Tocharian and Samoyed: on the question of Uralic substrate influence in Tocharian

Warries, A.R.

1 Introduction

This dissertation is about the hypothesis that early Tocharian was in contact with a form of Uralic. In this introduction I will first give an overview of the overarching context in which the research of this dissertation has been conducted in section 1.1., followed in section 1.2 by some further introductory information on the Samoyed languages and their place within Uralic. The archaeological background of the main research question will be provided in section 1.3, and the methodological background of my linguistic research in section 1.4. Finally, section 1.5 will give an overview of the structure of the dissertation and introduce the most important research questions that require an answer as intermediate steps toward the final, overarching conclusion.

1.1 A Uralic piece to the Tocharian puzzle

Tocharian is in many ways an outlier among the Indo-European languages. On the eastern edge of the Indo-European linguistic area, it does not have any apparent close connections with other branches and shows an array of both archaic and innovative features. The Tocharian language group contains two closely related languages: Tocharian A (or "Agnean", "East Tocharian") and Tocharian B (or "Kučean", "West Tocharian"). The speakers of Tocharian A and B lived along the northern Silk Road in the Tarim Basin until their language disappeared around 1000 CE. The oldest Tocharian B texts date to around 400 CE, so that we have only a narrow window of approximately 600 years on this language. Tocharian A attestations begin approximately 250 years later. Further limiting our knowledge of the language is the fact that the contents of the sources is overwhelmingly Buddhist literature, and that most manuscripts are damaged. The great majority of texts are more properly characterized truly as fragments. There are approximately 9000 fragments of Tocharian B texts and 2000 of Tocharian A texts (Peyrot 2013: 1-2).

Much of the prehistory of the Tocharians is still shrouded in mystery, especially as we cast our view farther back in time. When the earliest attested texts were written around 400 CE, both Tocharian languages were already in the Tarim Basin as separate entities, and they were in contact with neighbouring languages like Middle Chinese, early Turkic and Middle Iranian languages such as Khotanese (e.g., Lubotsky & Starostin 2003; Peyrot 2018b; Dragoni 2023). The contact with Iranian languages stretches even farther back in time to the (pre-)Proto-Tocharian period, as has been demonstrated on the basis of prehistoric borrowings from Old Iranian into Tocharian (Bernard 2023). Interactions with Iranian continued in the form of contacts with Khotanese in the Tarim Basin itself (Dragoni 2023). Loanwords from another, unknown language have been related to the

language of the BMAC, which is otherwise only known in the form of loanwords into Indo-Iranian (Lubotsky 2001; Bernard 2023).

Beyond that, not much more is known with certainty about the early prehistory of the ancestors of Tocharian. However, since they spoke an Indo-European language, they must derive from the Proto-Indo-European speech community, which originated in the Pontic-Caspian Steppe (e.g., Mallory 1989; Anthony 2013; Anthony & Ringe 2015; Allentoft 2015). The potential route of the Tocharians from the Proto-Indo-European homeland to the Tarim Basin has been partially charted on the basis of archaeological and genetic evidence (e.g., Mallory & Mair 2000; Kroonen, Barjamovic & Peyrot 2018: 7–9), but much uncertainty still remains (see especially Mallory 2015).

A pertinent part of this discussion is the linguistic relationship between Tocharian and the other Indo-European languages. Anatolian is considered by a growing consensus among scholars to have been the first branch of Indo-European to split off from the protolanguage (see, e.g., Sturtevant 1933: 30, Kloekhorst 2008: 7–11; Oettinger 2013/14), and Tocharian has received a similar treatment due a number of lexical and grammatical quirks. This has given rise to the "Tocharian second" hypothesis. According to the Tocharian second hypothesis, it is thought that Tocharian was the second branch of Indo-European to split after Anatolian, based on a variety of features that appear to set Tocharian apart from the other non-Anatolian Indo-European languages (e.g., Schmidt 1992, Winter 1997, Carling 2005).

There is no consensus about the validity and applicability of each individual feature (for discussions, see Malzahn 2016; Peyrot 2019b; Friis 2024: 20–36), but if Tocharian was indeed the second branch to split from Proto-Indo-European, any Tocharian feature that happens to not be found in Anatolian could in principle represent an archaism in Tocharian that was lost or altered in the other branches of Indo-European. The directionality of a change cannot always be determined with certainty, however, and in such cases, we must remain cautious of assuming either Tocharian archaism or Tocharian innovation until a scenario can be presented that accords better with either option.

In general, many aspects of Tocharian that make it stand out are considered to be particularly innovative rather than archaic. A number of such features that appear unusual from an Indo-European perspective have been associated with language contact, particularly often with Uralic substrate influence (e.g., Krause 1951; Kallio 2001; Bednarczuk 2015; Peyrot 2019a). Uralic languages were historically spoken in southern Siberia along part of the route the Tocharians may have taken to the Tarim Basin, so that a demonstration of Tocharian-Uralic contact can provide a further anchor point for our understanding of early Tocharian prehistory.

