Univerzita Karlova

Diplomová práce

Ústav obecné lingvistiky

Diplomová práce

Bc. Jan Židek

# Tocharian Loanwords in Chinese

Tocharské výpůjčky v čínštině

Praha 2017 *vedoucí* Ronald Kim, Ph.D., *konzultant* doc. Mgr. Lukáš Zádrapa, Ph.D.

I would like to express my utmost gratitude towards my thesis supervisor Ronald Kim, Ph.D. and consultant doc. Mgr. Lukáš Zádrapa, Ph.D., I would also like to thank everyone who took care of my academic needs throughout all those years of my university studies, especially Mgr. Jan Bičovský, Ph.D. who made me complete my bachelor studies and Mgr. Jakub Maršálek, Ph.D. who taught me basics of Classical Chinese, essentially pointing me in the direction of this work. I would also like to thank my mother for the material support and Buddha-like patience. Also, I hereby thank all my colleagues and friends who supported me in uncountable ways.

Prohlašuji, že jsem diplomovou práci vypracoval samostatně, že jsem řádně citoval všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 28.5.2017, .....

## Abstract

This work was created to review the evidence for lexical borrowing from the Tocharian languages to the Chinese languages. The used methodology relies on lexical lists, previous etymological findings, linguistic typology and anthropological input. For preparatory data manipulation, a set of semi-automatic scripts has been created. Presented is a qualitative research based on previous findings assisted by raw data. The outcome of this work should be testable findings which could be extracted to a computer processable form.

## Abstrakt

Tato práce byla vytvořena za účelem revize důkazů lexikálního vypůjčování z tocharských jazyků do jazyků čínských. Užitá metodologie spočívá na lexikálních seznamech, předchozích etymologických zjištěních, lingvistické typologii a antropologických informacích. Pro předzpracování dat byla vytvořena sada poloautomatických skriptů. Předkládán je kvalitativní výzkum založený na předchozích zjištěních, podpořený přímými daty. Výstupem této práce by měla být testovatelná, která lze extrahovat do počítačem zpracovatelné formy.

## Keywords

Tocharian, Chinese, loanwords, historical linguistics, reconstruction

## Klíčová slova

tocharština, čínština, výpůjčky, historická lingvistika, rekonstrukce

# Index

# Abbreviations and notation conventions

## Languages and families

- CT – Common Tocharian
- IE - Indo-European language family
- KS – Khotanese Saka
- MC – Middle Chinese
- ModJ – Modern Japanese
- ModK – Modern Korean
- ModM – Modern Mandarin
- ModV – Modern Vietnamese
- OC – Old Chinese
- PC – Proto-Chinese
- PIA – Proto-Indo-Aryan
- PIE – Proto-Indo-European
- PII – Proto-Indo-Iranian
- postPIE – post-Proto-Indo-European
- prePIE – pre-Proto-Indo-European
- prePST – pre-Proto-Sino-Tibetan
- PST – Proto-Sino-Tibetan
- PT – Proto-Tocharian
- PTB – Proto-Tibeto-Burman
- SKR - Sanskrit
- ST – Sino-Tibetan language family
- TA – Tocharian A
- TB – Tocharian B
- TC – Tocharian C
- WT – Written Tibetan

## Morphology

Morphemic transcription used generally follows the Leipzig glossing rules, modified and expanded by notation common in historical linguistics needed for transcription of diachrony and disambiguating plus sign. A special notation for infixes and reduplication is not used. An explicit mark of compounding is used.

- 1 – first person
- 2 – second person
- 3 – third person
- ACT – agens / nomen agentis marker
- CAUS – causative
- DEF – deference (speaker → hearer) marker
- DIM – dimunitive
- FEM – feminine
- LOC – locative
- NACT – nomen actionis
- NOM – nominative case
- PL – plural
- PST – past tense
- PSV – passivization marker
- RED - reduplication

- SG – singular
- TERM – terminative marker
- - morphemic boundary
- V verb
- N noun
- = boundary between a word and cliticon
- ° word interrupted at sub-morphemic division (for whatever reason)
- + hypothetical compounding or idiom-coining
- # word boundary as part of morphonology
- . syllable boundary[1]

## General

- L1 – mother tongue
- L2 – foreign language
- IL – interlanguage (intermediate stage when learning a language)
- JB – *Jianbo*, bamboo and silk script
- JGW – *Jiaguwen*, oracle bone script

