Turkic Studies Journal Special Issue,
2 (2026) 77-92
Journal homepage: https:tsj.enu.kz

Article

4TB: A new tool to study Tocharian A – Old Uyghur parallels

M.V. Vyzhlakov

University of Vienna, Vienna, Austria
ORCID: https://orcid.org/0000-0002-7813-6640
(E-mail: maksim.vyzhlakov@univie.ac.at)

ARTICLE INFO

Keywords:
Central Asian Buddhism, Maitreya, Maitreyasamiti-Nāṭaka, Maitrisimit nom bitig, Old Turkic, Old Uyghur, parallel corpus, Silk Road cultures, Tocharian A, translation corpus.

IRSTI 16.21.21

DOI: *https://doi.org/10.32523/2664-5157-2026-2SI-77-92*

ABSTRACT

The paper introduces 4TB, a web-based corpus developed to support research on Tocharian A and Old Uyghur. Its aim is to collect and align Tocharian A fragments with their Old Uyghur (and later Sanskrit) parallels, particularly the Tocharian A *Maitreyasamiti-Nāṭaka* and its Old Uyghur translation *Maitrisimit nom bitig*. Despite extensive scholarship, no complete editions of these texts are currently published. A Tocharian A edition is in preparation and will serve as the basis for the corpus, while 4TB focuses on assembling Old Uyghur parallel fragments as a foundation for a future full edition of the *Maitrisimit nom bitig*. Compared to existing digital resources, Tocharian is relatively well represented, whereas Old Turkic corpora remain limited, making even a partial parallel corpus a meaningful contribution. The paper discusses key problems of corpus design and development. Transliteration is omitted for both languages, while transcription is handled differently. For Tocharian A, it follows established conventions; for Old Uyghur, all data are normalized into a unified transcription system, with limited correction of outdated forms. Additional challenges arise from inconsistent transcription practices in the literature. Translations are provided as in the source publications (Russian for Tocharian A, mainly German for Old Uyghur), with plans to add automatic English translations. The corpus is structured around tokens as the smallest unit, but alignment operates on higher-level units due to the lack of reliable word-to-word correspondence. Because sentence segmentation is problematic, a flexible unit called a “passage” is introduced. Passages are grouped into “passage groups” to account for manuscript variation and are then aligned across languages, allowing for discrepancies such as omissions and additions. This approach preserves textual coherence and differs from standard KWIC-based corpus models. The corpus offers several core features. Dictionaries cover all lemmatized tokens and link forms and spellings with grammatical annotation. A concordance

Funded by the European Union

displays passages together with their parallels. Search tools support queries by spelling, form, lemma, and grammatical features, including cross-linguistic searches. Editing tools allow modification of texts and lexical data, with semi-automatic lemmatization and potential for further expansion. Overall, 4TB provides a scalable framework that can be used in future corpus projects. It facilitates cross-linguistic research on Tocharian and Turkic languages and improves access to the material for non-linguistic specialists such as Buddhologists.

М.В. Выжлаков
Венский университет, Вена, Австрия
ORCID: https://orcid.org/0000-0002-7813-6640
(E-mail: maksim.vyzhlakov@univie.ac.at)

4TB: новый инструмент для изучения тохарских А и древнеуйгурских параллельных текстов

Аннотация. В статье представлен 4TB – онлайн-корпус, разработанный для исследования тохарского А и древнеуйгурского языков. Его цель – собрать и выровнять тохарские А фрагменты с их древнеуйгурскими (а в дальнейшем и санскритскими) параллелями, прежде всего тохарский *Maitreyasamiti-Nāṭaka* и его древнеуйгурский перевод *Maitrisimit nom bitig*. Несмотря на обширную исследовательскую традицию, полные издания этих текстов до сих пор не опубликованы. В настоящее время готовится редакция тохарской версии, которая послужит основой для корпуса, тогда как 4TB сосредоточен на сборе древнеуйгурских параллельных фрагментов как основе для полного издания *Maitrisimit nom bitig* в будущем. По сравнению с существующими цифровыми ресурсами тохарские тексты представлены относительно полно, тогда как корпусы древнетюркских текстов остаются неполными, что делает даже частичный параллельный корпус значимым вкладом. В статье рассматриваются ключевые проблемы проектирования и разработки корпуса. Транслитерация для обоих языков не используется, тогда как транскрипция реализована по-разному. Для тохарского А она следует устоявшимся конвенциям, тогда как для древнеуйгурского все данные нормализованы в рамках единой системы транскрипции с исправлением некоторых устаревших форм. Дополнительные трудности связаны с неоднородностью транскрипционных систем в научной литературе. Переводы приводятся в соответствии с языками исходных публикаций (русский для тохарского А, преимущественно немецкий для древнеуйгурского), планируется добавление автоматического английского перевода. Корпус структурирован вокруг токенов как минимальных единиц, однако выравнивание

Received 25 February 2026. Revised 25 March 2026. Accepted 26 April 2026. Available online 30 June 2026.

For citation: M.V. Vyzhlakov 4TB: A new tool to study Tocharian A – Old Uyghur parallels // Turkic Studies Journal. 2026. V. 8. № 2SI. P. 77-92. DOI: http://doi.org/10.32523/2664-5157-2026-2SI-77-92

фрагментов осуществляется на более высоком уровне из-за отсутствия надежного пословного соответствия. В связи с проблематичностью сегментации по предложениям вводится более гибкая единица – «пассаж». Пассажи объединяются в «группы пассажей» для учета различий между манускриптами и затем выравниваются между языками, допуская расхождения, такие, как пропуски и добавления. Такой подход сохраняет связность текста и отличается от стандартных корпусных моделей, основанных на формате KWIC. Корпус включает ряд основных функций. Словари охватывают все лемматизированные токены и связывают формы и написания с грамматической разметкой. Конкорданс отображает пассажи вместе с их параллелями. Поисковые инструменты поддерживают запросы по написанию, форме, лемме и грамматическим признакам (последнее включает межъязыковой поиск). Инструменты редактирования позволяют изменять тексты и лексические данные, включая полуавтоматическую лемматизацию, и имеют потенциал для дальнейшего развития. В целом 4TB представляет собой масштабируемую платформу, которая может быть использована в будущих корпусных проектах. Он способствует межъязыковым исследованиям тохарских и тюркских языков и расширяет доступ к материалу для нелингвистов, включая буддологов.