A particular relationship with the easternmost branch of Uralic, the Samoyed branch has been advanced by Peyrot (2019a). Seeing as early Samoyed was spoken in an area of

southwestern Siberia almost directly north of the Tarim Basin (e.g, Janhunen 1998: 457), this would fit well in geographical terms. Closer investigation of the contact hypothesis is the topic of this dissertation. I will scrutinize the various arguments and make explicit what we are comparing and what we can base ourselves on, in order to further our understanding of this hypothetical language contact and to see if part of the Tocharian puzzle can indeed be filled in with Uralic pieces.

1.2 Samoyed and the other Uralic languages

There are nine uncontroversial branches of Uralic: Saami, Finnic, Mordvin, Mari, Permic, Mansi, Khanty, Hungarian, and Samoyed. According to the traditional phylogeny, Uralic is as a generally consistent westward-branching, language family. In this model, Samoyed, the easternmost branch, is more distant to all the other branches than the other branches are to one other (e.g., Janhunen 1981; 2000: 59–60). However, according to recent scholarship, the old model of binary splits cannot be taken for granted (K. Häkkinen 1984; Salminen 2002), and no new model has been accepted as an adequate replacement yet (see Aikio 2022a: 3–4). The phylogenetic position of Samoyed thus remains debated, like that of Tocharian.

While the phylogenetic reality of a primary split between Samoyed and Finno-Ugric is uncertain, it remains useful to differentiate between these two groups, and it is possible to use the term Finno-Ugric to refer to all branches of Uralic except for Samoyed (cf. e.g., Grünthal et al. 2022: 492). Further grouped together, perhaps in a true genealogical sense, are the "Ugric" languages Mansi, Khanty and Hungarian, of which Mansi and Khanty are together referred to as "Ob-Ugric". It is debated to what extent the similarities between these languages are due to shared innovation or due to parallel innovation and later convergence.

The Samoyed branch itself is usually divided into eight languages: Nganasan, Tundra Enets, Forest Enets, Tundra Nenets, Forest Nenets, Selkup, Kamas, and Mator. Of these, Kamas and Mator are extinct. Especially Mator is attested very incompletely and we have onlly little information on morphology or syntax (Helimski 1997 is a comprehensive treatment of the Mator material). Kamas was documented more extensively, and its grammar is known much better (see, e.g., Klumpp 2022). Selkup is subdivided into dialects in up to five main groups: northern (Taz) Selkup, central (Tym) Selkup, southern Selkup with the Middle Ob and Upper Ob dialects, and Ket Selkup (see Klumpp & Budzisch 2023: 898 with references). Tundra and Forest Nenets are closely related, but distinct, and the situation with Tundra and Forest Enets is of a similar nature (see, e.g., Mus 2023 on Nenets and Khanina & Shluinsky 2023 on Enets for more details).

The Nenets and Enets languages together with Nganasan are conventionally known as the northern Samoyed languages, while Selkup, Kamas, and Mator make up the

southern Samoyed languages. Kamas and Mator are sometimes referred to together as Sayan Samoyed, so named for the Sayan Mountains where they were spoken (Janhunen 1998: 458–459). These subgroupings are perhaps areal rather than genealogical, and the true phylogenetic structure of the Samoyed branch is still uncertain. The divide between northern and southern Samoyed languages is sometimes taken as a true phylogenetic divide, but it has also been argued that Nganasan and Mator, the northernmost and the southernmost language, represent particularly early offshoots (Janhunen 1991; Janhunen 1998: 459). The issue requires further research.

The traditional position of Samoyed as the first branch to split off from Proto-Uralic has led to its image as the odd one out, or "the Anatolian of Uralic" from an Indo-European point of reference. However, several important aberrant features of Samoyed concern the lexicon, and there it at least partially concerns Samoyed-specific innovations. For example, the word for 'hand' can be reconstructed as PU **käti* (illustrated in 1. below), but Samoyed instead has PS **utå* (in 2. below). This cannot be regarded as a meaningful isogloss, however, since PS **utå* is indicated by its phonology to be innovative: the Samoyed vowel combination **u-å* does not occur in inherited vocabulary (see subsection 3.6.1 for more details). The aberrant word for 'hand' thus has no bearing on the phylogenetic position of Samoyed. Many other differences in the basic vocabulary of Samoyed and Finno-Ugric may well be of a similar nature. The reputation of Samoyed as an outlier is also partially the result of its later documentation and the research history (Aikio 2022a: 4; Saarikivi 2022: 48).

1. PU **käti* > (e.g.) SaaN *giehta*, Fi. *käsi*, MdE *ked'*, MariE *kid*, Hu. *kéz*
2. PS **utå* > (e.g.) Ng. *d'ütü*, EnT *uδa*, NeT *ŋuda*, NeF *ŋŭta*, SkTaz *utį*, Mt. *uda*

A further striking difference between the Samoyed and Finno-Ugric languages can be found in the numeral system. Only the words for 'two (2)' and 'seven (7)' seem to be shared between Samoyed and the other Uralic languages, with the Samoyed word for 'ten (10)' additionally corresponding to 'five (5)' elsewhere. This is illustrated in Table 1.1.

Table 1.1: The Samoyed and Finno-Ugric numeral systems. Especially among the Finno-Ugric languages, there are a number of unclear phonological correspondences (see Aikio 2022: 25 for the reconstructions). The formations of '8' and '9' vary even within Finno-Ugric.

