



Universiteit
Leiden

The Netherlands

Like dust on the Silk Road: an investigation of the earliest Iranian loanwords and of possible BMAC borrowings in Tocharian

Bernard, C.B.A.S.

Citation

Bernard, C. B. A. S. (2023, April 20). *Like dust on the Silk Road: an investigation of the earliest Iranian loanwords and of possible BMAC borrowings in Tocharian*. Retrieved from <https://hdl.handle.net/1887/3594385>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3594385>

Note: To cite this publication please use the final published version (if applicable).

Chapter 1: Introduction

This PhD thesis presents an investigation of contacts between the Tocharian languages and an Iranian language I tentatively named Old Steppe Iranian, as well as BMAC languages. In this introduction, I will first introduce both the Tocharian and the relevant Iranian languages, mainly Old Steppe Iranian (§1); then discuss the state of the art of the research in this particular field (§2) and introduce and delimit the problems of Tocharian–Iranian contacts to be addressed and present my research questions (§3). After this, I will discuss the methodology of the present work (§4), describe the structure of the thesis (§5) and finish with a short presentation of the notation of stress in Tocharian B (§6).

1.1 Tocharian and Iranian

Tocharian is a branch of the Indo-European language family. This branch is known through two extinct languages: Tocharian A and Tocharian B, which were spoken in the north of the Tarim Basin, today part of Xinjiang region in the Northwest of China. These languages are known to us from texts found in the beginning of the 20th century by British, German, French, Russian and Japanese expeditions (cf. Pinault 1989: 7; Fellner 2007). The speakers of Tocharian, during the historical period and as far as we know, were Buddhist.

A large number of texts in multiple languages were found in the Tarim Basin, among them the Tocharian manuscripts (see map p. 250). The Tocharian texts were found along the northern rim of the Taklamakan desert, in several places around Kuča, Yānqí and the famous city of Turfan. About 10,000 Tocharian B manuscript fragments and about 2,000 Tocharian A manuscript fragments have been preserved (Peyrot 2015a).

Linguistically, Tocharian B can be divided into three chronological stages: Archaic Tocharian B, Classical Tocharian B and Late Tocharian B (Peyrot 2008). Tocharian A texts do not show such internal variation (see, however, Itkin 2002). While only Tocharian B fragments have been found in the region of Kuča, Tocharian B and Tocharian A fragments have been found side by side in the regions of Yānqí and Turfan (Peyrot 2010: 133). No unanimously recognized Tocharian A borrowing could be identified in Tocharian

B, but Tocharian A has many Tocharian B loanwords, usually, if not exclusively, from (later) Classical Tocharian B (cf. Peyrot 2010: 139, with references).

Most of the Tocharian texts are of Buddhist content, but a trove of documents, that is, non-literary texts, were also found, including everyday documents such as laissez-passers for caravans (Pinault 1987), and also inscriptions on walls of grottoes (Ching & Ogihara 2020), etc. The texts can be dated to the 5th to the 9th or 10th centuries A.D. for Tocharian B, and to the 7th to the 10th centuries for Tocharian A. It is thought that Tocharian A and B were not mutually intelligible (Pinault 2002a: 245).

Iranian languages are a branch of Indo-European that was spoken in a huge area, spanning from Bulgaria to Siberia, including parts of China, the entire Western part of Central Asia, the Iranian plateau, and Eastern Anatolia. Nowadays, Iranian languages are still found from Eastern Anatolia and the Caucasus to China and Pakistan. For more information on the major Iranian languages, see for example the *Compendium Linguarum Iranicarum* (Schmitt 1989). Three chronological stages of the Iranian languages can be distinguished:

1. Old Iranian: Avestan (documented), Old Persian (documented), Scythian (indirect, scarce documentation);
2. Documented Middle Iranian: Middle Persian, Parthian, Bactrian, Sogdian, Khwarezmian, Khotanese, Tumshuqese;
3. New Iranian: New Persian, Balochi, Kurmanji, Sorani, Yidgha, Munjī, Pashto, Yaghnobi, Wakhi, etc.

The diversity and dialectal features of the Middle Iranian languages show that there must have existed more than the three Old Iranian languages listed above. Likewise, New Iranian languages show that there must have existed more than seven Middle Iranian languages. Most, if not all of the documented Middle Iranian languages were spoken in Central Asia. Many of these languages were spoken in the West of Central Asia, in regions too far removed from the Tarim Basin for our purposes. However, some, such as Bactrian and Khotanese, were spoken in regions close to the Tocharian languages.