Ключевые слова: Maitreyasamiti-Nāṭaka, Maitrisimit nom bitig, древнетюркский язык, древнеуйгурский язык, корпус переводов, культуры Великого Шёлкового пути, Майтрея, параллельный корпус, тохарский А язык, центральноазиатский буддизм.

М.В. Выжлаков
Вена университеті, Вена, Австрия
ORCID: https://orcid.org/0000-0002-7813-6640
(E-mail: maksim.vyzhlakov@univie.ac.at)

4TB: Тохар А және көне ұйғыр параллель мәтіндерін зерттеуге арналған жаңа құрал

Аннотация. Мақалада Тохар А және көне ұйғыр тілдерін зерттеуге арналған 4TB онлайн-корпусы таныстырылады. Оның мақсаты – Тохар А тіліндегі мәтін үзінділерін, көне ұйғыр (болашақта санскрит тілімен) тіліндегі сәйкес нұсқаларымен салыстыра отырып жинақтау. Зерттеу негізінен тохар тіліндегі «Maitreyasamiti-Nāṭaka» мәтіні мен оның көне ұйғыр тіліндегі аудармасы «Maitrisimit nom bitig» шығармасына арналған. Бұл мәтіндердің зерттелу тарихы терең болғанына қарамастан, олардың толық ғылыми басылымдары әлі күнге дейін жарияланбаған. Қазіргі уақытта корпустың негізін құрайтын тохар нұсқасының редакциясы дайындалуда. 4TB жобасы болашақта «Maitrisimit nom bitig» мәтінінің толық басылымын әзірлеуге негіз болатын, көне ұйғыр тілінде сәйкес мәтін үзінділерін жинақтауға бағытталған. Қолданыстағы сандық ресурстармен салыстырғанда, тохар мәтіндері салыстырмалы түрде толық ұсынылған, ал көне түркі мәтіндерінің корпустары әлі де толық емес күйде қалып отыр. Бұл жағдай ішінара жасалған сәйкес мәтіндер корпусының өзін де маңызды ғылыми үлеске айналдырады. Мақалада корпусты жобалау мен әзірлеуге қатысты негізгі мәселелер қарастырылады. Екі тіл үшін де транслитерация қолданылмайды, ал транскрипция әртүрлі қағидаттар

негізінде жүзеге асырылады. Тохар А тілі қалыптасқан ғылыми конвенцияларға сәйкес берілсе, көне ұйғыр тілі бірыңғай транскрипция жүйесінің аясында біріздендіріліп, кейбір ескірген формалары түзетілген. Қосымша қиындықтар негізінен ғылыми әдебиеттегі транскрипция жүйелерінің біркелкі еместігіне байланысты. Аудармасы түпнұсқа жарияланымдардың тілдеріне сәйкес беріледі (тохар А – орыс тілі, көне ұйғыр тілі – неміс тілі), ал болашақта ағылшын тіліне автоматты аударуды енгізу жоспарлануда. Корпус, ең кіші бірлік – токендерге негізделген, алайда сөзбе-сөз сәйкестіктің болмауы себебінен, фрагменттерді сәйкестендіру жоғары деңгейде жүзеге асырылады. Сөйлемдегі сегментацияның күрделілігіне байланысты «пассаж» деп аталатын икемді бірлік енгізіледі. Пассаждар манускрипттер арасындағы айырмашылықтарды ескеру мақсатында «пассаж топтарына» біріктіріліп, кейін тілдер арасында сәйкестендіріледі. Ол сөздердің түсіп қалуы мен қосылуы сияқты мәтіндік ауытқуларда ескеріледі. Мұндай тәсіл мәтіннің тұтастығын сақтай отырып, оны KWIC форматындағы стандартты корпус модельдерінен ажыратады. Корпус бірқатар негізгі функцияларды қамтиды. Сөздіктер барлық лемматизацияланған токендерді қамтып, олардың формаларын грамматикалық белгілеумен байланыстырады. Конкорданс пассаждарды параллель мәтіндерімен бірге ұсынады. Іздеу құралдары жазылым, форма, лемма және грамматикалық белгілер бойынша сұраныстарды қолдайды (соңғысы тілдер арасындағы іздеуді де қамтиды). Редакциялау құралдары мәтіндік және лексикалық деректерді өзгертуге мүмкіндік береді, онда жартылай автоматты лемматизация да бар және оны әрі қарай жетілдіруге мүмкіндік береді. Жалпы алғанда, 4TB – болашақ корпус жобаларына бейімделетін платформа. Ол тохар және түркі тілдерін салыстырмалы зерттеуге мүмкіндік беріп, материалдарды тіл мамандары ғана емес, әсіресе буддологтарға қолжетімді етеді.

Кілт сөздер: Maitreyasamiti-Nāṭaka, Maitrisimit nom bitig, көне түркі тілі, көне ұйғыр тілі, аудармалар корпусы, Ұлы Жібек жолы мәдениеттері, Майтрея, параллельді корпус, тохардың А тілі, Орталық Азия буддизмі.