	Samoyed	Finno-Ugric
2	**kitä*[(a)]	~ **kAktA*
3	**nakur*	≠ **kolmi/kulmi/kurmi*
4	**tättə*	≠ **neljä*
5	**sə̂mpə̂laŋkə̂*	**wij(i)t(t)i*
6	**mə̂ktut*	≠ **kuw(V)t(t)i*
7	**säjt³wə*	~ **ćäjćimä/ćäććimä*
10	**wüət*	≠ **loka*

(a) The correspondence between FU **-kt-* and PS **-t-* is regular, but the vowels are unclear, even within Finno-Ugric.

On the basis of this discrepancy, Janhunen (2000: 60–61) has argued that Proto-Uralic did not have these numerals at all, and that they were created in Samoyed and Finno-Ugric independently. Since the Finno-Ugric numerals are related, this would be a significant shared innovation. However, as pointed out by Aikio (2002: 31), the Samoyed numerals for 'three (3)', 'four (4)' and 'six (6)' can be identified as innovative within Samoyed on the basis of their phonology. The second-syllable **u* of **nakur* and **mə̂ktut* would be difficult to derive from Proto-Uralic, and the internal **-kt-* of **mə̂ktut* should have regularly developed to PS **-t-* if this word were inherited from Proto-Uralic (see also 3.5.2). Similarly, the geminate of **tättə* should have been simplified to a singleton PS **-t-*. Furthermore, 'five (5)' looks to be a compound of some sort and may be a relatively recent creation, displacing or replacing the original word for 'five (5)' as found in the other Uralic languages as that came to mean 'ten (10)'. The Samoyed numeral system is thus demonstrably innovative, while the Finno-Ugric languages may at least in part preserve a more archaic situation (Aikio 2002: 31). As in the case of the word for 'hand', this does not directly support an early split of Samoyed from Proto-Uralic, it only points to strong lexical influence from other languages on Samoyed in its position at the eastern periphery of the Uralic linguistic area (Aikio 2002: 31–32).

The case for an early split of Samoyed is better made on the basis of examples like the word for 'hare', which contains an additional suffix in the Finno-Ugric languages that is absent in Samoyed. The Proto-Samoyed form is *ńåmå* from PU **ńoma* (in 3.), while other Uralic languages reflect **ńoma-la* instead (in 4.) (Janhunen 2007: 221).

3. PU **ńoma* > PS **ńåmå* > e.g., Ng. *ńomu*, NeT *nyawa* 'hare'
4. PU **ńomala* > e.g., SaaN *njoammil*, MdE *ńumolo*, Hu. *nyúl* 'hare'

A potential phonological argument may be found in the reflexes of PS **o* opposed to FU **u* in roots of the shape **C{o/u}Ci*, depending on which is the archaic state (see the discussion in 3.6.3). A systematic investigation of all the arguments may result in the conclusion that was Samoyed indeed the first to split off, but this remains a task of future research.

The Uralic language family is spread out over a large area of Eurasia, from the Scandinavian peninsula in the west to the Yenisei River in the east. Along their entire distribution the Uralic languages border on Indo-European languages, which have influenced and continue to influence each branch of Uralic in various ways. Contact between Saami and Germanic or between Finnic and both Germanic and Baltic is well-established and can be extensively shown with both ancient and modern loanwords (e.g., Aikio 2006b; Jakob 2023). There has also been much contact between Uralic languages and Indo-Iranian languages, dating far back to some of the earliest stages of Uralic that can be reconstructed (Holopainen 2019). On the easternmost fringes of the Uralic language family, the Samoyed languages have not been impacted as heavily with Indo-Iranian influence, although some loanwords have been suggested (Janhunen 1983, see also Holopainen 2019). It is as yet unclear how the relatively low number of Iranian loanwords in Samoyed is to be understood in relation to its phylogenetic position. Loan contacts with early Tocharian have also been proposed, both into Samoyed specifically (Janhunen 1983, Napol'skikh 2001, Kallio 2004) or even into Proto-Uralic (Bjørn 2022). It is difficult to verify the Tocharian origins of many of these words, however, as will be discussed in chapter 9 of this dissertation.

The last word on the early prehistory of Uralic has not yet been written, and the search for the precise location of the Proto-Uralic homeland continues (recently Grünthal et al. 2022; Saarikivi 2022; J. Häkkinen 2023). However, the discussion has narrowed to the region ranging from the western slopes of the southern or central Ural Mountains to west or central south Siberia. Genetic evidence supports the Siberian origins of Uralic (Tambets et al. 2018, Saag et al. 2019, Childebayeva et al. 2024). Archaeological correlates for Uralic and the dispersal of the Uralic languages have recently been sought in the Seima-Turbino phenomenon of around 2200–2000 BCE (dates per Marchenko et al. 2017), which is spread out across much of the known southern range of the Uralic linguistic area (Kallio 2006: 16–17, Parpola 2012: 156–160; Grünthal et al. 2022, and Childebayeva et al. 2024 for a genetic perspective).

