with a first-syllable back vowel *waśki would be a possible reconstruction for Mordvinic, if there is an explanation for the loss of word-initial *w-. The word presents a further morphological problem: most Mordvinic words ending in *-ke/-kä* are derivatives. Conclusion: Mordvinic only has a phonologically and structurally ambiguous cognate, with irregular correspondences, and therefore unlikely to descend from *wäśkä.

Mari: The Mari item (Hill) *-waž* ~ (Meadow) *-wož* is attested as the second component of compounds, e.g., *kərtńi-waž* 'iron ore'. As the internal cluster **śk* yields regularly Mari (*)šk, as in *mośki- 'wash' > Mari *mušk-a-*, a development *wäśkä > *-wVšk- > *-waž would, in principle, be possible in view of the suffixal position of the element. Even so, the velar vowel *a ~ o* remains unexplained. Alternatively, the Mari item could be a borrowing from Pre-Proto-Hungarian (*)*waš.* A third possibility, though not very likely, would be that Mari *-waž* has no relation to the Uralic words for 'metal, copper', in which case it is probably identical with the Mari regular noun *waž* 'root' (as has also been proposed). Conclusion: the Mari forms have an irregular and possibly secondary connection with *wäśkä.

Permic: The Permic reflexes are also attested only in compounds: Udmurt *az-veś* = Komi *ez-iś* 'silver', Udmurt *uz-veś* = Komi *oz-iś* 'tin'. (Compounds in origin, though not synchronically: neither piece exists as an independent word or has a recoverable meaning in the modern languages.) In view of the reduced shape of second elements of compounds, it is difficult to reconstruct its exact earlier form, but the sequence *-wVś is clearly identifiable and can very well represent a reduced trace of *wVśkä. Conclusion: the Permic forms do not contradict the reconstruction *wäśkä, though the vowel qualities remain ambiguous.

Khanty: Khanty *wăx* (with dialectal variants) < ? *wăg has a multiply irregular relationship with the other Uralic words for 'metal'. Neither the vowel *ă nor final consonant *-x* < ? *-g is compatible with

the reconstruction *wäśkä, nor with a velarized variant of the latter. As a loanword, the Khanty item is also present in Northern Mansi in the shape *wox*. Conclusion: The Khanty data probably represents an etymon separate from the item for 'metal' in the other Uralic languages. In view of its form it is probably a loanword, but apparently not from any Uralic language.

Mansi: Mansi *āt-wes* ~ *ɛt-küš* (with dialectal variants) 'tin, lead' is most probably a loanword from the corresponding Permic items, cf. also the words for 'silver' in Permic and Hungarian. The origin of the consonant *k* in the variant *-küš* remains unexplained. Conclusion: Mansi has probably lost any direct inherited reflex of the Uralic word for 'metal', but has regained a trace of it from Permic.

Hungarian: Hungarian *vas* [våš] (~ Old Hungarian *vos*) presupposes earlier *waskV < *waśkV, with the same development of the internal cluster as in *mos* [moš] 'to wash' < *moskV- < *mośki-. The back vocalism in the Hungarian item remains unexplained, and, in spite of the similarity of the vowel of the initial syllable there can be no direct connection with Finnic *waski. There is, however, a possibility that Hungarian *waš was the source of Mari *-waž. Another trace of Uralic *wäśkä is present in Hungarian as the obscured latter component of *ez-üst* 'silver', which (according to EWU 346; MSzFE), is a borrowing from Permic (further transmitted also to Ossetic). The back-vowel word *vas* 'iron' is a base of *ezüst*, etymologically a compound, with vowel harmony fronting the vowel. (A Permic origin is plausible given the medieval location of Hungarian just west of the Urals near the middle Volga.) Conclusion: Hungarian *waśkV is clearly connected with *wäśkä but the back vocalism remains unexplained and may imply interference of a third language.

Samoyedic: The Nenets forms (Tundra) *yesya* ~ (Forest) *wyesya* as well as Enets *bese* point unambiguously to the Proto-Samoyedic shape *wesä, which is a regular reflex of Proto-Uralic *wäśkä. Selkup *këzə* (with *w > *k*), Kamas *baza*, and Mator †*baze* ~ †*bese* also do not contradict this reconstruction. Nganasan *basa* has the exceptional second-syllable vowel *a*, which, however, must be secondary, since this vowel is normally attested in inflected forms of stems originally ending in *a/*ä > Nganasan *u*/*ï*, including the genitive plural of nouns, e.g., *kopa 'skin' > *kuxu* : NOM PL *kubu-q* : GEN PL *kuba-q*, and the aorist stem of verbs, e.g., **kata-* 'to kill' : *kotu-* : AOR *koda-q-*. The exact age of *a* in these cases is unknown, but it is most probably a Nganasan innovation, and it clearly signals the former presence of a suffixal element, probably *j. It may also involve morphological restructuring, and in some cases it can signal borrowing. The initial-syllable vowel *a* of Nganasan *basa* is congruous with several other etyma, e.g., *wekana 'sturgeon' > Nganasan *bakunu.* Conclusion: all Samoyedic languages are compatible with the Proto-Samoyedic reconstruction *wesä. (In a variant reconstruction, which is based on the assumption that Uralic *ä and *e are preserved as distinct in Nganasan, the reconstruction would be written *wäsä, but this has no immediate bearing to the present discussion.)

Proto-Uralic: Proto-Uralic *wäśkä is well enough reflected in three or four branches to support a PU reconstruction: it is the only option for Samoyedic, the most likely option for Saamic, and a possible option for Permic and Finnic. It is unlikely or impossible for Hungarian, Mari and Mordvinic, where the modern forms do not regularly reflect the PU form, though they appear to be related somehow to the same etymon. Aberrant back-vowel forms in Hungarian, Mari and Mordvinic may be due to inter-branch borrowing, phonological reduction in compounds, or other factors. Khanty is the only branch that does not seem to have any trace of the Uralic word.

To summarize, while PU had only a single root for metals, CU had others, acquired from I-I: words for 'tin' and probably 'silver'; 'gold' has a complex history of separate borrowings into the individual Uralic branches from early post-Proto-Indo-Iranian (Holopainen 2019:232). Some of the CU items have been replaced in individual branches by later loans, e.g., 'gold' from Germanic in Finnic and Saamic (probably replacing original I-I loans still found in Mordvin). (For these and others see Viitso 2012.) By 4,200 BP or later, CU communities had ample opportunities to acquire terminology for the key materials and artifacts of the metal trade. ST sites cluster densely along the major western Siberian waterways and

the Altai-Sayan foothills, making it probable that most languages of the western Siberian forest zone had some words for the important ST materials and artifacts and aspects of the trade traffic (see Map 1). Metalworking was important in the Akozino-Mälar and Anan'ino economies (Parpola 2012) and the existence of a metal terminology must be posited whether or not the words survive.

What is strongly diagnostic for the PU homeland is the absence of PIE terminology for metals and metal artifacts in PU: bronze production and casting near the Urals goes back to the Fatyanovo-Balanovo culture, and had PU been spoken anywhere near there it would have acquired IE, perhaps Baltic-looking, terms for metals.

## S3. Pre-Uralic IE languages and the Uralic homeland

This supplement deals with interactions and possible CU contacts in the three proposed PU homeland arenas: the middle Volga; the eastern Urals or western Siberia more generally; and the upper Yenisei.

S3.1. *Volga:* The Corded Ware cultures developed, after the Yamnaya spread reached Europe, from blended Late PIE language and IE genes with central European farming cultures, and beginning about 4800 BP the easternmost Corded Ware flank began to spread east from the upper Dnieper along the Oka and Volga and eventually east to the Urals. This was the Fatyanovo culture (also Balanovo in its eastern range, near the Urals). Its language was Pre-Balto-Slavic in its western range, and in its eastern range it could have been divergent eastern Balto-Slavic, a sister to Balto-Slavic (i.e., Para-Baltic), or the budding ancestor to a separate IE branch which has not survived. For convenience we refer to the eastern Fatyanovo language(s) as Para-Baltic.

Pre-Baltic speech survived only in the western forest-steppe and the forest north of there, where the 4.2 ka event may have brought increased rainfall. In this region, from the Dnieper to the Oka, etymologically Baltic river names are numerous (Toporov and Trubačev 1962). Also in this area there is evidence of Baltic substratal effects on Russian dialects (Andersen 1996, 2003); also here, much later, the remnant Baltic-speaking Galindians (Old Russian *goljad'*) survived to early historical times (map: Koryakov 2007). These things mean that some or all of the Fatyanovo area west of the Oka was demonstrably Baltic-speaking in prehistoric to early historical times.

East of approximately the Oka-Volga confluence, where the 4.2 ka event brought drought, Para-Baltic speech became entirely extinct without leaving a trace. Across the former Fatyanovo and Abashevo ranges, new cultures appeared: in the west, along the upper to middle Volga and Oka, the Textile Ware (also known as Textile Ceramic, Netted Ware) culture; to its east, along the middle Volga, the Akozino-Mälar culture, and then, along the Volga bend and the Kama, the Anan'ino culture (map: Lang 2018: 204; dates: Lavento and Patrushev 2015). The latter two were bronze-working societies, and the Anan'ino culture had access to Ural mines and produced bronze. Akozina-Mälar and Anan'ino bronze items, especially spear and axe heads, are found in Textile Ware sites and in as far west as Scandinavia. The early Textile Ware culture did not use bronze, but its later phases did.

These societies were almost certainly Finno-Ugric-speaking. The Permic branch appears to have originated in the Anan'ino culture, probably on the lower Kama (Bartens 2000: 10-11; Belyx 2009).[11] The spreads of Saamic and Finnic (§2.2) emanated from the western Textile Ware range; Mordvin is within it; and in what is now Russian-speaking northwest Russia, a number of Finnic-like, transitional Finnic-Mordvin, Saamic-like, and Mari-like or Permic-like languages were spoken until the early middle ages (Rahkonen 2013; Saarikivi 2006).[12] We propose that the appearance of these Volga cultures marks the initial spread of the Uralic branch ancestors.[13] There was no Uralic contact with Pre-Baltic during this

[11] The archaeological Anan'ino culture was geographically larger than the plausible range of one language. Other Anan'ino languages may have included ancestral Mari, but have otherwise not survived. They were probably a mix of now-extinct indigenous languages, now-extinct unattested branches of Uralic, and now-extinct Para-Baltic.

[12] The evidence for these affiliations comes from toponyms and from ethnonyms and personal names recorded in medieval Russian chronicles. An implication of these findings is that the initial Uralic spread created additional languages which would have become ancestors to additional branches had they survived, but their speakers have shifted to other Uralic languages, pruning the branch structure down to the elementary nine. Similar pruning probably occurred all along the early Uralic range.

[13] Honkola et al. (2013) relate Uralic divergence events to climate developments, but do not include the 4.2 ka event in their model, and they propose what is by now mostly considered an excessively early date for the Finno-Ugric/Samoyedic split and excessively late dates for the Finnic-Saamic and internal Saamic divergences. They do not stipulate an absolute chronology for the Indo-Iranian episode. They assume a middle Volga PU homeland and steady population growth around the middle Volga driving language spreads. Indo-European prehistory, and the question of why only Uralic and not IE expanded in the critical time frame, are not in the scope of their model.

period; Baltic loans occur, in good numbers, in Finnic as a result of direct contact with early Baltic (probably from the otherwise unattested North Baltic branch: Kallio in press) and probably as a substratum after the secondary westward spread of Finnic began; Saamic has some, probably acquired from Finnic; and Mordvin has a few (Grünthal 2012; Junttila 2012; Aikio 2012).[14] East of the Oka, the economically important metalworking Para-Baltic-speaking societies were succeeded, with no substratal effects, no loanwords, and no survival of toponyms, by economically similar Finno-Ugric-speaking societies. Either Para-Baltic moved away or went extinct before Finno-Ugric languages appeared in the area, or the sociolinguistic situation was such that Para-Baltic was replaced by Finno-Ugric in a clean shift. Eastern Fatyanovo and especially Balanovo are thought to have absorbed or mingled with indigenous Volosovo hunter-fishers, whose languages must therefore have taken in early Para-Baltic terms for domesticates, wheeled transport, wool, etc., but no trace remains of these. Thus, whether directly or indirectly, Finno-Ugric speakers along the Volga had no contact with Indo-European speech prior to the Indo-Iranian episode.

A variant of this scenario is suggested by Koivulehto (1999, 2000, 2001), who proposes a loanword stratum in Finno-Ugric languages borrowed from Northwest IE (NW IE), an intermediate IE branch ancestral to Italic-Celtic, Balto-Slavic, and Germanic (on current knowledge this would be the language of the early Corded Ware culture). The idea of NW IE is generally accepted by Uralicists (e.g., Kallio 2015), though a number of Koivulehto's etymologies have been rejected and of those that remain most are not unambiguously NW IE but could equally well be from Pre-Baltic or Pre-Germanic (Holopainen 2021). Few words remain as necessarily NW IE, probably not enough to exceed chance. But this vocabulary has been compiled and assessed in a framework that assumes NW IE had to be ancestral to Balto-Slavic and Germanic. If eastern Fatyanovo and Balanovo spoke Para-Baltic, then that is the likely source of the NW IE words and it rebalances the ambiguous ones. The NW IE words, then, are worth a close reassessment as possibly Para-Baltic.

S3.2. *Urals:* What happened east of the Urals is less known. Early Ugric and Samoyedic speakers must have been involved in the trade routes between the Urals and the Altai. The Ugric languages have some terms for horses and horse culture from an early unknown source (Appendix 3), which indicates that their branch homelands were in the south, close to the steppe. Horses were domesticated very early, by culturally and presumably linguistically non-IE people, on the northern Kazakh steppe (Outram et al. 2009). The earliest historical information on Hungarian places it on or near the steppe in today's Bashkiria (southern Urals), and (given the standing northward spread pattern and the general lack of southward spreads in Uralic)[15] the entire Ugric group may have originated near the southern Urals (probably on the eastern side: Fodor 2001: 150, 2016: 218). Matveev (1962: 292-297) notes that substrate toponyms (especially hydronyms) in the Ob area prove that Khanty and Mansi are latecomers there; Smirnov (2012, 2013) describes a stratum of ancient Ugric toponyms in the southern and central Urals. Taken together, these document a northward spread from an earlier southern location. Matveev and Korenchy (1972: 39) consider the Ob-Ugric northward spread to have been due to pressure from the Turkic-speaking steppe to the south. There were Ob-Ugric speakers just west of the Urals in the late middle ages, and both Korenchy and Hajdú and Domokos (1978: 143-147) consider the Kama area to have been the Proto-Ugric homeland, and Hajdú and Domokos consider the eastward spread to have been due to pressure from expanding Novgorod and Muscovy. If the Uralic homeland is placed west of the

[14] Although no Pre-Baltic loanwords can be shown in Common Uralic, it is not clear when and where the Pre-Baltic contact episode of Finnic began, as the earliest loanwords in Finnic and Saami point to Proto-Balto-Slavic phonology and these few loans clearly predate the North Baltic stage. There are possible examples of such early loanwords also in Mordvin, Mari, and Permic (see Koivulehto 2006; Kallio 2009; Junttila 2016 ). However, the issue requires further research, and it is clear that the majority of the Baltic loanwords in Finnic are from North Baltic.