Acknowledgements

I would like to express my sincere gratitude to Hannes Fellner, Ilya Itkin, Irina Nevskaya, and Jens Wilkens for their advice on the project; to the CEToM and VATEC teams for allowing me to use their data; and, last but not least, to Sergey Malyshev for allowing me to use his edition and for testing the stability and usability of my web interface.

This project has received funding from the European Union's Horizon 2023 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101150017. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Introduction

In this paper, I present a new web-based text corpus, 4TB: *The Transmission and Transformation of Texts in the Tarim Basin,* which aims to contribute to Tocharian and Old Turkic studies, as well as to research on Central Asian Buddhism. The corpus is being developed as part of

my personal postdoctoral project at the University of Vienna, under the supervision of Prof. Hannes Fellner.

The initial aim of the corpus is to bring together all texts in Sanskrit and in the Old Uyghur variety of Old Turkic that are parallel to the preserved Tocharian A fragments[1]. The most important and extensive among these is, of course, the Old Uyghur *Maitrisimit nom bitig*, a translation of the Tocharian *A Maitreyasamiti-Nāṭaka* (hereafter *Maitrisimit* and MSN, respectively), an extensive 27-chapter Buddhist drama about the future Buddha Maitreya and currently the main focus of the corpus.

The overarching idea behind the 4TB corpus is to facilitate cross-linguistic research and the reconstruction of both the Tocharian and Turkic versions. It could also simplify access to the material in one language for specialists in the other, or even for non-linguists such as Buddhologists.

Materials and research methods

The materials for the corpus originate from nine Tocharian A manuscripts from Shorchuk, Yanqi, and Kocho, while the Old Uyghur component is represented by manuscripts from Hami (Tömürti and Närnasi) and Turfan (Sängim and Murtuk). All these locations are situated within the present-day Xinjiang Uyghur Autonomous Region of the People's Republic of China, or along the northeastern edge of the Tarim Basin. The editions of these manuscripts, which serve as the direct source of the corpus data, will be discussed in the next section.

It should be noted that the processing of the Old Uyghur component is divided into two phases. First, the Hami version fragments are added; the Turfan material will be introduced later.

As for the methodology, while there are many ways to categorize a text corpus[2], I would like to highlight the following core characteristics of 4TB, which motivate the methodological considerations and the wide range of user-friendly tools described in the following sections:

– 4TB is translation/parallel corpus;

– it is historical but not diachronic, in the sense that it is not designed to analyze language development, even though some variation among the manuscripts may be attributed to it;

– it is annotated. At present, the feature-level mark-up is limited to grammatical annotation. The structural mark-up is treated in detail below;

– it is dynamic, meaning that textual additions and revisions will continue beyond the end of the project;

– it is noisy. "A noisy parallel corpus contains bilingual sentences that are not perfectly aligned or have poor quality translations" (Wołk, 2015: 171). To a certain extent, both of these aspects apply to the *MSN / Maitrisimit*. On the one hand, there are both technical (different fragments preserved) and substantive (e.g., different colophons) discrepancies. On the other hand, the Old Uyghur translation is not very precise per se, perhaps even intentionally so (Wilkens, 2023: 569).

Research background

The *MSN / Maitrisimit* has been of great importance for both Tocharian and Old Turkic studies ever since the connection between these texts was first proposed by Müller and Sieg (1916).

[1] The size of the Sanskrit-Tocharian A parallel corpus is relatively small. Its inclusion is planned for a later stage of the corpus development and will not be discussed in this article.

[2] In this categorization, I largely follow (Weisser, 2022: 90-98).

While a complete overview of the publications devoted to either the Tocharian or the Turkic version would be impractical, two key fragment concordances that laid the groundwork for the comparison of the *MSN* and the *Maitrisimit* should be mentioned, namely Tekin (Tekin, 1980: 11-12) and Pinault (Pinault, 1999: 193-205). This line of research, often focusing on fragment identification and textual reconstruction, has since been further developed in a number of studies, including (Burlak, Itkin, 2004), (Geng et al., 2004a), (Geng et al., 2004b), (Wilkens, 2008), (Semet, Äysa, 2014), (Peyrot, Semet, 2016), (Itkin et al., 2017), (Wilkens, 2023), (Itkin et al., 2025).

Despite the wealth of publications on the topic, there is still no complete edition of either the *MSN* or the *Maitrisimit*. The former is currently being addressed, as an edition of the Tocharian A version is being prepared by Sergey Malyshev (p.c.)[3]. It will serve as the basis for the Tocharian A component of the corpus. At the same time, the work will be made available on the 4TB website, either as a static page or as a downloadable file.

As for the *Maitrisimit*, the corpus cannot, of course, pursue such an ambitious goal. Its aim is to gather the parallel fragments and provide them with a range of research and editorial tools, thereby creating a useful resource for the preparation of a full edition.

The situation with regard to digital corpora is somewhat similar. CEToM (Malzahn et al., 2011) already contains almost the entire Tocharian textual corpus. Since the Old Turkic material is much more abundant, a complete corpus remains a task for the future. Nevertheless, an important foundation was laid by VATEC (Erdal et al., 2003), which includes the introduction and the first two chapters of the Hami manuscript. Apart from this, there has also been an attempt to create a collection of Orkhon inscriptions (Derin, Harada, 2021), although, as far as I am aware, the amount of processed material remains rather limited.

To sum up, even a partial edition of the *Maitrisimit* would constitute a significant contribution to the digital Old Turkic corpus.

Analysis

This section outlines various aspects of the corpus design and the methodological considerations associated with them.

Transliteration, transcription, and translation

The transliteration of Tocharian A is expected to be available only within the publication of Sergey Malyshev's edition of the *MSN*, without any technical or visual integration with the transcription. The transcription follows the conventions of the field, which are generally quite uniform.