In the present thesis, I will look at the oldest contacts between Tocharian and Iranian, which can be dated to the Old Iranian stage. I will also investigate Tocharian words possibly borrowed from the so-called BMAC language, which may form an even earlier layer of the Tocharian lexicon.

1.2 State of the art

Meillet (1913) was the first to recognize an Iranian loanword in Tocharian: the word *kuñcit* ‘sesame’. Lidén (1916) repeated the observation by discussing a few more words, at a very early stage of Tocharian studies.

The first systematic study was by Hansen in 1940, who brought together a number of Tocharian words which he considered to be of Middle Iranian origin. Many of the words are indeed of Middle or Old Iranian origin, although some others were incorrectly identified, and some were based on Tocharian ghost words, due to wrong readings of the texts. Even though it was not followed by any major work for some time, Hansen’s study opened the road to the critical analysis of Iranian loanwords in the Tocharian corpora.

After this, an important article was published by Werner Winter (1971). On the basis of the analysis of common phonological features, Winter suggests that a number of Iranian words in Tocharian are borrowed from Bactrian, such as TB *perne*, TA *parām* ‘glory, dignity; rank’. As Bactrian was only fragmentarily known at the time, the Bactrian source forms were not attested but had to be reconstructed. As Schwartz (1974) demonstrated, Winter’s attribution of this layer of loanwords to Bactrian was incorrect. Although there are Bactrian loanwords in Tocharian, like TB *akālk*, TA *ākāl* ‘wish’ and TB *mālo* ‘alcohol’, words like *perne* ‘glory, dignity; rank’ cannot belong to the Bactrian layer.

Schmidt (1985) showed that the words that Winter had misidentified as Bactrian are in fact of Old Iranian origin, and not Middle Iranian, as had been assumed up to that point. However, he did not provide a specific identification of this “Old Iranian” language, and limited himself to listing sound correspondences between Old Iranian and the Tocharian borrowings, thus expanding the correspondences established by Winter.

Although it has never been published, Lambert Isebaert’s 1980 PhD thesis, which discusses a great number of Indic and Iranian loanwords in Tocharian, deserves special mention. Based on a large corpus of words, and adding a considerable number of new ones, it systematically discusses the rendering in Tocharian of Indic and Iranian *a* and *ā*, which he considers to be “diaphones”.

Based on the new discoveries in Bactrian made by Sims-Williams (2002), Pinault (2002a) studied various strata of Iranian–Tocharian contacts, focusing on Bactrian–Tocharian contacts. He posits the following stages of Irani-

an loans: 1. Old Iranian, 2. Bactrian, 3. Sogdian and Khotanese. He qualifies Old Iranian as “unidentified languages, differing from Avestan and perhaps attributable to the ancestors of the Ossetes” (Pinault 2002a: 245).

A major systematic study following Isebaert, Schwartz, Schmidt and Pinault was that of Tremblay (2005). Tremblay tried to systematically classify all Iranian loanwords in Tocharian known to him. He also tried to identify the Old Iranian language discovered by Schmidt (1985), and suggested that it is “Old Sakan”, that is, the ancestor language of Khotanese. He also assumed various stages of Middle Iranian loanwords in Proto-Tocharian, and of later borrowings.

An advantage of Tremblay’s theory that the Old Iranian elements in Tocharian derive from “Old Sakan” is that earlier stages of Khotanese are known to have been spoken in the region, and we thus would not need to add a new language to explain the Iranian elements in Tocharian. However, he did not take Schmidt’s (1985) findings into full account, and considered Proto-Iranian **j* to be reflected as **z* in Old Sakan, whereas Schmidt had already shown that the source of the Tocharian borrowings must have had the reflex **ḏ*. Peyrot (2018a) adduced further counterarguments against Tremblay’s identification, arguing that the Old Iranian language cannot be closely related to Khotanese. Another shortcoming of Tremblay’s paper is that he assumes Middle Iranian loanwords in Proto-Tocharian, but this is chronologically flawed, because Proto-Tocharian is certainly much older than Middle Iranian. Despite these shortcomings, Tremblay (2005) is usually used as a reference work in scholarship (e.g. Peyrot 2018a).

After Tremblay’s article came two studies by Peyrot (2015a; 2018a) and one by Isebaert (2019). Peyrot (2015a) follows Pinault’s (2002a: 245) classification of the loanword stages, and he discusses examples of each stage. Peyrot (2018a) discusses the discovery of the Iranian loanword *etswe* ‘mule’, which derives from Old Iranian **at̪ua-* (in my view originally from **at̪ua-tara-*, s.v. *etswe* and section 2.6.4.f). This later article is very important, as it is the first seriously argued discussion on the identification of the Old Iranian language in contact with Tocharian, claiming that it is not closely related to Khotanese, but is rather closer to the rest of Iranian languages. Isebaert (2019), based on the publications mentioned above, proposed two new Iranian etymologies, which I also discovered independently and presented at a conference, and are now to be found in this thesis (s.v. *entse*, *ñyātse*).