Where pre-Proto-Uralic should have been located before that is another open question. For the Samoyed branch especially, longstanding associations with parts of the area to the north of the Tarim Basin are assumed (Janhunen 2022). If the Proto-Uralic homeland should be located closer to the Ural Mountains, however, we must take into account the time needed for pre-Proto-Samoyed to move farther to the east. In this way,

the interpretation of the Uralic homeland interacts with the question of Tocharian-Uralic contacts and their location in time and space.

The split of Proto-Uralic used to be estimated around 4000 BCE, but recent scholarship has shifted the date closer to 2500 BCE (Kallio 2006; Grünthal et al. 2022; J. Häkkinen 2023). The shallower time depth is based in part on the general abandonment of the idea that Samoyed represents a significantly early offshoot to the family tree, and in part on a reappraisal of the loan contacts between early Uralic and early Indo-Iranian (Holopainen 2019).

The potential of loan relationships between Proto-Indo-European and Proto-Uralic has shaped many analyses of the positions of these linguistic entities in time and space (e.g., Mallory 1989: 147–149; Anthony 2007: 74–76; J. Häkkinen 2012; 2023), but a number of the arguments for such early contacts between the two protolanguages themselves should be treated with caution or might have to be abandoned altogether, in accordance with more recent insights (Simon 2020; Holopainen 2021; Aikio 2022a: 25–26). The dating and position of early Uralic is thus intimately connected with our ideas about early Indo-European, and in some ways dependent on it. As our understanding of the relationships between these two language families changes, so do our hypotheses regarding their ancient past. After all, the Uralic and Indo-European puzzles are part of the same, larger puzzle of Eurasian prehistory over the last few millennia.

1.3 Archaeological background

Our linguistic investigation into the past cannot be separated from the physical world. The speakers of the ancient languages reconstructed with the comparative method have left traces here and there in the archaeological record, and while an exact identification of a linguistic entity with an archaeological culture is far from straightforward, it is possible to find correlates that explain (aspects of) the relevant linguistic facts. Tocharian represents a branch of Indo-European on the eastern periphery of the language family. Since Tocharian is not known to form a sub-branch with any of the other Indo-European branches (Peyrot 2022: 83–84), its early prehistory needs to be considered separately starting immediately from its separation from Proto-Indo-European. In accordance with the Tocharian second hypothesis, an early spatial separation of Tocharian from the rest of the early Indo-European linguistic area might be expected.

According to current insights, the speakers of Proto-Indo-European were associated with the Yamnaya ("pit grave") Culture located in the Pontic-Caspian steppe. Tocharian presence in the Tarim Basin, far to the east, should thus be the result of a prehistoric migration from this area. The Yamnaya Culture dates to around 3300–2600 BCE (dates per Morgunova & Khokhlova 2013) and consists mainly of burial sites. With advances in

ancient DNA research, the associated genetic signatures have been uncovered as well (e.g., Allentoft et al. 2015, Damgaard et al. 2018). This provides additional insights into the connections between Yamnaya and related archaeological cultures. The southern Siberian Afanasievo Culture of around 3100–2500 BCE (dates per Poliakov et al. 2019) is closely related to the Yamnaya culture and shows a quite exact genetic match (Damgaard et al. 2018; Narasimhan et al. 2019). Dating to around 2800 BCE, there is evidence of an extension of Afanasievo southward into the Dzungar Basin north of the Tarim Basin (Zhang et al. 2021). The archaeological and genetic connections between Yamnaya and Afanasievo have led to an association between the Afanasievo Culture and the Tocharians (Mallory & Mair 2000; Anthony 2007; Kroonen, Barjamovic & Peyrot 2018, etc.), which we can refer to as the "Afanasievo-Tocharian" hypothesis.

If a direct descent of the Afanasievo Culture to people in the northern Tarim Basin could be shown, that would strengthen the Afanasievo-Tocharian hypothesis. Genetically, this connection is not evident. The characteristic Y-haplogroup represented by most individuals of the Afanasievo culture is R1b1a1a2. While the Shirenzigou site in the Tianshan mountains on the north-eastern side of the Tarim Basin (see Figure 1 below) also shows this Y-haplogroup, the relevant samples are dated to 200–100 BCE (Ning et al. 2019). This could indicate a continued southward movement from Afanasievo and thus represent a Tocharian migration to the northeast of the Tarim Basim, but with more than 2000 years, the difference in time is too great to be certain of a direct connection.

Furthermore, more ancient genetic samples from the Tarim Basin Xiaohe cemetery (see Figure 1 below) have not shown any particular association with Afanasievo. Analysis of some individuals from the oldest layers of Xiaohe dating to around 2000 BCE were more consistent with the ancient DNA of the Andronovo Culture, whose area extended from the west to the Altai area following both the Afanasievo Culture and its successor, the Okunevo Culture (Li et al. 2010; Mallory 2015: 44). These individuals had the Y-haplogroup R1a1a, rather than Afanasievo-associated R1b1a1a2, and also showed signs of extensive admixture with southern Siberian populations (Li et al. 2010). A more recent analysis of ancient Xiaohe genetic material instead characterized their overall genetic signature as autochthonous to the Tarim Basin, showing admixture between Ancient North Eurasians and Northeast Asians, rather than a connection with populations associated with Indo-European languages, so that a connection with early Tocharian becomes uncertain (Zhang et al. 2021).

