

Drastic demographic events triggered the Uralic spread

To appear in *Diachronica*

Riho Grünthal (University of Helsinki)

Volker Heyd (University of Helsinki)

Sampsa Holopainen (University of Helsinki)

Juha Janhunen (University of Helsinki)

Olesya Khanina (Institute of Linguistics, Russian Academy of Sciences)

Matti Miestamo (University of Helsinki)

Johanna Nichols (University of California, Berkeley; University of Helsinki; Higher School of Economics, Moscow)

Janne Saarikivi (University of Helsinki)

Kaius Sinnemäki (University of Helsinki)

Abstract:

The widespread Uralic family offers several advantages for tracing prehistory: a firm absolute chronological anchor point in an ancient contact episode with well-dated Indo-Iranian; other points of intersection or diagnostic non-intersection with early Indo-European (the Late PIE-speaking Yamnaya culture of the western steppe, the Afanasievo culture of the upper Yenisei, and the Fatyanovo culture of the middle Volga); lexical and morphological reconstruction sufficient to establish critical absences of sharings and contacts. We add information on climate, linguistic geography, typology, and cognate frequency distributions to reconstruct the Uralic origin and spread. We argue that the Uralic homeland was east of the Urals and initially out of contact with Indo-European. The spread was rapid and without widespread shared substratal effects. We reconstruct its cause as the interconnected reactions of early Uralic and Indo-European populations to a catastrophic climate change episode and interregionalization opportunities which advantaged riverine hunter-fishers over herders.

Keywords Uralic; Finno-Ugric; Indo-European; Yamnaya; Indo-Iranian; Siberia; Eurasia; Seima-Turbino, 4.2 ka event; linguistic homeland

Note: In-text references have not yet been fully deanonymized.

Drastic demographic events triggered the Uralic spread

(Contents, for convenience)

Main text (pp. 3-22)

- 1 Introduction
- 2 Prehistory and timeline
 - 2.1. The PU homeland
 - 2.2. Linguistic and extralinguistic chronology
 - 2.3. Early Uralic stages
 - 2.4. Timeline
3. The CU spreads
 - 3.1 Northward spreads
 - 3.2. The CU east-west dispersal
 - 3.3. Typology and the CU spread
 - 3.4. Early diversification in the east
 - 3.5. Genetic evidence
4. Sociolinguistics of post-catastrophe spreads
5. Conclusions
 - Summary of main points
 - Remaining questions

References (for both main text and supplements) (pp. 23-33)

Supplements (following main text; separately numbered pp. S1-S29)	Text reference:
S1 The Uralic family tree	(§1)
S2 Linguistic paleontology	(§2.2; 3.2.1fn)
S3 Pre-Uralic IE languages and the PU homeland	(§2.3)
S4 More on isolation by distance <i>R script for Supplement S4 – not included</i>	(§3.2.5, §3.2.1)
S5 The Finnic and Saamic secondary westward spreads	(§3.2.2)
S6 Post-CU diffusions with IBD effects	(§3.2.6)
S7 Typology and the CU spread	(§3.3)
S8 Early diversification in the east	(§3.4)
S9 Sociolinguistics of post-catastrophe spreads	(§4)

Numbering of tables in supplements: e.g. Table S2.3, Figure S4.1

(S = supplement, 2 = supplement number, 3 = table number within that supplement)

Appendices:

https://www.dropbox.com/s/26ut4rsmvxxm43/UralicSpread_appendixes%2Bsources2.xlsx?dl=0

1. Introduction

The Uralic language family, probably about 4500 years old, comprises over 30 languages which at one time or another have extended from the Yenisei River basin in western Siberia to Norway and from near the middle Volga to the Arctic Ocean. There was a window in time, from late medieval to early modern times, when this large region was populated almost exclusively by Uralic languages -- when Uralic northward expansion was absorbing the remaining non-Uralic languages in the far north just as Swedish, Russian, and Siberian Turkic were beginning their expansions in the west and south.¹ At present, except for the three national languages Finnish, Estonian, and Hungarian, most of the languages are those of minorities scattered through their former ranges. Most are endangered, and some are extinct. (A map of the modern ranges is: <https://bedlan.net/uralic/> A map of historical ranges is Grünthal & Salminen 1993.)

The Uralic family is divided into nine elementary branches (Table 1), all between about 2500 and 1000 years old internally (i.e. since their own divergence into individual languages), while their external ages (since the initial Uralic split) are all about 4000-4500 years.² (Here and below, in tables and prose we list languages from east to west, following the standing directionalities of language spread in northern Eurasia.) See Supplement 1 for tree diagrams and a full list of languages.³ Any valid family tree must distinguish these elementary branches from each other; what higher-level structure exists is debated (and discussed below and in Supplement S1). For the branch historical homelands see Map 1.

Table 1. The Uralic daughter branches, in approximately east to west order of branch homelands. Brackets = likely or known higher-level groupings. 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.

Samoyedic			
Hungarian			
Mansi	}	Ob Ugric	}
Khanty			
Permic			
Mari			
Mordvin	}	Western Uralic	
Finnic			
Saamic			

¹ Even then, Turkic expansions, chiefly Volga and Siberian Tatar, had absorbed Uralic speakers south of the Volga and in the southern Urals.

² This time depth makes them all genera by the definition of Dryer 1989 and 2013. The age is supported below.

³ Online supplements contain additional graphs, tables, data, and details.



Map 1. Linguistic and archaeological distributions c. 4200-4000 BP. Light shaded band across the entire area: recent and modern forest zone (steppe to south, tundra to north). Languages and cultures in chronological order: Fatyanovo, Balanovo, Abashevo: Pre-Baltic and Para-Baltic-speaking post-Corded Ware IE cultures. Poltavka, Sintashta: Indo-Iranian-speaking. Shapes with upward coarse parallel hatching: Seima-Turbino major sites. Labeled ovals (finer hatching): core locations of Uralic branch ancestors: Saa(mi), Fin(nic), M(or)d(vin), Ma(ri), Pe(rmic), Hun(garian), Man(si), Kh(anty), Sam(oyedic). Fin2 = later staging ground for Finnic. Heavy line along rivers: southern trade/travel route. Heavy dotted lines: northern route. Light, short dotted line: combined water and overland route across the Urals. (After Saarikivi in press., Lang 2018, Chernykh 2008.)

The traditional view of the family tree posits an initial bifurcation into Samoyedic vs. Finno-Ugric, a clade comprising the rest of the branches. The bifurcation is based primarily on the fact that Samoyedic has far fewer cognates with other branches than any two Finno-Ugric branches have with each other, which suggests a longer separate development of Samoyedic from the rest. The low cognate count is then due to losses over time. After that initial split, successive bifurcations within Finno-Ugric produced a west-branching family tree with Finnic and Saamic the last to separate from each other. Since, however, no or almost no shared innovative sound changes define Finno-Ugric or its subsequent bifurcations, a recent proposal sees the family as a star phylogeny or rake consisting of nine coequal branches. (For trees see Supplement S1.) Here and below, we speak of Finno-Ugric as the set of non-Samoyedic branches, without commitment as to whether it is a clade; and likewise for Ugric and Ob-Ugric.

Whether the tree is hierarchical or flat (rake) has implications for what cognate sets are accepted for PU. If PU first split into Samoyedic vs. Finno-Ugric, cognates accepted as PU must be attested in both branches, and that necessarily means in Samoyedic and at least one Finno-Ugric branch. This gives Samoyedic veto power over what is considered PU and no doubt produces false negatives (especially given the low number of cognates in Samoyedic). If the tree is flat, however, no one branch has veto power, and any reasonably diverse and dense attestation suffices; then there will be false positives. Recent and ongoing work by Ante Aikio (2013, 2014ab, 2015ab, in press) uses a criterion of sufficiently diverse attestation. On the traditional definition there are about 200 firm PU etyma (see Appendix 1); Aikio has from 500 to 700.

The large Uralic range is remarkable for a language family traditionally thought to have expanded from a Neolithic hunter-gatherer population.⁴ Other large spreads have occurred in northern Eurasia, but they have involved advances in food production (reindeer herding in the north), connections to Chinese imperial expansion (see Barfield 1989, Janhunen 2008, 2012), or, in the case of Indo-European, the advances in mobility, technology, economy, and network-extension mechanisms that impelled the Yamnaya culture across the western steppe (Anthony 2007:300-339, Parpola 2012, Anthony & Ringe 2015:208).

Even more remarkable is the absence of unambiguous early Indo-European (henceforth IE) loanwords in PU. The sweep of the Yamnaya culture, which spread late Proto-Indo-European (PIE) speech, from c. 5100 BP,^{5, 6} was a major economic, cultural, and demographic event that brought wheeled transport, domesticated horses, pastoralism, new forms of social organization, and new forms of exchange and wealth to the entire western steppe and nearby.⁷

⁴ Ancestral Uralic society is sometimes described as Mesolithic. The technology was Neolithic in that it included pottery, but there was no food production. Consistent with the archaeological evidence, the reconstructable Proto-Uralic vocabulary included a word for 'pot' but no terms for domesticates or foods prepared from them.

⁵ BP = before present, i.e. years ago. These are estimates in calendar years. All are approximate and expressed in rounded numbers. Our sources express prehistoric and protohistoric dates variously as BP, BC/AD, BCE/CE, BZ/AZ, or parts of millennia (e.g. last quarter of third millennium). We convert them all to BP. This is consistent with historical-linguistics practice, where language family ages are estimated in years ago. Dates not accompanied by a before/after abbreviation are in the present era (we use such dates only for historical times).

⁶ Or 5300, counting the predecessor Repin culture. The large spread begins with Yamnaya c. 5100 BP.

⁷ Geographical terms (all approximate): western steppe = steppe from eastern Romania to the Urals; its subparts: Pontic steppe (west; southeastern Europe, Ukraine, and Russia), Caspian steppe (east; lower Volga to Urals).

In its wake, the PIE language spread to central Europe and from there, as part of the wider Corded Ware culture complex, back eastward along the Volga to the Urals, where the cultural successors of the IE-speaking Fatyanovo culture mined and worked copper.⁸ Also remarkable is the tenacity of Uralic-speaking populations along the Volga. Speakers of the ancestral major branches of Uralic succeeded IE to occupy good farming and herding land along the major trade arteries (the Volga, Oka, Kama) and were never dislodged until the Russian expansion,⁹ while major language spreads and extinctions have taken place both north and south. One would expect to find at least remnants of IE speech along the Volga and to the north, accompanied by aspects of IE social organization, as local hunting, trapping, and fishing societies shifted to the language of wealth and power. Instead we have Uralic languages from the Volga north, with only recent Russian and Turkic overlays. Here we present a model of the Uralic spread that accounts for these things.

2. Prehistory and timeline

2.1. The PU homeland

Probably the majority view among Uralicists through the 20th century, and still widely taken as received view by archaeologists, is that the Proto-Uralic homeland was on or near the middle Volga and Volga bend. There are four current arguments for a PU homeland in that region:

(1) The region of highest phylogenetic divergence, or center of gravity: the projection on the ground of the highest-level divergence in the family tree (Sapir 1916, Dyen 1956). This criterion does not always obtain, but the success rate is good for directional spreads such as those of Austronesian (on the traditional view: Blust 1985, 2009) and Algonquian (Goddard 1996). Highest phylogenetic divergence requires a family tree with an initial split like that of the traditional Uralic tree. For a phylogeny like the Uralic rake proposal there is no single highest divergence.

(2) The region of greatest diversity, i.e. the greatest number of branches clustered together. In the Uralic literature that is usually identified as the Volga bend area, where Mordvin, Mari, and Permic are in close proximity.

Western Siberia = Urals to Yenisei. Lower Volga = the north-south extent in the Caspian steppe up to the Kama confluence (Astrakhan to Kazan); middle Volga = the east-west extent in the forest-steppe region (from the big bend at Kazan' to the Oka confluence at Nizhnii Novgorod; also the lower Oka and lower Kama); upper Volga = the more tortuous course from Nizhnii Novgorod east (and via portages to the Baltic).

⁸ Here and below, when we speak of a prehistoric population as "Uralic-speaking", "IE-speaking", "pre-Saamic-speaking", etc. we do not imply that the population was monolingual or that we know its speech situation with certainty. Rather, the language was the dominant one as determined by best methods (and usually justified here).

⁹ Though the rise of Volga Bulgar power (650-900) and the Tatar Golden Horde (1200-1500) pushed the southern Uralic frontier northward along the middle Volga. East of the Urals, the original Ugric range may have been in the south of the forest zone and not far from the steppe, as suggested by three pieces of evidence: early loans in Ugric languages dealing with horse culture and nomadic life; the persistence of the Indo-Iranian word for 'honeybee' in Hungarian, the only eastern Uralic language whose earliest known location was in honeybee range; and non-Ugric substratal toponyms in much of the Ob-Ugric territory (§§2.1, 3.2.7, and Supplements S2 and S3.2). If that is correct, the earliest Ob-Ugric core areas have not persisted but have shifted to Turkic, chiefly Siberian Tatar.

(3) About a dozen resemblant proto-roots found in PU and PIE, regarded variously as loans (Koivulehto 2001) or cognates (Helimski 2001b), are taken as evidence for proximity of PIE and PU.¹⁰

(4) The word for 'honeybee' (*Apis mellifera*), an Indo-Iranian borrowing, points to an origin west of the Urals; in Siberia the natural range of the honeybee does not extend to the forest zone.

The linguistic arguments against the middle Volga homeland are stronger than the arguments for it:

(1) The center of gravity for the traditional tree is in western Siberia where Samoyedic approaches or contacts the nearest Finno-Ugric branch (represented by Khanty), well to the east of the Urals. A relevant consideration even on the rake model is that diversification and geographical distancing arguably began earlier in western Siberia (§3.4).