## Special marks

- A > B – word A evolved into word B directly (inheritance)
- "X" – "meaning X"
- "A" ► "B" – *semantic shift* from meaning "A" to "B"
- A → B – borrowing of word A into L2 as word B
- A ⇝ B – indirect borrowing through intermediary
- A ↔ B – presupposed correlation between words
- /a/ – phoneme "a"
- [a] – allophone "a"
- *a – reconstructed speech sound "a"
- †a – incorrect word form, refuted reconstruction, projected descendant of possible preform
- ⁺a – amended/emended form
- **a – reconstructed pre-proto-form, dubious form or a projection; used also for Baxter-Sagart MC abstraction on attested forms
- A+B – word derived in language by compounding A-B, possibly a morphological adaptation process

## Symbols

- *α ambiguous front or central vowel in reconstruct
- *C any consonant (reconstructed)
- *D dental/alveolar plosive (reconstructed)
- *H undetermined PIE "laryngleal" (any of $h_1$ , $h_2$ or $h_3$ )
- *K velar consonant (reconstructed)
- *M plain voiced plosive (reconstructed)
- *MA aspirated (murmured) voiced plosive (reconstructed)
- *T unvoiced plosive (reconstructed)
- *V any vowel (reconstructed)
- D dental plosive
- R resonant (sonorant)
- L liquid (r-l sound)

[1] When transcribing words, the morphemic analysis follows the Leipzig glossing rules.

- [#] character omission due to technical restrictions, see corresponding number in the attachment Omissions

### Chinese transcription

Please note that most romanisations of modern Chinese before the mid-1980s[2] use Wade-Giles transcription. These were not emended when quoted. In other places, Hànyǔ Pīnyīn is used consistently where needed.

### Chinese characters usage

Unless referring to PRC-related entities (people, places, *Putonghua* usage), traditional characters have been used.

### Translations

Unless explicitly stated otherwise, all text translations were done by the author for this work.

[2] Probably following the ISO adoption of Pinyin in ISO 7098:1982.

# 1. Introduction

"My hovercraft is full of eels."
(people never fully understand each other)

The original idea for this work was to suggest a new approach to doing computer linguistics on unprepared data in historical linguistics and expand on our knowledge of early Indo-European – Chinese contacts using the method. The idea was that by relaxing the requirements on data in certain stages of preparation and handling, time-consuming tasks can be avoided. The research proved this idea impossible to use for the data chosen.

The now famous Tocharians, thus called out of respect to the tradition and a need for continuity, are probably not the historical *Tokharoi Τόχαροι* (who in turn were *probably Yuèzhī* 月氏/月支, see corresponding entry in work) who they were originally identified as (Kim 2006:725), they in fact present a separate branch of Indo-Europeans. A mutual influence of Tocharians and (Indo-)Iranians played a large role in the development of their culture.[3] (e.g. Mallory & Adams 1997:591).

The Tocharians were people living in the northern part of the Tarim Basin in what now constitutes a part of *Xinjiang* (Kim 2006:725) 新疆 province in the North-West of People's Republic of China. It is unknown when the language became localised there. (Fortson 2010:401) Nearly all written sources date to their late period, between the sixth and eighth centuries; in the ninth century, the languages probably went extinct (Kim 2006:725) with the complete assimilation of the community to the newly arrived Old Uyghur culture[4] (Blažek & Schwarz 2008:113).

Why is a possible borrowing of items from the Tocharians, a culture completely unknown to a layman and even some linguists and, most importantly, to many sinologists, so important that it deserves a coherent revision? No grandiose claims can be made. Still, it may help explain some details of the evolution of Central and East Asian cultures, where Persian, Turko-Mongol-Tungusic and Sino-Tibetan features are widely studied while other, ancient cultures, are largely left unnoticed by the majority of the scientific community.[5]

## 1.1. Brief history of the region

The early history is not well known. Later history is connected to the spread of Buddhism, conquest by Tangs and gradual Uygurisation.