[15] Modern Hungarian is an exception to the standing northward spread: the Pre-Hungarian speakers joined a Turkic confederacy, adopted the nomadic pastoralist lifestyle, and moved southward onto the steppe (and then westward, eventually to Eastern Europe).

Urals, it is natural to interpret the western Ob-Ugric speakers as a remnant population; if, as we argue, the homeland was east of the Urals, the western Ob-Ugric population is a secondary minor expansion. Rather clearly, the bulk of the Ob-Ugric expansion was to the north, and today's Khanty and Mansi populations could also be due to unpressured shift by northern foragers to the Ugric languages of the traders and food producers. Then today's Khanty and Mansi speakers represent the Ugric northern periphery and the result of language shift; the former core population, nearer to the steppe, has shifted to Siberian Tatar and/or Russian. The documented myth and religion among the Khanty and Mansi include elements of the steppe horse cult and the circumpolar Arctic bear cult, testifying to this dual cultural heritage.

S3.3. *Yenisei:* Around 5,000 BP, earlier than the beginnings of PU divergence, the Yamnaya-derived Afanasievo culture appeared in the Minusinsk Basin (upper Yenisei) and the nearby Altai foothills. Its exact source in the Yamnaya population and its route to the Yenisei are unknown. The culture and genes were Yamnaya-like (Rasmussen et al. 2015), but they were separated from the rest of the IE zone by the Kazakh steppe and western Siberian forest, neither of which had any IE population at the time. The draw is likely to have been the metal deposits of the Altai-Sayan, which would have been of interest to the copper-using Yamnaya culture but which they could not have known of directly. The language of the Afanasievo culture is widely assumed to have been an IE variety, most often specifically ancestral to Tocharian, a set of two related written languages plus a third (indirectly attested) from the Tarim basin of Xinjiang attested 800-1600 BP (=400-1200 CE), a distance of ~1,000 km and ~3,000 years from Afanasievo (Mallory 2015, the source of all statements about Tocharian here). There is no linguistic evidence for the connection, and some against it: Tocharian preserves native PIE terms for domesticated plants while there is no evidence for cultivation in the Afanasievo sites. What the two have in common is pre-I-I settlement east of the Kazakh steppe. Diagnostic cultural sharings are few, and plausible entry routes dubious. Evidence of under-attested early IE varieties at the periphery of the early historical IE range is not uncommon, and the most parsimonious analysis is probably to assume that Tocharian and Afanasievo are two such. PU *wäśkä 'copper, metal' and its Proto-Samoyedic form *wesä 'iron, metal' resemble Tocharian A *wäs*, B *yasa* 'gold', but a word of this shape is a more widespread ancient Wanderwort so the Tocharian-Samoyedic resemblance does not point to a specific direct connection.

Later, with the spread of Indo-Iranian across the western and Kazakh steppes, the I-I Andronovo culture appears in the Minusinsk area as well, supporting an I-I role in the ST phenomenon.

S3.4. *Implications for the Uralic homeland:* IE speakers first brought the epoch-making technological advance of wheeled transport, as well as similarly influential cultural and economic practices, to the rest of western Eurasia, and any language in contact (direct or mediated) with early IE should have had a stratum of IE loans referring to these phenomena. If the PU homeland was anywhere near the upper Yenisei we would expect to find borrowed IE terms for domesticated animals, wheels, and other important IE technology in PU as loans from the Afanasievo language, but there are none. The I-I loans in Samoyedic are probably fewer than would be expected if there was direct contact between the Andronovo culture (a Sintashta successor, found across the northern Kazakh steppe, and almost certainly Indo-Iranian-speaking) and Pre-Proto-Samoyedic. If the Uralic homeland was anywhere from the middle Volga to the western Urals there should be an early Baltic-like or Para-Baltic stratum of such loans in PU, but there are none.

To summarize from the IE side, we would have expected to find divergent or transitional Pre-Baltic and Para-Baltic languages along the Volga, descending from the Fatyanovo culture (which was probably Pre-Baltic-speaking) and its offshoot the Abashevo culture (also probably Pre-Baltic or Para-Baltic) along the northern periphery of the Pontic and Caspian steppes (map: Anthony 2007:379; both were ultimately Yamnaya descendents and proximately part of the Corded Ware agricultural cultures of central to northern Europe), but there are none. There might well also have been surviving IE languages, distinct from any other branch, indigenous to the western Ural area prior to the Russian colonial spread, and distinct IE branches entrenched along the Seima-Turbino routes.

From this negative loanword evidence in circumstances where loanwords should have been plentiful it must be concluded that the PU homeland was east of the Urals and out of contact with PIE or

early IE. In fact, in Uralic the expected stratum of IE loans for the salient IE technological and cultural advances is the I-I loan stratum of CU. Thus, the I-I contact episode is the earliest reconstructable IE contact for Uralic, and it affected CU and not PU. Ancestral Samoyedic, in CU times, was not in contact (or only barely in contact) with I-I.

## S4 More on isolation by distance (IBD)

S4.1. Testing for isolation by distance in the Uralic languages

As was mentioned in §3.2, isolation by distance is a general phenomenon of decreasing numbers of shared traits as geographic distance between related populations increases (see the introductory overviews in Holman et al. 2007; Haynie 2014). Earlier work by non-linguists used geographical distances between languages and language families in interpreting genetic patterning (e.g., Sokal 1988). Holman et al. (2007) is the first application to historical linguistic evolution, using simulation of typological features. Several studies have applied IBD to historical linguistic themes, for instance, to modeling the language history of Melanesia (e.g., Donohue et al. 2012; Lansing et al. 2007; Padilla-Iglesias et al. 2020), to the Bantu expansion (de Filippo et al. 2012), to expansion from Africa (Hunley et al. 2012), and to correlations of linguistic genealogical diversity with ecological factors (Cardillo et al. 2015; Honkola et al. 2018).

Isolation-by-distance (IBD) effects should produce more or less monotonic frequency clines, with highest frequencies occurring near the origin point and frequencies tapering off with increasing distance from that center. On this scenario, Proto-Uralic had some set of lexemes that were carried by the CU-speaking population as it moved away from the homeland. Movement most often took the form of groups of speakers who originated near the frontier of the CU-speaking territory moving away into nearby non-PU-speaking lands. They carried only that part of the lexicon that was known in their peripheral dialect, and on migration they probably proceeded to borrow words from non-Uralic-speaking neighbors, temporarily accelerating the rate of loss of CU vocabulary. Generations later, some of their descendants moved out again, with another step up in vocabulary loss. Meanwhile, their sister CU speakers closer to the center also borrowed words, but from their PU-speaking neighbors. Unless those words happened to contain sounds that were affected by dialect sound changes that are now branch-identifying correspondences, those individual intrafamily borrowings would now, millennia later, be indistinguishable from native terms. Etymological nativization (adaptation to the borrowing language's phonology and phonotactics of words borrowed from a sister language: Aikio 2007) and contact-induced retention (Supplement S1; Seržant, in press) can be presumed, further braking loss of native vocabulary. In speech communities with a center of diffusion, innovative words, formations, or sound changes can spread so widely from the center that the unchanged native forms no longer exist except in a few far peripheral communities. Finally, since a distant frontier community is smaller than the whole mass of central communities, any one individual can have more impact on the norm or perceived norm of a peripheral community than would be the case in the center, where one idiosyncracy rarely ousts a regular form. The net effect of faster loss toward the frontier and reinforced unity closer to the center is a more or less monotonic dropoff, i.e., IBD, measurable as a negative correlation between the frequency of native items and distance from the center.

In the case of the Uralic dispersal our contention is that the initial dispersal was so rapid that IBD effects did not occur (or hardly occurred), so that numbers of inherited words in today's languages chiefly reflect post-dispersal processes. Therefore, we need to test for monotonicity among the branches, more precisely in the reconstructed branch ancestors. IBD effects, if present, should be visible in both native PU vocabulary and I-I borrowings, since those borrowings occurred so early that they can be reconstructed to the individual branch ancestors. Since the homeland cannot have been west of the Urals (§2.1), we expect to find peak frequencies in the eastern part of the range. However, since the center of original expansion might have been closer to the center of the PU range, we need to test for monotonic decreases with distance from various centers.

We used the Mann-Kendall trend test to determine whether there is significant monotonicity in numbers of PU cognates or numbers of I-I loans with distance from the starting point of expansion. The Mann-Kendall trend test evaluates whether there is a monotonic trend in a time series. Our data does not represent time series (e.g., series of 12 months in successive years) but rather one set of distances in

kilometers from a potential point of origin; therefore, the trend test can be used for testing monotonic trend in our data as well.[16]

We used the following five key points for alternative distance calculations, all of them important river cities along the major water routes in the south of the attested range and therefore convenient proxies for possible PU expansion centers: Minusinsk (upper Yenisei, in the center of Proto-Samoyedic internal dispersal), Novosibirsk (upper Ob, near Seima-Turbino sites, in the historical Ugric range and the Pre-Proto-Samoyedic range posited by Janhunen 1998:457), Chelyabinsk (southeastern Ural foothills, close to major Sintashta sites and a candidate for the early Ugric center we posit in §3.2 and Supplement S3); Kazan' (at the Volga-Kama confluence and the Volga bend, and close to both Mari and Permic homelands), and Nizhnii Novgorod (at the Volga-Oka confluence), in a likely Mordvin or Finno-Saamic-Mordvin homeland and a likely center of Finnic migration to its (secondary) staging ground near the Gulf of Finland. Kazan' and Nizhnii Novgorod are west of the Urals and therefore precluded as PU homelands. Distances from the potential points of origin were calculated as overland great circle distances between published coordinates for city locations. The distances were computed using the online service provided by Veness (2002-2020). The distances are presented in Table S4.1 and the city coordinates (latitudes and longitudes in decimal) are presented in the caption of Table S4.2.[17] Recall that the high rate of PU traits in Samoyedic is artifactual; for this reason Samoyedic was not used when testing for IBD effects on the Uralic data but only for testing for IBD effects for loanwords from Indo-Iranian.[18]

The results of the Mann-Kendall trend tests are presented in Table S4.2. The null hypothesis H0 is that there is no trend and the alternative hypothesis H1 is that there is a negative trend, that is, the farther the language is from the potential point of origin the fewer PU traits or I-I loans the language has. None of the trends were significant and most were far from being significant. In addition, the sign of the trend was positive in all tests except for when testing Indo-Iranian loans using Kazan' or Nizhnij Novgorod as points of origin. These results strongly suggest that there is no evidence for IBD effects from the most likely points of origin of PU.

[16] For computing the Mann-Kendall trend test we used the package trend (Pohlert 2020) in the open-source programming environment R (R Core Team 2020).

[17] In Table S4.1 the overland distances to Saamic are smaller than to Finnic in the three leftmost points of PU origin. The calculations are based on geographic distances on the surface of the Earth without constraining them to plausible traveling routes by land. Based on linguistic and archaeological evidence the most plausible route from alternative points of PU origin to Saamic is via Southern or Central Finland rather than across the Arctic Ocean (White Sea). To double-check the IBD effects, we recalculated distances from alternative points of PU origin to Saamic by taking the distance from the alternative points of PU origin to Finnic and adding the distance between Finnic and Saamic (280 km) to those distances. These recalculations did not affect the results, and therefore we report below the results of the IBD tests based on the distances in Table S4.1.

[18] We also performed the IBD test for the Uralic etyma including the Samoyedic branch. The results were very similar to when excluding Samoyedic: tau > 0 and p > 0.6 regardless of where the distances were calculated from. In addition, we performed the IBD test for the Indo-Iranian loans excluding the Samoyedic branch. The results were very similar to when including Samoyedic: tau > 0 in all but when calculating distances from Nizhnij Novgorod and p > 0.4 regardless of where the distances were calculated from.

**Table S4.1.** Overland great circle distances (in km) of Uralic branches from five possible points of PU origin. The points of origin are provided as latitude and longitude coordinates of the following cities as explained above: Minusinsk (latitude: 53.7, longitude: 91.68), Novosibirsk (55.05, 82.95), Chelyabinsk (55.15, 61.38), Kazan' (55.80, 49.11), and Nizhnij Novgorod (56.33, 44.01).

| Branch | Minusinsk | Novosibirsk | Chelyabinsk | Kazan' | Nizhnij Novgorod |
|---|---|---|---|---|---|
| Samoyedic | 471 | 332 | 1594 | 2306 | 2585 |
| Khanty | 1574 | 328 | 383 | 1099 | 1435 |
| Mansi | 1792 | 322 | 160 | 902 | 1247 |
| Hungarian | 2141 | 506 | 362 | 800 | 1143 |
| Permic | 2438 | 1107 | 656 | 300 | 559 |
| Mari | 2793 | 749 | 893 | 167 | 200 |
| Mordvin | 3068 | 1536 | 1124 | 410 | 259 |
| Finnic | 3755 | 2439 | 1989 | 1292 | 955 |
| Saamic | 3618 | 2393 | 1945 | 1306 | 1004 |

**Table S4.2.** Results of the one-tailed Mann-Kendall trend tests on PU etyma (excluding Samoyedic) and I-I loans (including Samoyedic). S is the Mann-Kendall test statistic which tells how strong the trend is and whether it is monotonically increasing (positive values) or decreasing (negative values).

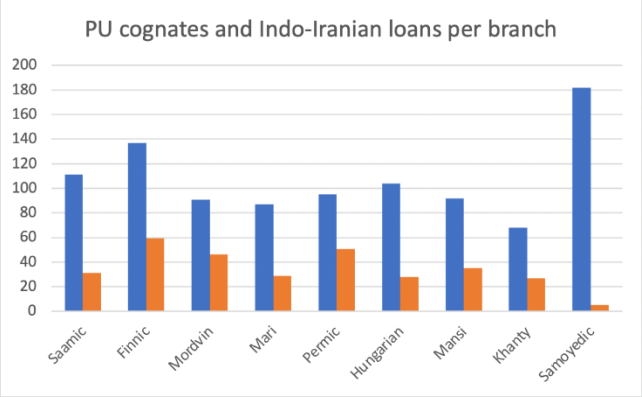
| | Uralic etyma | | | Indo-Iranian loans | | |
|---|---|---|---|---|---|---|
| Distance from | tau | p | S | tau | p | S |
| Minusinsk | 0.500 | 0.946 | 14 | 0.500 | 0.962 | 18 |
| Novosibirsk | 0.500 | 0.946 | 14 | 0.389 | 0.912 | 14 |
| Chelyabinsk | 0.357 | 0.867 | 10 | 0.167 | 0.699 | 6 |
| Kazan’ | 0.357 | 0.867 | 10 | -0.167 | 0.301 | -6 |
| Nizhnij Novgorod | 0.000 | 0.500 | 0 | -0.333 | 0.126 | -12 |

This outcome confirms the visual impression of Figure S4(a): the distribution of PU lexemes and I-I loans is not monotonic from any plausible PU center. (It would of course appear monotonic if the center of dispersal were placed in northern Estonia, the staging ground for Finnic; but this a known secondary location and highly unlikely for PU.)