(2) The criterion equating greatest diversity with greatest number of branches in the area has to our knowledge never been shown to reliably identify homelands of language families; perhaps it comes from a misunderstanding of the technical sense of highest phylogenetic diversity, which considers only primary branches. On grounds of sheer number of branches the western Middle Volga area where ancestral Finnic and Saamic were in proximity to Mordvin would appear to be an equally good candidate, as would northwestern Siberia where Khanty, Komi, and Forest Nenets are in proximity.¹¹ In fact those regions and the middle Volga are what are variously known as residual zones, accretion zones, or language sinks (Nichols 1992, 1997, Nichols and Rhodes 2018): areas where languages move in more readily than out and diversity increases over time.¹² Language sinks are unlikely proto-homelands.

(3) Despite careful attention by both Koivulehto and Helimski to systematicity in correspondences and PIE forms, the set of resemblant roots is not large enough to exceed chance, and the relatively wide ranges of both form and meaning reduce the probative value of the individual pieces (Nichols 2010, Nichols and Rhodes 2018, Simon 2020).¹³

¹⁰ Both the possibility of a much earlier PIE-PU genealogical relatedness and the putative loanwords are invoked by Anthony and Ringe 2015:206-207 (citing Koivulehto), and somewhat similarly Anthony 2007:96-97, Mallory 1989:149, as evidence for a PIE origin not far from the Volga bend, near where they judged the linguistic literature to place the PU homeland.

¹¹ An extreme example is the Austronesian family, of which all but one of the primary branches are found on Taiwan while the dozens or more branches of the other primary branch, Malayo-Polynesian, stretch across island Southeast Asia, Melanesia, Micronesia, and Polynesia (Blust 2009; the number of primary branches recognized varies).

¹² The Minusinsk basin and middle-upper Yenisei area in south central Siberia, in the likely Samoyedic homeland region, is a high-diversity language sink where different Samoyedic branches have neighbored with Yeniseian, Turkic, and Tungusic languages over time; and it is also the starting point for major northward spreads down the Taz and Yenisei, today illustrated by northern Samoyedic languages (Janhunen 2012a, Khanina et al. 2018). The western Middle Volga area hosts diversity and was the starting point for major northward spreads by Saamic and then northern Finnic. It must be that northward spreads are conditioned by ecological and economic factors that are orthogonal to those that create language sinks.

¹³ Of what we take to be the two statistically soundest recent quantitative tests, Kessler and Lehtonen 2006 (using a 100-item Swadesh-like wordlist) found no evidence for Indo-Uralic, and Kassian et al. 2015 (using a shorter wordlist) found evidence (but see the questions on their data and method in Kallio 2015, Kessler 2015, and Ringe 2015).

(4) The term for 'honeybee', like all I-I loans, was borrowed not into PU but into post-dispersal early Uralic; it is irrelevant for the PU homeland. It is borrowed into those branches whose branch homelands are in the honeybee range. (Supplement S2; Appendix 2.)

(5) The absence of loanwords in PU from the late PIE language of the Yamnaya spread or the pre-Balto-Slavic or Para-Baltic¹⁴ languages of the Fatyanovo culture or its descendents is telling (§1 above, §2.4, and Supplement S3). More generally, the reconstructed wordstock of PU points to a Neolithic technology and economy without food production (Janhunen 2020, 2009), which is unlikely for regions near the middle Volga or Volga bend as food-producing PIE and early IE-speaking groups had been a major cultural force in the nearby steppe and forest-steppe areas for nearly a thousand years before the PU dispersal.

(6) Other linguistic paleontology: A PU term is reconstructed for the Siberian stone pine or cedar pine (*Pinus cembra sibirica*), a food plant found only in Siberia. (Supplement S2, on this and other ecological terms; Appendix 1.) Importantly, PU had only a single undifferentiated term for 'metal', inconsistent with propinquity to the southern Urals, while the early post-dispersal stage we call Common Uralic (§2.3 below) used loans and derivation to create a more elaborate terminology consistent with involvement in the bronze trade (see Supplement S2).

(7) A Volga homeland makes the long Samoyedic movement to its branch homeland in the area around the middle Ob to middle Yenisei area implausible, and that movement bucks the generally east-to-west trajectory of north Eurasian language spreads (Janhunen 2014, 2012).

The conclusion is that there is no linguistic evidence in favor of a middle Volga homeland, or any homeland west of the Urals, and strong evidence in favor of a western Siberian homeland.

2.2. Linguistic and extralinguistic chronology

Early Uralic prehistory needs to be placed in the context of three temporally and geographically overlapping important events:

(a) *The 4.2 ka event* (as it is termed in the earth sciences), a global climate development from 4200 to 3900 BP that caused drought in the low and mid latitudes of continental interiors, wetter conditions in the northwest of Eurasia, and global cooling. Civilizations collapsed in the Near East and Southwest Asia and the Caspian and Kazakh steppes suffered extreme drought while the westernmost steppe and eastern Europe may have had more rainfall than usual. Temperatures were cooler overall. (Perşoiu et al. 2019, Helama and Oinonen 2019; Dalfes et al. 1997.)

(b) *The Seima-Turbino Transcultural Phenomenon* (henceforth ST), an archaeological complex marked by distinctive bronze artifacts, especially symbols of power such as spearheads and axe heads, found across many archaeological cultures and along major waterways from the Altai to Scandinavia. In particular, tin from Altai mines made possible large-scale bronze production, with the forging and casting done there as well as in the southeastern Ural area and

¹⁴ Here and below, *Pre-* labels any unspecified stage prior to *Proto-* (as that is reconstructed from attested daughter languages) and *Para-* labels an immediate sister to an existing branch. Thus Pre-Baltic is any stage prior to (depending on whether Baltic is regarded as a clade) Proto-Baltic or Proto-Balto-Slavic; and Para-Baltic is a sister to all of Baltic or Balto-Slavic (see also Supplement 3).

Southwest Asia. The most recent radiocarbon dates place ST between c. 4200 and 3900 BP, and somewhat later west of the Urals along the Volga (Marchenko et al. 2017, Krause et al. 2019, Chernykh 2008).

ST is evidently the archaeological signature of a waterborne trade network (Barfield 2009, Nichols and Rhodes 2018) that brought metal from Ural and Sayan mines and metal artifacts from Ural forgeries westward to Europe. Its heyday coincides rather closely with the 4.2 ka event. The trade network itself probably existed long before the Bronze Age and demonstrably continued into the middle ages as the Bulgar and Viking trade routes.

(c) The Indo-Iranian contact episode. Approximately 4000 years ago the Finno-Ugric branch ancestors absorbed vocabulary from the Indo-Iranian (henceforth I-I) branch of Indo-European (Holopainen 2019). I-I figured prominently in the development and spread of bronze technology from the vicinity of the southern Urals in this time frame. Proto-I-I is well reconstructed and dated on linguistic grounds, including very early stages of daughter languages that are well attested in writing, and the time and general location of its spread are well established archaeologically. The Poltavka archaeological culture and its Potapovka and Sintashta-Petrovka descendents (See Map 1) were culturally, and almost certainly linguistically, Indo-Iranian (Kuzmina 2007); this is one of the clearest linguistic signatures known to archaeology, short of inscriptional evidence, and its time frame needs to be stipulated as an absolute date (albeit an approximate one) in any chronology of Uralic.

The inventory of Indo-Iranian words differs from branch to branch in Uralic, and etyma come from time frames ranging from Pre-Proto-Indo-Iranian to Proto-Iranian to early Iranian. This shows that I-I interacted not with a single Proto-Uralic but with an incipiently differentiated early Uralic, over some extent of time and some extent of space. In view of this distribution, the I-I contact episode cannot be regarded as a single clade-defining event and therefore as establishing the reality of a unitary Finno-Ugric branch. What it does establish is the time of the initial Uralic divergence: it occurred before 4000 BP but not long before, as the I-I loans entered at the branch protolanguage level or not long thereafter, they enter the early Uralic branches separately, and the internal evolution of the daughter branches began after 4000 BP as shown by the application of branch-specific sound laws to the I-I material.

The Samoyedic branch lacks the I-I stratum almost entirely. This, together with its low number of cognates, may point to an early and fairly clean separation of Proto-Samoyedic from the rest of the family, as was widely assumed in 20th century Uralic studies. On the other hand, the retention in Samoyedic of much PU inflectional morphology and the regular phonological evolution of its surviving native vocabulary suggest that that separation did not precede the I-I episode by long. The spread of Finno-Ugric could have been simultaneous with the separation of Samoyedic or later; their different histories may be due to different directionalities and geographies as much as to different chronologies. The geography of a reconstructed PU homeland needs to provide for a clean break and exit of Proto-Samoyedic and a rapid spread of Finno-Ugric to bring it into I-I contact in a spatially differentiated but structurally homogeneous form.

The 4.2 ka event and ST are nearly simultaneous and the I-I episode occurred within that time range, and we suggest they are causally connected. During the drought herders saw their traditional subsistence falter and fail, and the populations of herds and probably herders were reduced by famine. There is evidence of overgrazing on the steppe in late Yamnaya times

(Anthony 2007:330), half a millennium earlier, so it must be assumed that by 4200 BP the entire grazable Pontic-Caspian steppe was inhabited to carrying capacity, with no leeway except what could be gained by warfare. To the east, the alternative was to seek water sources at the steppe periphery. Herders from the almost certainly Indo-Iranian-speaking Poltavka culture crowded into the river valleys around the Urals, where they established fixed year-round settlements (Anthony 2007:371-411, especially 389-391, describing permanent settlement as a strategy for claiming access among traditionally mobile societies facing diminishing resources). Competition and warfare were intense and militarization increased.

Around the southern Urals, IE-speaking societies found economic security in bronze production and bronze trade. The Poltavka- and Abashevo-derived Sintashta culture of the southeastern Urals developed bronze manufacture into a major cottage industry, and c. 3900 BP invented the chariot and developed chariot warfare, solidifying Iranian-speaking domination of the entire Caspian steppe (Lindner 2020). ST is the visible signature of this strategy: hoards, burials, and other finds along the routes of what had long been a waterborne trade network and which now quickly adapted and began to carry bronze from Ural and Altai mines. ST was a symptom of a broader process of interregionalization that brought expanded trade networks and an expanded inventory of trade goods to a widening range of markets (Frachetti 2008:47-67). Bronze-producing societies in the southern Ural region established trade connections as far afield as Southwest Asian cities (Anthony 2007:389-393, 418-427).

A background condition is the presence of endemic bubonic plague (*Yersinia pestis*) and salmonella (*Salmonella enterica*) on and near the steppe (Rascovan et al. 2019; Andrades et al. 2016, Rasmussen et al. 2015; Key et al. 2020). Both of these bacteria had undergone major genealogical diversifications beginning in the Neolithic, when the denser village and urban populations and proximity to livestock favored their transmission. Bubonic plague became virulent, that is, transmissible by fleas rather than requiring direct contact, somewhat later but still probably by 5000 years ago (Spyrou et al. 2018). Both diseases would have presented particular risks in the denser conditions and year-round presence of livestock that accompanied the drought. Mobile and less dense hunter-fisher populations would have been much less affected.

In the model proposed here, early Uralic speech spread rapidly along the waterborne trade network north of the steppe, expanding as part of the same interregionalization as ST. Uralic speakers were the prospectors, miners, boatsmen, trade managers, procurers, and first settlers of trading posts at major river confluences; the Indo-Iranian-speaking Sintashta culture and its successors financed trade and prospecting and developed markets. Uralic-speaking trading post settlements became well entrenched and demographically strong along the trade routes before the pastoral steppe populations recovered from the drought, allowing Uralic-speaking populations to dominate the forest-steppe and forest zones thereafter.

2.3. Early Uralic stages

Some previous work has suggested that PU was the language of ST, but it is essential to distinguish PU, dated to about 4500 BP, from the language of 4200-3900 BP when IE retreated and Uralic speech spread. We define PU as the linguistic system that can be reconstructed by applying the comparative method to the Uralic daughter languages, and the unified speech community that spoke it. The divergence of PU into incipient daughter branches, and

application of changes only branch-internally, brought about the end of PU and the beginning of the separate evolution of the daughter branches. We use the term Common Uralic (CU) to refer to the speech community and the language from the time of the initial divergence to the point, some centuries later, when the daughter dialects ceased to be mutually intelligible. PU began diverging before ST arose, and it had only a minimal terminology for metals, probably only a single term for 'copper', the metal that was mined and worked near the Urals prior to the ST phenomenon. CU was involved in ST and appears to have had a richer terminology for the important ST trade items and materials (Supplement S2).

2.4. Timeline

A timeline for Uralic-relevant ethnolinguistic events along the upper and middle Volga, c. 6000-3000 BP, is as follows (Anthony 2007, Lang 2018, Nordqvist and Heyd 2020; see Supplement S3).

(a) c. 6000-3500(?) BP: The Volosovo culture of settled hunting-fishing societies is found along the entire Oka-Volga-lower Kama and well to the north. Several of its cultural features persist in ethnographically documented societies of the north European forest zone, but they are not specific enough to suggest any particular language identity. A territory that large must have contained a number of different societies and languages and probably from more than one language family. Much of the Volosovo range has been Finno-Ugric-speaking in historical times, but this is due to parallel northward spreads (§3.1) from different branch centers rather than continuing an ancient ethnolinguistic unity.

(b) 5100 BP: The Yamnaya culture, almost certainly late PIE-speaking,¹⁵ spreads rapidly across the Caspian and Pontic steppes. It adapted wheeled transport and expanded by a mix of language shift, migration, and population growth. Contemporaneous with the Yamnaya spread, the Afanasievo culture, Yamnaya-like and likely speaking an early IE language, appeared in the Minusinsk basin (upper Yenisei) and across the Altai foothills.

Neither Yamnaya nor Afanasievo has left any detectable loan vocabulary in PU.