As for Old Uyghur, given the scope of the corpus, as well as the sheer volume of work involved, transliteration is unfortunately not provided. Moreover, the data from (Tekin, 1980), which is available only in transliteration, is converted into transcription for the sake of uniformity. The same normalization is applied in cases where elements of transliteration occur within transcription in the source literature (e.g., *q̈* or *q̣*). An exception is made where one element of the digraph representing *ŋ* is damaged or restored. In such cases, it is written as *ṅġ* (for example, *ṅ[ġ]* instead of *[ŋ]*).

Working with transcription presents additional challenges. First, the source literature employs a variety of transcription systems, which must be standardized. I have chosen to adopt

[3] Sergey Malyshev has indicated that he is willing to share his edition via personal communication (email: sjerjozha@yandex.ru).

the system used in (Wilkens, 2021: V, VI). Another issue concerns the marking of damaged text with italics, a common convention in the literature. This is rarely recognized by OCR tools and may therefore be lost during the digitization process. It may also disappear when users copy text from a browser into a text editor. Unlike other editorial symbols, italicized text requires additional programming logic both for storage and display[4].

Finally, a large portion of the source publications, especially those from the 1980s and 1990s, employ outdated transcriptions of many lexemes. It is beyond the scope of the present project to update all such instances. However, I aim to correct the most obvious cases, such as the choice between *ö* and *ü* or *o* and *u,* based on (Wilkens, 2021). Moreover, these can be addressed in a semi-automatic manner (cf. the *Results* section below).

The translations are provided in the language of the source publication: for Tocharian A, this is Russian; for Old Uyghur, it is primarily German and, in some cases, English. Technically, each passage (cf. the next subsection) in the corpus has three translations "slots" (English, German, Russian). There are plans to provide an automatic English translation for all passages, with explicit indication where it has not been revised.

Corpus architecture

The smallest unit of the corpus is the token, which may correspond to a word, a word group[5], a number, an (original) punctuation mark, or an editorial symbol indicating a lacuna or a torn edge of a leaf. All other editorial symbols, such as "(?)" or caesura marks, are technically treated as part of a neighboring token, while additional information (e.g., the number of missing symbols or lines) is recorded in token notes. Lemmatized tokens are linked to entries in the respective dictionaries (see below), from which they also receive their glosses.

Ideally, parallel Tocharian A and Old Uyghur tokens would be linked to improve comprehension and analysis of the parallel edition. While this is technically possible, it is currently not feasible given the limited resources available. Instead, I aim to facilitate working with the text by providing as much lemmatization and glossing as possible.

As a result, the token cannot serve as the basic unit of text alignment and comparison; rather, this role must be assigned to a higher-level structure, such as a sentence, which contains tokens. This, however, raises the problem of text segmentation, a common issue in corpora of this kind[6].

Both CEToM and VATEC adhered to the line division of the underlying manuscripts, which appears to be the most natural approach for non-parallel historical corpora (also allowing for the straightforward presentation of both transcription and transliteration). Since 4TB omits transliteration and is designed to present and process two aligned texts, this approach is not suitable.

Segmenting texts into sentences may seem a more natural solution in theory, but in practice it proves to be a rather non-trivial task, especially if one aims to do it automatically[7]. Punctuation in Old Uyghur is not well understood (Erdal, 2004: 41), and punctuation in Tocharian is used mainly to mark poetic fragments and is otherwise quite sparse.

[4] Notably, VATEC did not use italics either.

[5] Mainly used in the case of multiword personal names.

[6] Cf., for instance, (Kenning, 2010: 490) and (Lefer, 2020: 269). It should be noted that both segmentation and alignment are usually handled in corpus linguistics using specialized software. However, given the extremely complex nature of the source texts, I do not consider such tools applicable in this case.

[7] Cf. similar problems in (Derin, Harada, 2021: 135).

Under these circumstances, I have opted for a vaguer unit, the "passage". In general, this follows the sentence segmentation found in translations, which is sometimes based on the original punctuation, as well as on the tendency of both Old Uyghur and Tocharian toward SOV word order (Erdal, 2004: 426; Burlak, Itkin, 2013: 395). Otherwise, I have attempted to divide each passage into either simple or complex sentences, avoiding compound sentences.

As a rule of thumb, a change of subject is treated as a signal to begin a new passage, unless we are dealing with an unambiguous subordinate clause. Consequently, sentences such as *"X said: <direct speech>"* are *split into "X said"* and *"<direct speech>"*[8]. Even very short compound sentences (*"Music sounds, the earth trembles")* or those containing explicit repetition of the subject (*"X does this, X does that")* are handled in the same way.

As an additional, somewhat arbitrary measure, I have also chosen to divide lists (such as the 32 marks of the Buddha), treating each item as a separate passage in order to better highlight potential discrepancies between the parallels.

It should be noted that there are differences among the manuscripts of both the *MSN* and the *Maitrisimit*. The corpus aims to record these as well. From a technical standpoint, the simplest solution is to create two (or more) versions of a given passage and to group them within what is termed a "passage group"[9]. Thus, most of the text in either language consists of simple passage groups containing a single passage shared by all manuscripts, alongside much rarer passage groups containing manuscript variants.

These passage groups are organized into a chapter, and chapters into texts. While the number of chapters is identical across the different language versions, this is not the case for the number of passage groups and passages, due to discrepancies in the original content in either of version, differences in segmentation, or simply the varying amount of preserved text. As a result, it is not possible to present the parallel view just putting two columns along each other; more complex alignment is needed instead.

To address this problem (and also to enable sophisticated searching), passage groups from different languages are linked into the parallel groups. The overall visual structure of the parallel text can be schematically shown as follows:

TA text	**OU text**	
Passage group 1	Passage group 1	Non-parallel (e.g., colophons)
Passage group 2	Passage group 2	Parallel group 1
	Passage group 3	Lacuna in TA or addition in OU
Passage group 3	Passage group 4	Parallel group 2
Passage group 4		Lacuna or omission in OU

[8] And "<direct speech>" is further divided into separate passages where necessary.