Tremblay (2007) discussed spread of Buddhism in the Serindia, a region combining Northwestern Afghanistan with Turkestan, in the first half of the first millennium CE, consisted of a part of Western Iran, Bactria, Sogdiana, Ferghana, Kashgar, Khotan where Iranian speakers lived, Aqsu, Kucha, Agni, Turfan where Tocharian speakers lived, Loulan[6] with unknown vernacular, northern steppes where Xiongnu, Turks, Mongols and Tungusic speakers lived. Various religions coexisted there, of which Buddhism is of central interest. Sogdians seem to have helped spread the religion, no substantial attestation of their belief in Sogdiana has been uncovered, the state religion was Mazdaism. The buddhism in Tocharian territories seems to have been widespread among speakers of various languages. In between the first and third centuries CE, the kingdoms of the region adopted Buddhism, with translations came Bactrian and Saka borrowings and the Kharosthi and Brahmi scripts. Parts of the

[3] E.g. the development of writing – Khotanese Saka seems to have a nearly identical system (see Wilson 2005). Some sources consider Yuezhi a conglomerate that includes Tocharians. Whether real Tocharians were Iranian people is also a matter of debate.
[4] As my supervisor pointed out, this should not be understood as Turkic speakers not being already present.
[5] For the extent to which the influence of Tocharian culture seems to have extended, see Secondary literature in the Method section.
[6] Kroraina, cf. Tocharian C.

region became Tang protectorate at the end of the eighth century. Tocharian had influence on translation into Turkic languages. When the Uyghurs became the dominant power in the region (763-1008), they converted to Manicheism as part of their Anti-Chinese policy.

At least parts of Tocharian domain were ruled by the Chinese Tang dynasty from 648 to 790s CE (e.g. Ching 2011:64).

## 1.2. Delimitation of Tocharian and Chinese for the purpose of this study

As is widely known among linguists, the so-called Chinese language is actually a set of (in fact many) related languages and dialects. What is less widely known among non-specialists is that a similar situation is in Tocharian. While the name implies a single language, it is in fact at least two different (possibly three, developmental stages aside) languages.

In contrast to Tocharian, which belongs to the Indo-European family (Fortson 2010:401), Chinese is a branch of the Sino-Tibetan languages (e.g. SIL International 2017a).[7]

The developmental stages of Chinese are differentiated differently by different authors, they are simplified here as: (pre/post)Proto-Chinese, Old Chinese, (Early and Late) Middle Chinese and Modern; where most of the modern varieties are derived from Middle Chinese with the exception of Min varieties (Norman 1988:228-229). Like all languages with many speakers, even the Old Chinese is expected to have had dialects. (Schuessler 2006:6-7).

The Chinese languages central to this study are the Early Middle Chinese[8] and late Old Chinese[9], which together correspond roughly to early[10] Classical Chinese[11]. As with every reconstructed language, distinction between subsequent stages is often impossible to make and it is exactly the intermediate stage that is of interest for direct contact between speakers of Tocharian and Chinese.

Middle Chinese, as is usually reconstructed from written sources[12], is not to be understood at the direct ancestor of modern varieties, since it is a kind of *koiné* – an approximation or amalgam of dialects. (Schuessler 2006:1) Still, the rough approximation serves the purpose of this work.

---

[7] I do not believe there is any need to dispute that nowadays. Macrogroups are not proven to be of any relevance to genetic affiliation, and contact has not been proven either. It is my stern belief that while this is and will be untestable, current language families are the biggest groups that make sense in relation to history.

[8] To be taken here as a stage more or less ending with the beginning of Tocharian written records.

[9] In older sources and those following a non-updated terminology of Karlgren, Old Chinese is referred to as *Archaic Chinese*, which would seem to be a good translation of the indigenous term *Shang Gu Hanyu* 上古漢語, however, it is not in line with comparative linguists' terminology and may be misunderstood as meaning the archaic Chinese script form and practices. It will therefore not be used here. Confusingly enough, Middle Chinese is sometimes called Ancient Chinese, against after Karlgren. Yet more confusion may stem from my own usage of postOC, which could mean anything from the Western Han to the beginning of Tang and should be basically what a comparative linguist educated in IE languages would probably understand it as, against Schuessler's (e.g. 2016) term postOC which seems to be only the part after the end of Eastern Han as he reconstructs Later Han (LH) forms separately from OC, ONW (could be taken to be part of MC), and general MC without referring to the postOC in his work (Schuessler 2006). I have not used LH here to refer to any period so as not to make matters worse by making someone misunderstand it as the short-lived Later Han of the Five Dynasties, whose language would undoubtedly fall under MC.