(c) Middle Bronze Age, 4800-3900 BP: The Fatyanovo, Balanovo, and Abashevo cultures, ultimately Yamnaya descendents via the central European Corded Ware culture, spread east from the Western Dvina/Daugava and middle Dnieper along the forest-steppe belt. Fatyanovo (4800-4200) extended mostly along the upper and middle Volga (and its tributaries) and in the west ultimately far north to Lake Ladoga; Balanovo was a Fatyanovo extension east along the middle Volga and to the Kama confluence; Abashevo, a Fatyanovo descendent (4200-3900),

¹⁵ More precisely, **spreading** late PIE if not necessarily **speaking** it as home language. Before the Yamnaya spread, more than one language must have been spoken by the mostly herding cultures of the steppe and steppe periphery; the range is too large for a single pedestrian language. The rapid Yamnaya spread indicates language shift, so pre-Yamnaya cultures must have adopted Yamnaya customs and economy and begun to enter the Yamnaya social networks before becoming primarily IE-speaking. However it happened, Yamnaya spread only one language – late PIE – across the steppe and into southeastern and central Europe. The only surviving trace of other languages is terms for crop plants and a few others acquired from the farming cultures of eastern Europe (Iversen & Kroonen 2018).

extended along the Volga Bend (and south of the Volga to the middle Don catchment) and the lower Kama and the Belaya to the southern Urals. These were farming and herding societies (Abashevo mostly herding) which made and used metal implements. Fatyanovo was very likely Pre-Baltic-speaking (and almost certainly so in its western range). There is no trace of contact with any of these languages in the PU lexicon.

Midway through this phase, PU began to diverge.

(d) 4200-3900 BP: The 4.2 ka event and S-T; late in this period, the I-I episode; see §2.2.

The I-I episode was the first evidence of early Uralic contact with any IE language. By this time the CU varieties had begun to diverge and separate, and they borrowed I-I words separately.

(e) 3900 BP and later: Rebound. With climate amelioration, agriculture recovers across the northern steppe periphery. Indo-Iranian languages come to be spoken across the entire Pontic-Caspian steppe and into the Kazakh steppe. The center of bronze production and trade shifts west and the ST complex comes to an end, perhaps because the new opportunities presented by recolonization of lands abandoned during the drought made warfare and displays of power less necessary. Bronze artifacts exported to eastern Europe and Scandinavia now originate chiefly along the western Urals and near the Volga bend.

The ethnolinguistic composition of the middle Volga and southwestern Ural areas has changed: formerly Indo-European and probably Pre-Balto-Slavic- or Para-Baltic-speaking prior to the 4.2 ka event, it became durably Finno-Ugric-speaking after that (Supplement S3) and remained so until marginalized in the Russian colonial expansion.

3. The CU spreads

3.1 Northward spreads

The large northern extent of the attested Uralic range is secondary, reached after the initial Uralic dispersal, as dialects and daughter languages of Uralic languages spread northward from the branch ancestor¹⁶ homelands located in the southern part of the early Uralic range (Map 1; s.a. Saarikivi 2020), eventually to reach the Arctic Ocean coast. (Such northward spreads are a recurrent pattern in northern Eurasia, though not in North America: Nichols and Rhodes 2018.)

These northward spreads are useful for comparison to the primary CU spread, as they give us the geographically closest and most comparable known cases of hunter-gatherer spreads. The process may have begun early but proceeded slowly enough that in several cases the far north was Uralicized only in recent centuries (Aikio 2012, Helimski 2001a). The present-day northernmost languages, the Saamic and Samoyedic branches, exhibit exotic substratal

¹⁶ We use the term *branch ancestor* to refer to the protolanguages of the nine branches (Table 1) and their presumed speech communities. (Saarikivi in press) Branch ancestor homelands, or branch homelands, are the places from which the branches have expanded and spread northward. Sometimes the homelands are supported by archaeological and/or toponymic evidence (strongest for Saamic and Finnic: Aikio 2012, Lang 2018, Saarikivi & Lavento 2012, Saarikivi in press). The branch homelands are likely to have been settled in the initial spread, with the exception of Saamic and Finnic, whose points of dispersal are known to have been secondary.

vocabulary and/or grammatical and phonological effects, and in most places local northern toponymy and vocabulary for tundra flora and fauna include words of non-Uralic origin, showing that today's northernmost languages were the frontier languages in the spread (Aikio 2012, Helimski 2001a, Saarikivi 2006, in press) and that not only PU and CU speakers but also branch ancestor speakers were unfamiliar with the tundra ecology and needed to borrow terms for it. The rate of northward spread appears to have been accelerated by the emergence of reindeer herding in the last two millennia and (possibly, though debated) the development of industrial-scale fur trades in which Uralic speakers were trappers and procurers for Scandinavian and Russian traders (Aikio 2012, Helimski 2001a). Since resources in tundra landscapes are sparse and patchy, survival there requires larger ranges per capita, larger-scale mobility, and spatially more extensive social and economic networks, and correspondingly the ranges of the northernmost languages have spread out widely in this environment.

The branch homelands were in forest lands, and the northward spreads eventually moved into tundra ecologies. These northernmost phases are quite likely to have involved primarily language shift, as sparsely distributed northern hunter-gatherers shifted to the languages of the denser and more economically advanced populations to the south. (Unpressured shift from the language of a sparse foraging society to the language of a denser food-producing society, especially if the latter is also technologically advanced, is a common though not universal development where foraging and food-producing economies are in contact: Güldemann et al. 2020:30-32.) When tundra populations adopted reindeer herding and/or became procurers in the European fur trade, individual languages became influential and spread widely (North Saami and especially Komi and Tundra Nenets).

Judging from the reconstructed and partly attested recent history of Saamic and Samoyedic groups (Aikio 2012, Khanina et al. 2018), northward spread proceeded unevenly, responding to fluctuations in climate, economic and demographic pressures, technological advances, and other factors. The northward spread of the Samoyedic branch proceeded in spurts, most probably driven by advances in reindeer herding (Khanina and Gusev in preparation). In the more recent northward spreads of Finnic and Permic, important factors were the adoption of agriculture and stockbreeding and their concomitant population increases. (The scale of the agriculture, which was slash-and-burn or floodmeadow, was small, and the livestock often amounted to two or three cows per household for dairy products. Fishing and hunting were important in the diet. This economic scale persisted in rural households in parts of central and northern Finland well into the 20th century.)

Northward spread likely involved a mix of small-group migration and language shift, and was chaotic locally and in the short term but northward overall. The process gave rise to a modern stratigraphy in which the more southerly language or languages lack recent substratal effects and have more compact ranges while the northern languages have recent substratal effects, especially terminology for arctic phenomena, and larger ranges. In their overall gradual pace, substratal effects, and expansion primarily from the frontier, these northward spreads are different from what can be reconstructed for the initial Uralic spread.

3.2. The CU east-west dispersal

In contrast to the northward spreads, the initial Uralic spread was almost entirely east-west in directionality, with daughter branches taking root along most of the east-west extent of

the Volga and probably along the middle and upper Tobol, Irtysh, and Ob and the upper Yenisei (Map 1). This spread appears to have been rapid, largely without substratum, and with minimal evidence of frontier expansion and isolation by distance. It was westward overall; only the Samoyedic branch probably did not take part in the westward spread, and may have moved eastward. Evidence in favor of these claims includes the following points.

3.2.1. Minimal isolation by distance (IBD) effects in vocabulary. Isolation by distance (IBD; also called autocorrelation) refers to the general phenomenon of finding decreasing numbers of shared traits as geographic distance between related populations increases (Holman et al. 2007; Haynie 2014: 344–345, 349–350). Strong IBD effects suggest a slow and steady movement away from the center of expansion owing to ongoing exchange between neighboring populations. Minimal IBD effects, on the contrary, suggest the spread from the center of expansion was rapid. In historical linguistics IBD has been applied to modeling language history in several continents, for instance, to modeling the Bantu expansion (de Filippo et al. 2012), modeling linguistic diversity of Japonic languages (Huisman et al. 2019), and evaluating the language history in Melanesia (Hunley et al. 2008; see Supplement S4 for more information).

To the extent that Uralic branch ancestors spread slowly we should thus find monotonically increasing linguistic distance from language to language as their geographic distance from the center of expansion increases. However, we do not find this kind of regular IBD effects. Figure 1 shows the numbers of Proto-Uralic reconstructed etyma retained in each daughter language.^{17,18} It also shows the numbers of early Indo-Iranian (I-I) loans that entered the early Uralic branch ancestors c. 4000 BP (§2.2 above; source for the items: Holopainen 2019). The I-I stratum was borrowed early enough that the items reconstruct to the protolanguage of each Uralic branch, so that stratum should have undergone IBD attrition at rates similar to those of native Uralic vocabulary. IBD effects should show up as a more or less monotonic dropoff from an origin point; the dropoff should be unidirectional if the origin point was near the present edge of the range, or bidirectional if it was in the center or of the range.

¹⁷ Used here are PU cognates from the major languages and varieties for which lexical documentation is adequate for meaningful comparison, and I-I loans for those same languages.

¹⁸ Holopainen 2019 and Appendixes 1-3 here.

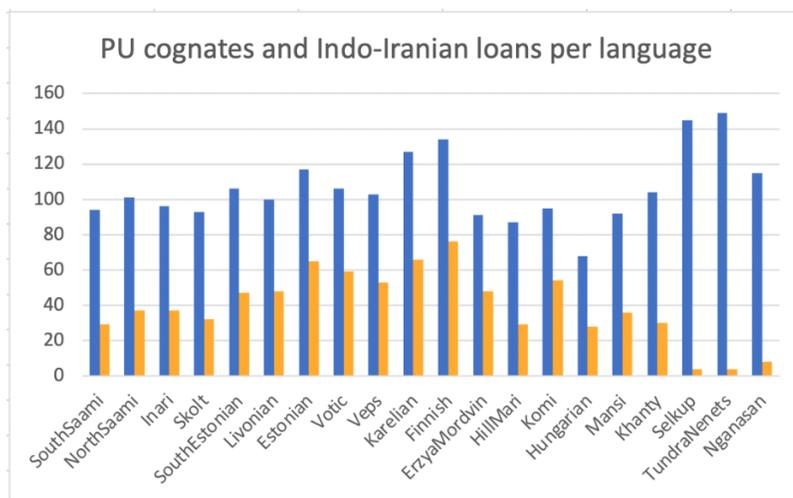


Figure 1. Numbers of Uralic (blue) and Indo-Iranian (orange) etyma retained per language.

Languages in Figure 1 are shown in geographical ordering, west to east (by branch) and then south to north (within branches). There is no evidence of IBD effects (this is also true if, following traditional ordering, Finnic is listed before Saamic and Hungarian before Khanty). (For statistical tests of monotonicity on languages and branches see Supplement S4.) If anything, retention rates are highest at the far west of the Uralic range, a highly unlikely origin point for the initial spread; these figures undoubtedly reflect peripheral archaism in the case of PU vocabulary, early isolation of branches in the east (§3.4), and a concentration of I-I contacts west of the Urals (so that the Ugric languages have few I-I words and the Samoyedic ones next to none).¹⁹ They may also reflect the larger numbers of daughter languages in the Saamic and Finnic branches, as that raises the probability that a PU or I-I item will be attested in the branch.²⁰ For other factors see Supplement S4.

PU words are identified, following the traditional phylogeny with its initial split of Samoyedic from the rest (§1 and Supplement S1), as those having a reflex in Samoyedic and one other branch, as well as regular sound correspondences. There are about 200 such words (Appendix 1), each with a reflex in Samoyedic by definition and a more scattered distribution over the other branches. Since they are always present in one or more Samoyedic languages, the Samoyedic languages have artifactually high frequencies.

3.2.2. Branch homelands and attested or reconstructed core areas are in the southern parts of the current branch ranges, along the Volga or nearby (Map 1). These homelands are

¹⁹ A regular decrease from west to east, if one exists, should be visible within Finno-Ugric if all of the reconstructable Finno-Ugric vocabulary could be surveyed. A definitive compilation of currently accepted Finno-Ugric etymologies does not yet exist, but a good approximation should be extractable from the ongoing Proto-Uralic project of Ante Aikio when that is complete (see Aikio 2013, 2014ab, 2015).

²⁰ Samoyedic also has a fairly large number of daughter languages, but several of them went extinct before they could get good lexical documentation.

identified on criteria such as lexical comparison, toponymy, archaeological evidence, and early historical evidence, and they represent the general area from which the branch later spread to its larger historical range (Saarikivi in press). Branch homelands are clearest for Mordvin and Permian; Mari was somewhere in the vicinity. For the Ob-Ugric languages they are uncertain, but probably in the southern parts of the historical ranges, and for Hungarian somewhat south of Ob-Ugric in Western Siberia. For Finnic and Saamic the centers of later spread are secondary locations, reached by westward spreads from nearer to the middle Volga (Supplement S5).

3.2.3. Each branch has undergone a number of separate innovations since splitting from the rest; in some cases there are as many as 20 or 30 such branch-specific innovations prior to its internal breakup or even more, and at least as many since then in the individual daughter languages (Saarikivi in press). This points to longer periods of independent evolution from the emergence of branch ancestors (when Pre-Permian, etc. emerged as distinct languages) to their internal breakups (e.g. of Proto-Permian into Komi and Udmurt) and from that to the present, as compared to the short CU period from dispersal to emergence of branch ancestors established in §3.2.1. In addition, the general uncertainty about the higher branching structure of Uralic is due to scarcity or lack of diagnostic shared changes, which in itself indicates a rapid separation of the initial branches.

3.2.4. Peripheral archaisms are substantial and well preserved (Author 2019), some in both the far east (Samoyedic) and the far west (Saamic, Finnic) and some only in the far west. These include the well-preserved inflectional paradigms of Samoyedic, Finnic, and Saamic; the preservation of the ancient two-syllable root structure also in Samoyedic, Finnic, and Saamic; the survival of the dual number only in Samoyedic, Ugric, and Saamic; preservation of the sequence of numerals 1-9 with some modifications in Saamic and Finnic²¹; and the high lexical retention rates in Saamic and Finnic. Peripheral retention of archaisms, a common occurrence in dialect geography, is consistent with rapid full separation, as it indicates that inherited material was prone to be replaced by innovations diffusing from post-spread centers of innovation and this was more likely to occur closer to such a center. The central branches of Uralic, especially Permian and Mari, have less PU vocabulary and more changes in the design of inflectional paradigms than Finnic or Saamic because the entire Finno-Ugric range ceased to function as a single dialect-geographical area early and these central languages formed a local interactive area whose innovations did not spread to the far peripheries. Besides, unlike the peripheral dialects, the ones in the center could continue to be in contact with each other which made them more prone to lexical differentiation based on an ideological urge for a separate linguistic identity (see Ellison & Miceli 2017). This is not relevant for phonological or grammatical retentions, but could be so for lexical ones, including the numerals. Some criss-crossing isoglosses in Volga languages based on presumably substrate lexical influence point in this direction.