One can also cite Kim (1999), a work that tries to establish an absolute and relative chronology of Tocharian sound changes based on loanwords, some of these uncertain, some others well-known.

A related issue is BMAC loanwords. BMAC, namely, the Bactriana-Margiana Archaeological Complex denotes a group of coherent archaeological findings from a vast region in Northern Afghanistan and Turkmenistan. The BMAC language (or languages) are unknown, or at least, have not been directly documented. Lubotsky (2001) has demonstrated that a great number of BMAC loanwords can be identified in Indo-Iranian. Pinault (2002, and especially 2006) further applied this observation to the Tocharian vocabulary.

1.3 Research issues

The previous scholarship mentioned above answers many questions but raises a number of new ones, because not all the discoveries and not all the findings have led to clear, defined results, and many facts remain unclear or doubtful. The most important research issues seem to be:

1. Are the words belonging to the Old Iranian stage from one single language, as Schmidt thought, or multiple ones, as found elsewhere in the literature?
2. If the Old Iranian layer is one single language, can its features be established? Is it homogeneous? How can we identify the words belonging to this stage and layer?
3. Are there any BMAC loanwords in Tocharian? Did they come directly from a BMAC language, or through, for instance, Iranian?

The goal of my thesis is to address these questions. The first research question will be answered in chapter 2: I will start with the working hypothesis that there was one single language, and analyze all previously identified loanwords in order to test this hypothesis. I will try to establish features of this language or these languages, if possible, and see if we can draw a coherent picture of the source language. This will permit me to answer the second research question, and establish a specific methodology in order to identify Old Iranian loanwords in Tocharian (see section 2.6 below).

Chapter 3 will be dedicated to the third research question: are there any BMAC loanwords in Tocharian? I will start with the working hypothesis that

there were indeed BMAC borrowings in Tocharian. I will analyze a selection of such possible loanwords, among which some rejected Old Iranian loanwords, and I will investigate whether they share common features. This will lead me to conclusions about the specific BMAC language or dialect in contact with Tocharian, and on the type of contact these two might have had on each other.

1.4 Methodology

In order to address the two first questions above, concerning the Old Iranian stage of Tocharian loanwords, I will base myself on the following methodology.

Working along similar lines as Schmidt (1985) and Koivulehto (e.g. 2001), I will first select the words which were already recognized as being of Old Iranian origin. If these words have specific traits that are coherent within their subgroup, I will take these traits as representing either features of the source language, or as Tocharian adaptations of these features, thus, building on the feature analysis of Schmidt (1985).

Since the donor language is not otherwise known, I always compare any of these features with what we know of Proto-Iranian, and with other Iranian languages. In this way, the donor language(s) of these Iranian words in Tocharian is surrounded by Proto-Iranian on one side, and by Tocharian on the other, and it becomes more easy to both reconstruct the source words and understand some of its features. In order to identify what is specific to the source language, or source languages, I will focus on where it differs from Proto-Iranian. The features that are identified will be used throughout the etymological study part of chapter 2, in order to judge existing and to find new etymologies.

This will lead to a double process: first, starting with a smaller group of secure words, namely, a group of evident etymologies recognized in scholarship, I propose a series of sound correspondences. These sound correspondences will lead me to check and control those already established etymologies, in order to secure them even further. They also permit me to add new etymologies, which, after being checked against those correspondences, might yield new correspondences. These new correspondences will permit me to reject some etymologies, consider some as doubtful, and add new ones. Thus, *ainsi de suite*. The resulting etymologies are subdivided into four

categories: 1) plausible cases in 2.2; 2) possible cases in 2.3; 3) difficult cases in 2.4; and finally 4) rejected cases in 2.5.

For the BMAC words in Tocharian (chapter 3), the method will be somewhat different since it will in a certain sense be based on negative results: I discuss here words which according to my analysis cannot be of Old Iranian origin, but nevertheless appear to be old, for instance because they show regular correspondences between Tocharian B and Tocharian A and are thus reconstructible for Proto-Tocharian, alongside words which were already proposed to be of BMAC origin in previous scholarship (e.g. in Pinault 2006). My goal will be to analyze these words and to determine if they share common features. Based on the coherence of these features, I hope to be able to gather more BMAC words, primarily among the words that I discard as Old Iranian loanwords, but which still show features of early borrowing.