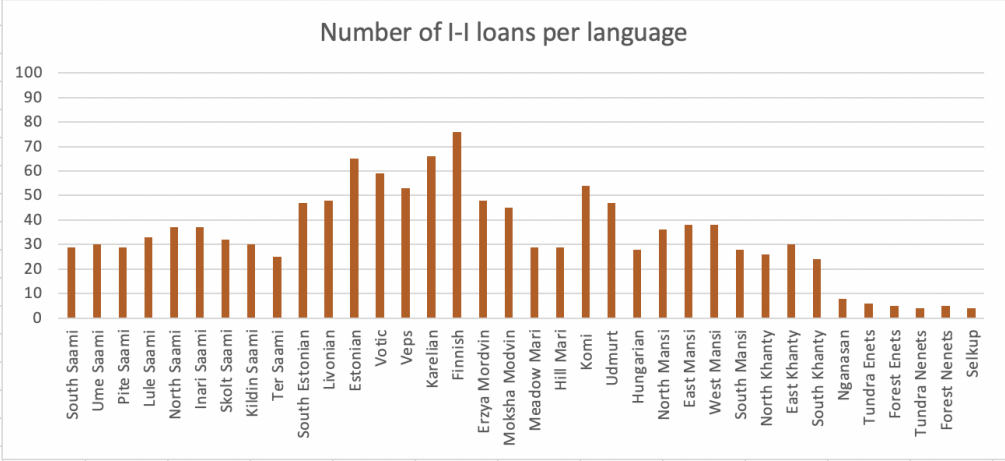
Much the same results were obtained by testing the individual languages, shown in Figures S4(b-c) and also Figure 1 in the main text.

**Figure S4.** Numbers of Proto-Uralic (blue) and Indo-Iranian (orange) etyma retained per branch or language. The figures for the Samoyedic languages (the rightmost three bars) are artifactually inflated by the definition of Proto-Uralic (§3.1). (a) Branches. (b) Languages including the Khanty and Mansi varieties (information available only on I-I loans), on a smaller scale. (c) Sharings: the total numbers of retained etyma that are shared with other languages.

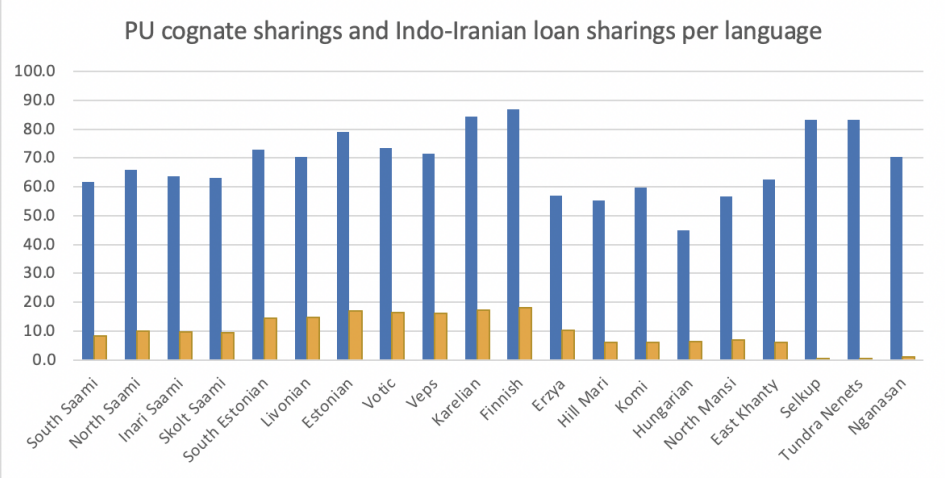
*(Figures on next page.)*




(S4a) Branches.




(S4b) I-I loans only. Detailed breakdown including for Khanty and Mansi. (This breakdown not available for PU cognates.)




(S4c) Numbers of sharings per language. (Not calculated for branches.)

Figures S4ab show counts similar to Figure 2 of the main text, with different breakdowns. The general profile is the same for all counts. The artifactually high figure for Samoyedic is exaggerated on the branch count (Fig. S4a) compared to the count of languages (Figure 2 in main text; see §1 there).

Figure S4c measures not numbers of retentions but numbers of shared retentions, i.e., numbers of PU cognates shared with the other languages. For each language, for each word in the PU list (Appendix 1) it determines whether that word is also found in each other language. For example, Finnish and Mansi both have reflexes of PU *elä- 'live', neither has PU *jasi 'chilly weather', Finnish has *ojwa 'head' but Mansi does not, and Mansi has *ipsi- 'sell' but Finnish does not. Of these the first, where the word is found in both languages, is a Finnish-Mansi sharing; the other three are not. This procedure is repeated across the entire database: for each of the 19 pairs of our 20 languages, it considers each PU word and asks whether it is present in both languages. For Finnish and Mansi there are 66 such pairs; this is in the mid-range. (Total sharings for each language are shown in Figures S8ab below, and their extent of integration is color-coded.)

This calculation measures the integration of each language with the others. For example, the Finnic languages (South Estonian through Finnish in the figure) are generally well integrated, and the Saamic languages are fairly well integrated; the others are less so, and Hungarian is quite low. The Samoyedic languages are in the high range, but by a much lesser extent than the high Samoyedic values for retentions (Figure S4a above and Figure 2 in the main text). This shows that measuring integration partly offsets the artifactual high PU count for Samoyedic, suggesting that the overall integration of Samoyedic is in fact low.

For all of these graphs there is a peak in Finnic (involving Finnish, Karelian, and Estonian), the opposite of what IBD would predict. Factors that correlate, probably causally, with the number of retentions or sharings are (* = statistically significant or nearly so):

- *East-west location, with more retentions and more integration in the west.
- *Number of daughter languages in the branch. Finnic, with seven daughters in the sample, is highest; Saamic has four in the sample and five others, and is second highest. Mansi has four but has low totals. Samoyedic has seven, but several went extinct before gaining adequate documentation. More daughters means more chances of attestation of an etymon.
- Amount of etymological work published; time and quantity of attestation. The correlation with attestation by itself is uneven: Finnic (with high frequencies) has probably received the most work, and Finnish and Estonian have been written since the 16th century; but Hungarian (fragments from c. 1055, literature from the 14th century) and Komi (writing from the 14th century) are low. The history of literacy and the quantity and quality of descriptive and academic work are undoubtedly relevant but we have not tried to compare them. All of these factors are bound up with geography: history of writing and documentation, accessibility to scholarship, and history of literacy developed earlier in Europe than in Siberia.
- Extralinguistic contingency: Samoyedic has experienced strong contact effects and isolation from the rest, though the specifics of its prehistory remain mysterious. Hungarian separated relatively early and has spent much of its existence in the different linguistic and cultural context of the steppe.

## S4.2. R script

The R script used here is as follows. Links to script as text file and data spreadsheet:

10.5281/zenodo.6345559 (same as for this Supplements file)

```
#####################################
######    Basic information    #######
#####################################

# R code for Supplement S4 of the article "Drastic demographic events
triggered the Proto-Uralic spread"
# Date of last update: 12 May 2021
```

```
##################################
######    Preliminaries    #######
##################################

### Load necessary packages to R session
# If they are not installed yet, install them first
# pkgs = c("trend")
# install.packages(pkgs, repos = "https://cloud.r-project.org")
library(trend)
library(readxl)


### Read data into the session
# IBD data for languages
ibd_lang <- read_excel("PU_I-I_2021-data.xlsx",
                       sheet = 2)

# IBD data for branches
ibd_br <- read_excel("PU_I-I_2021-data.xlsx",
                     sheet = 3)


##################################
######        Tests        #######
##################################

### Correlation of Uralic etyma and I-I loans
# Excluding Samoyedic
cor.test(ibd_lang$PU_etyma[1:22], ibd_lang$II_Loans[1:22],
         method="k")

# Including Samoyedic
cor.test(ibd_lang$PU_etyma, ibd_lang$II_Loans,
         method="k")



### IBD for retentions of Uralic traits in branches
## Including Samoyedic
# Create the matrix
ur.all = matrix(ncol=3, nrow=5)
colnames(ur.all) = c("tau", "p", "S")
rownames(ur.all) = colnames(ibd_br)[2:6]

# Perform one-sided Mann-Kendall calculations.
# H0: no monotonic trend; H1: negative monotonic trend
# Data is ranked for the test in terms of distances from
# a point of origin.
for(i in 1:nrow(ur.all)){
  mk = mk.test(ibd_br$PU[order(ibd_br[,i+1])],
               alternative = "less")
  # assign tau-values
  ur.all[i,1] <- round(mk$estimates[[3]],3)
  # assign p-values
  ur.all[i,2] <- round(mk$p.value,3)
  # assign S-values
```

```
  ur.all[i,3] <- mk$estimates[[1]]
}
ur.all


## Excluding Samoyedic
# Create the matrix
ur.NoSam = matrix(ncol=3, nrow=5)
colnames(ur.NoSam) = c("tau", "p", "S")
rownames(ur.NoSam) = colnames(ibd_br)[2:6]

# Perform one-sided Mann-Kendall calculations.
# H0: no monotonic trend; H1: negative monotonic trend
# Data is ranked for the test in terms of distances
# from a point of origin.
ibd_br.NoSam = ibd_br[ibd_br$Branches !="Samoyedic",]
for(i in 1:nrow(ur.NoSam)){
  mk = mk.test(ibd_br.NoSam$PU[order(ibd_br.NoSam[,i+1])],
               alternative = "less")
  # assign tau-values
  ur.NoSam[i,1] <- round(mk$estimates[[3]],3)
  # assign p-values
  ur.NoSam[i,2] <- round(mk$p.value,3)
  # assign S-values
  ur.NoSam[i,3] <- mk$estimates[[1]]
}
ur.NoSam

### IBD for Indo-Iranian loans
## Including Samoyedic
# Create the matrix
ii.all = matrix(ncol=3, nrow=5)
colnames(ii.all) = c("tau", "p", "S")
rownames(ii.all) = colnames(ibd_br)[2:6]

# Perform one-sided Mann-Kendall calculations.
# H0: no monotonic trend; H1: negative monotonic trend
# Data is ranked for the test in terms of distances
# from a point of origin.
for(i in 1:nrow(ii.all)){
  mk = mk.test(ibd_br$II_Loans[order(ibd_br[,i+1])],
               alternative = "less")
  # assign tau-values
  ii.all[i,1] <- round(mk$estimates[[3]],3)
  # assign p-values
  ii.all[i,2] <- round(mk$p.value,3)
  # assign S-values
  ii.all[i,3] <- mk$estimates[[1]]
}
ii.all

## Excluding Samoyedic
# Create the matrix
ii.NoSam = matrix(ncol=3, nrow=5)
colnames(ii.NoSam) = c("tau", "p", "S")
rownames(ii.NoSam) = colnames(ibd_br)[2:6]

# Perform one-sided Mann-Kendall calculations.
```

```
# H0: no monotonic trend; H1: negative monotonic trend
# Data is ranked for the test in terms of distances
# from a point of origin.
for(i in 1:nrow(ii.NoSam)){
  mk = mk.test(ibd_br.NoSam$II_Loans[order(ibd_br.NoSam[,i+1])],
               alternative = "less")
  # assign tau-values
  ii.NoSam[i,1] <- round(mk$estimates[[3]],3)
  # assign p-values
  ii.NoSam[i,2] <- round(mk$p.value,3)
  # assign S-values
  ii.NoSam[i,3] <- mk$estimates[[1]]
}
ii.NoSam



####################################
### Saamic distances recalculated ###
####################################

ibd_br.saa = ibd_br
# Distance from Finnihc to Saamic is 281.2km using
# https://www.movable-type.co.uk/scripts/latlong.html
# Update Saamic distances
ibd_br.saa[ibd_br.saa$Branches == "Saamic",2:6] <-
  ibd_br.saa[ibd_br.saa$Branches == "Finnic",2:6] + 281


### IBD for retentions of Uralic traits in branches
## Including Samoyedic
# Create the matrix
ur.all.saa = matrix(ncol=3, nrow=5)
colnames(ur.all.saa) = c("tau", "p", "S")
rownames(ur.all.saa) = colnames(ibd_br.saa)[2:6]

# Perform one-sided Mann-Kendall calculations.
# H0: no monotonic trend; H1: negative monotonic trend
# Data is ranked for the test in terms of distances from
# a point of origin.
for(i in 1:nrow(ur.all.saa)){
  mk = mk.test(ibd_br.saa$PU[order(ibd_br.saa[,i+1])],
               alternative = "less")
  # assign tau-values
  ur.all.saa[i,1] <- round(mk$estimates[[3]],3)
  # assign p-values
  ur.all.saa[i,2] <- round(mk$p.value,3)
  # assign S-values
  ur.all.saa[i,3] <- mk$estimates[[1]]
}
ur.all.saa


## Excluding Samoyedic
# Create the matrix
ur.NoSam.saa = matrix(ncol=3, nrow=5)
colnames(ur.NoSam.saa) = c("tau", "p", "S")
rownames(ur.NoSam.saa) = colnames(ibd_br.saa)[2:6]
```

```
# Perform one-sided Mann-Kendall calculations.
# H0: no monotonic trend; H1: negative monotonic trend
# Data is ranked for the test in terms of distances
# from a point of origin.
ibd_br.NoSam.saa = ibd_br.saa[ibd_br.saa$Branches !="Samoyedic",]
for(i in 1:nrow(ur.NoSam.saa)){
  mk = mk.test(ibd_br.NoSam.saa$PU[order(ibd_br.NoSam.saa[,i+1])],
               alternative = "less")
  # assign tau-values
  ur.NoSam.saa[i,1] <- round(mk$estimates[[3]],3)
  # assign p-values
  ur.NoSam.saa[i,2] <- round(mk$p.value,3)
  # assign S-values
  ur.NoSam.saa[i,3] <- mk$estimates[[1]]
}
ur.NoSam.saa

### IBD for Indo-Iranian loans
## Including Samoyedic
# Create the matrix
ii.all.saa = matrix(ncol=3, nrow=5)
colnames(ii.all.saa) = c("tau", "p", "S")
rownames(ii.all.saa) = colnames(ibd_br.saa)[2:6]

# Perform one-sided Mann-Kendall calculations.
# H0: no monotonic trend; H1: negative monotonic trend
# Data is ranked for the test in terms of distances
# from a point of origin.
for(i in 1:nrow(ii.all.saa)){
  mk = mk.test(ibd_br.saa$II_Loans[order(ibd_br.saa[,i+1])],
               alternative = "less")
  # assign tau-values
  ii.all.saa[i,1] <- round(mk$estimates[[3]],3)
  # assign p-values
  ii.all.saa[i,2] <- round(mk$p.value,3)
  # assign S-values
  ii.all.saa[i,3] <- mk$estimates[[1]]
}
ii.all.saa

## Excluding Samoyedic
# Create the matrix
ii.NoSam.saa = matrix(ncol=3, nrow=5)
colnames(ii.NoSam.saa) = c("tau", "p", "S")
rownames(ii.NoSam.saa) = colnames(ibd_br.saa)[2:6]

# Perform one-sided Mann-Kendall calculations.
# H0: no monotonic trend; H1: negative monotonic trend
# Data is ranked for the test in terms of distances
# from a point of origin.
for(i in 1:nrow(ii.NoSam.saa)){
  mk = mk.test(ibd_br.NoSam.saa$II_Loans[order(ibd_br.NoSam.saa[,i+1])],
               alternative = "less")
  # assign tau-values
  ii.NoSam.saa[i,1] <- round(mk$estimates[[3]],3)
  # assign p-values
  ii.NoSam.saa[i,2] <- round(mk$p.value,3)
```

```
  # assign S-values
  ii.NoSam.saa[i,3] <- mk$estimates[[1]]
}
ii.NoSam.saa
```