²¹ The numerals 'eight' and 'nine' actually reflect a subtractive pattern 'less by two' resp. 'less by one', presumably an areal phenomenon with parallel etymological structure in Mordvinic, Mari and Permian (Aikio 2020b, E. Itkonen 1973:336-339, Parpola 1999:198-199).

3.2.5. The distribution of Indo-Iranian loans displays no IBD effects, as Figure 1 shows. The number of PU etyma retained per language correlates highly significantly (*Kendall's tau* = 0.632; $p = 0.0005$) with the number of I-I etyma if the Samoyedic languages are omitted from the count. If Samoyedic is included, the disparity between its artifactually high PU counts and the low I-I counts singlehandedly destroys the correlation (*Kendall's tau* = 0.188; $p = 0.254$). For branches (Supplement S4) there is no significant correlation either with or without Samoyedic, because the number of branches is low. All of this means that the different numbers of I-I etyma in the different languages and branches should not be ascribed to different local intensities of contact between early Uralic and early I-I; the most parsimonious solution is that I-I etyma counts reflect overall vocabulary evolution in the branches and languages, including their different rates of change and loss. (This pertains to just the ratios of I-I loans and PU words. The different inventories of I-I etyma in each branch show that borrowing proceeded individually in the branches.) Samoyedic is an exception, with very low numbers of I-I words out of proportion to its artifactually high numbers of PU words, indicating that it was not in the steppe sphere of interaction. (See §3.4 below; also Figure S3b-c.) However, a secure determination awaits reconstruction and analysis of a full PU vocabulary without veto power by Samoyedic.

Otherwise, trends are clearest and most meaningful when it is branches that are compared. When individual languages are compared, the high number of Finnic languages in the sample give that branch disproportionate impact; and the ordering in which languages are listed within branches can appreciably change the slope and strength of correlations. (For Finnic, ordering south-to-north vs. north-to-south puts Finnish closer to the left edge of the graph vs. near the middle.) North-south position within branches is caused primarily by northward spreads (§2.1), which are irrelevant to the initial spread, and there is no consistency between branches as to whether north or south retains the highest numbers of etyma.)

3.2.6. Reconstructable inflectional morphology points to a fairly unified picture across the branch ancestors (Aikio 2020; Saarikivi in press; Janhunen 1982, 2000, 2009; Kulonen 2001; Majtinskaja 1993). In contrast, developments in individual branches have produced some major reanalyses and reshaping of paradigms. As with sound changes (above), the early similarity and subsequent branch-internal changes point to a short period from initial PU dispersal to branch ancestor formation and longer periods of independent evolution of branches. Thus, a minimally divergent set of future branch ancestors spread out rapidly, at which point branch-internal evolution began (§3.2.3 above and Supplement S6).

3.2.7. Only the Samoyedic branch gives evidence of substantial substratal effects arguably dating to the time of the initial Uralic dispersal, and the general lack of initial substrata in the other branches is an unexpected finding given the linguistic geography: the more western languages have spread farthest from the homeland and have had more opportunities for contacts and more need to borrow terms for non-Siberian phenomena. The expanse from east of the Urals to the upper Volga harbors a number of different languages today and should have contained no fewer in early times, so one might expect different effects in every language. In contrast, there is a modest body of words in the Volga and Western Uralic languages for broadleaf trees and items of agriculture and stockbreeding, some containing novel phoneme

combinations that were impermissible in PU and brand the words as foreign (Häkkinen 2009:37-40, Aikio 2015:44-47, Zhivlov 2015). All are found in more than one Uralic branch. The set suggests borrowings shared among the sister Uralic languages (and part of a broader Middle Volga areality in grammar and lexicon) and not the pervasive effects we find from tundra languages in Saamic (see §3.1) or the large and varied non-Uralic vocabulary of Samoyedic (Saarikivi in press). Thus the only evidence for substratum at the shortly post-PU level comes from Samoyedic, where a good deal of PU vocabulary was lost and replaced by non-Uralic loans with un-Uralic canon shapes and phonotactics (Helke in progress), probably beginning shortly post-PU and continuing to the time of Proto-Samoyedic divergence (Aikio 2020a).²² A counterargument is that the Uralic/un-Uralic contrast may be anachronistic, comparing the young non-Uralic vocabulary (reconstructable only to Proto-Samoyedic, c. 2000-2500 years ago) with PU forms of native vocabulary.

Elsewhere in the family, later substrata are evident in individual languages or subbranches, often connected with northward spreads.²³ The Ugric languages have some important exotic vocabulary, such as terms for horses, horse breeding, and nomadic culture, but whether these point to a substratum or are ordinary loans is less clear. The question of substrates in Ugric has had relatively little research.

Just what happened in Samoyedic is debated. On a traditional family-tree model, where the initial split is seen as a bifurcation into Samoyedic vs. Finno-Ugric, the issue can be seen as vocabulary that is present in Finno-Ugric but absent in Samoyedic and replaced there from another language or languages (the traditional view). Alternatively, it can be asked whether it is not Finno-Ugric that has lost and replaced vocabulary, and the putative exotic vocabulary of Samoyedic is in fact native Uralic vocabulary lost in Finno-Ugric. This issue largely evaporates on the rake model (Supplement S1), where Samoyedic is one of nine initial branches and there is no single branch whose lexical differences are crucial to PU reconstruction. Parallel massive vocabulary replacement across the other eight branches is so unlikely that Samoyedic must be seen as different and as having lost vocabulary.²⁴ A view from a different perspective, applying in either case, is that no drastic loss and replacement has occurred in Samoyedic; rather,

²² The Samoyedic replacement is sometimes referred to as relexification, but that term has a more specific meaning in creolization studies, where it refers to taking the phonological forms of words from a lexifier (superstrate) language but the grammar, including semantic structure and lexical classes of words, from the substrate language (e.g. Lefebvre 1998). There is no evidence of anything like creolization in the history of Samoyedic, and no evidence that Samoyedic word grammar or grammar in general is un-Uralic; as noted, Samoyedic inflectional morphology reflects PU morphology well, unlike the usual development in creolization.

²³ The Ob-Ugric languages have non-native vocabulary in common, much of it with irregular correspondences indicating separate borrowing by early Khanty and Mansi (Sipos 2002, Saarikivi in press, Author 2019). Saamic and Finnic place names have a number of non-native elements (nearly half of the high-frequency elements in Finnish place names) and the two branches have about 220 unique shared cognate roots of unknown etymology, at least some of which are likely to be substratal; but this vocabulary appears to have been acquired not in the initial dispersal but in the Saamic secondary staging area in southeastern Finland and the Northern Dvina basin of northern Russia. (See Aikio 2012, 2004, Saarikivi 2004ab, 2006.)

²⁴ Atkinson et al. 2008 suggest that languages resulting from more branching retain fewer lexical items from the protolanguage (see also Ellison & Miceli 2017 for a possible explanation based on a sociocultural preference for dissimilar forms in individuals who are bilingual in related languages). This could indeed explain why Samoyedic with fewer branches kept more Proto-Uralic lexemes than Finno-Ugric, but crucially it could not explain why Finno-Ugric languages share so many cognates between themselves.

Samoyedic separated early enough that it has had more time for vocabulary attrition. The good preservation of inflectional morphology in Samoyedic argues against a very early separation. So far the issue remains open.²⁵

3.3. Typology and the CU spread

The morphosyntactic typology of Uralic is distinctive in western Eurasia. A number of typological properties are eastern-looking overall, fitting comfortably into northeast Asia, Siberia, or the North Pacific Rim. These include (Supplement S7): low finiteness, high inflectional person, high part-of-speech flexibility in the lexicon, possible evidence of non-accusative alignment, and base intransitivity. A number of traits are cross-linguistically infrequent. Despite the evolutionary disadvantage, or fragility in contact situations, suggested by the low frequency, they are persistent in the family, either as morphemes or as types. They include (Supplement S7): a dual number category; negative auxiliary verb; differential object marking with unusual stability in the conditions on object marking; a contrast of subjective vs. objective conjugation; and personless pronoun stems. Taken together, these various traits are consistent with an origin of PU in the eastern part of its range and a rapid initial spread with minimal contact influence, so that they all remained firmly in place as the dispersing CU branch ancestors settled in among new neighbors.

3.4. Early diversification in the east

Though the higher-level branching structure in the Uralic tree is debated, it is clear that if there was any early branching and/or any early geographical separation it began in the east of the PU range. Evidence for an initial binary split of Samoyedic vs. Finno-Ugric includes the substantial vocabulary replacement in Samoyedic and only there (§§1, 2.2) and the derivational patterns shared within Finno-Ugric but absent in Samoyedic (Supplement S1). The near-total absence of Indo-Iranian loans in Samoyedic (2.2.1), though not a clade-defining event, is evidence of early geographical separation of Samoyedic from the rest and its distance from Indo-Iranian-speaking groups. Other possible evidence is a two-step sound change arguably shared by Samoyedic and Ugric, suggesting a very early branching (Supplement S8); numbers of

²⁵ Cases where a daughter language preserves morphology well but loses vocabulary massively are not frequent, but they are reported. Comrie 1988, 1989, 2000 shows that Haruai (Piawi family, New Guinea) is a sister of Hagahai (Piawi) and not of Kobon (Kalam family), despite massive vocabulary sharing with Kobon, on the evidence of shared morphological paradigms with Hagahai. Green 2003 shows that Murrinh-Patha, long considered an isolate, is a close sister of Ngan'gityemerri making up a Southern Daly family (northern Australia) on the strength of parallel auxiliary verb morphology and despite near-absence of lexical cognates. Comrie suggests that the Piawi discrepancy could have arisen as a result of word taboos compensated by loans from a neighboring language, whose cumulative effect in a small speech community might have been substantial. Green leaves the question open for Southern Daly. Miceli 2015 describes the Pama-Nyungan family as languages which share phonology and grammar, but very little lexicon (and many of the potential cognate forms are nearly identical, so they could as well be loanwords and not inherited from the common protolanguage). She suggests that conscious efforts of bilingual speakers on keeping languages apart (in particular in case of related languages) may have contributed to extremely high levels of lexical divergence in Pama-Nyungan, attested simultaneously with morphosyntactic convergence (see also Ellison & Miceli 2017 for a more general approach to the same phenomenon). In the case of lexical discrepancy between Samoyedic and the rest of Uralic, the drivers of lexical divergence would have been bilinguals in (Pre-)Proto-Samoyedic and some other Common Uralic variety(/ies).

retained PU etyma are generally higher in the west than in the east, as are inter-branch lexical sharings of PU etyma, suggesting that earlier there was less inter-branch integration in the east (Supplement S8). This could be due to the river geography: while the Volga offers direct east-west connections from near the Urals to eastern Europe, the major rivers in western Siberia flow north or northwest and there are no short overwater connections between the Altai-Sayan and the Urals. (Nonetheless, the locations of ST sites indicate that rivers were major transportation routes in Siberia: Map 1.)

Thus there is evidence for early distancing of Samoyedic and the eastern branches more generally from the rest of the family and early internal distancing in western Siberia. If the early phylogenetic diversification was anything other than a rake, the Samoyedic branch was the first to split off, followed by Ugric.

3.5. Genetic evidence

Rapid spread of a small population and expansion of its subgroups through language shift should have produced exactly the genetic picture we see in Uralic (Tambets et al. 2018): There is a detectable pan-Uralic component showing that movement of people was involved to at least some extent in the Uralic spread; and Uralic-speaking populations are everywhere similar to their neighbors, so much so that in some of the westernmost groups the original Uralic genetic component is invisible or nearly so, while in Samoyedic populations the Uralic component is the one that is shared with neighbors. (See also Balanovsky 2019, Saag et al. 2019, Ilumäe et al. 2016, Lamnidis et al. 2018.)

4. Sociolinguistics of post-catastrophe spreads

Little is known of typical sociolinguistic consequences of post-catastrophe spreads (Supplement S9). Catastrophes such as the Plague of Justinian, the Black Death, and prolonged severe droughts have often resulted in language spreads, shifts, and extinctions, and the shifting and the social turmoil of catastrophes may sometimes have led to decomplexification of the surviving language(s) (as is expected when an expanding language absorbs an appreciable number of adult L2 learners: Trudgill 2011). Most modern Finno-Ugric languages are in fact phonologically and morphologically 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, only partly due to post-Proto-Samoyedic developments; and most Saamic languages are among Eurasia's most complex, due to post-Proto-Saamic phonological developments (Supplement S9; Nichols 2019). Thus it is possible, but not necessary, that early Finno-Ugric has undergone some decomplexification as might be expected of an inter-ethnic trade language. The effect, if real, is subtle, however, and the general lack of substratal effects at the branch ancestor level is a stronger argument and one that speaks against impact of shifting speakers on CU grammar.

5. Conclusions

The following are the main conclusions drawn here:

- Proto-Uralic originated east of the Urals and out of contact with Proto-Indo-European.
- The Uralic spread took place rapidly and for the most part without substratal effects.
- The simultaneous 4.2 ka event and the ST transcultural phenomenon, and the Indo-Iranian contact episode within their time frame, explain the Uralic spread and situate it in space and time.
- Uralic spread with ST trade along the rivers that were the main avenues of communication and transport.
- The I-I contact episode gives a reliable absolute date for the Uralic divergence and dispersal: not long before 4000 BP.
- What was involved in ST and I-I was not Proto-Uralic but Common Uralic, the set of still mutually intelligible but separate and separately evolving varieties that emerged from the initial diversification of Proto-Uralic.
- Proto-Uralic has a number of eastern typological features suggesting an eastern origin.
- Proto-Uralic has some rare features that are stable in the family, indicating that, while early Uralic must have expanded via language shift, the shifting population had minimal impact on PU grammar and vocabulary.
- The non-pastoral, non-agricultural, sparsely distributed CU populations suffered less from the 4.2 ka event and were able to recover from it faster than the denser stockbreeding IE populations. As a result, early Uralic varieties replaced IE-speaking populations along the Volga and near the Urals.
- The early history of the Samoyedic branch is mysterious. It retains a low absolute number of PU etyma, but until more PU reconstruction and statistical analysis are done it is not known whether the number of etyma is significantly low.
- Northward spreads, a standing pattern in northern Eurasia, gave the Uralic language family its large northern extent. An important contributor to these spreads was language shift. As a result, apart from what are now national languages, Uralic languages became known to linguistics and ethnography primarily as languages of hunter-gatherers and northern people.