[9] Terms such as "paragraph" or "section" seem even less appropriate here than "sentence" instead of "passage".

A practical implementation of the scheme can be seen in Fig. 1.

Fig. 1. Screenshot of a fragment of the parallel edition. The Old Uyghur text is on the left, the Tocharian A one is on the right. Parallel groups are marked with a green border. Work in progress.

Рис. 1. Скриншот фрагмента параллельной редакции. Древнеуйгурский текст – слева, тохарский А – справа. Параллельные группы отмечены зелёной рамкой. В стадии разработки.

1-сур. Параллельді редакцияның фрагментінің скриншоты. Сол жақта – көне ұйғыр мәтіні, оң жақта – тохар А мәтіні. Параллель топтар жасыл жақтаумен белгіленген. Әзірлеу сатысында.

In my opinion, this representation of the parallel edition is well suited to demonstrating discrepancies between the source and the translation, drawing attention to places where correction or further reconstruction is possible, and allowing each version to be read as a coherent text[10]. In this respect, 4TB differs from the main approach in corpus linguistics, which typically favors the KWIC (Key Word in Context) format for analyzing and presenting data (cf. Bonelli, 2010: 18-19). I consider this approach justified, since the 4TB corpus is relatively small.

Dictionaries and grammatical annotation

The corpus includes dictionaries covering all lemmatized tokens in the included languages. The Tocharian A one is based on that of CEToM, while the Old Uyghur one is based on (Wilkens, 2021).

Each dictionary is organized as a list of lemmata. Each dictionary entry (i.e., lemma page) contains linked forms, and each form is represented as a list of spellings. The visual presentation of entry information largely follows the CEToM model, with nominal and verbal forms arranged in predefined declension and conjugation tables. Forms that cannot be accommodated in these tables are listed separately below.

[10] Bearing in mind, of course, that the corpus contains fragments rather than complete texts.

The organization of grammatical annotation at the form level broadly follows the current CEToM model. Each form is associated with a set of key-value pairs (e.g., “case: nominative; number: singular”). This approach works well for inflectional languages such as Tocharian A and B. Unlike CEToM, however, the system does not permit multiple values for a single key (e.g., “case: accusative, nominative”). Instead, such cases are represented by independent forms (e.g., “case: accusative”, “case: nominative”, “case: accusative/nominative”).

The inventory and naming of grammatical categories in the Tocharian A component largely coincide with those used in CEToM. For Old Uyghur, I developed a separate system based primarily on (Erdal, 2004), although its overall structure was likewise inspired by CEToM. Thus, rather than distinguishing nouns and adjectives as separate parts of speech, 4TB employs a broader category “nominal”, with subclasses such as “substantive” and “adjective”.

This highest level of categorization is the only area in which Tocharian A and Old Uyghur annotations are partially merged. Although the question of unified versus language-specific annotation in multilingual corpora is a well-known issue in corpus linguistics (Lefer, 2020: 270–271), this choice is motivated primarily by practical constraints rather than theoretical considerations; the same applies to the other solutions listed in this section. Likewise, imposing a new terminology on either Tocharologists or Turkologists is not among the aims of 4TB.

At the same time, the corpus partly adopts the approach used in VATEC. It derives grammatical annotation from the explicit presence of affixes (thus, for example, a suffixless nominative singular noun remains untagged)[11] and is better suited to an agglutinative language such as Old Uyghur. For this reason, the “nominative” is relabeled as the “zero case”, and instead of the traditional “singular/plural” distinction, the system distinguishes “zero singular”, “explicit singular” (mainly for pronouns), and (explicit) “plural”.

Although this terminology may seem unconventional, it allows the combination of the CEToM “key: value” model with the VATEC “unmarked category” approach to grammatical annotation. It also avoids the laborious task of disambiguating nominative forms from casus indefinitus forms, among others.

The internal structure of the dictionaries (that is, the criteria determining what is grouped under a single lemma) likewise depends on the source material. The Tocharian A dictionary largely follows CEToM, whereas the Old Uyghur one reflects the decisions made in (Wilkens, 2021)[12]. Exceptions are nevertheless made in certain cases, especially for pronouns, whose forms in (ibid.) tend to appear as independent entries.

In addition, for Old Uyghur I have automatically generated hypothetical forms that can be used for automatic lemmatization (see below). Unused forms will be hidden once the corpus is published.

Results

Currently, the corpus is closed to the public and is expected to be released by the end of this year at *https://4tb.univie.ac.at/*. Open access will be provided both for the text editions and the concordance and search tools (see below). The final state of its data at the end of

[11] In fact, VATEC does not even allow filtering by nominative forms, since there is no nominative suffix; filtering by singular is likewise limited to certain forms.

[12] For example, hendiadyses and analytical constructions are not treated as independent lemmata. By contrast, some infinitives receive separate entries both in (Wilkens, 2021) and in the corpus, despite generally being treated as part of the verbal conjugational paradigm; this usually reflects a higher degree of lexicalization.

the project (March, 2027) will be stored in TEI-XML format in the digital repository of the University of Vienna *(https://phaidra.univie.ac.at/)*. After that, the corpus will continue to be developed and updated for at least another year.

The corpus will also provide possibility for converting and downloading the content of corpus pages in TEI-XML or JSON format.

The dynamic corpus will be maintained, updated, and/or expanded for as long as time and resources permit. If its maintenance is no longer possible, it will be converted into a static website based on the digital infrastructure of the University of Vienna.

Apart from the preliminary version of the parallel edition, the corpus currently includes the following tools:

Concordance

Dictionary entries also provide concordance links for each form, as well as for the entire lemma (i.e., for all forms associated with it). The concordance view includes both passages in the queried language containing the respective form or lemma and their parallels in the other language. Since the languages are linked only at the passage level, only the matching tokens in the queried language are highlighted.