Thus, apart from the rather late Shirenzigou evidence of 200-100 BCE (Ning et al. 2019), the trail from the Afanasievo culture in the direction of the Tarim Basin only extends to around 2800 BCE in the Dzungar Basin (Zhang et al. 2021). Linguistic evidence for early Tocharian in the Dzungar basin has been roughly estimated to around 1000 BCE, almost two millennia after Afanasievo genetics enter the area, in the form of contacts of Proto-Tocharian with Old Iranian coming from the steppe (Bernard 2023: 251–252).

Whether the Proto-Tocharian speakers in the Dzungar Basin at that time represent the initial Afanasievo migration of 2800 BCE is uncertain, however. They might also represent a later, otherwise undetected migration into the area from the west, or a migration from the northern Afanasievo area later than the one around 2800 BCE currently evidenced by the genetic record.

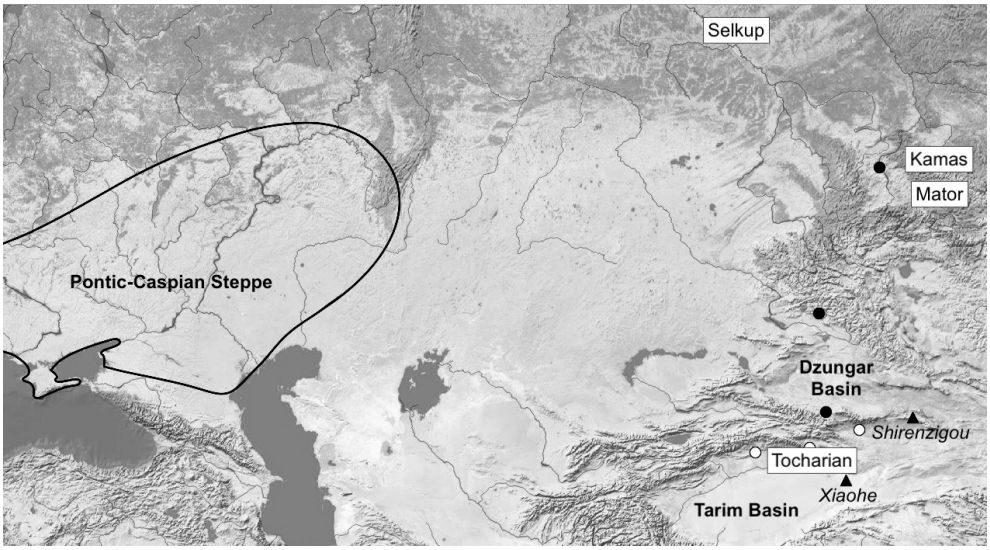

Figure 1: Geographical locations of Afanasievo sites sampled by Zhang et al (2021). marked with black dots; main find spots of Tocharian texts marked with white dots; Yamnaya Culture area approximated with a black line; location of other relevant elements marked with labels.

One great advantage that the Afanasievo-Tocharian hypothesis has is that it accounts for the phylogenetic distance between Tocharian and Indo-Iranian: with its early date starting around 3300 BCE, the establishment of the Afanasievo culture clearly precedes the Indo-Iranian migration to the steppe around 2000 BCE (Kroonen, Barjamovic & Peyrot 2018: 7–9). While this makes the Afanasievo culture the best fit for Tocharian on the basis of currently available evidence, the match is not in all aspects perfect one. Kroonen et al. (2022: 31) point out that Tocharian A *āre** 'plough (n.)' shows that Tocharian shares in the semantic innovation of the root PIE *h_2erh_3- from 'grind, crush, break up' to 'plough'. This development is found in all other branches where this root is attested except for Anatolian, which preserves the original, non-agricultural meaning (Kloekhorst 2008: 9–10). Interestingly, TA *āre** appears to be an independent formation within Tocharian (Peyrot 2018b: 262–263), unrelated to the instrument noun *h_2erh_3-*tro*- found elsewhere. Still, the specialized meaning of the root could point to the agricultural character of pre-Proto-Tocharian society. The etymology for TB *pīsäl*, TA *psäl* 'chaff', proposed by Kroonen et al. to derive from a PIE **peis*- 'grind, thresh' (2022: 16) also

suggests continued agriculture in pre-Proto-Tocharian society. Since there is no evidence for farming associated with the Afanasievo Culture, Kroonen et al. (2022: 33) argue that it may be not as good a match for early Tocharian as commonly believed. On the other hand, many other agricultural terms in Tocharian were adopted more recently from neighbouring languages (Peyrot 2018b), so that earlier pre-Proto-Tocharian agriculture was perhaps more limited or different from the farming practices of the eventual Tocharians themselves.