Based on my analysis of the features of loanwords from Old Iranian and the so-called BMAC language, it will become easier to determine to which of these two sources a given loanword more plausibly belongs. Also, it will become easier to recognize Middle Iranian loanwords, since these do not conform to the features identified for the oldest layers of borrowings. In this thesis, I do not discuss Middle Iranian loanwords systematically, but I have added a few discussions on Middle Iranian loanwords where this seemed relevant to the discussion of an Old Iranian loanword. Further, I discuss some Middle Iranian loanwords that have been proposed to be of Old Iranian origin, or could perhaps be argued to be so.

Of course a very crucial element in the analysis is presented by the semantics: the meaning of the word should not be too far removed from the meanings we find in Iranian languages, in the sense that there should not be too many steps between an attested Iranian meaning and the attested Tocharian meaning. There is no objective method to determine whether an etymology is acceptable or not, apart from one very important element, which is the existence of parallels permitting to confirm that a specific semantic directionality is possible. Whether or not parallels exist, it is important as well to ensure that a specific semantic shift makes sense, both in general terms and in the specific cultural setting in which it is supposed to have taken place.

1.5 Structure

The thesis is structured as follows. First, in chapter 2, Old Iranian loanwords in Tocharian are discussed, and then, in chapter 3, possible BMAC words. The thesis ends with a conclusion.

In chapters 2 and 3, I will follow a similar structure for most loanwords: first a more or less short presentation of the Tocharian meaning, and then a discussion of the etymologies previously proposed in the literature, if there are any. Then follows a critical analysis of the etymology, and, if need be, I will propose my own etymology. At the end, I present a short conclusion summarizing all the elements described above.

At the end of the two chapters I add a linguistic discussion of the findings based on the etymological discussions of the relevant chapter. The discussion of chapter 2 will consider all the sound changes, the isoglosses and the semantic shifts that can be attributed to the Old Iranian stage of loanwords. These will be compared to a number of selected Iranian languages, in order to refine the conclusions one can make on that Old Iranian stage. For chapter 3, on BMAC loanwords, I will follow a similar structure, although somewhat shorter.

My work is exclusively of a linguistic nature, that is, I base both my research and my conclusions on linguistic observations. However, beyond linguistics, there is much more to say, on the basis of many elements described in this thesis. I thus chose to venture beyond pure linguistics in the conclusion of the thesis (p. 250), and I propose a (pre-)historical scenario of how and where Tocharians and ancient Iranians came into contact.

1.6 Spelling of Tocharian stress

A very important part of the discussion in the next chapters is on questions relative to Tocharian B stress and Proto-Tocharian reconstructed stress. I will depart from the assumption that Tocharian B preserves Proto-Tocharian stress as a rule, and that this stress was lexical, and try to verify this assumption. For this, it is necessary to understand how Tocharian B stress was spelled.

While there is no clear spelling of the stress in Tocharian A and in Archaic Tocharian B (Peyrot 2008), stress is indicated in Classical and Late Tocharian B in the following way. When stressed, the Tocharian B phoneme /a/ is written as <ā> (there is no phonological length in Tocharian B), but as <a>

when unstressed. The Tocharian B phoneme /ə/ is written as ⟨ä⟩ when unstressed, and as ⟨a⟩ when stressed. There is thus sometimes a spelling ambiguity between unstressed /a/ and stressed schwa /ə/, which are both spelled ⟨a⟩. This ambiguity is usually solved by either etymology or by variants of the same word. Indeed, either an archaic spelling or a suffixed or an inflected form, such as the plural ending, can confirm that the vowel was originally a schwa /ə/ or /a/. For example Tocharian B *yasar* ‘blood’ is phonologically /yəsar/ rather than /yasár/, as can be deduced from its plural *ysāra* /ysára/. Besides, Tocharian B words could never be accented on the final syllable (Krause 1971: 11), so that there is no doubt about the stressed syllable in disyllabic Tocharian B words.

In some cases, we find the phonemes *i* and *u* spelled as ⟨ī⟩ and ⟨ū⟩. This can sometimes indicate stress, for example in TB *kamartīke* ‘ruler’, TB *kuñcīt* ‘sesame’, where the stress is also known from other data. This is however not always the rule, for instance in the Tocharian B spelling *āktīke* ‘wonderful, astonishing’, where the stress can obviously not be on both the first and the second syllable. As a rule, stress is only systematically marked on the Tocharian B phonemes *a* and *ə*, but this will be enough to be able to make a number of deductions concerning both Tocharian and Iranian stress.