## S4.3. Language and branch metadata. See Table S4.3.

**Table S4.3.** Uralic languages used here, and their branches, Glottolog codes, and coordinates. (Coordinates follow Glottolog except that for the national languages we use the coordinates of the capital city and describe the standard variety. For West, East, and North Mansi coordinates are from http://www.endangeredlanguages.com/lang/ (*Ugric is at least an areal cluster and perhaps a branch; see Supplement S1. Generic: this is the variety used to represent the entire branch (e.g., when PU cognacy data is attributed only to branches).

| Language | Branch | Glottocode | Latitude | Longitude | Variety |
|---|---|---|---|---|---|
| Nganasan | Samoyedic | ngan1291 | 73.14 | 86.21 | |
| Tundra Nenets | Samoyedic | nene1249 | 66.18 | 71.02 | |
| Selkup | Samoyedic | selk1253 | 59.94 | 82.30 | |
| South Khanty | Ugric* | sout3226 | 59.50 | 69.09 | |
| East Khanty | Ugric* | east2774 | 61.25 | 73.35 | Generic |
| North Khanty | Ugric* | obdo1234 | 66.31 | 67.09 | |
| South Mansi | Ugric* | sout3253 | 58.44 | 65.57 | |
| West Mansi | Ugric* | cent2322 | 61.14 | 62.40 | |
| East Mansi | Ugric* | cent2322 | 62.00 | 67.60 | |
| North Mansi | Ugric* | nort3271 | 62.65 | 63.15 | Generic |
| Hungarian | Ugric* | hung1274 | 46.91 | 19.66 | standard |
| Komi | Permic | komi1267 | 64.05 | 54.95 | Zyrian |
| Hill Mari | Mari | west2392 | 56.22 | 46.57 | Hill (West) |
| Erzya | Mordvin | erzy1239 | 52.85 | 45.39 | |
| Finnish | Finnic | finn1318 | 62.00 | 25.00 | standard |
| Karelian | Finnic | kare1335 | 65.17 | 30.87 | |
| Veps | Finnic | veps1250 | 60.34 | 34.79 | |
| Votic | Finnic | voti1245 | 59.38 | 28.62 | |
| Estonian | Finnic | esto1258 | 58.43 | 24.73 | standard |
| Livonian | Finnic | livv1244 | 57.57 | 22.03 | |
| South Estonian | Finnic | sout2679 | 57.85 | 27.00 | |
| Skolt Saami | Saamic | skol1241 | 68.83 | 29.72 | |
| Inari Saami | Saamic | inar1241 | 69.42 | 27.87 | |
| North Saami | Saamic | nort2671 | 68.72 | 22.11 | |
| South Saami | Saamic | sout2674 | 62.88 | 13.70 | |

## S5. The Finnic and Saamic secondary westward spreads

For nearly a millennium, from the initial CU spread to not long before 3,000 BP, the western frontier of the Uralic family (archaeologically, the western frontier of the Textile Ware cultures) lay along the upper Volga and Oka. The spreads of the Saamic and Finnic languages then gradually brought that frontier to the Gulf of Finland and nearby. Saamic moved along what Lang (2018: 310, 2015) terms the Northwest Passage (upper Volga to Ladoga to southeastern Finland; the idea but not the term is from Parpola 2012), reaching southeastern Finland c. 3,000 BP, and some time after that spread to central Scandinavia perhaps as early as 1800 BP (200 CE), by then probably with livestock (Lang 2018; Piha 2018) and forming the basis for the development of South Saami. The basis for the development of later northern and eastern Saamic had spread across northern Fennoscandia by 500-700, as hunter-gatherers. These movements may have proceeded via two routes: overwater in the south but overland, north of the Gulf of Bothnia, in the north.

Finnic spread separately, starting slightly later. Moving along waterways of what Lang (2018: 310) terms the Southwest Passage (middle-upper Oka to the south coast of the Gulf of Finland), initially in small groups of non-farmers and then in larger farming groups that built fortified settlements and eventually assimilated Baltic and Germanic groups, it took ancestral Pre-Finnic speakers over half a millennium to reach the Baltic coast, absorbing Baltic and then Germanic linguistic influence. Movement continued to the west, south, and then north, bringing ancestral Finnish to the southeast of Finland in the early centuries CE (c. 2000-1500 BP) and Karelian to southern Karelia slightly later. Finnish then began spreading north in Finland and Karelian in Karelia, displacing or absorbing Saamic speakers and spreading slash-and-burn agriculture to formerly hunter-gatherer lands, around 1100 BP = 900 CE (probably as the onset of the Medieval Warm Period made agriculture productive in the north) (Lang 2018; Saarikivi 2004ab).

Thus, this western spread proceeded in two local spurts, the first after nearly a millennium's hiatus with a stable frontier around the Oka, and the second almost a millennium after that. The second spurt appears to have been driven primarily by climate factors. As of c. 1,500 BP there was an unbroken (though sparse) Western Uralic population from the Oka to the Baltic Sea area; it is now split and partly absorbed by the Slavic and later Russian expansion. This was a different process from the rapid initial Uralic spread. It produced no durable initial settlements that are also branch homelands. It produced substratal effects in both Saamic and Finnic. Northern Estonia and southeastern Finland are conventionally called homelands (Saarikivi in press), but they were more nearly staging areas where frontiers halted temporarily and where dialect divergences began and gained the status of distinct languages.

## S6. Post-CU diffusions with IBD effects

Some of the reshapings of morphological paradigms do show east-west effects. We review two clear cases here: pronoun stems and suffix ordering.

S6.1. Pronoun stems

In the Samoyedic branch, most clearly in Tundra Nenets, and in Hungarian, personal pronouns inflect for case but there is no single pronoun stem carrying person-number and no set of case suffixes. Rather, the stem carries case or case-like information, and person-number is carried in the suffix, which usually contains or is identical to the possessive suffixes (Table S6.1). In contrast, in the westernmost languages there is a single person-number stem which takes the same case suffixes as nouns do (Table S6.2). Intermediate languages have person marking in some of the pronoun case endings of a single-stem pronoun paradigm. The westernmost such language, Veps, has it in only one case and only in the singular in northern and central varieties (Table S6.3) but in most or all oblique cases, singular and plural, in southern varieties (Grünthal 2015:276; Zajceva 1981:169, 234; 1993).

**Table S6.1.** Hungarian partial pronoun paradigm, and a noun for comparison. Person-number element underlined.

| | 1sg | 2sg | Noun |
|---|---|---|---|
| Nominative | én | te | ház 'house' |
| Accusative | eng-<u>em</u>(-et) | tég-<u>ed</u>(-et) | ház-at |
| Dative | nek-<u>em</u> | nek-<u>ed</u> | ház-nak |
| Inessive | benn-<u>em</u> | benn-<u>ed</u> | ház-ban |

**Table S6.2.** Finnish partial pronoun paradigm, and a noun for comparison

| | 1sg | 2sg | Noun |
|---|---|---|---|
| Nominative | minä | sinä | maito 'milk' |
| Accusative | minu-t | sinu-t | (syncretic with genitive)[19] |
| Genitive | minu-n | sinu-n | maido-n |
| Partitive | minu-a | sinu-a | maito-a |
| Adessive | minu-lla | sinu-lla | maido-lla |

**Table S6.3.** Veps (Finnic) partial pronoun paradigm (person element underlined) and a noun for comparison (Grünthal 2015:275, 62)

| | 1sg | 2sg | Noun |
|---|---|---|---|
| Nominative | minä | sinä | hebo 'horse' |
| Genitive | minu-n | sinu-n | hebo-n |
| Illative | minu-hu-<u>in</u> | sinu-he-<u>iž</u> | hebo-he |

The Nenets-Hungarian type is archaic. Few noun cases can be reconstructed for PU, and what are now oblique case suffixes mostly descend from accreted postpositions that carried person-number

[19] Or, with DOM, syncretic with nominative. The Finnish accusative ending *-t* and Hungarian *-d* are cognate.

suffixes agreeing with the object. In the western branches the person suffixes are increasingly lost or less often included in the first place, an IBD distribution.

S6.2. Suffix ordering

The reconstructable order of case and possessive suffixes on nouns is Case-Possessive, reflected consistently in Samoyedic and Saamic-Finnic-Mordvin, e.g., Finnish *talo-ssa-ni* (house-INESSIVE-1sg) 'in my house'. Elsewhere the order has shifted to Possessive-Case, e.g., Hungarian *ház-am-ban* (house-1sg-INESSIVE) 'in my house', in the entire paradigm in Ugric but in fewer and fewer cases farther west (so Permic and Mari have a mix of Case-Possessive and Possessive-Case order, e.g., Mari *kniga-m-yn* [book-1sg-GENITIVE], *kniga-šte-m* [book-LOCATIVE-1sg]), which is a sign of the distribution following a regular IBD pattern. The switch has two explanations: (a) Most of the noun case endings are not PU suffixes but secondary accretions of postpositions, which attached following the inherited possessive suffixes. (b) The change was evidently triggered by close contact with Turkic languages, where the order is Possessive-Case. Either way the change occurred well after the initial Uralic dispersal (the first Turkic contacts came in the 7th century). These later developments have no bearing on the question of IBD effects in the original dispersal. (For the history see Nichols 1973.)

## S7. Typology and the CU spread

Two historically telling distributions of typological properties in Uralic are covered in this supplement: salient typologically eastern properties of Uralic languages, all of them traceable to PU (S7.1-S7.5); and cross-linguistically infrequent traits inherited from Proto-Uralic, most of them fairly stable in Uralic (SS 7.6-7.10).

S7.1. Low finiteness (Shagal et al. 2019)

European languages, and especially western European languages, use finite verbs (with or without subordinating conjunctions, depending on the construction) in many types of subordination and complementation. Languages to the east use nonfinites such as infinitives, participles, verbal nouns, and converbs, all with increasing frequency farther east. Even the westernmost Uralic languages, the Finnic and Saamic ones, use nonfinites much more frequently than their Indo-European neighbors Swedish, German, Latvian, and Russian. The eastern languages use nonfinites with high frequencies similar to those of their Turkic neighbors. Thus Uralic preserves an eastern cast overall, while internally to the family individual languages adapt in the direction of their neighbors without changing fully to western profiles.

S7.2.High inflectional person (Nichols 2017)

In PU and conservative languages, person is marked on verbs (argument indexation), nouns (possessive marking), and adpositions (object indexation); in most of Europe it is marked only on verbs. The conservative Uralic personal pronouns do not have the same case suffixes and stem shapes as nouns, the lexical class with which they share most syntactic properties. Unlike nouns, and rather like inflectional affixes, they have a uniform shape using rhyme and/or alliteration which serves to identify them as a paradigmatic set and to echo inflectional categories such as number and case (see S6.1 above and S7.10 below). In all of these respects, person resembles an inflectional category rather than a lexical one.

S7.3. High part-of-speech flexibility in the lexicon (Janhunen 2001, 2020)

Many PU roots are reconstructible as flexible noun/verbs (traditional Uralicist term: nomina-verba); in addition, in many modern Uralic languages there is ready neutralization of the noun-verb contrast in predicate function, where nouns take verbal person-number and TAM suffixes directly or can be used without either verbal suffixes or a copula (for the analysis of this phenomenon as neutralization, see Beck 2013). Noun-verb flexibility, whether at the root or the lexeme level, is a Pacific Rim property, reflected clearly in Tagalog and other Philippine languages and in Salishan and Wakashan languages among others (Foley 2017; Nichols 2016).[20]

S7.4. Non-accusative alignment

The synchronic alignment types of Uralic languages are solidly accusative, with the sole exception of limited ergativity in Eastern Khanty (Filchenko 2007:410-413; Kulonen 1989), but there are several cases where one or another small corner of one or another Uralic language displays ergativity

[20] An orthogonal distributional pattern is a tendency for POS flexibility to decrease in strong contact situations (Foley 2017). In the Turkic and Mongolic families, which have histories of spread by language shift and whose daughters are contact languages par excellence, flexibility is very low, especially in languages on or near the steppe. In Tungusic it is considerably lower in Manchu than in Evenki and Even at the northern periphery. In modern Uralic languages it is low in Finnish and Hungarian and high in Kildin Saami, Mansi, and Nganasan. – Beck 2013 shows that "noun-verb flexibility" is a misnomer, which does not affect our observation of the distribution.

(Janhunen 2020:383-386; cf. also Havas 2003)[21]. A number of these involve parallel or cognate forms (and not just the abstract functional pattern of ergativity). Since ergativity is prone to be lost and not easily gained (Maslova and Nikitina 2008; Nichols 2003:295), taken together these could plausibly be survivals of once more pervasive ergative patterning, and they are more likely to point to inheritance than to substratum or other contact. Ergative alignment is rare in central northern Eurasia, but found at the peripheries (Basque in the west, the Caucasus in the southwest, and Eskimoan and Chukotkan in the far east). If the minor patterns of Uralic are indeed surviving traces, the ancestor of Proto-Uralic may have brought an intrusive eastern pattern to the interior of the continent.

The implications may be more general than ergativity. In daughter languages the accusative has tended to be lost or syncretized with other cases. A case paradigm can be reconstructed for PU (Aikio in press: §1.4.3), but only in the singular; in the plural only nominative and genitive can be reconstructed (Janhunen 2020). Possibly the singular case endings were added to a plural suffix in the other cases, but if not, this was number-based split alignment (accusative in the singular, neutral in the plural). In the objective conjugation (see §2.2.9), the subject agreement markers were the same as the noun possessive markers. These and other patterns reviewed by Havas (2008) and Janhunen (2020) suggest tendencies away from canonical accusative marking and/or secondary development of the PU case system; Havas describes PU as *prenominative.* In addition to alignment, these patterns have implications for locus of marking: in the plural – in noun inflection and the objective conjugation of verbs – PU inflection was head-marking, and head-marking morphology is a north Asian and North American feature.