Since the palaeolinguistic evidence against an association with the Afanasievo Culture is so far limited to the semantic innovation of a single root, *$*h_2erh_3$-*, and potentially the etymological identification of TB *pīsäl*, TA *psäl* 'chaff', a complete dismissal of the Afanasievo-Tocharian hypothesis is probably premature. However, it is good to keep in mind that not all the evidence fits perfectly. It is theoretically possible that there were other impulses of migration from the western steppe to southern Siberia that are simply not (yet) evidenced at this moment in time, and that we therefore cannot weigh against the Afanasievo Culture. The Afanasievo-Tocharian hypothesis thus remains the most appropriate explanation for the presence of Tocharian in the Tarim Basin at the present moment, albeit in part only due to a lack of known alternatives.

This dissertation is concerned with the Afanasievo-Tocharian hypothesis in two ways. First, a Tocharian-Uralic contact hypothesis can only be understood if the ancestors of the Tocharians migrated via southern Siberia, like individuals coming from the (pre-)Yamnaya Culture to form the Afanasievo Culture. Evidence for Tocharian-Uralic linguistic contact would thus support this migration route. Second, the time depth of any linguistic contact in southern Siberia should match the time depth of the migration through southern Siberia. This latter point will remain difficult to specify in absolute terms, but nevertheless needs to be taken into account. If linguistic contacts between the ancestor of Tocharian and speakers of Uralic could be demonstrated, that would provide a further link on the migration of the Tocharians, locating them to the north of the Dzungar Basin at some point in the past, at least in partial accordance with the "Afanasievo-Tocharian" hypothesis.

1.4 Methodological considerations

Demonstrating a prehistoric language contact situation is a complicated matter. The traces that language contact can leave behind are varied, and may apply to any aspect of the linguistic system. Furthermore, since languages will change over time in any case, it is difficult to be certain whether particular changes were indeed due to influence from another language, or simply result from internal linguistic changes. Aside from these difficulties, in a prehistoric language contact situation, the arguments are necessarily based on a comparison between the reconstructed stages of two languages, which can

only be accessed indirectly. Due diligence must be paid to the chronologies of changes in the two languages separately to avoid anachronistic comparisons.

Broadly speaking, language contact may leave two types of traces: loanwords and structural influence. The establishment of language contact in the distant past is the clearest in the first case. Borrowed lexical material can constitute a relatively unambiguous trace of influence of a donor language, on the recipient. If the existence of loanwords can be proven to a satisfactory degree, a contact hypothesis in even the deep past may be widely accepted. For example, this is the case for the early contacts between Indo-Iranian and especially Finno-Ugric Uralic (Holopainen 2019), and for the early contacts between Tocharian and Iranian (Bernard 2023).

Without the concrete evidence from loanwords, prehistoric language contact is not easy to prove. The absence of loanwords does not, however, provide decisive proof against prehistoric contact (Thomason 2001: 91–92). Absence of evidence is not evidence of absence, and moreover, there are certain circumstances that can lead to a dearth of loanwords. For instance, it may be the case that the loanwords that were there have been replaced by newer (loan)words since they entered the language. This is especially likely when the contact would have taken place in the very distant past. In the case of an incompletely attested language like Tocharian, an additional problem could that the relevant words are simply not found in the available sources, either due to accident, or because they belonged to a register that was not written down. For example, words in the semantic domain of local plants and animals, where borrowings from a substrate language are likely (Thomason & Kaufman 1988: 39), may have fallen out of use due to changes in the environment when the Tocharians migrated into the Tarim Basin and settled there.

More significantly, not all types of linguistic contact are likely to leave traces in the form of loanwords. In the case of substrate interference as caused by language shift (Thomason & Kaufman 1988; Thomason 2001), the main effects of language shift may be found in the phonology and the grammar, rather than in the lexicon. A group of people that shift from one language to another, the target language, may take with them a foreign pronunciation and grammatical patterns in their version of the target language. A number of features native to the original language of a shifting population may end up being introduced into the speech community at large, thus altering the target language (Matras 2009: 57–58).

When loanwords are almost entirely absent, our investigation of language contact can only be carried out by looking at structural features. Thomason has put forward a set of requirements in the form of questions, according to which such a contact hypothesis can be tested (Thomason 2001: 93–95). Based on the result of this test, we may evaluate the prehistoric contact hypothesis at hand.

1) Is there a plausible historical scenario according to which the two speech communities were in contact, and was the type of contact conducive to contact induced change in one direction or the other? In this case, we do not have access to any information on the prehistorical contact in the case of the Tocharian-Uralic contact hypothesis, so that this particular question cannot be answered in detail. We can only argue that contact between the two groups was plausible (see section 1.3), but the exact nature of this plausible contact remains unknown.

2) Are there multiple features in the receiving language under investigation that can be reasonably connected to the suspected source language? If the two languages share only a single feature, the probability that this similarity came about as a result of chance is higher, and the argument that they resulted from a contact situation between precisely these languages is correspondingly not compelling. The more features can be shown to be shared, the more convincing the argument for contact-induced change will be.

3) Is the feature innovative in the target language, i.e., not inherited from its ancestor? If one or more of the features under investigation turn out to have been originally present in the target language before contact, it cannot have been the result of contact. Whether a feature is archaic or innovative in the target language can only be established by comparison with related language, preferably outside the same linguistic area.