[10] My term, means a span from the beginning of extensive written records to the beginning of Tang rule.

[11] Term is used here to refer to a written form with its own grammar, largely unchanged during its usage (until the fall of Qing dynasty, that is, beginning of 20th century). S. Starostin uses a set of his own terms where Classical Chinese would be one of the stages of Old Chinese. The use here basically agrees with note that it refers to written form only. The word early is meant here to differentiate it from Literary Chinese which might mean this period's written language, or the whole of premodern written language using the rules established in this era. Like many other sinologist terms, it is confusing and is therefore being confused often, that is one of the reasons I have tried to restrict myself to generic Proto-Chinese, Old Chinese and Middle Chinese with very broad intersecting periods.

[12] Baxter & Sagart (2014a; 2014b) stress that the form they give is not to be understood as a reconstruction. Certainly it is not one in the terms of comparative method, but since it is an abstraction of rules attested indirectly

In ideal situation, one would be only taking into account actually attested varieties of the languages in question. While that was to be done when the work was proposed, detailed study of material proved that is course is probably impossible. The timespan to be probed is based on Chinese periodisation to a larger extent than on the Indo-European one. It should contain maximal number of possible cognates with a certain amount of surety.

## 1.3. Tocharian

Tocharian is a centum[13] branch of the Indo-European (IE) language family (for precise positioning in the family see Mallory & Adams 1997:552-556).

Tocharian is subdivided into two languages, Tocharian A, and Tocharian B, or in older terms East and West Tocharian, respectively (Fortson 2010:402). Both languages are known from textual evidence, no spoken form survived to this day. The common endonym for the Tocharians, if any, is not known. (Kim 2012:725) The language of Kucha, TB was probably called *kᵤśiññe*, "Kuchean", and TA possibly *ārśi-käntwā* or "language of Agni".

In recent times, the so-called Tocharian C, started to be recognized as a third language of the branch, although it is attested only as a part of glosses in Prakrit from Kroraina.[14] "It consists of over thousand personal names and about one hundred other words." (Mallory 2015:6)

The texts in TA are linguistically homogenous, which leads some scholars to believe it was no longer spoken by the time of its writing, on the other hand TB shows variation; this may in fact show a possible diglossia. Of note is that the languages are mutually unintelligible (Kim 2006:725).

### 1.3.1. Writing system

The writing system is commonly referred to as *slanting* or *Turkestani*[15] *Brahmi* or *Gupta*[16]. It is an abugida[17] derived from its surrounding contemporaries, although from which is still a matter of some debate.

While very interesting, it plays little role in this work since the digitization of manuscripts by experts is being done using transliteration/transcription. The script itself does not have an officially appointed Unicode range up to this date[18] and even has some need to externally supply as advanced typesetting

in the period, it *is* a reconstruction similar to that which has been applied to Classical Latin. To mark that these forms are considered only a "representation of the information given" in MC (Baxter & Sagart 2014b:1), I have marked them accordingly with "dubious" marker.

[13] The Centum-Satem isogloss of (Post-)Proto-Indo-European was originally seen as a West-East (resp.) dialect split. With the discovery of the Tocharian branch, the concept shifted to a convenient grouping of languages that underwent some common changes. In Prague, it is generally not seen as a dialect division in the *Stammbaum* framework. I don't see much reason to reject the idea under the framework developed after the *Wellentheorie*. While this view should have no influence on interpretation of PIE data, it may be seen as possibly not in line with the mainstream and I feel that author's views of this kind should be spelled out so as not to become an external variable in data analysis.

[14] The sometimes mentioned possibility of connection to the fragmentarily attested Gutian language seems quite obscure.

[15] Specifically North Turkestani, as opposed to the variant used for writing Khotanese Saka.

[16] Gupta being shorthand for Brahmi from the times of Guptas. And Turkestan as one of the modern names for the general area where Tarim Basin is located.

[17] True abugidas of the Indic type would have ideally all the shapes of letter unchanged by an added diacritic. The Tocharian is more like Ge'ez in this respect for certain consonants – some characters include what could graphically be understood as a diacritic in conjunction with other characters, so they have to change shape. No irregularities of the Thai-type are there (inline "diacritics" and other features that effectively change the script to a non-linear one). There is a large number of ligatures.