S7.5. Base intransitivity (in the terms of Nichols et al. 2004)

In Uralic languages, the simplest form in a derivational paradigm of causal and non-causal verbs is often intransitive, and semantic causatives are often derived (with causative suffixes). Aikio (in press) reconstructs three causative suffixes but only one detransitivizing suffix, *-w-, glossed 'stative/automative passive' (Aikio in press:§1.4.5.D), which implies an aktionsart (actionality) category as much as a valence change, while the causatives are clearly valence-changing. The modern reflexes of the stative/automative passive are functionally diverse, often lexicalized or frozen, and not primarily means for deriving valence pairs, while the causative suffixes are usually dedicated causatives, productive, and valence-deriving. Derived causatives are also common in Siberia and nearby, but, in contrast, western European languages often derive the non-causal by reflexivization (Table S7.1). In addition, in Uralic languages as in Siberian languages (though not universally among causativizing languages), verb sets with prototypically animate S/O (such as 'fear' : 'frighten') are more prone to be causativizing than those with prototypically inanimate S/O (such as 'boil') (Grünthal and Nichols 2016; Grünthal et al. 2021).

**Table S7.1.** Examples of causativization (Erzya Mordvin) and decausativization (Spanish), with the relevant derivational morphology underlined.

| | | 'sit down' | 'seat (someone), have sit, let sit' |
|---|---|---|---|
| (a) | Erzya | oza-ms | oza-vto-ms |
| | | sit-INF | sit-CAUS-INF |
| (b) | Spanish | senta-r=se | senta-r |
| | | sit-INF=REFL | sit-INF |

[21] Janhunen (2020:385) mentions the Finnish genitive used in nominalizations (*minu-n teke-mä-ni* [1sg-GEN do-NMLZ-PX.1sg] 'the one done by me') and converb constructions, and the objective conjugation of languages such as Hungarian, where the person-number marker is based on possessive suffixes. In all of these the subject of a transitive verb is in an oblique case.

All of these traits suggest that PU entered the western Eurasian region from the eastern part of its range, and stemmed ultimately from a typologically eastern linguistic population. (A negative verb and personless pronoun roots, just discussed below as cross-linguistically rare traits, also have eastern distributions.)

Typologically rare traits, again inherited and most of them stable in the family, are covered in the following sections. Subsequently to the initial CU spread they have some tendency to be eroded under contact, especially in the west, but their tenacity is still notable. (Most of these fall into the type of infrequent and stable features in the terms of Greenberg 1978:75-76.)

S7.6. Dual

A dual number category is found in Samoyedic, Ob-Ugric, and Saamic, and can be reconstructed to PU since a suffix *-kV is found those three branches. However, since the suffix resembles the numeral 'two' *kektä~käktä, it is conceivably an independent analogical innovation and not unambiguously PU. Based on our preliminary search, a dual is found in under 25% of the world's languages and is inherently recessive, tending to be unstable and/or inconsistent in the language families and areas that do have it; this makes it an unlikely independent development and a likely retention in the three Uralic branches**.** The history of the dual in IE languages appears to be one of steady erosion (and by now mostly loss). In Saamic at least the dual has undergone some formal renewal, testifying to vitality, a sharp difference from its Germanic neighbors. Otherwise, the dual is lost in European Uralic. Its interest here lies in the fact that it was present in the protolanguage at all, rather than being acquired in the process of spread.

S7.7. Negative verb

A negative auxiliary verb is found in most Uralic languages, and a PU root *e- is reconstructable (Aikio in press) (some daughter languages also have others). Often it is a finite verb which takes person-number and TAM marking and takes an underspecified nonfinite form of the lexical verb (known as *connegative*), e.g., Finnish *e-n puhu* (NEG-1sg speak.CNG) 'I don't speak'. This construction is reconstructable to PU and is most consistently retained in Samoyedic; elsewhere there is a tendency to lose finite properties of the negative verb, with tense and even person shifting to the lexical verb (Aikio in press: §1.5). Worldwide, as our counts below show, only about 12% of the languages have a negative verb or put person marking on the negative element. Frequencies for both peak in North Asia, western North America, and (to a lesser extent) Mexico-Central America. Thus the Uralic negative construction is a typologically rare structure and associated with the North Pacific Rim.

We drew a sample of 400 languages stratified by genera, trying to include at least one language per genus for as many genera as possible and then, additionally, sampling some genera more densely. We determined the proportions of languages, genera, and stocks that have negative verbs or put person marking on the negative morpheme. The sample is that of Miestamo (2005), plus most of the languages Dryer (2013) coded as having or potentially having a negative verb and several more added by us; the number of genera recognized grew between 2005 and 2013 and again in our survey as the languages we added to Miestamo's sample are more recently published or reclassified. Negative verb is defined as in Miestamo (2005:81-82): the finite element of the negated clause is the negative marker (either an auxiliary verb or a higher verb taking a clausal complement). Putting person on the negative morpheme means that the negative morpheme inflects for person: person (or person-number) indexes are attached to the negative morpheme by affixation, cliticization, or stem change such as ablaut or tone change. (Mere adjacency of negative and person markers in the inflected verb does not count. If person markers are clitics and positioned relative to phrase or clause boundaries, as with second-position clitics, and in some or all instances of negation the negative morpheme is clause-initial and the person clitic mechanically follows it, we did not count it. In these cases, person and negation are adjacent, but the negative morpheme cannot be said to be inflecting for person.)

Table S7.2 shows the proportions of languages, genera, and stocks that have a negative verb or put person marking on the negative element, continent by continent. We calculated this for all languages and then for all except Uralic, to see whether Uralic was singlehandedly responsible for high frequencies in Central and North Asia (the continent where all Uralic languages except Hungarian are found; for continent definitions see Bickel et al. 2022). The most general outcome is that Central and North Asia, Western North America, and Mexico-Central America tend to have high frequencies – for both constructions, at all three levels, and with or without Uralic (Mexico-Central America is less consistent but still follows the trend). These three continents are almost always above the mean and often more than one standard deviation above it (while the other continents are usually below the mean, except that Africa is usually just at the mean). Removing Uralic weakens the position of Central and North Asia for genera and stocks, but strengthens it for Mexico-Central America. These three continents define the North Pacific Rim language population, showing that negative verbs and person on negation are features of that population. Uralic does not create that distribution but follows it.

**Table S7.2.** Proportions of languages, genera, and stocks with negative verbs and person marking on the negative morpheme. Top: all languages; bottom: excluding Uralic. Yellow = 1 s.d. or more above the mean; light yellow = very close (within 10% of 1 s.d.). Genera following WALS (Dryer and Haspelmath 2013). Stocks following AUTOTYP (Bickel et al. 2022).

| | **No. languages with:** | | | | **% genera with any:** | | | | **% stocks with any:** | | | | **% yes per stock** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Continent | NegVerb | | PersNeg | | NegVerb | | PersNeg | | NegVerb | | PersNeg | | NegVer | PersNeg |
| | N | % | N | % | N | % | N | % | N | % | N | % | | |
| Africa | 69 | 0.10 | 76 | 0.09 | 49 | 0.12 | 53 | 0.13 | 32 | 0.08 | 36 | 0.11 | 0.10 | 0.10 |
| Europe-Caucasus | 21 | 0.04 | 23 | 0.04 | 16 | 0.06 | 17 | 0.06 | 7 | 0.07 | 7 | 0.07 | 0.07 | 0.07 |
| Central & N Asia | 25 | 0.36 | 33 | 0.30 | 21 | 0.33 | 24 | 0.29 | 13 | 0.10 | 13 | 0.07 | 0.36 | 0.36 |
| S & SE Asia | 42 | 0.07 | 43 | 0.09 | 27 | 0.11 | 29 | 0.10 | 9 | 0.09 | 10 | 0.06 | 0.07 | 0.10 |
| Australia | 20 | 0.00 | 26 | 0.04 | 18 | 0.00 | 21 | 0.05 | 16 | 0.00 | 16 | 0.04 | 0.00 | 0.04 |
| New Guinea & Oceania | 45 | 0.09 | 57 | 0.02 | 31 | 0.1 | 41 | 0.02 | 27 | 0.05 | 34 | 0.02 | 0.09 | 0.02 |
| W North America | 36 | 0.22 | 39 | 0.21 | 33 | 0.21 | 34 | 0.24 | 25 | 0.11 | 26 | 0.20 | 0.21 | 0.20 |
| E North America | 17 | 0.06 | 17 | 0.06 | 16 | 0.06 | 16 | 0.06 | 14 | 0.06 | 14 | 0.06 | 0.06 | 0.06 |
| Mexico-Central America | 22 | 0.14 | 25 | 0.16 | 20 | 0.15 | 22 | 0.18 | 10 | 0.11 | 11 | 0.11 | 0.13 | 0.15 |
| South America | 53 | 0.08 | 60 | 0.05 | 45 | 0.09 | 48 | 0.06 | 40 | 0.07 | 42 | 0.05 | 0.08 | 0.05 |
| | | | | | | | | | | | | | | |
| Total | 350 | | 399 | | 276 | | 305 | | 193 | | 209 | | | |
| Mean | | 0.12 | | 0.11 | | 0.12 | | 0.12 | | 0.08 | | 0.08 | 0.12 | 0.11 |
| s.d. | | 0.10 | | 0.09 | | 0.09 | | 0.09 | | 0.03 | | 0.05 | 0.10 | 0.10 |
| high (+1 sd) | | 0.22 | | 0.20 | | 0.21 | | 0.21 | | 0.11 | | 0.13 | 0.22 | 0.22 |
| | | | | | | | | | | | | | | |
| Same, minus Uralic | | | | | | | | | | | | | | |
| Africa | 69 | 0.10 | 76 | 0.09 | 49 | 0.12 | 53 | 0.13 | 32 | 0.08 | 36 | 0.11 | 0.10 | 0.10 |
| Europe-Caucasus | 20 | 0.04 | 23 | 0.05 | 15 | 0.06 | 15 | 0.06 | 7 | 0.07 | 7 | 0.07 | 0.07 | 0.07 |
| Central & N Asia | 19 | 0.21 | 25 | 0.16 | 16 | 0.19 | 18 | 0.11 | 12 | 0.10 | 12 | 0.07 | 0.19 | 0.17 |
| S & SE Asia | 42 | 0.07 | 43 | 0.09 | 27 | 0.11 | 29 | 0.10 | 9 | 0.09 | 10 | 0.06 | 0.07 | 0.10 |
| Australia | 20 | 0.00 | 26 | 0.04 | 18 | 0.00 | 21 | 0.05 | 16 | 0.00 | 16 | 0.04 | 0.00 | 0.04 |
| New Guinea & Oceania | 45 | 0.09 | 57 | 0.02 | 31 | 0.10 | 41 | 0.02 | 27 | 0.05 | 34 | 0.02 | 0.09 | 0.02 |
| W North America | 36 | 0.22 | 39 | 0.21 | 33 | 0.21 | 34 | 0.24 | 25 | 0.11 | 26 | 0.20 | 0.21 | 0.20 |
| E North America | 17 | 0.06 | 17 | 0.06 | 16 | 0.06 | 16 | 0.06 | 14 | 0.06 | 14 | 0.06 | 0.06 | 0.06 |
| Mexico-Central America | 22 | 0.14 | 25 | 0.16 | 20 | 0.15 | 22 | 0.18 | 10 | 0.11 | 11 | 0.11 | 0.13 | 0.15 |
| South America | 53 | 0.08 | 60 | 0.05 | 45 | 0.09 | 48 | 0.06 | 40 | 0.07 | 42 | 0.05 | 0.08 | 0.05 |
| | | | | | | | | | | | | | | |
| Total | 343 | | 391 | | 270 | | 297 | | 192 | | 208 | | | |
| Mean | | 0.10 | | 0.09 | | 0.11 | | 0.10 | | 0.08 | | 0.08 | 0.10 | 0.09 |
| s.d. | | 0.07 | | 0.06 | | 0.06 | | 0.07 | | 0.03 | | 0.05 | 0.06 | 0.06 |
| high (+1 sd) | | 0.17 | | 0.15 | | 0.17 | | 0.17 | | 0.11 | | 0.13 | 0.16 | 0.15 |

## S7.8. Differential object marking

Most modern Uralic languages use differential object marking (DOM; also known as *unmarked object*) of a common eastern Eurasian type: there is an accusative case, but it is used only of definite or

specific objects; and the default marking of objects is no case suffix (so it is identical to the nominative). This pattern can probably be reconstructed to PU (Havas 2008; Janhunen 2020; Aikio in press). DOM in itself is not infrequent cross-linguistically; DOM or other restriction on overt case marking of objects is expected in languages with object case marking (Sinnemäki 2014), and DOM is particularly frequent in Eurasia (Sinnemäki 2014; Bickel et al. 2014, Bickel et al. 2015). What is distinctive about DOM in Uralic is its stability: cross-linguistically, DOM itself is usually fairly stable in families, but the conditioning factors on it are not (Sinnemäki 2014). In Eurasia, animacy dominates strongly to the south and southeast of Uralic (in the Indian peninsula and in Southeast Asia) and somewhat also west of Uralic (in Europe). Definiteness and specificity dominate in the Uralic-speaking areas and in the Caucasus. In Turkic languages definiteness is also important, but animacy appears to be overall more important than in the languages of the Uralic family. The conditioning factors of DOM seem to be areal but in Uralic they are also persistent within the family. Most Uralic languages, atypically, preserve the same conditioning factors of definiteness or specificity (Hungarian has lost it and uses unrestricted overt accusative marking).

Uralic DOM is furthermore part of a larger pattern involving general inhibition of accusative case marking, number-based split alignment, and locus of marking (above). This makes DOM better integrated with the rest of Uralic grammar than is usual; in other languages it is often a lone pattern. Perhaps this accounts for its stability in Uralic.

S7.9. Contrast of subjective vs. objective conjugation

Objective conjugation is the Uralicist term for indexation of the object (as well as the subject) on the verb. It is found in the eastern branches (Samoyedic, Khanty, Mansi, Hungarian) and Mordvin; endings are reconstructable but whether it existed in PU is debated (Aikio in press: his Table 1.6, §1.4.4). In most of the languages it takes the form of person-number indexation of the subject but only number indexation of the object. Exceptions are Mordvin, which indexes both person and number of both arguments, though with hierarchical effects involving second person; and Hungarian, which registers the presence of an object but does not index its properties. Verbal object indexation is entirely lost in Saamic and Finnic, likely due to long contact with Baltic and Germanic languages, which lack it; and also in Permic and Mari. Object indexation is rare in western Eurasia, and where it does occur (chiefly in Basque, West Caucasian languages, and Kartvelian languages) it involves full person-number indexation; object indexation also emerges, again for both person and number, from pronoun cliticization in Balkan and Romance languages. In some Balkan languages this is true agreement, where the indexed argument can be doubled by an overt noun or pronoun in the clause. It is also true agreement in most Uralic languages. Worldwide, it is probably more often pronominal agreement, where doubling with an overt object is ungrammatical. Thus object indexation in general, and its Uralic form more specifically, are persistent infrequent features in Uralic.