A KWIC representation of the concordance is not planned at this stage.

Search tools

The corpus provides a user interface that supports the following types of searches:

– by exact spelling
– by partial spelling
– by form (all spellings)
– by lemma (all spellings of all forms)
– by grammar features

Each type supports multiword queries, as well as the display of parallel passages for the results. In addition, grammatical searches can be performed simultaneously across multiple languages. For example, querying verbal forms, such as the optative in Tocharian A and the volitional in Old Uyghur, will return only those parallel pairs in which the Tocharian A passage contains the queried optative form and the Old Uyghur one the queried volitional form. This does not guarantee that the matched words themselves are directly parallel; in other words, the search is not fully precise. However, in the absence of direct word-to-word alignment, this appears to be the most practical solution.

Editing tools

The corpus allows users to register and obtain special permissions granting access to the editing tools. Although 4TB is currently limited to three languages, it has the potential both for further development (i.e., enhancing the existing data) and for expansion, whether by adding further *Maitrisimit* fragments or even entirely new texts and dictionaries. In other words, the corpus can host additional projects and enable other researchers to develop corpora of a similar type without having to create their own tools.

At present, the available editing tools include:

– interfaces for creating and deleting passages. The former allows multiple passages to be added and parsed simultaneously, with an option for automatic lemmatization.

– an interface for editing the content of passages. It allows adding, editing, and deleting tokens, notes, passage markers, and translations, as well as lemmatization (either manually, by entering a form identifier, or semi-automatically, by selecting a form from a dropdown list with preloaded options). The selected lemma link can be applied to all tokens with a given

spelling throughout the corpus. Moreover, bulk editing of tokens with identical spelling is available, which is particularly useful when updating Old Uyghur transcriptions.

– an interface for merging and splitting passages

– an interface for grouping and ungrouping passages within a parallel group

– interfaces for adding, editing, and deleting lemmata and their associated forms and spellings. In the case of Old Uyghur, it is possible to pre-generate or re-generate forms (with the required grammatical annotation), provided that the user enters the core part of speech and subclass in the lemma level input.

– an interface for adding, editing, and deleting grammatical categories used in form annotation

– interfaces for editing the bibliography

The introduction of additional common corpus tools, such as word frequency lists and n-grams, as well as higher-level tools for adding entirely new texts or even subcorpora, is under consideration and will depend on the interest and needs of Tocharian and Old Turkic specialists. Any feedback would be highly appreciated.

Conclusion

The 4TB corpus demonstrates that a passage-based, dynamically expandable architecture can serve as a viable solution for fragmentary and non-uniform historical texts. By prioritizing aligned passages over isolated keyword contexts, the corpus enables more coherent reading and more precise philological comparison between Tocharian A and Old Uyghur materials. At the same time, the article highlights persistent challenges in corpus construction, including segmentation, alignment without one-to-one correspondence, and the integration of heterogeneous annotation systems. Despite these limitations, 4TB provides a scalable framework for future development that can potentially be used by other researchers for their own corpus projects. Ultimately, the article demonstrates that 4TB has strong potential not only to facilitate cross-linguistic research for scholars of the Tocharian and Turkic languages, but also to improve access to the field for non-linguistic specialists such as Buddhologists.

References

Bonelli E.T., 2010. Theoretical overview of the evolution of corpus linguistics. The Routledge Handbook of Corpus Linguistics. Editors: Anne O'Keeffe and Michael McCarthy. 1st ed. p. cm. (Routledge handbooks in applied linguistics). P. 14–28.

Burlak S.A., Itkin I.B., 2004. Tokharskij tekst A 446: yeshchë odna rukopis' tokharskoy versii Maitreyasamiti-Nāṭaka. Voprosy Jazykoznanija. 3. P. 24–35.

Burlak S.A., Itkin I.B., 2013. Tokharskie yazyki. Yazyki mira. Reliktovye indoevropeyskie yazyki Peredney i Tsentral'noy Azii. Red. koll.: Yu.B. Koryakov, A.A. Kibrik. Moscow: Academia. P. 386–485.

Geng Sh., Laut J.-P., Pinault G.-J., 2004a. Neue Ergebnisse der Maitrisimit-Forschung. Zeitschrift der Deutschen Morgenländischen Gesellschaft. 154. P. 347–369.

Geng Sh., Laut J.-P., Pinault G.-J., 2004b. Neue Ergebnisse der Maitrisimit-Forschung (II): Struktur und Inhalt des 26. Kapitels. Studies on the Inner Asian Languages. 19. P. 29–94 + Plates III–XIII.

Erdal M., 2004. A Grammar of Old Turkic. Vol. Central Asia 3. Handbook of Oriental Studies 8. Leiden: Brill. 575 p.

Erdal M., Gippert J., Röhrborn K., Zieme P., Nevskaya I., Knüppel M., Özertural Z., Taube J., 2003. Vorislamische Alttürkische Texte: Elektronisches Corpus. [Electronic resource]. Available at: https://vatec2.fkidg1.uni-frankfurt.de/ (Accessed: 29.03.2026).

Derin M.O., Harada T., 2021. Universal Dependencies for Old Turkish. In Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021). Sofia, Bulgaria. Association for Computational Linguistics. P. 129–141.

Itkin I.B., Kuritsyna A.V., Malyshev S.V., 2017. Tocharian A text THT 1331 and the "Höllenkapitel" of the "Maitrisimit nom bitig": some more remarks. Tocharian and Indo-European studies. 18. P. 71–81.