The following are questions that remain open:

- Can the Northwest Indo-European of Koivulehto 2000 be identified with Para-Baltic as that is defined here? If so, can possible loan etyma from it be identified in PU or CU?

Indicators might include irregular correspondences in the treatment of early Baltic loans in Uralic branches.

- Can evidence for a Para-Uralic be identified? Indicators would include irregular correspondences in what are now identified as PU etyma.
- What are the limits and parameters of variation in numbers of PU etyma retained in the various Uralic branches, on definitions of PU cognacy other than the traditional one (reflex in Samoyedic and at least one other)? Is the low retention figure for Samoyedic significantly lower than those of others?
- Where, when, and how did Pre-Samoyedic begin to differentiate from Finno-Ugric?

References (Anonymized)

DTB = Adams 2013

EWU = Benkő ed. 1993

EST = Sevortjan et al. 1978ff.

MdWb = Paasonen 1990-1999

MszFE = Lakó ed. 1967-1978

SW = Janhunen1977

- Aalto, Pentti. 1959. Ein alter Name des Kupfers. *Ural-Altäische Jahrbücher* 31 (Gedenkband Julius von Farkas): 33–40.
- Adams, Douglas W. 2013. *A Dictionary of Tocharian B: Revised and Greatly Enlarged*. (Leiden Studies in Indo-European 10.) Amsterdam: Rodopi.
- Aikio, Ante. 2002. New and old Samoyedic etymologies. *Finnisch-Ugrische Forschungen* 56.9–67.
- Aikio, Ante. 2004. An essay on substrate studies and the origin of Saami. In *Etymologie, entlehnungen und wntwicklungen*, ed. Irma Hyvärinen, Petri Kallio, and Jarmo Korhonen, 187–214.
- Aikio, Ante. 2006. New and old Samoyedic etymologies, 2. *Finnisch-Ugrische Forschungen* 59:1-3.9–34.
- 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. Grünthal & Kallio, eds., 63-117.
- Aikio, Ante. 2013. Studies in Uralic etymology I: Saami etymologies. *Linguistica Uralica* 49.161–174.
- Aikio, Ante. 2014a. Studies in Uralic etymology II: Finnic etymologies. *Linguistica Uralica* 50.1–19.
- Aikio, Ante. 2014b. Studies in Uralic etymology III: Mari etymologies. *Linguistica Uralica* 50.81–94.
- Aikio, Ante. 2015a. Studies in Uralic etymology IV: Ob-Ugric etymologies. *Linguistica Uralica* 51.1–20.
- Aikio, Ante. 2015b. The Finnic ‘secondary e-stems’ and Proto-Uralic vocalism. *SUSA/JSFOu* 95.25–66.
- Aikio, Ante. 2018. Uralic Etymological Dictionary. (Näyte tulevan sanakirjan käsikirjoituksesta.) Lecture handout, Finno-Ugrian Society, February 2018. Helsinki.
- Aikio, Ante. c. 2020a. Proto-Uralic. In *The Oxford guide to the Uralic languages*, ed. Marianne Bákro-Nagy, Johanna Laakso, and Elena Skribnik, Oxford: Oxford University Press. (In press.)
- Aikio, Ante. 2020b. Uralic etymological dictionary. Draft version of entries A-C.
https://www.academia.edu/41659514/URALIC_ETYMOLOGICAL_DICTIONARY_draft_version_of_entries_A_C.
- Andersen, Henning. 2003. Slavic and the Indo-European migrations. In *Language contacts in prehistory: Studies in stratigraphy*, ed. Henning Andersen, 45–76. Amsterdam-Philadelphia: Benjamins.
- Andersen, Henning. 1996. *Reconstructing prehistoric dialects: Initial vowels in Slavic and Baltic*. Berlin: Mouton de Gruyter.
- Andrades Valtueña, Aida, Mittnik, A., Key, F. M., Haak, W., & Krause, J. 2017. The Stone Age plague and its persistence in Eurasia. *Current Biology* 27.3683-3691.
- Anonymous. 2018. Poltavkinskaja kul'tura. Wikipedia.
 ru.wikipedia.org > wiki > Полтавкинская культура
- Anthony, David. 2020. Ancient DNA, mating networks, and the Anatolian split. In Serangeli & Olander, eds., 21-53. Leiden: Brill.
- 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.
- Anthony, David, & Ringe, Don. 2015. The Indo-European homeland from linguistic and archaeological perspectives. *Annual Review of Linguistics* 1.199-219.

- 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 Skaftø Jensen, Jens E. Mogensen & Lene Schøsler (eds.), *Historical Linguistics 2003: Selected Papers from the 16th International Conference on Historical Linguistics, Copenhagen, 11–15 August 2003* (Current Issues in Linguistic Theory 257), 1–19. Amsterdam: John Benjamins.
- Atkinson, Qentian D., Meade, Andrew, Venditti, Chris, Greenhill, Simon J. & Pagel, Mark. 2008. Languages evolve in punctuational bursts. *Science* 319(5863). 588.
- Balanovsky, Oleg 2019. Evolution of the Y-chromosomal pool in East Europe. Presented at Contacts: Archaeology, genetics, languages. Joining forces to shed light on early contacts (4000 BC – 1000 AD) between Indo-European and Uralic speakers. Suomenlinna/Helsinki.
- Barfield, Thomas. 2009. Introduction: Frontiers and border dynamics. In B. K. Hanks & K. M. Linduff, eds., *Social Complexity in Prehistoric Eurasia: Monuments, Metals, and Mobility*, 235–240. Cambridge: Cambridge University Press.
- Barfield, Thomas. 1989. *The perilous frontier: Nomadic empires and China, 221 BC to AD 1757*. London: Blackwell.
- Bartens, Raija. 2000. *Permiläisten kielten rakenne ja kehitys*. Helsinki: Finno-Ugrian Society.
- Beck, David. 2013. Unidirectional flexibility and the noun-verb distinction in Lushootseed. In *Flexible word classes*, ed. Jan Rijkhoff and Eva van Lier, 185–220. Oxford: Oxford University Press.
- Belyx, S. K. 2009. *Problema raspada prapermskojazykovoj obščnosti*. Iževsk: Udmurt University.
- Benkő, Lóránd, 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, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Riessler, Lennart Bierkandt, Fernando Zúñiga, and John B. Lowe. The Autotyp database, release version 0.1.0.
<https://github.com/autotyp/autotyp-data>
- Cardillo, M., L. Bromham, and Simon J. Greenhill. 2014. 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 *Early contacts between Uralic and Indo-European: Linguistic and archaeological considerations*, ed. Christian Carpelan, Asko Parpola, and Petteri Koskikallio, 55–150. Helsinki: Suomalais-Ugrilainen Seura.
- Carpelan, Christian, Parpola, Asko, & Koskikallio, Petri, eds. (2001). *Early Contacts between Uralic and Indo-European: Linguistic and archaeological considerations*. (SUST 242) 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.
- Comrie, Bernard. 2000. Language contact, lexical borrowing, and semantic fields. In *Languages in contact*, ed. Dicky G. Gilbers, John Nerbonne, and Jos Schaeken, 73–86. Amsterdam: Rodopi.
- Comrie, Bernard. 1989. Genetic classification, contact, and variation. In *Synchronic and diachronic approaches to linguistic variation and change*, ed. Thomas J. Walsh, 81–93. (
- Comrie, Bernard. 1988. Haruai verb structure and language classification in the Upper Yuat. *Language and Linguistics in Melanesia* 17.140–160.
- Corbett, Greville. 2007. Canonical typology, suppletion, and possible words. *Language* 83:1.8–42.
- Corbett, Greville G. 2015. Morphosyntactic complexity: A typology of lexical splits. *Language* 91:1.145–193.

- Dalfes, H. Nüzhet, George Kukla, and Harvey Weiss, eds. 1997. *Third Millennium BC climate change and Old World collapse*. Berlin-New York: Springer.
- De Filippo, Cesare, Koen Bostoen, Mark Stoneking, and Brigitte Pakendorf. 2012. Bringing together linguistic and genetic evidence to the Bantu expansion. *Proceedings of the Royal Society B* 279(1741): 3256–3263. <https://doi.org/10.1098/rspb.2012.0318>.
- 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. 1989. Large linguistic areas and language sampling. *Studies in Language* 13.257–292.
- Dryer, Matthew S. and Martin Haspelmath, eds. *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library, 2013. Leipzig: Planck Institute for Evolutionary Anthropology. Online: <http://wals.info>. Accessed 2-28-20.
- Dyen, Isidore. 1956. Language distribution and migration theory. *Language* 32.611–626.
- Ellison, T. Mark & Miceli, Luisa. 2017. Language monitoring in bilinguals as a mechanism for rapid lexical divergence. *Language* 93:2.255–287.
- Evans, Nicholas, & McConvell, Patrick. (1998). The enigma of Pama-Nyungan expansion in Australia. In R. Blench & 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.
- Frachetti, Michael D. 2008. *Pastoralist landscapes and social interaction in Bronze Age Eurasia*. Berkeley-Los Angeles: University of California Press.
- Fodor, István. 2001. Az uráli őstörténet és a régészet. *Ünnepi könyv Keresztes László tiszteletére. Folia Uralica Debreceniensia* 8. 143–162.
- Fodor, István. 2016. Nyelvek, népek, térképek: Néhány észrevétel Juha Janhunen hipotéziséhez. *Archaeologiai Értesítő* 141 (1). 217–229.
- Foley, William A. 2017. Structural and semantic dependencies in word class membership. In N. Enfield, ed., *Dependencies in language: On the causal ontology of linguistic systems*, 179-195. Berlin: Language Science Press.
- Goddard, Ives. 1996. The west-to-east cline in Algonquian dialectology. In *Actes du vingt-cinquième Congrès des algonquistes*, ed. William Cowan, 187–211. Ottawa: Carleton University.
- Green, Ian. 2003. The genetic status of Murrinh-Patha. In *The NonPpama-Yyungan languages of northern Australia*, ed. Nicholas Evans, 125–158. Canberra: Research School of Pacific and Asian Studies, Australian National University.
- 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. 2015. *Vepsän kielioppi*. Helsinki: Suomalais-ugrilainen Seura.
- Grünthal, Riho. 2012. Baltic loanwords in Mordvin. Grünthal and Kallio, 297–343.
- Grünthal, Riho and Johanna Nichols. 2016. Transitivity/detransitivity typology and language family history. *Lingua Posnaniensis* 58:2.11–31.
- Grünthal, Riho and Petri Kallio, eds. 2012. *A linguistic map of prehistoric northern Europe*. (SUST 266.) Helsinki: Société Finno-Ougrienne.
- Grünthal, Riho, Tapani Salminen, and Pirkko Numminen. 1993. Geographical distribution of the Uralic languages. Helsinki: Finno-Ugrian society.
- Güldemann, Tom, Patrick McConvell, and Richard A. Rhodes. 2020. Hunter-gatherer anthropology and language. In *The language of hunter-gatherers*, ed. Tom Güldemann, Patrick McConvell, and Richard A. Rhodes, 3–48. Cambridge: Cambridge University Press.