4) Conversely, we must determine that the feature was present in the source language before contact, i.e., that it is inherited from the ancestor of the source language. If a feature is innovative in the source language, it cannot have been passed on to the target language in the distant past before it had come into existence.

In this dissertation, I will attempt to answer these questions for the hypothesis that pre-Proto-Tocharian has undergone contact induced change due to early Uralic, particularly pre-Proto-Samoyed. The choice to focus on the Samoyed branch of Uralic rests on previous research pointing to contact of early Samoyed with early Tocharian (Janhunen 1983, Kallio 2001, Kallio 2004, Peyrot 2019a). Comparison with pre-Proto-Samoyed also provides a bounded linguistic entity, defined at the most recent end by the reconstruction of Proto-Samoyed, and most distantly by Proto-Uralic. This is entirely parallel to the way that pre-Proto-Tocharian is bounded by Proto-Tocharian and Proto-Indo-European. Since the absolute time depths involved are unclear, it is possible that pre-Proto-Uralic could provide a potential alternative candidate for early contact with pre-Proto-Tocharian, but at present, a reconstruction of pre-Proto-Uralic in any detail is not yet possible. A contact hypothesis between pre-Proto-Tocharian and pre-Proto-Uralic is thus difficult to properly investigate in the same way that a contact hypothesis between pre-Proto-Tocharian and pre-Proto-Samoyed can be investigated. It is furthermore possible that early contacts between Tocharian and Uralic involved an extinct branch of Uralic, either a para-Samoyed or perhaps even a para-Uralic branch. We have no direct access to either, so that pre-Proto-Samoyed must serve as the only

available proxy at this moment. That being said, it is in some cases impossible to meaningfully distinguish between Proto-Uralic and early Samoyed and I will use the label Uralic/Samoyed to reflect this.

The question in (1) above has been satisfied as much as currently possible: contact between an early form of Tocharian and early Uralic/Samoyed can be understood in accordance with what we know about the respective language groups and associated archaeological and genetic information (see section 1.3). Since further elaboration on this point is not feasible due to the great time depth and the state of the non-linguistic evidence at this moment, the linguistic investigation may instead inform our understanding of the prehistoric contact scenario if it is judged strong enough.

The breadth of features investigated in this dissertation is in accordance with (2) and will be further elaborated in the next subsection. In order to establish whether the features are really shared between the receiving and source languages, reconstructions of the relevant pre-Proto-Samoyed and pre-Proto-Tocharian features must be undertaken independent from one another. This goes for every aspect of the comparison: phonology, morphology, and syntax. Only then will it be possible to say if the independent histories of the two language families suggest an intersection and an interaction at some point in time in accordance with points (3) and (4).

1.5 Research questions

The main aim of this dissertation is to refine our understanding of the Tocharian-Uralic/Samoyed contact hypothesis. I do this by investigating a number of interrelated research questions, all variations on a common theme.

1) Can the significant innovations in Tocharian consonant and/or vowel phonology be understood as the result of influence from Uralic/Samoyed?

Tocharian phonology is quite innovative from the Indo-European point of view, and substrate influence from Uralic/Samoyed has been considered as one possible driving factor in the more significant changes in both the consonant system (Kallio 2001; Bednarczuk 2015; Peyrot 2019a) and the vowel system (Peyrot 2019a; Warries 2022). In order to provide a well-informed view on the validity of these comparisons, it is necessary to understand the developments of the Tocharian and Samoyed phonological systems independently from one another. Only then can we try to establish whether the phonological systems showed significant overlap in their organization at some point in the past, which could indicate a convergence, and therefore, potential contact. To this end, chapters 2 and 3 treat the relative chronologies of sound changes of Tocharian and Samoyed independently. The final comparison is provided in chapter 4.

The historical grammar of Tocharian is full of controversial issues, and in order to arrive at the relative chronology necessary for a comparison with Samoyed, all the changes that affected the system need to receive an interpretation. Without a detailed understanding how both the vowel and consonant system changed through time, and how the two interacted, any external comparison would be problematic. Important secondary research questions here on the Tocharian side are: i) How should the development of the Tocharian vowel system be understood in relation to palatalization? ii) Which umlaut developments can be reconstructed for Proto-Tocharian, and how do they affect our understanding of the changes in the vowel system? iii) What are the regular developments of PIE *$\bar{o}$ and *eh_2 in Tocharian? iv) When did the reflexes of PIE *o and *$\bar{e}$ merge? v) How can the development of PIE *d in Tocharian be understood? These questions are addressed in chapter 2, embedded in an overview of the other relevant Tocharian sound changes.

On the Samoyed side, the same detailed view on the diachronic development of the phonology is necessary. Thus, various unsolved or controversial issues that affect our understanding of the development of the Samoyed phonological system through time need to be treated and evaluated. Particularly important questions are: i) How can the discrepancy between Samoyed *o and Finno-Ugric *u in stems of the type *$C\{o/u\}Ci$ be interpreted in a relative chronology of Samoyed sound changes? ii) How can the differentiation of PU *a-a into PS *$å$-$å$ and *a-$ə̂$ be understood in pre-Proto-Samoyed? iii) Can the differentiation of PU *$ę$ into PS *$ę$ and *$į$ be dated relative to other changes? iv) What are the developments of roots with the shape PU *$CVwi$ and *$CVɣi$? v) Which sound changes are expected to have caused paradigmatic alternations, and to what extent do we see traces of this? Chapter 3 covers these questions along with the necessary context of the other Samoyed sound changes to provide a complete picture.