Itkin I.B., Kuritsyna A.V., Wilkens J., Nugteren H., 2025. THT-fragments of Maitreyasamiti-Nāṭaka: Current state of the topic and some new identifications. Acta Orientalia Academiae Scientiarum Hungaricae. 1 (78). P. 85–113.

Lefer M.-A., 2020. Parallel Corpora. Magali Paquot, Stefan Th. Gries (eds.). A Practical Handbook of Corpus Linguistics. Springer. P. 257–282.

Kenning M.-M., 2010. What are parallel and comparable corpora and how can we use them? The Routledge Handbook of Corpus Linguistics. editors: Anne O'Keeffe and Michael McCarthy. 1st ed. p. cm. Routledge handbooks in applied linguistics. P. 487–500.

Malzahn M., Braun M., Fellner H.A., Koller B., 2011. A Comprehensive Edition of Tocharian Manuscripts. [Electronic resource]. Available at: https://cetom.univie.ac.at/ (Accessed: 29.03.2026).

Müller F.W.K., Sieg E., 1916. Maitrisimit und 'Tocharisch'. Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften. P. 395–417.

Peyrot M., Semet A., 2016. A comparative study of the beginning of the 11th act of the Tocharian A Maitreyasamitināṭaka and the Old Uyghur Maitrisimit. Acta Orientalia Hungarica. 69. P. 355–78.

Pinault G.-J., 1999. Restitution du Maitreyasamiti-Nāṭaka en tokharien A: Bilan provisoire et recherches complémentaires sur l'acte XXVI. Tocharian and Indo-European Studies. 8. P. 189–240.

Semet A., Äysa A., 2014. Prophezeiung über die Maitreya-Geburt. Neues zum 11. Kapitel der uighurischen Maitrisimit nom bitig. Aysima Mirsultan. Mihriban Tursun Aydın. Erhan Aydın (Hrsg.): Eski Türkçeden Çağdaş Uygurcaya. Mirsultan Osman'ın Doğumunun 85. Yılına Armağan. Konya. P. 221–249.

Tekin Ş., 1980. Maitrisimit nom bitig. Die uigurische Übersetzung eines Werkes der buddhistischen Vaibhāṣika-Schule. 1. Teil: Transliteration, Übersetzung, Anmerkungen. Schriften zur Geschichte und Kultur des Alten Orients. Berliner Turfantexte. IX. Berlin: Akademie-Verlag. 264 p.

Weisser M., 2022. What corpora are available? Anne O'Keeffe and Michael McCarthy (Eds.). The Routledge Handbook of Corpus Linguistics, second edition. Routledge. P. 89–102.

Wilkens J., 2008. Maitrisimit und Maitreyasamitināṭaka. Aspects of research into Central Asian Buddhism. In memoriam Kōgi Kudara. Edit. by Peter Zieme. Silk Road Studies 16. Turnhout: Brepols. P. 407–433.

Wilkens J., 2021. Handwörterbuch des Altuigurischen. Altuigurisch – Deutsch – Türkisch. Herausgegeben von der Akademie der Wissenschaften zu Göttingen. Göttingen: Universitätsverlag. 929 p.

Wilkens J., 2023. Einige Beobachtungen zu Übersetzungstechnik der altuigurischen Maitrisimit. Journal of Old Turkic Studies. 2 (7). P. 553–571.

Wołk K., 2015. Noisy-parallel and comparable corpora filtering methodology for the extraction of bi-lingual equivalent data at sentence level. Computer Science. 2 (16). P. 169–184.

References

Bonelli E.T., 2010. Theoretical overview of the evolution of corpus linguistics. The Routledge Handbook of Corpus Linguistics. Editors: Anne O'Keeffe and Michael McCarthy. 1st ed. p. cm. (Routledge handbooks in applied linguistics). P. 14–28.

Burlak S.A., Itkin I.B., 2004. Tokharskij tekst A 446: yeshchë odna rukopis' tokharskoy versii Maitreyasamiti-Nāṭaka. [Tocharian Text A 446: Another manuscript of the Tocharian version of the Maitreyasamiti-Nāṭaka] Voprosy Jazykoznanija. 3. P. 24–35). [in Russian].

Burlak S.A., Itkin I.B., 2013. Tokharskie yazyki [Tocharian languages]. Yazyki mira. Reliktovye indoevropeyskie yazyki Peredney i Tsentral'noy Azii. Red. koll.: Yu.B. Koryakov, A.A. Kibrik [Languages of the world: relict Indo-European languages of Western and Central Asia. Editorial Board: Yu.B. Koryakov, A.A. Kibrik]. Moscow: Academia. P. 386–485. [in Russian].

Geng Sh., Laut J.-P., Pinault G.-J., 2004a. Neue Ergebnisse der Maitrisimit-Forschung. Zeitschrift der Deutschen Morgenländischen Gesellschaft [New results of Maitrisimit research. Journal of the German Oriental Society]. 154. P. 347–369. [in German].

Geng Sh., Laut J.-P., Pinault G.-J., 2004b. Neue Ergebnisse der Maitrisimit-Forschung (II): Struktur und Inhalt des 26. Kapitels [New Results of Maitrisimit Research (II): Structure and content of chapter 26]. Studies on the Inner Asian Languages. 19. P. 29–94 + Plates III–XIII. [in German].

Erdal M., 2004. A Grammar of Old Turkic. Vol. Central Asia 3. Handbook of Oriental Studies 8. Leiden: Brill. 575 p.

Erdal M., Gippert J., Röhrborn K., Zieme P., Nevskaya I., Knüppel M., Özertural Z., Taube J., 2003. Vorislamische Alttürkische Texte: Elektronisches Corpus [Pre-Islamic Old Turkic texts: Electronic corpus]. [Electronic resource]. Available at: https://vatec2.fkidg1.uni-frankfurt.de/ (Accessed: 29.03.2026). [in German].