S7.10. Personless pronoun stems

The archaic Uralic personal pronoun paradigms (S6.1 above) have multiple case-suppletive stems whose lexical content is case-like and does not include person. Head-marking languages rather often have generic pronoun stems that take person inflection as their only marking of person (Nichols 2017), but the combination of that kind of person marking with case marking by lexical stems is very rare. The case marking is what has created the Uralic system, as postpositions inflected for person accreted to some generic or personless stem to function as case markers. This history suggests that Pre-Proto-Uralic may have been more consistently head-marking than any daughter language is; if so, that is another eastern trait.

S7.11. Conclusions

All of these features are of typological and historical interest and were brought across northwestern Eurasia in the Uralic spread. Sections S7.1-S7.5, S7.7, and S7.10 describe features that are eastern in type and support an origin of PU in the linguistic population to the east of the Uralic range.

S7.6-S7.10 describe cross-linguistically infrequent features that are inherited from PU and spread with the CU spread. Of these, S7.6, S7.8, and S7.9 describe features that are not eastern but also not western (where the prehistoric western type can be estimated from substratal influences on western frontier IE and Uralic languages, and from Basque and what can be gleaned bout Etruscan). Despite a variety of contact effects around the Uralic periphery, and known substrata in the northward spreads of at least Saamic and Samoyedic, the features covered here have remained sufficiently stable to easternize and Uralicize the typological profile of the northern part of western Eurasia.

## S8. Early diversification in the east

A sequence of early sound changes is arguably shared between Samoyedic and Ugric: change of PU *s to a sound symbolized *L, which has varied reflexes but regular correspondences; followed by merger of PU *š and *s to yield a new *s (Zhivlov 2018). This is a shared ordered sequence, a very strong piece of evidence, but Zhivlov shows that in Mansi it occurred before the Indo-Iranian borrowings while in Khanty it occurred after them (so it affects them). He concludes that the sequence of changes could have been an areal phenomenon rather than inherited. We also note that, since the I-I borrowings affected each of the nine major branches separately, they could have reached Mansi later than Khanty; this could be assessed if we knew the early locations of Proto-Mansi and Proto-Khanty and the source(s) of their I-I loans. Figure 1 in the main text shows major clusters of Seima-Turbino archaeological sites along all three of the Tobol, Irtysh, and Ob', any of which might have hosted or drawn both Ugric and Indo-Iranian speakers, making it difficult to identify a region where ancestral Mansi might have been out of contact with I-I while ancestral Khanty was in contact.

The numbers of retained PU etyma (Figure 2 in main text and Figures S4(a-c)) are higher in the west than in the east (except for artifactually high Samoyedic), which may be consistent with earlier separation times in the east. Alternatively, it could reflect peripheral archaism in the west and/or the known areal effects in the central languages, where early in their histories there was contact among the languages and with non-Uralic neighbors, producing lexical and grammatical innovations which replaced inherited words and structures. Background factors are the long attestation of Finnic, the more extensive research history of Finnic and Saamic, and the greater numbers of well-attested daughter languages in both, all of which are expected to raise frequencies of attested cognates.

Inter-branch and inter-language lexical sharings of PU etyma (Figure S8; see also Figure S4c) are high for Saamic and especially Finnic, high between Khanty and Mansi, and otherwise low for Ugric and especially Hungarian. These can reflect the same factors as above, and also long-term close interaction between Finnic and Saamic and between Khanty and Mansi (which are known to have been in close contact at least in protohistorical and historical times). They indicate that, except for the close Khanty-Mansi interaction, the Ugric and Samoyedic languages were less connected to each other than was the case for the languages along the Volga (other than Mari, which is less connected). This state of affairs began early, with the initial dispersal, and continued at least until the branch-internal dispersals.

| | Saamic | Finnic | Mordvin | Mari | Permic | Hungaria | Mansi | Khanty | Samoyedic |
|---|---|---|---|---|---|---|---|---|---|
| Saamic | – | 87 | 58 | 54 | 62 | 41 | 55 | 63 | 105 |
| Finnic | | – | 72 | 69 | 69 | 53 | 66 | 74 | 130 |
| Mordvin | | | – | 59 | 51 | 40 | 44 | 52 | 88 |
| Mari | | | | – | 55 | 41 | 47 | 48 | 85 |
| Permic | | | | | – | 45 | 58 | 60 | 93 |
| Hungarian | | | | | | – | 49 | 48 | 65 |
| Mansi | | | | | | | – | 81 | 91 |
| Khanty | | | | | | | | – | 101 |
| Samoyedic | | | | | | | | | – |

(S8.a) Branches

| | Skolt Saami | Inari Saami | North Saami | South Saami | South Estonian | Livonian | Estonian | Votic | Veps | Karelian | Finnish | Erzya Mordvin | Mari | Permic | Hun-garian | Mansi | Khanty | Selkup | Tundra Nenets | Ngana-san |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Skolt Saami | – | 91 | 91 | 78 | 58 | 55 | 65 | 59 | 64 | 73 | 74 | 50 | 45 | 54 | 36 | 46 | 55 | 73 | 73 | 60 |
| Inari Saami | | – | 95 | 83 | 57 | 54 | 64 | 58 | 62 | 73 | 74 | 49 | 47 | 55 | 37 | 48 | 56 | 73 | 74 | 59 |
| North Saami | | | – | 85 | 60 | 57 | 68 | 61 | 65 | 76 | 78 | 52 | 48 | 57 | 37 | 49 | 57 | 76 | 78 | 64 |
| South Saami | | | | – | 57 | 53 | 64 | 59 | 61 | 73 | 75 | 47 | 48 | 53 | 38 | 45 | 51 | 69 | 75 | 59 |
| SouthEstonian | | | | | – | 93 | 104 | 95 | 87 | 99 | 104 | 60 | 57 | 59 | 44 | 55 | 58 | 85 | 84 | 70 |
| Livonian | | | | | | – | 97 | 89 | 84 | 96 | 99 | 59 | 55 | 59 | 44 | 52 | 56 | 83 | 79 | 71 |
| Estonian | | | | | | | – | 102 | 95 | 110 | 115 | 65 | 61 | 63 | 46 | 55 | 62 | 93 | 94 | 80 |
| Votic | | | | | | | | – | 91 | 101 | 105 | 58 | 53 | 59 | 44 | 52 | 62 | 84 | 86 | 76 |
| Veps | | | | | | | | | – | 101 | 103 | 55 | 52 | 55 | 43 | 51 | 58 | 80 | 83 | 71 |
| Karelian | | | | | | | | | | – | 127 | 69 | 65 | 65 | 50 | 64 | 71 | 99 | 102 | 89 |
| Finnish | | | | | | | | | | | – | 70 | 67 | 66 | 52 | 66 | 73 | 103 | 107 | 92 |
| Erzya Mordvin | | | | | | | | | | | | – | 59 | 51 | 40 | 44 | 52 | 75 | 70 | 58 |
| Mari | | | | | | | | | | | | | – | 55 | 41 | 47 | 48 | 74 | 69 | 57 |
| Permic | | | | | | | | | | | | | | – | 45 | 58 | 60 | 82 | 77 | 62 |
| Hungarian | | | | | | | | | | | | | | | – | 49 | 48 | 56 | 57 | 47 |
| Mansi | | | | | | | | | | | | | | | | – | 81 | 79 | 75 | 61 |
| Khanty | | | | | | | | | | | | | | | | | – | 89 | 82 | 68 |
| Selkup | | | | | | | | | | | | | | | | | | – | 115 | 92 |
| TundraNenets | | | | | | | | | | | | | | | | | | | – | 103 |
| Nganasan | | | | | | | | | | | | | | | | | | | | – |

(S8.b) Languages. Boxes enclose branches and the Ugric possible branch (dashed box): from left, Saamic, Finnic, Ugric, Samoyedic. Conventions as for (a).

**Figure S8.** Shared retentions of PU etyma. Yellow = ≥ 1 s.d. above the mean; blue = ≥1 s.d. below the mean. Samoyedic languages are included for completeness, though their totals are artifactually high (gray); colors in the rest of the graph are based on mean and standard deviation calculated excluding Samoyedic.

## S9 Sociolinguistics of post-catastrophe spreads

It is widely assumed that the medieval bubonic plague pandemic (the Black Death in Europe) had sociolinguistic consequences that may have affected grammar, but there seems to have been no systematic comparison. (For views on whether the consequences of the Black Death simplified Norwegian inflection, see Askedal 2005; Jahr 1999, 2008; Mæhlum 2000; Olthoff 2017; Sandøy 2004. For coastal European languages see Johansson 1997.)

Clearer cases are those in which a catastrophe affecting one language affords an expansion opportunity to another. Other factors are often also involved, so that the catastrophe is a major contributor to changes but not the sole one. One factor in the seventh-century spread of Arabic across the Near East and North Africa was that these areas had suffered drastic population losses in the Plague of Justinian (5th–6th centuries), while Arabia had not. This gave Arab armies a numerical advantage; but other factors included recent border conflicts between Byzantium and Persia, and the Arab policy of taxing only non-Muslims, which encouraged conversion and consequent language shift. In the Balkan Peninsula, the spread of Slavs and Avars was partly favored by local population losses during the Plague of Justinian; but other factors included a preceding cold period which drew farmers (including Slavs) southward (Lindstedt and Salmela 2020). In the North American Great Basin, a prolonged and severe drought during the medieval period drove out the farming Fremont Culture and led to a sweep of Uto-Aztecan-speaking desert foragers across the area (Madsen and Rhode eds. 1994). The 4.2 ka drought may have depopulated the western Great Plains, after which ancestral Algonquian expanded eastward, eventually to dominate a large part of the North American Fur Road (Hill 2004). In Australia, after a millennia-long arid period largely depopulated the interior desert, climate amelioration enabled Pama-Nyungan speakers to recolonize the area; but other factors probably included intensification of plant and especially seed usage among Pama-Nyungans (Evans and McConvell 1998; Lourandos 1997). In North Africa, a 3000-year northward shift of the monsoon belt c. 10,500 BP turned the Sahara Desert into a grassland with a number of archaeological sites as people moved in from all directions; midway in the event, livestock were introduced to North Africa. When the monsoon belt shifted back c. 7300 BP the population again became very sparse except in the Nile valley, where there is evidence of conflict over increasingly limited resources (Kuper and Kröpelin 2006). What must have been a linguistically diverse population with a mix of hunter-gatherer and pastoral economies became primarily Afroasiatic-speaking and pastoral. In some of these cases a catastrophe affected one of two neighboring populations much more than the other.

Such spreads have generally reduced the linguistic diversity of the affected areas, replacing members of more than one language family with the one successor. At least some of the cases may have involved language shift (of the decimated population to the successor language), and language shift could well have decomplexified the spreading successor language (as is expected when an expanding language absorbs an appreciable number of adult L2 learners: Trudgill 2011). Most Uralic languages are in fact less complex than the general northern Eurasian level and comparable to the languages that have undergone large spreads (German, Spanish, Turkish, Yakut, Mongolian); but the Samoyedic languages, which are notably archaic at least in their morphology, are among the most complex (Nichols 2019). Most Saamic languages have high complexity because of their many unpredictable noun stem alternations, which are post-Proto-Saamic. South Saami, where those have not developed, has low complexity. The complexity of Samoyedic languages is partly post-Proto-Samoyedic (and due to stem alternations) but partly inherited.

Thus there is no evidence of decomplexification in PU, but possible evidence of it in Finno-Ugric.

**Abbreviations** (used in Supplements; those also used in the main text are given there)

CNG Connegative (the form of the lexical stem following the negative auxiliary verb)
PX Possessive suffix (followed by person-number abbreviation)

Others follow the Leipzig glossing conventions. Protolanguage and language family names as in main text.

## Abbreviations to the references

EWU = Benkő ed. 1993
MdWb = Paasonen 1990-1999
MszFE = Lakó ed. 1967-1978
SW = Janhunen 1977
UEW = Rédei 1986-1989

**References** (for Supplements only)


Aikio, Ante. 2007. Etymological nativization of loanwords: A case study of Saami and Finnish. Ida Toivonen and Diane Nelson (eds.), *Saami linguistics*, 17–52. Amsterdam: Benjamins.

Aikio, Ante. 2012. An essay on Saami ethnolinguistic prehistory. In Riho Grünthal and Petri Kallio (eds.), *A linguistic map of prehistoric northern Europe,* 63-117. Helsinki: Société Finno-Ougrienne.

Aikio, Ante. 2015. The Finnic 'secondary e-stems' and Proto-Uralic vocalism. *Suomalais-Ugrilaisen Seuran Aikakauskirja / Journal de la Société Finno-Ougrienne* 95. 25–66.

Aikio, Ante. In press. Proto-Uralic. In Bakró-Nagy, Marianne, Johanna Laakso and Elena Skribnik, (eds.), *The Oxford Guide to the Uralic Languages* Oxford: Oxford University Press.

Andersen, Henning. 1996. *Reconstructing prehistoric dialects: Initial vowels in Slavic and Baltic*. Berlin: Mouton de Gruyter.

Andersen, Henning. 2003. Slavic and the Indo-European migrations. In Andersen, Henning (ed.), *Language contacts in prehistory: Studies in stratigraphy*, 45–76. Amsterdam-Philadelphia: Benjamins.

Anthony, David W. 2007. *The horse, the wheel, and language: How Bronze Age riders from the Eurasian steppes shaped the modern world*. Princeton, NJ: Princeton University Press.

Askedal, John O. 2005. Typological reflections on loss of morphological case in Middle Low German and in the Mainland Scandinavian languages. In Michael Fortescue, Eva Skafte Jensen, Jens E. Mogensen and Lene Schøsler (eds.), *Historical linguistics 2003: Selected Papers from the 16th International Conference on Historical Linguistics, Copenhagen*. 11–15. Amsterdam: John Benjamins.

Bartens, Raija. 2000. Permiläisten kielten rakenne ja kehitys [The structure and development of the Permic languages]. Helsinki: Suomalais-Ugrilainen Seura.

Beck, David. 2013. Unidirectional flexibility and the noun-verb distinction in Lushootseed. InJan Rijkhoff and Eva van Lier (eds.), *Flexible word classes*, 185–220. Oxford: Oxford University Press.

Belyx, S. K. 2009. *Problema raspada prapermskoj jazykovoj obščnosti.* [The split of Proto-Permic.] Iževsk: Udmurt University.

Benkő, Lórand (ed.). 1993. *Etymologisches Wörterbuch des Ungarischen*. Budapest: Akadémiai Kiadó.

Bickel, Balthasar, Taras Zakharko, Lennart Bierkandt and Alena Witzlack-Makarevich. 2014. Semantic role clustering: An empirical assessment of semantic role types in non-default case assignment. *Studies in Language* 38(3). 485–511.