- Hajdú, Péter 1969. Finnougrische Urheimatforschung. *Ural-Altische Jahrbücher* 41.
- Hajdú, Péter 1975. Sukulaisuuden kielellistä taustaa. In Péter Hajdú (ed.), *Suomalais-ugrilaiset*. Helsinki: Finnish Literature Society. 11–51.
- Hajdú, Péter. 1975. *Finno-Ugrian languages and peoples*. London: André Deutsch.
- Hajdú, Péter & Domokos, Péter. 1978. *Uráli nyelvrokonaik*. Budapest: Tankönyvkiadó.
- Helke, Jaakko. In progress. MA thesis, University of Helsinki.
- Häkkinen, Jaakko. 2009. Kantauralin ajoitus ja paikannus: perustelut puntarissa. *Journal de la Société Finno-Ougrienne* 92: 9–56.
- Häkkinen, Kaisa. 2001. Prehistoric Finno-Ugric culture in the light of historical lexicology. Carpelan et al. eds. 169-186.
- Häkkinen, Kaisa. 1996. *Suomalaisten esihistoria kielitieteen valossa*. Helsinki: Suomalaisen Kirjallisuuden Seura.
- Hammarström, Harald & Forkel, Robert & Haspelmath, Martin. 2019. Glottolog 4.1. Jena: Max Planck Institute for the Science of Human History.
(Available online at <http://glottolog.org> Accessed 2020-03-09.)
Uralic family entry: <https://glottolog.org/resource/languoid/id/ural1272>
- Havas, Ferenc. 2003. A tárgy tárgyában: Mondattipológiai fontolgatások. In: Beatrix Oszkó and Mária Sipos, eds., *Uráli TÁRGYALÓ. Budapesti Uráli Műhely III, 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>.
- Helama, Samuli and Markku Oinonen. 2019. Exact dating of the Meghalayan lower boundary based on high-latitude tree-ring isotope chronology. *Quaternary Science Reviews*, online: <https://doi.org/10.1016/j.quascirev.2019.04.013>
- Helimski, Eugene. 2003. Areal groupings (sprachbünde) within and across the borders of the Uralic language family: A survey. *Nyelvtudományi Közlemények* 100.156–167.
- Helimski, Eugene. 2001b. Early Indo-Uralic linguistic relationship: Real kinship and imagined contacts. Carpelan et al., eds, 235-263.
- Helimski, Evgenij. 2001a. Ural'cy i ix predšestvenniki: Belye pjatna na ètnoistoričeskoj karte severnoj evrazii i ural'skie jazyki. *Congressus nonus internationalis fenno-ugristarum IV*, *Linguistica* 1.332–336.
- Helimski, E. (Хелимский, Е. А.). 1982. Древнейшие венгенско-самодийские языковые параллели. Москва: Наука.
- 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. 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. Daniel Abondolo, ed., *The Uralic languages*, 327–357. London-New York: Routledge.
- Huisman, John L. A., Asifa Majid, and Roeland van Hout. 2019. The geographical configuration of a language area influences linguistic diversity. *PLOS One* 14:6.e0217363. <https://doi.org/10.1371/journal.pone.0217363>.
- Hunley, Keith, Michael Dunn, Eva Lindström, Ger P. Reesink, Angela Terrill, Meghan Healy, George Koki, Françoise R. Friedlaender, and Jonathan S. Friedlaender. 2008. Genetic and linguistic coevolution in northern Island Melanesia. *PLOS Genetics* 4:10.e1000239. <https://doi.org/10.1371/journal.pgen.1000239>.
- Ilumäe, A.-M. et al. 2016. Human Y chromosome haplogroup N: A non-trivial time-resolved phylogeography that cuts across language families. *American Journal of Human Genetics* 99. 163–173.
- Itkonen, Erkki. 1973. Zur Geschichte des Partitivs. *Finnisch-ugrische Forschungen* 40.279-339.
- Iversen, Rune and Guus Kroonen. 2017. Talking Neolithic: Linguistic and archaeological perspectives on how Indo-European was implemented in southern Scandinavia. *American Journal of Archaeology* 121:4.511–525.
- 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 & 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. 2020. Grammaticalization in Uralic as viewed from a general Eurasian perspective. In Andrej Malchukov & Walter Bisang, eds., *Handbook of grammaticalization*.
- Janhunen, Juha. 2014. A legkeletibb uráliak. *Nyelvtudományi Közlemények* 110. 7-30.
- Janhunen, Juha. 2012a. Etymological and ethnohistorical aspects of the Yenisei. *Studia Etymologica Cracoviensia* 17.67-87.
- Janhunen, Juha. 2012b. The expansion of Tungusic as an ethnic and linguistic process. In *Recent advances in Tungusic linguistics*, ed. Andrej L. Malchukov and Lindsay J. Whaley, 5–16. Harrassowitz: Wiesbaden.
- Janhunen, Juha. 2009. Proto-Uralic -- what, where, and when? *The Quasquicentennial of the Finno-Ugrian Society*, 57-78. Helsinki: Suomalais-Ugrilainen Seura.
- Janhunen, Juha. 2001. Indo-Uralic and Ural-Altaic: On the diachronic implications of areal typology. In *Early contacts between Uralic and Indo-European: Linguistic and archaeological considerations*, ed. Christian Carpelan, Asko Parpola, and Petteri Koskikallio, 207–220. Helsinki: Finno-Ugrian Society.
- Janhunen, Juha. 2000. Reconstructing Pre-Proto-Uralic typology: Spanning the millennia of linguistic evolution. In A. Nurk, T. Palo, & T. Seilenthal, eds., *Congressus Nonus Internationalis Fenno-Ugristarum, Pars I: Orationes plenariae & orationes publicae*. pp. 59-76. Tartu: CIFU.
- Janhunen, Juha. 1998. Samoyedic. In *The Uralic languages*, ed. Daniel Abondolo, 457–479. London: Routledge.
- Janhunen, Juha. 1982. On the structure of Proto-Uralic. *Finnisch-Ugrische Forschungen* 44:1-3. 23-42.

- Janhunen, Juha. 1981. Uralilaisen kantakielen sanastosta. *Journal de la Société Finno-ougrienne* 77:9.219–274.
- Janhunen, Juha. 1977. *Samojedischer Wortschatz: Gemeinsamojedische etymologien*. Helsinki: Helsinki University.
- 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 and Kallio, 261–296.
- Kallio, Petri. In press. Substrates in Finnic. In *Substrate languages in northern Europe: Case studies and methodological perspectives*, ed. Ante Aikio and Santeri Palviainen, Berlin: de Gruyter-Mouton.
- Kallio, Petri. 2018. Yhteisitämerensuomalainen sanasto (Common Finnic vocabulary). <https://sanat.csc.fi/wiki/Etymologiadata:YSuS>
- Kallio, Petri. 2015. The language contact situation in prehistoric northeastern Europe. Robert Mailhammer, Theo Vennemann, & Birgit Anette Olsen, eds., *The linguistic roots of Europe*, 77–102. Copenhagen: Museum Tusulanum Press, University of Copenhagen.
- Kallio, Petri. 2006. Suomen kantakielen absoluuttista kronologiaa. *Virittäjä* 110.2–25.
- Kallio, Petri. 2004. Tocharian loanwords in Samoyed? *Mémoires de la Société Néophilologique de Helsinki* 63 (Etymologie, Entlehnungen und Entwicklungen: Festschrift für Jorma Koivulehto zum 70. Geburtstag): 129–137.
- Kessler, Brett and Annukka Lehtonen. 2006. Multilateral comparison and significance testing of the Indo-Uralic question. In *Phylogenetic methods and the prehistory of languages*, ed. James Clackson, Peter Forster, and Colin Renfrew, 33–42. Cambridge: McDonald Institute for Archaeological Research.
- Khanina, Olesya, Koryakov, Yuri & Shluinsky, Andrey. 2018. Enets in space and time: a case study in linguistic geography. *Finnisch-Ugrische Mitteilungen* 42: 109-135.
- Khanina, Olesya, and Valentin Gusev. In progress. The northward spread of Samoyedic.
- Koivulehto, Jorma. 2001. The earliest contacts between Indo-European and Uralic speakers in the light of lexical loans. Carpelan et al., eds., 235-263.
- Koivulehto, Jorma. 2000. Finno-Ugric reflexes of Northwest Indo-European and early stages of Indo-Iranian. In *Proceedings of the eleventh annual UCLA Indo-European conference*, ed. Karlene Jones-Bley, Martin E. Huld, and Angela Della Volpe, 21–44. Washington, DC: Institute for the Study of Man.
- Korenchy, Éva. 1972. *Iranische Lehnwörter in den obugrischen Sprachen*. Budapest: Akadémiai Kiadó.
- Korhonen, Mikko. 1981. *Johdatus lapin kielen historiaan*. Helsinki: Suomalaisen Kirjallisuuden Seura.
- Kortlandt, Frederik. 2019. On the reconstruction of Proto-Uralic. In Santeri Junttila & Juha Kuokkala, eds., *Petri Kallio rocks: Liber semisaecularis*, 11-14. Helsinki.
- Koryakov, Yuri. 2007. Slavjane i ix sosedi v VII-VIII vv. In: Platonov, S. F., *Učebnik russkoj istorii. Ch. 3: Russkie slavjane i ix sosedi*. <https://history.wikireading.ru/92803>
also: <https://en.wikipedia.org/wiki/Galindians#/media/File:Slav-7-8-obrez.png>
- Krause, R., A. V. Epimaxov, E. B Kuprijanova et al. 2019. Petrovskie pamjatniki bronzovogo veka: Problemy taksonomii i xronologii. *Arxeologija, ètnografija i antropologija Evrazii* 47:1.54–63.
- Kristiansen, Kristian. 2020. The archaeology of Proto-Indo-European and Proto-Anatolian: Locating the split. In Serangeli and Olander, 157–165. Leiden: Brill.
- Kulonen, Ulla-Maija. 2001. Über die Deklination der Personalpronomina in der finnisch-ugrischen Grundsprache. In A. Turk, T. Palo, & T. Seilenthal, eds., *Congressus Nonus Internationalis Fenno-Ugristarum*, 6.178-182. Tartu.
- Kulonen, Ulla-Maija. 1989. *The passive in Ob-Ugrian*. Helsinki: Finno-Ugrian Society.

- Kuper, Rudolph and Stefan Kroepelin. 2006. Climate-controlled Holocene occupation in the Sahara: Motor of Africa's evolution. *Science* 313.803–807.
- Kuzmina, Elena E. 2007. *Origin of the Indo-Iranians*. (Ed. J. P. Mallory.) Leiden: Brill.
- Lakó, György, ed. 1967-1978. *A magyar szókészlet finnougor elemei*, vols. 1-3. Budapest: Akadémiai Kiadó.
- Lamnidis, Thiseas, Kerttu Majander, et al. 2018. Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe. *Nature Communications* 9:5018.
- Lang, Valter. 2018. *Läänemeresoome tulemised*. 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. 2001. Textile ceramics in Finland and on the Karelian isthmus: Nine variations and figure on a theme of C. F. Meinander. PhD Dissertation, University of Helsinki.
- Lavento, Mika and V. S. Patrushev. 2015. Razvitie i xronologija tekstil'noj keramiki na Srednej i Verxnej Volge: Kritičeskij vzgljad vozmožnostej xronologii po ¹⁴C-, AMS-, i tipologičeskomu metodam (First appearance and chronology of textile ceramics in the Middle and Upper Volga region: Critical comparison of conventional ¹⁴C, AMS, and typological chronologies. *Povolžskaja arxeologija* 2:12.160–188.
- Lefebvre, Claire. 1998. *Creole genesis and the acquisition of grammar: The case of Haitian Creole*. Cambridge: Cambridge University Press.
- Lehtinen, Jyri. c. 2021. Uralic phonological phylogeny. PhD dissertation in progress, University of Helsinki.
- Lehtinen, Jyri, Honkola, Terhi, Korhonen, Kalle, Syrjänen, Kai, Wahlberg, Niklas, & Vesakoski, Outi. 2014. Behind family trees: Secondary connections in Uralic language networks. *Language Dynamics and Change* 4.189-221.
- Ligeti, Lajos. 1941. A kínai átírásos barbár nyelvi glosszák kérdése. *Nyelvtudományi Közlemények* 51: 174–207.
- Lindner, Stephan. 2020. Chariots in the Eurasian steppe: A Bayesian approach to the emergence of horse-drawn transport in the early second millennium BC. *Antiquity* 2020. <https://doi.org/10.15184/aqy.2020.37> (accessed July 1, 2020)
- Lindstedt, Jouko, & Salmela, Elina. In press. Migrations and language shifts as components of the Slavic spread. In T. Klír & V. Boček (Eds.), *Language contact and the early Slavs*. Heidelberg: Winter.
- Lourandos, Harry. (1997). *Continent of Hunter-Gatherers*. Cambridge: Cambridge University Press.
- Madsen, David & David Rhode, eds. 1994. *Across the west: Human population movement and the expansion of the Numa*. Salt Lake City: University of Utah Press.
- Majtinskaja, K. E. (1993). *Istoriko-sopostavitel'naja morfologija finno-ugorskix jazykov*. Moscow.
- 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.
- Mallory, James P. 1997. Poltavka culture. In J. P. Mallory and Douglas Q. Adams, *Encyclopedia of Indo-European culture*. London: Fitzroy Dearborn.
- Marchenko, Z. V., S. V. Svyatko, V. I. Molodin, A. E. Grishin, and M. P. Rykun. 2017. Radiocarbon chronology of complexes with Seima-Turbino type objects (Bronze Age) in southwestern Siberia. *Radiocarbon* 59:5.1381–1397.
- Maslova, Elena and Tatiana Nikitina. 2008. Stochastic universals and dynamics of cross-linguistic distributions: The case of alignment types. MS.

- Mathieson, Iain, Iosif Lazaridis, Nadin Rohland et al. 2015. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528.499–503.
- Matveev, Aleksander Konstantinovich. 1962. Az ugor népek ősi letelepülési helyei. *Nyelvtudományi Közlemények* 64. 289–297.
- Maurits, Luke. 2018. A new Bayesian phylogeny for the Uralic languages. Invited lecture, University of Helsinki.
- Miceli, Luisa. 2015. Pama-Nyungan. In C. Bowern & B. Evans (eds.), *The Routledge Handbook of Historical Linguistics*, 704–725. Oxon & New York, Abingdon: Routledge.
- Michalove, Peter A. 2002. The classification of the Uralic languages: Lexical evidence from Finno-Ugric. *Finnisch-Ugrische Forschungen* 57:1-3.58–67U
- 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. In Jahr, Ernst H. (ed.), *Språkkontakt: Innverknaden frå nedertysk på andre nordeuropeiske språk* [Language contact: Low German influence on other north European languages], 87–94. København: Nordisk Ministerråd.
- Napol'skix, V. V. 1997. *Vvedenie v istoričeskiju uralistiku*. Izhevsk: RAN: Ural'skoe otdelenie: Udmurtskij institut istorii, jazyka i literatury.
- Nichols, Johanna. 2020. Canonical complexity. In *The complexities of morphology*, ed. Peter Arkadiev and Francesco Gardani, 163–192. Oxford: Oxford University Press.
- Nichols, Johanna. 2019. Grammatical complexity in Saami: A historical-typological view. Presented at SAALS 4, Uppsala.
- 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. 2018. Non-linguistic conditions for causativization as a linguistic attractor. *Frontiers in Psychology* 8. 2356.
- Nichols, Johanna. 2017. Person as an inflectional category. *Linguistic Typology* 21:3.387-456.
- Nichols, Johanna. 2016. Verb-based and noun-based languages. Poster presented at SLE 49, Leiden.
- Nichols, Johanna. 2010. Proof of Dene-Yeniseian relatedness. In James Kari & Ben A. Potter, eds., *The Dene-Yeniseian Connection*, 266-278. Fairbanks: Alaska Native Language Center.
- 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. 1997. Modeling ancient population structures and movement in linguistics. *Annual Review of Anthropology* 26.359–384.
- Nichols, Johanna. 1992. *Linguistic diversity in space and time*. Chicago: University of Chicago Press.
- Nichols, Johanna. 1973. Suffix ordering in Proto-Uralic. *Lingua* 32:3.227-238.
- Nichols, Johanna, and Richard A. Rhodes. 2018. Vectors of language spread at the central steppe periphery: Finno-Ugric as catalyst language. In Guus Kroonen & Rune Iversen, eds., *Digging for words*, 58-68. Oxford: BAR Publishing.
- Nichols, Johanna and Tandy Warnow. 2008. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass* 2:5.760–820.
- Nichols, Johanna, David A. Peterson, and Jonathan Barnes. 2004. Transitivity and detransitivizing languages. *Linguistic Typology* 8:2.149–211.
- Nordqvist, Kerkko and Volker Heyd. 2020. The forgotten child of the wider Corded Ware family: The Russian Fatyanovo culture in context. *Proceedings of the Prehistoric Society* 86.65-93.