2) Does the potential for phonological influence from Uralic/Samoyed on Tocharian extend to the accent?

The place of the accent cannot be determined for Tocharian A, but Tocharian B shows a general second-syllable accent, unless the second syllable is also the final syllable. This differs from the first-syllable accent of Uralic/Samoyed. However, it is possible that the Tocharian B accentuation constitutes a later innovation that obscures an earlier system with first-syllable accent, which shifted to a second-syllable accent. Such an accent shift is assumed by Marggraf (1970: 21), Peyrot (2013: 507–515) and Jasanoff (2015), although the details of their analyses differ on very important points. No earlier first-syllable accent is assumed in other theories on the prehistory of the Tocharian accent (Ringe 1987; Winter 1993, 1994b; Malzahn 2010 *passim*), and the topic as a whole has received a number of different interpretations, which sometimes overlap only partially.

The place of the accent in prehistoric Tocharian must include an explanation for not only the eventual accentuation in Tocharian B, but also rise of accented epenthetic vowels and the loss of earlier second-syllable vowels inherited from Proto-Indo-European. In order to better understand the prehistory of the Tocharian accent, I will attempt to provide an answer to the following connected secondary research questions: i) In which phonological environments did epenthesis occur regularly? ii) How can the accented epenthetic vowels of Tocharian B be understood? iii) How did syncope interact with the accent? iv) How did the accent interact with vowel contractions? v) How can the irregularly accented verbal categories of Tocharian B be analysed in a framework with phonologically regular accentual developments? vi) How is the accentuation of the secondary cases in Tocharian B to be interpreted? In the background is the question of how the accent of Proto-Tocharian should be reconstructed, and which effects in Tocharian A can provide any indications. Can the usual assumption that the Proto-Tocharian accent was like the Tocharian B accent with innovations on the Tocharian A side only be substantiated? These questions are addressed in chapter 5.

3) Can the Tocharian case system be understood as the result of influence from Uralic/Samoyed?

In particular, the Tocharian local cases display an innovative agglutinative character, and include in their ranks new cases like the perlative. The connection of the agglutinative Tocharian case system with some type of substrate influence has a long history (Krause 1951; K.H. Schmidt 1990; Thomas 1994, Bednarczuk 2015, Peyrot 2019a), but there is disagreement as to which substrate language should be considered as the source (e.g., Akao 2020 argues for Turkic influence). The case systems of Tocharian and Uralic/Samoyed are discussed in chapter 6.

4) Can the Tocharian participial system be understood as the result of influence from Uralic/Samoyed?

This aspect of the Tocharian-Uralic/Samoyed contact hypothesis is new. Tocharian has participles for the present and the past, and also gerundives and a privative. A number of these participles are neither active nor passive, but rather display contextual orientation, which is atypical from an Indo-European perspective. Similar types of participles are found in Samoyed, and in Uralic languages more generally. In chapter 7, I compare the participial system of Tocharian to the participial systems that are found in Uralic/Samoyed.

5) Can the innovative Tocharian use of pronoun suffixes attached to the finite verb be understood in comparison with the Uralic/Samoyed objective conjugation?

Similarities between Tocharian and Uralic in this respect have been pointed out before (Peyrot 2019a; Georg 2023), but it is also clear that there are many differences between the two language groups. In chapter 8, I will investigate whether these can be reconciled. Importantly, the age of the objective conjugation in Uralic is a controversial question (see 8.3 for more details), leading to a natural secondary research question on this topic: can the objective conjugation be reconstructed for Proto-Uralic, and in what form? This question needs to be addressed before a comparison with Tocharian can be attempted.

6) Finally, in chapter 9, I treat potential loanwords from early Tocharian into early Samoyed and/or Uralic to see if there is any confirmation of the contact hypothesis in the form of lexical borrowings.

Loanwords can be used to support the information gleaned from our discussions of Tocharian grammatical innovations have a great potential to bolster the Tocharian-Uralic/Samoyed contact hypothesis. Unfortunately the evidence remains rather meagre (cf. Janhunen 1983; Napol'skikh 2001; Kallio 2004).

I evaluate the various features on both a case-by-case basis and with consideration for the aggregate. If substrate influence from early Uralic/Samoyed indeed applied to early Tocharian, that carries with it implications for our understanding of the prehistory of the languages on both sides of the comparison. The aim of this dissertation is to better inform our understanding of the past of the Tocharians and their placement in time and space. In order to carry out this study, a detailed consideration of both language families is necessary, spanning the different topics, from phonology to verbal morphology and participial syntax. The underlying details on both sides will be presented at each stage to make the comparison as transparent as possible. My hope is that this will not only contribute to our understanding of the Tocharian-Uralic/Samoyed contact hypothesis, but that the individual discussions that resulted from this investigation can contribute to our understanding of the historical grammar of both Tocharian and Samoyed, and highlight uncertain aspects where further investigation is needed.