Bickel, Balthasar, Alena Witzlack-Makarevich, and Taras Zakharko. 2015. Typological evidence against universal effects of referential scales on case alignment. *Scales and hierarchies: A cross-disciplinary perspective,* eds. Ina Bornkessel-Schlesewsky, Andrej L. Malchukov, and Marc Richards, 7–44. Berlin: De Gruyter Mouton.

Bickel, Balthasar, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Riessler, Lennart Bierkandt, Fernando Zúñiga and John B. Lowe. 2022. The Autotyp database, release version 1.0.0.
Zenodo: https://doi.org/10.5281/zenodo.5931509
GitHub: https://github.com/autotyp/autotyp-data/tree/v1.0.0
(accessed Feb. 5, 2022)

Cardillo, M., L. Bromham and Simon J. Greenhill. 2015. Links between language diversity and species richness can be confounded by spatial autocorrelation. *Proceedings of the Royal Society B* 282. 2986.

Carpelan, Christian and Asko Parpola. 2001. Emergence, contacts, and dispersal of Proto-Indo-European, Proto-Uralic, and Proto-Aryan in archaeological perspective. In Christian Carpelan, Asko Parpola and Petteri Koskikallio (eds.), *Early contacts between Uralic and Indo-European: Linguistic and archaeological considerations*, 55–150. Helsinki: Suomalais-Ugrilainen Seura.

Chernykh, E. N. 2008. The "Steppe Belt" of stockbreeding cultures in Eurasia during the early metal age. *Trabajos de prehistoria* 65(2). 73-93.

DeFilippo, Cesare, Chiara Barbieri, Mark Whitten, Sununguko Wata Mpoloka, Ellen Drofn Gunnarsdóttir, Koen Bostoen, and Brigitte Pakendorf. 2011. Y-chromosomal variation in sub-Saharan Africa: Insights into the history of Niger-Congo groups. *Molecular Biology and Evolution* 28:3.1255–1269.

Dellert, Johannes. 2019. *Information-theoretic causal inference of lexical flow*. Berlin: Language Science Press.

Donohue, Mark, Tim Denham and Stephen Oppenheimer. 2012. Uncoupling inheritance and diffusion: A lexical-based methodology detects social distance. *Diachronica* 29(4). 502–522.

Dryer, Matthew. 2013. Negative morphemes. In Matthew Dryer and Martin Haspelmath, eds., *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.
http://wals.info/chapter/112

Dryer, Matthew S. and Martin Haspelmath (eds.). 2013. *The world atlas of language structures online.* Munich: Max Planck Digital Library. Leipzig: Max Planck Institute for Evolutionary Anthropology. Online: http://wals.info. Accessed 2-28-20.

Ellison, T. Mark and Miceli, Luisa. 2017. Language monitoring in bilinguals as a mechanism for rapid lexical divergence. *Language* 93(2). 255–287.

Evans, Nicholas and Patrick McConvell. 1998. The enigma of Pama-Nyungan expansion in Australia. In R. Blench and M. Spriggs (eds.), *Archaeology and language II: Archaeological data and linguistic hypotheses,* 174–191. London-New York: Routledge.

Filchenko, Andrey Y. 2007. *A grammar of Eastern Khanty.* Ph.D. dissertation, Rice University.

Friedrich, Paul. 1970. *Proto-Indo-European trees: The arboreal system of a prehistoric people*. Chicago: University of Chicago Press.

Fodor, István. 2001. Az uráli őstörténet és a régészet. [Uralic prehistory and archaeology.] *Ünnepi könyv Keresztes László tiszteletére. Folia Uralica Debreceniensia* 8. 143–162.

Fodor, István. 2016. Nyelvek, népek, térképek: Néhány észrevétel Juha Janhunen hipotéziséhez. [Languages, peoples, maps: Some observations on the hypthesis of Juha Janhunen.] *Archaeologiai Értesítő* 141(1). 217–229.

Foley, William A. 2017. Structural and semantic dependencies in word class membership. In Nicholas J. Enfield, (ed.), *Dependencies in language: On the causal ontology of linguistic systems,* 179–195. Berlin: Language Science Press.

Greenberg, Joseph H. 1978. Diachrony, synchrony, and language universals. In Joseph H. Greenberg (ed.), *Universals of human language, vol. 1: Method and theory*, 61–92. Stanford: Stanford University Press.
Grünthal, Riho. 2012. Baltic loanwords in Mordvin. 2012. In Riho Grünthal and Petri Kallio, eds. *A linguistic map of prehistoric northern Europe* (SUST 266)*,* 297–343. Helsinki: Suomalais-Ugrilainen Seura.
Grünthal, Riho. 2015. *Vepsän kielioppi*. [A grammar of Veps.] Helsinki: Suomalais-Ugrilainen Seura.
Grünthal, Riho and Johanna Nichols. 2016. Transitivizing/detransitivizing typology and language family history. *Lingua Posnaniensis* 58:2. 11–31.
Grünthal, Riho, Heini Arjava, Jyri Lehtinen, and Johanna Nichols. 2021. Basic causative patterns in Uralic: Retention and renewal in grammar and lexicon. In Luraghi, Silvia and Elisa Roma (eds.), *Valence over time: Diachronic perspectives on valence patterns and valence orientation*. Berlin: Mouton de Gruyter.
Hajdú, Péter 1969. Finnougrische Urheimatforschung. Ural-Altaische Jahrbücher 41.
Hajdú, Péter 1975a. Sukulaisuuden kielellistä taustaa. [On the linguistic background of genetic relationships.] In Péter Hajdú (ed.), *Suomalais-ugrilaiset [Finno-Ugrians],* 11-51. Helsinki: Finnish Literature Society-SKS.
Hajdú, Péter. 1975b. *Finno-Ugrian languages and peoples*. London: André Deutsch.
Hajdú, Péter and Domokos, Péter. 1978. *Uráli nyelvrokonaink*. [Our Uralic linguistic relatives.] Budapest: Tankönyvkiadó.
Häkkinen, Jaakko. 2007. Kantauralin murteutuminen vokaalivastaavuuksien valossa [The dialect formation from Proto-Uralic in the light of vowel correspondences]. MA thesis, University of Helsinki.
Häkkinen, Jaakko. 2009. Kantauralin ajoitus ja paikannus: Perustelut putarissa. [The dating and placing of Proto-Uralic: Weighing the evidence.] *J Sociéte Finno-Ougrienne* 92.9–56.
Häkkinen, Kaisa. 1996. *Suomalaisten esihistoria kielitieteen valossa.* [The prehistory of the Finns in the light of linguistics]. Helsinki: Finnish Literature Society-SKS.
Häkkinen, Kaisa. 2001. Prehistoric Finno-Ugric culture in the light of historical lexicology. In Carpelan, Christian, Asko Parpola and Petri Koskikallio (eds.), *Early contacts between Uralic and Indo-European: Linguistic and archaeological considerations,* 169-186. Helsinki: Suomalais-Ugrilainen Seura.
Hammarström, Harald, Robert Forkel and Martin Haspelmath. 2019. Glottolog 4.1. Jena: Max Planck Institute for the Science of Human History.
http://glottolog.org (ccessed 2020-03-09.)
Uralic family entry: https://glottolog.org/resource/languoid/id/ural1272
Havas, Ferenc. 2003. A tárgy tárgyában: Mondattipológiai fontolgatások. [On the object of the object: Syntactic typological considerations.] In Beatrix Oszkó and Mária Sipos (eds.), *Uráli TÁRGYaló. Budapesti Uráli Műhely* 3. 7-44. Budapest: MTA Nyelvtudományi Intézet.
Havas, Ferenc. 2008. Unmarked object in the Uralic Languages. A diachronic typological approach. *Linguistica Uralica* 44. 1–33.
Haynie, Hannah J. 2014. Geography and spatial analysis in historical linguistics. *Language and Linguistics Compass* 8(8):344–357. https://doi.org/10.1111/lnc3.12087.
Helimski, E. A. 1982. *Drevnejšie vengersko-samodijskie jazykovye paralleli.* Moscow: Nauka.
Helimski, Eugene. 2003. Areal groupings (Sprachbünde) within and across the borders of the Uralic language family: A survey. *Nyelvtudományi Közlemények* 100. 156–167.
Hill, Jane H. 2004. Language spread among hunter-gatherers: The North American evidence. Presented at The Ecology of Language: A Symposium in Honor of Murray B. Emeneau on his 100th Birthday. University of California, Berkeley, February 2004.
Holman, Eric W., Christian Schulze, Dietrich Stauffer and Søren Wichmann. 2007. On the relation between structural diversity and geographical distance among languages: Observations and

computer simulations. *Linguistic Typology* 11(2). 395-423. https://doi.org/10.1515/LINGTY.2007.027.
Holopainen, Sampsa. 2021. On the question of substitution of palatovelars in Indo-Iranian loanwords into Uralic. *JSFOu* 98
Holopainen, Sampsa. 2019. *Indo-Iranian borrowings in Uralic: Critical overview of the sound substitutions and distribution criterion.* PhD dissertation, University of Helsinki
Honkola, Terhi, Kalle Ruokalainen, Kai J. J. Syrjänen, Unni-Päivä Leino, Ilpo Tammi, Niklas Wahlberg and Outi Vesakoski. 2018. Evolution within a language: Environmental differences contribute to divergence of dialect groups. *BMC Evolutionary Biology* 18. 132.
Honkola, Terhi, Outi Vesakoski, Kalle Korhonen, Jyri Lehtinen, Kai Syrjänen and Niklas Wahlberg. 2013. Cultural and climatic changes shape the evolutionary history of the Uralic languages. *Journal of Evolutionary Biology* 26. 1244–1253.
Honti, László. 1998. Obugrian. In Daniel Abondolo (ed.), *The Uralic languages*, 327–357. London-New York: Routledge.
Hunley, Keith, Claire Bowern and Meghan Healy. 2012. Rejection of a serial founder effects model of genetic and linguistic coevolution. *Proceedings of the Royal Society of London, B* 279. 2281–2288.
Itkonen, Terho. 1997. Reflections on Pre-Uralic and the "Saami-Finnic protolanguage". *Finnisch-Ugrische Forschungen* 54. 229–266.
Jahr, Ernst H. 1999. Sociolinguistics in historical language contact: The Scandinavian languages and Low German during the Hanseatic period. In Ernst H. Jahr (ed.), *Language Change: Advances in historical sociolinguistics,* 119–139. Berlin: Mouton de Gruyter.
Jahr, Ernst H. 2008. Historical sociolinguistics: The role of Low German language contact in the Scandinavian typological shift of the late Middle Ages. In Geirr Wiggen, Tove Bull and Marit A. Nielsen (eds.), *Språkhistorie og språkkontakt: Festskrift til Ernst Håkon Jahr på 60-årsdagen 4. mars 2008* [Language history and language contact: Festschrift for Ernst Håkon Jahr on the occasion of his 60th birthday on 4 March 2008], 246–257. Oslo: Novus Forlag.
Janhunen, Juha. 1977. *Samojedischer Wortschatz: Gemeinsamojedische etymologien*. Helsinki: Helsinki University.
Janhunen, Juha. 1981. Uralilaisen kantakielen sanastosta. [On the vocabulary of the Uralic protolanguage]. *Journal de la Société Finno-ougrienne* 77:9.219–274.
Janhunen, Juha. 1998. Samoyedic. In Daniel Abondolo (ed.), *The Uralic languages*. 457–479. London: Routledge.
Janhunen, Juha. 2000. Reconstructing Pre-Proto-Uralic typology: Spanning the millennia of linguistic evolution. In *Congressus Nonus Internationalis Fenno-Ugristarum, Pars I: Orationes plenariae and orationes publicae*, ed. Anu Nurk, Triinu Palo, and Tõnu Seilenthal, 59–76. Tartu: Eesti Fennougristide Komitee, Tartu Unversity.
Janhunen, Juha. 2001. Indo-Uralic and Ural-Altaic: On the diachronic implications of areal typology. In Christian Carpelan, Asko Parpola, and Petteri Koskikallio, eds., *Early contacts between Uralic and Indo-European: Linguistic and archaeological considerations*, 207–220. Helsinki: Suomalais-Ugrilainen Seura.
Janhunen, Juha. 2009. Proto-Uralic -- what, where, and when? *The Quasquicentennial of the Finno-Ugrian Society,* 57-78. Helsinki: Suomalais-Ugrilainen Seura.
Janhunen, Juha. 2020. Grammaticalization in Uralic as viewed from a general Eurasian perspective. In Walter Bisang and Andrej Malchukov, eds., *Grammaticalization scenarios from Europe and Asia.* Berlin: De Gruyter.
Johansson, Christer. 1997. *A view from language: Growth of language in individuals and populations*. Lund: Lund University Press.
Junttila, Santeri. 2012. The prehistoric context of the oldest contacts between Baltic and Finnic languages. In Grünthal, Riho and Petri Kallio, eds., *A linguistic map of prehistoric northern Europe* (SUST 266)*,* 261–296. Helsinki: Suomalais-Ugrilainen Seura.