- 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, H. 1990-1999. *H. Paasonens Mordwinisches Wörterbuch I–VI*. Zusammengestellt von Kaino Heikkilä. Bearbeitet und herausgegeben von Martti Kahla. Lexica Societatis Fenno-Ugricae XXIII. Helsinki: Finno-Ugric Society.
- Padilla-Iglesias, Cecilia, Erik Gjesfjeld & 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. Grünthal & Kallio, eds., 119–184.
- Parpola, Asko. 1999. Varhaisten indoeurooppalaiskontaktien ajoitus ja paikannus kielellisen ja arkeologisen aineiston perusteella. In Paul Fogelberg, ed., *Pohjan poluilla: Suomalaisten juuret nykytutkimuksen vuokaan*, 180-206. Helsinki: Societas Scientiarum Fennica.
- Perşoiu, Aurel, Monica Ionita, and Harvey Weiss. 2019. Atmospheric blocking induced by the strengthened Siberian high led to drying in West Asia during the 4.2ka bp event – a hypothesis. *Climate of the Past* 15.781–793.
- 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.
- Rascovan, Nicolás, Sjögren, K.-G., Kristiansen, K., Nielsen, R., Willerslev, E., Desnues, C. et al. 2019. Emergence and spread of basal lineages of *Yersinia pestis* during the Neolithic decline. *Cell* 176.1-11.
- Rasmussen, Simon, Allentoft, M. E., Nielsen, K., Nielsen, R., Kristiansen, K. et al. 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.
- Róna-Tas, András and Árpád Berta. 2011. *West Old Turkic: Turkic loanwords in Hungarian*. Wiesbaden: Harrassowitz.
- Ross, Malcolm D. 1988. *Proto Oceanic and the Austronesian languages of western Melanesia*. Canberra: Australian National University.
- Rybatzki, Volker. 1994. Bemerkungen zur türkischen und mongolischen Metallterminologie. *Studia Orientalia* 73: 193–251.
- Saag, Lehti, Margot Laneman, Liivi Varul et al. 2019. The arrival of Siberian ancestry connecting the Eastern Baltic to Uralic speakers further east. *Current Biology* 29. 1–11.
- Saarikivi, Janne. In press. Uralic intermediate protolanguages: A descendent reconstruction. In Johanna Laakso et al., eds., *The Oxford handbook of Uralic languages*. Oxford: Oxford University Press.
- Saarikivi, Janne. 2006. Substrata uralica: Studies on Finno-Ugric substrate in northern Russian dialects. PhD Dissertation, University of Helsinki.
- Saarikivi, Janne. 2004. 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. 2004. Über das saamische Substratnamengut in Nordrussland und Finnland. *Finnisch-Ugrische Forschungen* 58:1-3.187–214.

- Saarikivi, Janne, & Lavento, Mika. 2012. Linguistics and archaeology: A critical view of an interdisciplinary approach with reference to the prehistory of northern Scandinavia. In Charlotte Damm & Janne Saarikivi, eds., *Networks, Interaction, and Emerging Identities in Fennoscandia and Beyond*, 177-216. 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. Kiznetsovoj*, 44-55. Moscow: Moscow University Press.
- Salminen, Tapani. 2001. The rise of the Finno-Ugric language family. Carpelan et al. eds., 3858-396.
- Sammallahti, Pekka. 1988. Historical phonology of the Uralic languages. In *The Uralic languages*, ed. Denis Sinor, 478–554. Helsinki: Suomalais-Ugrilainen Seura.
- 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 & 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.
- Sapir, Edward. 1916. Time perspective in aboriginal American culture: A study in method. In *Selected writings of Edward Sapir in language, culture, and personality*, ed. D. G. Mandelbaum, 389–467. Berkeley-Los Angeles: University of California Press.
- Serangeli, Matilde and Thomas Olander, eds. 2020. *Dispersals and diversification: Linguistic and archaeological perspectives on the early stages of Indo-European*. Leiden: Brill.
- Seržant, Ilja A. In press. The Circum-Baltic area. In *The Slavonic languages*, ed. Jan Fellerer and Neil Bermel. Oxford: Oxford University Press.
- Sevortjan, È. V. et al., eds. 1978-2002-. *Ètimologičeskij slovar' tjurkskix jazykov*, vols. 1-7-. Moscow: Nauka.
- Shagal, Ksenia, Max Wahlström, and Johanna Nichols. 2019. (Non)finiteness in clause combining: A typological survey. Presented at ALT 13, University of Pavia.
- Simon, Zsolt. 2020. Urindogermanische Lehnwörter in den uralischen und finno-ugrischen Grundsprachen: Eine Fata Morgana? *Indogermanische Forschungen* 125:239-266.
- Sipos, Mária. 2002. Az obi-ugor alapnyelv lexikális innovációi I. *Nyelvtudományi Közlemények* 99:7-56.
- Smirnov, O. V. 2012. Substratnaja mansijskaja toponimija na territorii bylovo proživanija mansi. *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). *Voprosy onomastiki* 2013:2.7-59.
- Sokal, Robert R. 1988. Genetic, geographic, and linguistic distances in Europe. *PNAS* 85.1722–1726.
- Spyrou, Maria A., Tukhbatova, R. I., Wang, C.-C., Andrades Valtueña, A., & Krause, J. 2018. Analysis of 3800-year-old *Yersinia pestis* genomes suggests Bronze Age origin for bubonic plague. *Nature Communications* 9.2234.
- 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.
- Tambets, Kristiina, Bayazit Yunusbayev, Georgi Hudjashov, Anne-Mai Ilumäe, Siiri Roots et al. 2018. Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. *Genome Biology* 19.139.
- Toporov, V. N. and O. N. Trubačev 1962. *Lingvističeskij analiz gidronimov verxnego Podneprov'ja*. 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 *A linguistic map of prehistoric northern Europe*, ed. Riho Grünthal and Petri Kallio, 185–200. Helsinki: Finno-Ugrian Society.
- Warnow, Tandy. 2018. *Computational phylogenetics: An introduction to designing methods for phylogeny estimation*. Cambridge: Cambridge University Press.
- Wessa, P. 2020. Free statistics software, Office for Research and Development, version 1.2.1. <https://www.wessa.net/>
- 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*. Leningrad: Nauka.
- Zajceva, M. I. 1993. Vepsskij jazyk. V. N Jarceva, Ju. S. Eliseev, and K. E. Majtinskaja, eds., *Jazyki mira: Ural'skie jazyki*, 36-48. Moscow: Nauka.
- Zeleneev, Ju. A. and O. V. Zelencova. 2012. Средневековая мордва по археологическим данным (Medieval Mordva based on archaeological data). *Povolžskaja arxeologija* 2012:1.37–47.
- Zhivlov, Mixail A. 2014. Studies in Uralic vocalism III. *Journal of Language Relationship* 12.113-148.
- Zhivlov, Mikhail A. 2015. Neindoevropskij substrat v finno-volžskix jazykax. *X Čtenija pamjati S. A. Starostina*.
Online:
https://www.academia.edu/11689161/https://www.academia.edu/11689161/Неиндоевропейский_субстрат_в_финно-волжских_языках
- Zhivlov, Mixail A. 2018. Istoričeskaja fonetika i vnutrennjaja klassifikacija ural'skix jazykov. Presented at XIII tradicionnye čtenija pamjati S. A. Starostina, 22-23.3.2018. Moscow: Institut vostočnyx kul'tur i antičnosti RGGU.

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

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

The family is now increasingly being dated at about 4500 years old based on the following evidence: (1) The Indo-Iranian contact episode (§2.2), firmly dated at about 4000 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 sufficient to produce distinct daughter languages from one ancestor. By this reckoning the internal divergence of Proto-Uralic began no later than 4500 BP.² The traditional and still widely held view is that the family is older, perhaps 6000 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 4500 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. 2000-2500 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-drastring 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, for which 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:

a. 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 the more recent chronology.

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

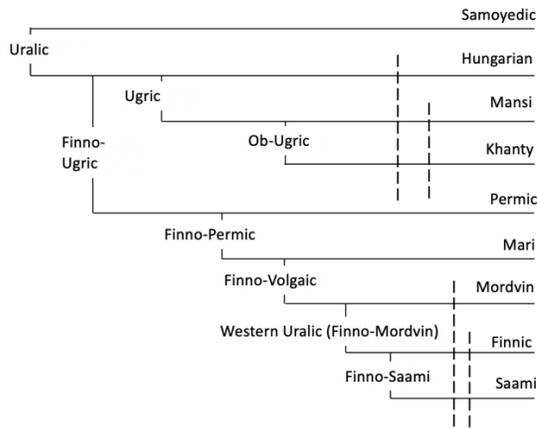


Figure S1.1. Traditional west-branching family tree (after Syrjänen et al. 2013:327, Korhonen 1981:27). For dashed lines see text below.

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

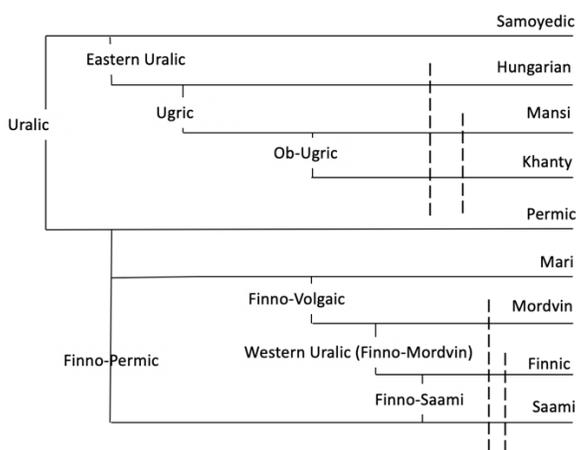


Figure S1.2. Same traditional tree with upper branching following J. Häkkinen 2007, 2009. Dashed lines as for Figure S1.1.

b. A flat structure (a.k.a. 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. Dashed lines as above.

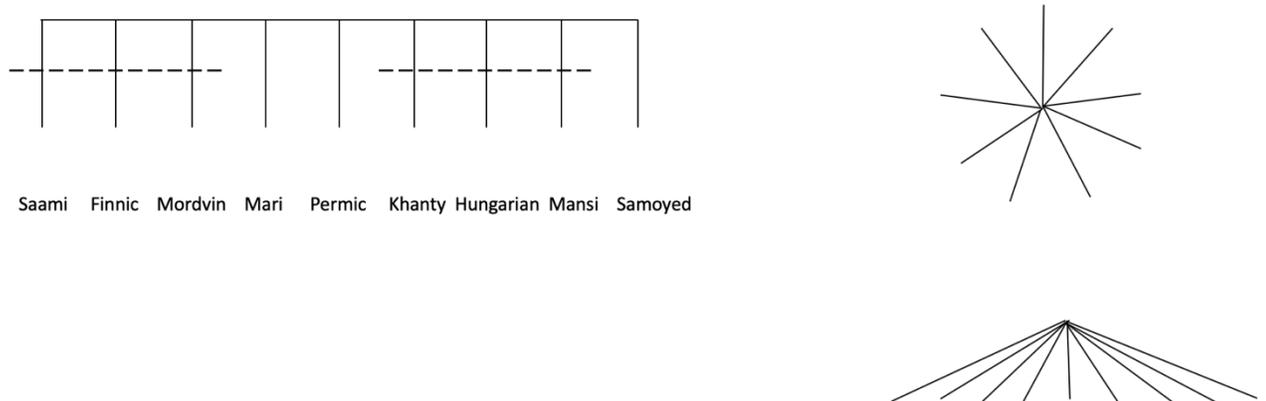


Figure S1.3. Rake tree (left) and star phylogenies (right), showing simultaneous separation of all branches. The three diagrams are phylogenetically equivalent. Dashed lines as for Figure S1.1.

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, Maurits 2018 (these three using cognate lexical material), Dellert 2018 (using very large numbers of resemblant but not necessarily cognate lexemes); Author subm. (using typological characters, with or without cognates). Higher-level structure along fairly traditional lines can also be inferred from Dellert 2018 and perhaps Author subm., 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 FU *ńoma-la 'hare' ((Hungarian nyúl, North Saami njoammil) (Janhunen 2009:67). Samoyedic has only unsuffixed Proto-Samoyedic *ńama. Also, Finno-Ugric shares the numerals '3', '4', and '6', absent from Samoyedic.

Ugric vs. Finno-Permic: The numeral '2', PU *kekta, is one of only two numerals reconstructable for PU. In Finno-Permic, '2' 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. '7' 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: Finnic 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 *i and *a. In these two sets Finnish, North Saami, and Erzya Mordvin have back vowels corresponding to Mari front vowels:

<u>PU</u>	<u>Finnish</u>	<u>North Saami</u>	<u>Erzya</u>	<u>Mari</u>
*ñili 'arrow'	nuoli	--	*nali > nal	nöl(ə)
*ipti '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 *i-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 'ten' (Finnish *kymmenen*, Erzya *kemen'*) 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, 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).

⁴ Unless the elative in the possessive declension preserves a lone trace of Permic *-s-ta (IN-ELA): Udmurt *busy* 'field' : *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.

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 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 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 & 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 relationship.⁵

There is no single standard way to reflect intra-family contact episodes in family trees. Above we have adopted 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 Finno-Saamic, or all of Western Uralic, are regarded as 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.

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

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).