Derin M.O., Harada T., 2021. Universal Dependencies for Old Turkish. In Proceedings of the Fifth Workshop on Universal Dependencies (UDW, Syntax Fest 2021). Sofia, Bulgaria. Association for Computational Linguistics. P. 129–141.

Itkin I.B., Kuritsyna A.V., Malyshev S.V., 2017. Tocharian A text THT 1331 and the "Höllenkapitel" of the "Maitrisimit nom bitig": some more remarks. Tocharian and Indo-European studies. 18. P. 71–81.

Itkin I.B., Kuritsyna A.V., Wilkens J., Nugteren H., 2025. THT-fragments of Maitreyasamiti-Nāṭaka: Current state of the topic and some new identifications. Acta Orientalia Academiae Scientiarum Hungaricae. 1 (78). P. 85–113.

Lefer M.-A., 2020. Parallel Corpora. Magali Paquot, Stefan Th. Gries (eds.). A Practical Handbook of Corpus Linguistics. Springer. P. 257–282.

Kenning M.-M., 2010. What are parallel and comparable corpora and how can we use them? The Routledge Handbook of Corpus Linguistics. Editors: Anne O'Keeffe and Michael McCarthy. 1st ed. p. cm. (Routledge handbooks in applied linguistics). P. 487–500.

Malzahn M., Braun M., Fellner H.A., Koller B., 2011. A Comprehensive Edition of Tocharian Manuscripts. [Electronic resource]. Available at: https://cetom.univie.ac.at/ (Accessed: 29.03.2026).

Müller F.W.K., Sieg E., 1916. Maitrisimit und ‘Tocharisch’ [Maitrisimit and ‘Tocharian’]. Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften [Proceedings of the Royal Prussian Academy of Sciences]. P. 395–417. [in German].

Peyrot M., Semet A., 2016. A comparative study of the beginning of the 11th act of the Tocharian A Maitreyasamitināṭaka and the Old Uyghur Maitrisimit. Acta Orientalia Hungarica. 69. P. 355–78.

Pinault G.-J., 1999. Restitution du Maitreyasamiti-Nāṭaka en tokharien A: Bilan provisoire et recherches complémentaires sur l’acte XXVI [Restoration of the Maitreyasamiti-Nāṭaka in Tocharian A: Provisional assessment and additional research on act XXVI]. Tocharian and Indo-European Studies. 8. P. 189–240. [in French].

Semet A., Äysa A., 2014. Prophezeiung über die Maitreya-Geburt. Neues zum 11. Kapitel der uighurischen Maitrisimit nom bitig [Prophecy of the birth of Maitreya: New findings on chapter 11 of the Uyghur Maitrisimit nom bitig]. Aysima Mirsultan. Mihriban Tursun Aydın. Erhan Aydın (Hrsg.): Eski Türkçeden Çağdaş Uygurcaya. Mirsultan Osman’ın Doğumunun 85. Yılına Armağan. Konya. P. 221–249. [in German].

Tekin Ş., 1980. Maitrisimit nom bitig. Die uigurische Übersetzung eines Werkes der buddhistischen Vaibhāṣika-Schule. 1. Teil: Transliteration, Übersetzung, Anmerkungen. [Maitrisimit nom bitig. The Uyghur translation of a work of the Buddhist Vaibhāṣika school. Part 1: Transliteration, translation, notes.] Schriften zur Geschichte und Kultur des Alten Orients, Berliner Turfantexte [Writings on the history and culture of the Ancient Orient, Berlin Turfan Texts]. IX. Berlin: Akademie-Verlag. 264 p. [in German].

Weisser M., 2022. What corpora are available? Anne O’Keeffe and Michael McCarthy (Eds.). The Routledge Handbook of Corpus Linguistics. Second edition. Routledge. P. 89–102.

Wilkens J., 2008. Maitrisimit und Maitreyasamitināṭaka. Aspects of research into Central Asian Buddhism. In memoriam Kōgi Kudara. Edited by Peter Zieme. Silk Road Studies 16. Turnhout: Brepols. P. 407–433.

Wilkens J., 2021. Handwörterbuch des Altuigurischen. Altuigurisch – Deutsch – Türkisch. Herausgegeben von der Akademie der Wissenschaften zu Göttingen [Concise dictionary of Old Uyghur. Old Uyghur – German – Turkish. Published by the Göttingen Academy of Sciences]. Göttingen: Universitätsverlag. 929 p. [in German].

Wilkens J., 2023. Einige Beobachtungen zu Übersetzungstechnik der altuigurischen Maitrisimit [Some observations on the translation technique of the Old Uyghur Maitrisimit]. Journal of Old Turkic Studies. 2 (7). P. 553–571. [in German].

Wołk K., 2015. Noisy-parallel and comparable corpora filtering methodology for the extraction of bi-lingual equivalent data at sentence level. Computer Science. 2 (16). P. 169–184.

Information about the author:

Maksim Vladimirovich Vyzhlakov, PhD, Postdoctoral Researcher, Department of Linguistics University of Vienna, 3a Sensengasse, 1090 Vienna, Austria.

Scopus ID: 57219216180

Автор туралы мәлімет:

Максим Владимирович Выжлаков, PhD, лингвистика кафедрасының постдокторанты, Вена университеті, Зензенгассе 3а, 1090 Вена, Австрия.

Scopus ID: 57219216180

Сведения об авторе:
Максим Владимирович Выжлаков, PhD, постдокторант кафедры лингвистики, Венский университет, Зензенгассе 3а, 1090 Вена, Австрия.
Scopus ID: 57219216180

Conflict of Interest.
There is no conflict of interest related to this article.
Мүдделер қақтығысы.
Мақалаға байланысты мүдде қақтығысы жоқ.
Конфликт интересов.
Нет конфликта интересов, связанного со статьей.