Junttila, Santeri. 2016. Die baltisch-slawische Frage im Lichte der alten Baltischen Lehnwörter des ostseefinnischen. *Baltistica* LI 2. 217–238.
Kallio, Petri. 2006. Suomen kantakielten absoluuttista kronologiaa. [On the absolute chronology of the protolanguages of Finnish.] *Virittäjä* 110. 2–25.
Kallio, Petri. 2009. Stratigraphy of Indo-European loanwords in Saami. In Tiina Äikäs (ed.), *Mattut – maddagat: The Roots of Saami ethnicities, societies and spaces / places*. 30–45. Oulu: Giellagas Institute at the University of Oulu.
Kallio, Petri. 2015. The language contact situation in prehistoric northeastern Europe. In Robert Mailhammer, Theo Vennemann, and Birgit Anette Olsen (eds.), *The linguistic roots of Europe*, 77–102. Copenhagen: Museum Tusculanum Press, University of Copenhagen.
Kallio, Petri. In press. Substrates in Finnic. In Ante Aikio and Santeri Palviainen (eds.), *Substrate languages in northern Europe: Case studies and methodological perspectives*, Berlin: de Gruyter-Mouton.
Khanina, Olesya, Yuri Koryakov and Andrey Shluinsky. 2018. Enets in space and time: A case study in linguistic geography. *Finnisch-Ugrische Mitteilungen* 42. 109-135.
Koivulehto, Jorma. 1999. Zur indogermanisch-germanischen Kontinuität in der Nachbarschaft der Finnougrier. In Klaas Ph. Ruppel (ed.), *Verba Mutuata*, 341–358. Helsinki: Suomalais-Ugrilainen Seura.
Koivulehto, Jorma. 2000. Finno-Ugric reflexes of Northwest Indo-European and early stages of Indo-Iranian. In Karlene Jones-Bley, Martin E. Huld, and Angela Della Volpe, eds., *Proceedings of the Eleventh Annual UCLA Indo-European Conference*, 21–44. Washington, DC: Institute for the Study of Man.
Koivulehto, Jorma. 2001. The earliest contacts between Indo-European and Uralic speakers in the light of lexical loans. In Christian Carpelan, Asko Parpola, and Petteri Koskikallio, eds., *Early Contacts between Uralic and Indo-European: Linguistic and Archaeological Considerations*, 235–263. Helsinki: Suomalais-Ugrilainen Seura.
Koivulehto, Jorma. 2006. Wie alt sind die Kontakte zwischen Finnisch-Ugrisch und Balto-Slavisch? In *The Slavicization of the Russian north: Mechanisms and chronology*, ed. Juhani Nuorluoto, 179–196. Helsinki: Helsinki University Press.
Korenchy, Éva. 1972. *Iranische Lehnwörter in den obugrischen Sprachen*. Budapest: Akadémiai Kiadó.
Korhonen, Mikko. 1981. Johdatus lapin kielen historiaan [Introduction to the history of the Saami languages]. Helsinki: Finnish Literature Society-SKS.
Koryakov, Yuri. 2007. Slavjane i ix sosedi v VII-VIII vv. [The Slavs and their neighbors in the 7th-8th centuries.] In S.F. Platonov, *Učebnik russkoj istorii. Ch. 3: Russkie slavjane i ix sosedi.* https://history.wikireading.ru/92803
Kulonen, Ulla-Maija. 1989. *The passive in Ob-Ugrian*. Helsinki: Suomalais-ugrilainen Seura.
Kuper, Rudolph and Stefan Kroepelin. 2006. Climate-controlled Holocene occupation in the Sahara: Motor of Africa's evolution. *Science* 313. 803–807.
Lakó, György (ed.). 1967-1978. A magyar szókészlet finnougor elemei [Finno-Ugric elements of the Hungarian vocabulary], 1-3. Budapest: Akdémiai Kiadó.
Lang, Valter. 2015. Formation of Proto-Finnic – an archaeological scenario from the Bronze Age/Early Iron Age. In Harri Mantila, Kaisa Leinonen, Sisko Brunni, Santeri Palviainen, and Jari Sivonen, eds., *Congressus Duodecimus Internationalis Fenno-Ugristaru, Oulu 2015 Plenary Papers*, ed. 63–84. Oulu: University of Oulu.
Lang, Valter. 2018. *Läänemeresoome tulemised*. [Finnic be-comings.] Tartu: Tartu Ülikooli Kirjastus.
Lansing, J. Stephen, Murray P. Cox, Sean S. Downey, Brandon M. Gabler, Brian Hallmark, Tatiana M. Karafet, Peter Norquest, John W. Schoenfelder, Herawati Sudoyo, Joseph C. Watkins and Michael F. Hammer. 2007. Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proceedings of the National Academy of Science of the U.S.A.* 104. 16022–16026.
Lavento, Mika and Valerij S. Patrushev. 2015. Razvitie i xronologija tekstil'noj keramiki na Srednej i Verxnej Volge: Kritičeskij vzgljad vozmožnostej xronologii po $^{14}$C-, AMS-, i tipologičeskomu

metodam. [Evolution and chronology of textile ceramics on the Middle and Upper Volga: Critical survey of $^{14}$C, AMS, and typologically based methods.] *Povolžskaja arxeologija* 2:12.160–188.
Lehtinen, Jyri, Terhi Honkola, Kalle Korhonen, Kai Syrjänen, Niklas Wahlberg and Outi Vesakoski. 2014. Behind family trees: Secondary connections in Uralic language networks. *Language Dynamics and Change* 4. 189–221.
Lindstedt, Jouko and Elina Salmela. 2020. Migrations and language shifts as components of the Slavic spread. In Tomas Klír, Vít Boček, and Nicolas Jansens (eds.), *New perspectives on the early Slavs and the rise of Slavic.* Heidelberg: Winter.
Lourandos, Harry. 1997. *Continent of hunter-gatherers*. Cambridge: Cambridge University Press.
Madsen, David and David Rhode (eds.). 1994. *Across the West: Human population movement and the expansion of the Numa*. Salt Lake City: University of Utah Press.
Mallory, James P. 2015. *The problem of Tocharian origins: An archaeological perspective*. Sino-Platonic Papers, 259. Philadelphia: Department of East Asian Languages and Civilizations, University of Pennsylvania.
Maslova, Elena and Tatiana Nikitina. 2008. Stochastic universals and dynamics of cross-linguistic distributions: The case of alignment types. MS.
https://pdfs.semanticscholar.org/9303/91fcaeeb7e30c02dc983c30c52e28a93e832.pdf?_ga=2.90513048.329825150.1585696309-1325946959.1585696309
Matveev, Aleksander Konstantinovich. 1962. Az ugor népek ősi letelepulési helyei. [The ancient settlement areas of the Ugric peoples.] *Nyelvtudományi Közlemények* 64. 289–297.
Maurits, Luke. 2018. A new Bayesian phylogeny for the Uralic languages. Invited lecture, University of Helsinki.
Metsäranta, Niklas. 2020. Perityminen ja lainautuminen. Marin ja permiläisten kielten sanastontutkimusta [Inheritance and borrowing. On the etymology and lexicon of Mari and Permic languages]. PhD dissertation, University of Helsinki.
Miestamo, Matti. 2005. *Standard negation: The negation of declarative verbal main clauses in typological perspective*. Berlin: Mouton de Gruyter.
Mæhlum, Brit. 2000. Social catastrophes as explanation in historical linguistics: A historiographical and critical discussion of a central topic in the investigation of Norwegian language history. Ernst H. Jahr, ed., *Språkkontakt: Innverknaden frå nedertysk på andre nordeuropeiske språk* [Language contact: Low German influence on other north European languages], 87–94. København: Nordisk Ministerråd.
Nichols, Johanna. 1973. Suffix ordering in Proto-Uralic. *Lingua* 32(3). 227–238.
Nichols, Johanna. 2003. Diversity and stability in language. In *The handbook of historical linguistics*, ed. Brian Joseph and Richard Janda, 283–310. London: Blackwell.
Nichols, Johanna. 2016. Verb-based and noun-based languages. Poster presented at SLE 49, Leiden.
Nichols, Johanna. 2017. Person as an inflectional category. *Linguistic Typology* 21(3). 387–456 .
Nichols, Johanna. 2019. Why is gender so complex? New typological considerations. In *Grammatical gender and linguistic complexity*, ed. Francesca Di Garbo and Bernhard Wälchli, 63–92. Berlin: Language Sciences Press.
Nichols, Johanna, David A. Peterson, and Jonathan Barnes. 2004. Transitivizing and detransitivizing languages. *Linguistic Typology* 8(2). 149–211.
Nichols, Johanna and Tandy Warnow. 2008. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass* 2(5). 760–820.
Nichols, Johanna, Heini Arjava, Jyri Lehtinen and Riho Grünthal. In preparation. Morphosyntactic characters in phylogeny.
Olthoff, M. 2017. Transparency in Norwegian and Icelandic: Language contact vs. language isolation. *Nordic Journal of Linguistics* 40(1). 73–115.
Outram, Alan K., Natalie A. Stear, Robin Bendrey, Sandra Olsen, Alexei Kasparov, Victor Zaibert, Nick Thorpe and Richard P. Evershed. 2009. The earliest horse harnessing and milking. *Science* 323. 1332–1335.

Paasonen, Heikki. 1990-1999. *H. Paasonens Mordwinisches Wörterbuch* I–VI. Zusammengestellt von Kaino Heikkilä. Bearbeitet und herausgegeben von Martti Kahla. Lexica Societatis Fenno-Ugricae XXIII. Helsinki: Suomalais-Ugrilainen Seura.
Padilla-Iglesias, Cecilia, Erik Gjesfjeld and Ljucio Vinicius. 2020. Geographical and social isolation drive the evolution of Austronesian languages. *PLoS ONE* 15(12). e0243171. https://doi.org/10.1371/journal.pone.0243171.
Parpola, Asko. 2012. Formation of the Indo-European and Uralic (Finno-Ugric) language families in the light of archaeology: Revised and integrated "total" correlations. Riho Grünthal and Petri Kallio, eds., *A linguistic map of prehistoric northern Europe* (SUST 266)*,* 119–184. Helsinki: Suomalais-Ugrilainen Seura.
Piha, Minerva. 2018. Combining Proto-Scandinavian loanword strata in South Saami with the Early Iron Age archaeological material of Jämtland and Dalarna, Sweden. *Finnisch-Ugrische Forschungen* 64. 118–233.
Pohlert, Thorsten. 2020. trend: Non-Parametric Trend Tests and Change-Point Detection. R package version 1.1.2. https://CRAN.R-project.org/package=trend.
R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
Rahkonen, Pauli. 2013. *Southeastern contact area of Finnic languages in the light of onomastics*. PhD dissertation, University of Helsinki.
Rasmussen, Simon, Allentoft, Morten E., Nielsen, Kasper, Orlando, Ludovic, Sikora, Martin, Sjögren, Karl-Göran, ... and Willerslev, Eske. 2015. Early divergent strains of *Yersinia pestis* in Eurasia 5000 years ago. *Cell* 163. 571-572.
Rédei, Károly. 1986-1989. *Uralisches etymologisches Wörterbuch*. Budapest: Akadémiai Kiadó.
Ross, Malcolm D. 1988. *Proto Oceanic and the Austronesian languages of western Melanesia*. Canberra: Pacific Linguistics.
Saarikivi, Janne. 2004a. Palaeo-European substratum interferences in the western branches of Uralic? *Suomalais-Ugrilaisen Seuran Aikakauskirja / Journal de la Société Finno-Ougrienne* 90.187–214.
Saarikivi, Janne. 2004b. Über das saamische Substratnamengut in Nordrussland und Finnland. *Finnisch-Ugrische Forschungen* 58(1–3). 187–214.
Saarikivi, Janne. 2006. *Substrata uralica: Studies on Finno-Ugrian substrate in northern Russian dialects.* PhD Dissertation, University of Helsinki.
Saarikivi, Janne. In press. In Bakró-Nagy, Marianne, Johanna Laakso and Elena Skribnik (eds.) *The Oxford Guide to the Uralic languages* Oxford: Oxford University Press.
Salminen, Tapani. 2001. The rise of the Finno-Ugric language family. In Christian Carpelan, Asko Parpola, and Petteri Koskikallio, eds., *Early contacts between Uralic and Indo-European: Linguistic and archaeological cosiderations,* 385–396. Helsinki: Suomalais-Ugrilainen Seura.
Salminen, Tapani 2002: Problems in the taxonomy of the Uralic languages in the light of modern comparative studies. A. E. Kibrik, ed., *Lingvističeskij bespredel: Sbornik statej k 70-letiju A. I. Kuznetsovoj*, 44-55. Moscow: Moscow University Press.
Sandøy, Helge. 2004. Types of society and language change in the Nordic countries. In Britt-Louise Gunnarsson, Lena Bergström, Gerd Eklund, Staffan Fridell, Lise H. Hansen, Angela Karstadt, Bengt Nordberg, Eva Sundgrenand and Mats Thelander (eds.), *Language Variation in Europe:* Papers from the Second International Conference on Language Variation in Europe, ICLaVE 2, 53–76. Uppsala: Uppsala Universitet.
Seržant, Ilja A. 2022. The Circum-Baltic area. In Jan Fellerer and Neil Bermel (eds.), *The Slavonic languages*. Oxford: Oxford University Press, in press.
Shagal, Ksenia, Max Wahlström and Johanna Nichols. 2019. (Non)finiteness in clause combining: A typological survey. Presented at ALT 13, University of Pavia.
Sinnemäki, Kaius. 2014. A typological perspective on differential object marking. *Linguistics* 52(2). 281–313.

Smirnov, O. V. 2012. Substratnaja mansijskaja toponimija na territorii bylovo proživanija mansi. [Substratal Mansi toponymy on former Mansi lands.] *Voprosy onomastiki* 2012(1). 43-78.
Smirnov, O. V. 2013. K voprosu o permskoj toponimičeskom substrate v territorii Marij El i v bassejne srednego tečenija reki Vjatka (v svete ètničeskoj interpretacii arxeologičeskix kul'tur). [A Permic toponymic substratum in Mari El on the middle Vjatka (in the light of ethnic interpretation of archaeological cultures).] *Voprosy onomastiki* 2013(2). 7-59.
Sokal, Robert R. 1988. Genetic, geographic, and linguistic distances in Europe. *PNAS* 85. 1722–1726.
Syrjänen, Kai, Terhi Honkola, Kalle Korhonen, Jyri Lehtinen, Outi Vesakoski and Niklas Wahlberg. 2013. Shedding more light on language classification using basic vocabularies and phylogenetic methods: A case study of Uralic. *Diachronica* 30(3). 323–352.
Toporov, Vladimir N. and Oleg N. Trubačev 1962. *Lingvističeskij analiz gidronimov verxnego Podneprov'ja*. [Linguistic analysis of hydronyms of the upper Dnepr basin.] Moscow: AN.
Trudgill, Peter. 2011. *Sociolinguistic typology: Social determinants of linguistic structure and complexity.* Oxford: Oxford University Press.
Veness, Chris. 2002-2020. Latitude/longitude spherical geodesy tools. Available at the Movable Type Scripts website at http://www.movable-type.co.uk/scripts/latlong.html.
Viitso, Tiit-Rein. 2012. Early metallurgy in language: The history of metal names in Finnic. In In Grünthal, Riho and Petri Kallio, eds., *A linguistic map of prehistoric northern Europe* (SUST 266)*,* 185–200. Helsinki: Suomalais-Ugrilainen Seura.
Warnow, Tandy. 2018. *Computational phylogenetics: An introduction to designing methods for phylogeny estimation*. Cambridge: Cambridge University Press.
Ylikoski, Jussi. 2016. The origins of the Western Uralic *s*-cases revisited: Historiographical, functional-typological, and Samoyedic perspectives. *Finnisch-Ugrische Forschungen* 63.6–78.
Zajceva, M. I. 1981. *Grammatika vepsskogo jazyka: Fonetika i morfologija.* [Veps grammar: Phonetics and morphology.] Leningrad: Nauka.
Zajceva. M. I. 1993. Vepsskij jazyk. [Veps.] Viktorija N Jarceva, Jurij S. Eliseev, and Klara E. Majtinskaja (eds.), *Jazyki mira: Ural'skie jazyki,* 36–48. Moscow: Nauka.
Zhivlov, Mixail A. 2014. Studies in Uralic vocalism III. *Journal of Language Relationship* 12. 113–148.
Zhivlov, Mixail A. 2018. Istoričeskaja fonetika i vnutrennjaja klassifikacija ural'skix jazykov. [Historical phonetics and Uralic internal classification.] Presented at XIII tradicionnye čtenija pamjati S. A. Starostina, 22–23.3.2018. Moscow: Institut vostočnyx kul'tur i antičnosti RGGU.