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	<i>Pinus</i> spp.?		Samoyedic, Ugric	Any homeland
*pVwi	tree			All but Khanty, Mordvin, Saamic	Any homeland
*ńulka	fir	<i>Abies</i> spp.	Boreal forest	Selkup, Ob-Ugric, Permic, Mari	Any homeland
*kVwsi	spruce	<i>Picea</i> spp.	Boreal forest	All but Hungarian	Any homeland ⁶
*sįksa	Siberian pine (Siberian stone pine, cedar pine)	<i>Pinus sibirica</i> (<i>Pinus cembra sibirica</i>)	Western Siberia	Samoyedic, Ugric, Permic	Food plant. East of Urals
*pVjV	willow	<i>Salix</i> spp.	Widespread	All but Mansi, Hungarian, Mari	Any homeland
*kVji	birch	<i>Betula</i> spp.	Widespread	All but Ugric, Permic	Any homeland
*d'įmi	bird cherry (chokecherry, hackberry)	<i>Prunus padus</i>	North of steppe; to 60° in Siberia, to tundra in Scandinavia	All but N Samoyedic	Any homeland; food plant east of Urals.
*mura	cloudberry	<i>Rubus chamaemorus</i>	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 from the Urals east, 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 and Permic is barely within this range). 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: <http://www.luontoportti.com/suomi/en/puut/bird-cherry>). 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, '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).

Cloudberry is 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*),⁷ *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, '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 4000 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)⁸ 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, 1975. 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, *wäškä, reconstructs to PU (Table S2.2; Appendix 1), and that term has some phonological irregularities suggesting a post-PU areal Wanderwort: Aikio 2020:§1.6.2, 2015:42-43.

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 2.2 illustrates the semantic range and gives examples of such compounds.

⁷ Another Eurasian honey-making species, *Apis cerana*, has a south and southeast Asian range well south of Siberia.

⁸ 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.

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 **wäškä* is reconstructible with the ambiguous meaning 'copper, metal' as 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. 4200 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 4200 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 **wäškä* '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' TNenets ješa		
Khanty			wäx 'iron'; ...+wäx in 'copper', 'steel', 'silver'
Mansi			āt+wəs 'lead'
Hungarian		vas 'iron'	
Permic	Udm veś, azveś 'silver'		
Mari			
Mordvin	E uške, M uškä 'wire, chain',		
Finnic		Finn vaski 'copper, bronze'	
Saamic	NSaa veaiki 'copper'		

Table 2.3. Basic cognates descended from PU **wäškä* 'metal, copper'. Sources: MszFE 1: 169–170, 3: 675–677, SSA 3: 416, SW 175.

Saamic	<i>*weške</i> 'copper'
Finnic	<i>*waski</i> : <i>*waske-</i> 'copper'
Mordvinic	<i>uške</i> (~ <i>viške</i>) '(metal) chain'
Mari	(-) <i>waž</i> ~ (-) <i>wož</i> '(metal) ore'
Permic	- <i>veś</i> ~ - <i>iš</i> 'metal'
Khanty	<i>wäx</i> 'metal, iron'

Mansi	- <i>weś</i> ~ - <i>küš</i> 'metal'
Hungarian	<i>vas</i> 'iron', <i>ez-üst</i> 'silver'
Samoyedic	* <i>wesä</i> 'iron, metal'

Comments on the forms

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

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 **tälve* > 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 **waski* or **wäskä*.

Mordvinic: Mordvinic (Erzya) *uške* ~ *viškä* and (Moksha) *uškä* 'wire, chain' (MdWb 2485) could presuppose an earlier shape of the type **wVškä*, if the Erzya western dialect variant *viškä* 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 **wäskä* ~ **weskä* (> ? **üškä*) > *viške* ~ *uške*.

However, as Aikio 2015: 42 notes, E *viškä* 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 *uške*, M *uškä* cannot descend from PU **wäskä* is that, firstly, in inherited vocabulary a first-syllable **ä* should be represented 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 **waski* 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 **wäskä*.

Mari: The Mari item (Hill) *-waž* ~ (Meadow) *-wož* is attested as the second component of compounds, e.g., *kārñi-waž* 'iron-ore'. As the internal cluster **śk* yields regularly Mari (**š*)*šk*, as in **moški*- 'to wash' > Mari *mušk-a*-, a development **wäskä* > **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 (**š*)*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 **wäskä*.

Permic: The Permic reflexes are also attested only in compounds: Udmurt *az-weś* = Komi *ez-iś* 'silver', Udmurt *uz-weś* = 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 are 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 represent 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* ~ *εt-küş* (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 *-küş* 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* [vãš] (~ Old Hungarian *vos*) presupposes earlier *waskV < *waśkV, with the same development of the internal cluster as in *mos* [moš] 'to wash' < *moskV- < *moški-. 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; MSzFUE), 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/i, 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 4200 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-Mälär 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; and much later, from the Dnieper headwaters to the Moskva headwaters, the remnant Baltic-speaking Galindians (Old Russian *goljad'*) survived in 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 went entirely extinct without 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 (a.k.a. Textile Ceramic, Netted Ware) culture; to its east, along the middle Volga, the Akozino-Mälär 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).⁹ 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).¹⁰ We propose that the appearance of these Volga cultures marks the initial spread of the Uralic branch ancestors.¹¹ There was no contact with Pre-Baltic during this 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). 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 sociolinguistics 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 2016), 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 2018). 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

⁹ 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.

¹⁰ 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, paring the branch structure down to the elementary nine. This was probably true all along the early Uralic range.

¹¹ 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.

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)¹² 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 also 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 & Domokos 1978:143-147 consider the Kama area to have been the Proto-Ugric homeland, and Hajdú & 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 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 includes elements of the steppe horse cult and the circumpolar Arctic bear cult, testifying to this dual cultural heritage.

S3.3. *Yenisei*: Around 5000 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 (Allentoft 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 in the first millennium CE, a distance of ~1000 km and ~3000 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

¹² 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).

Afanasievo are two such. PU **wäskä* '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)

As was mentioned in Section 3.2, isolation by distance is a general phenomenon of decreasing number 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 idiosyncrasy 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 frequency of the 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 frequencies 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 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.¹³

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

¹³ 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).

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 S4.2.¹⁴ 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.¹⁵

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 S4.1. 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).

Distance from	Uralic etyma			Indo-Iranian loans		
	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

¹⁴ 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.

¹⁵ 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 S4.2. 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).

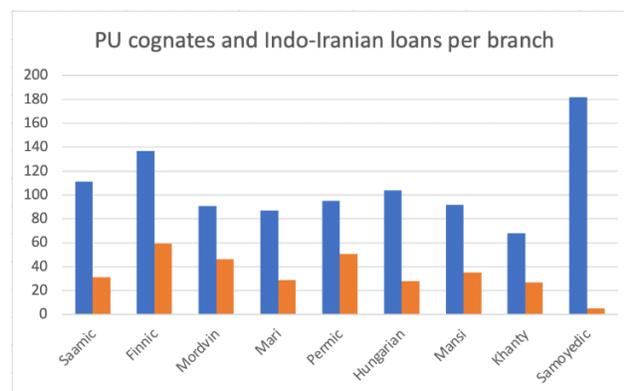
The results of the Mann-Kendall trend tests are presented in Table S4.2. The null hypothesis H_0 is that there is no trend and the alternative hypothesis H_1 is that there is a negative trend, that is, the farther the language is from the potential point of origin the fewer PU traits or I-I loans the 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.

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.)

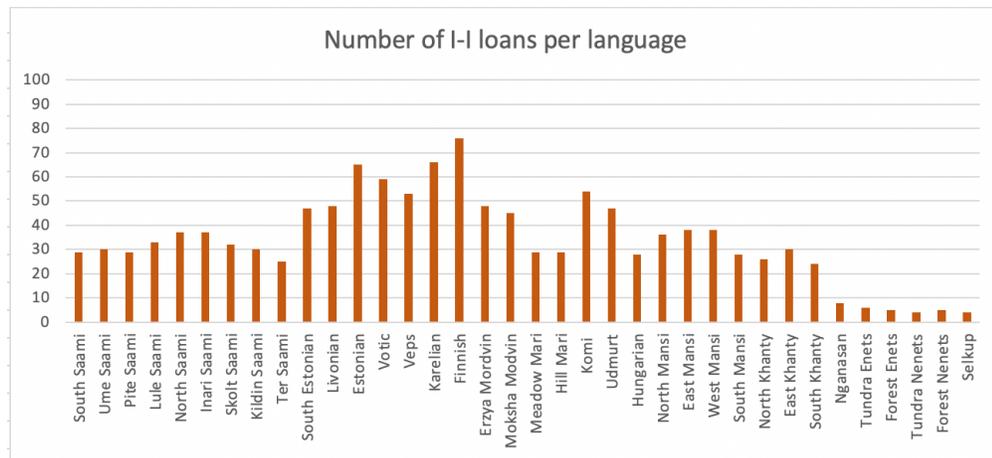
Much the same results were obtained by testing the individual languages, shown in Figures S4(b-c) and also Figure 1 in the main text.

Figure S4. Numbers of Proto-Uralic (blue) and Indo-Iranian (orange) etyma retained per 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.

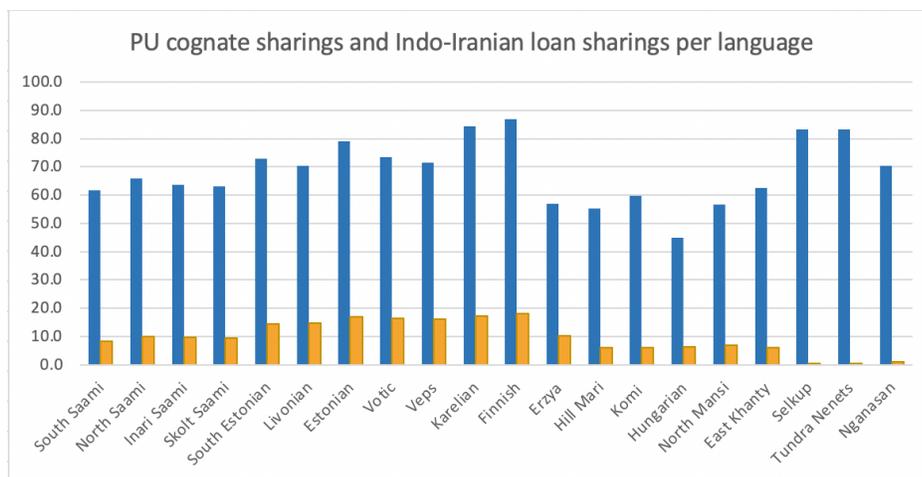
(a) Branches.



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



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



Figures S4ab show counts similar to Figure 1 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 1 in main text; see §1 for discussion).

Figure S4(c) 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 (Finnish through South Estonian 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 1 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.

S5. The Finnic and Saamic secondary westward spreads

For nearly a millennium, from the initial CU spread to not long before 3000 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. 3000 BP, and some time after that spread to central Scandinavia perhaps as early as 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 to the south and overland, north of the Gulf of Bothnia, to 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 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 900 (probably as the onset of the Medieval Warm Period made agriculture productive in the north) (Lang 2018, Saarikivi 2004a, b).

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. 1500 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.

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. Possessive 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	<u>minu-t</u>	<u>sinu-t</u>	(syncretic with genitive) ¹⁶
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	<u>minä</u>	<u>sinä</u>	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 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.

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 (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 Ugric 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 2021, 1973).

S7. Typology and the CU spread

Salient typologically eastern properties of Uralic languages, all of them traceable to PU, are these:

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.

High inflectional person (Nichols 2017). In PU and conservative languages, person is marked on verbs (argument indexation), nouns (possession), and adpositions (object indexation); in most of Europe

¹⁶ Or, with DOM, syncretic with nominative. The Finnish accusative ending *-t* and Hungarian *-d* are cognate.

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. In all of these respects, person resembles an inflectional category rather than a lexical one.

High part-of-speech flexibility in the lexicon. Many PU roots are reconstructible as nomina-verba (Janhunen 2001, 2020); 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).¹⁷

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 (Janhunen 2020:383-386; cf. also Havas 2003)¹⁸. 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 & 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 2020: §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 below), 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 *prenominal*. 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.

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

¹⁷ An orthogonal distributional pattern is a tendency for POS flexibility to decrease in strong contact situations (Foley 2017). In the Turkic and Mongolic families, whose histories are 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.

¹⁸ Janhunen (p. 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 subjective 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.

often derived (with causative suffixes). Aikio 2020 reconstructs three causative suffixes but only one detransitivizing suffix, **-w-*, glossed 'stative / automative passive' (Aikio 2020a:39), 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 (ex. S7). 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, Author subm.).

Table S7. 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- <u>ms</u> sit-INF	oza- <u>vto</u> -ms sit-CAUS-INF
(b)	Spanish	senta-r= <u>se</u> sit-INF=REFL	senta-r sit-INF

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, discussed below as cross-linguistically rare traits, also have eastern distributions.)

Typologically rare traits, again inherited and notably stable in the family, are the following. (These are infrequent and stable features in the terms of Greenberg 1978:75-76.) 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.

Dual. A dual number category is found in Samoyedic, Ob-Ugric, and Saamic, and can be reconstructed to PU as a suffix **-kV* is found in all 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.

Negative verb. A negative auxiliary verb is found in most Uralic languages, and a PU root **e-* is reconstructable (Aikio 2020) (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. Finn. *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 2020:§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 of Miestamo 2005 plus most of the languages Dryer 2013 coded as having or potentially having a negative verb and a few additional ones; the number of genera recognized grew between 2005 and 2013 and 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 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 & North Asia (the continent where all Uralic languages except Hungarian are found; for continent definitions see Bickel et al. 2017). The most general outcome is that Central & 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 often more than one standard deviation above the mean, and almost always above the mean (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 & 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.

Continent	No. languages with:				% genera with any:				% stocks with any:				% yes per stock	
	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	high	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

Table S7. 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. 2017).

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 2020). 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). 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.

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 2020: 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 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.

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.

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. (Map 1 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 1 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

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 where a catastrophe affecting one language affords an expansion opportunity to another. Other factors are often also involved, and 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 & Salmela in press). 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 & 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 & 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 & 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.