[18] The newest version of Unicode is 9.0, early drafts of 10 do not seem to include the outcome of discussion on the proposal to include both of the related "Turkestani" Brahmi scripts – Tocharian and Khotanese. To my knowledge, there are only two fonts in existence – one by L.Wilson who submitted the proposal and one by yours truly, which was for the most part lost in a series of unfortunate accidents.

some of the characters. The result Is that either a scholar chooses to create their own non-standard font for the indigenous script or simply uses the transcription when working with larger sets of data unless there is a serious reason not to.[19]

Aside from the native script, Manichean is also attested (Hitch 1993).

### 1.3.2. Phonology

Peyrot (2015) introduces a simplified version of phonology thusly:

No distinctive length for vowels. TA <ā, a, ä> stand for /a, ʌ, ə/, TB /á, ə, a/ə́/. No distinctive voicing or aspiration for consonants. <ṃ> mostly denotes /n/. <ts> denotes dental affricate, <c> palatal stop *or* affricate[20]. <ś> considered palatal sibilant, <ṣ> is considered a retroflex sibilant. <ly> denotes palatal lateral. <ñ> is used for palatal nasal. Heavy consonant clusters are present. In transliteration, <$_u$> is used for non-syllabic vowel.[21]

### 1.3.3. Tocharian (B) morpho-phonology

Tocharian languages belonged to the synthetic type meaning the morphology is quite rich. Fortson (2010:406-412) shortly surmises these characteristics: Nouns had these cases: Nominative, Oblique, Genitive, Instrumental, Perlative, Comitative, Allative, Ablative, Locative, Causative. The number distinction was in singular, dual, plural and paral – a number for natural pairs, with TB adding plurative[22] There is a masculine-feminine-neuter genter distinction. Verbs had three stems: present, preterite, subjunctive. The present stem is divided into 12 classes and forms present, imperfect, present participle. Subjunctive stem forms subjunctive and optative. Preterite forms preterite tense and pret. participle. The morphology is relatively complicated and is not a central topic in this work, since a large part of it is undisputed, it is commented on at respective entries where it is relevant and no critical overview seems necessary.

#### Tocharian morphosyntax

##### Suppletion

Suppletion is one of the very popular terms in the last few years.[23] The term describes a phenomenon where forms in a single paradigm are not derivable by standard means of the grammar, e.g. English *was/were/will be.*

As e.g. Juge (1999) notes, strong suppletion are those instances, where suppletion is indisputable, the paradigm was supplanted by a form of a different word, e.g. English *is/am*. Weak suppletion are those instances, where no synchronic means of inflection/derivation are apparent, yet the forms are historically related in a way that is to be expected if the paradigm was regular. There are borderline cases where exaptation happened and a form with a certain function in a paradigm shifted to another position in a certain word but not in others.

The linguistic usefulness of subsuming the weak cases under the term may be a controversial subject, computational linguistics, however, should have a simpler view on this matter. Suppletion in both its strong and soft kind serve as a large hindrance to both (semi-)automatic data processing and processing

---

Update 17/04/2017: while version 10 does not list Tocharian, the recently published roadmap to 9.0.1 Supplemental multilingual plane does include Tocharian tentatively at *11e00 – 11e67 (*Unicode Roadmap Committee & Unicode Consortium 2017).

[19] To my knowledge, there is no dictionary and/or longer text collection using the writing in its digitized form to this date (18/04/2017).

[20] For reasons of shown later, the affricate is chosen here to be the only interpretation.

[21] Note that both languages are written in the same script and transcribed/transliterated using the same set of graphemes – not all are useful for both languages, however: a, ā, ä, e, i, o, u, p, t, k, c, ts, w, r, l, ly, y, tś, ś, ṣ, s, n, ñ, m, ṃ. The graphemes are mostly self-descriptory.

[22] He uses the term in the sense of a distributiveness.

[23] E.g. the Comparative linguistics department of Charles University hosted a conference devoted solely to it in 2016 and special databases are being made.

by human scientist when making a language comparison since every person is prone to mistakes when dealing with very large set of data. When dealing with quantitative methods, suppletion is a source of inability to deal with a problem by algorithmic, that is, analytical means only.

Suppletion in Tocharian is unfortunately an extremely common phenomenon, making the language a very hard one to deal with by standard means.

### 1.3.4. Indo-European evolution into Tocharian

For reasons obvious from data analysis done on the pre-existing literature, I will only discuss the evolution of regular TB outcomes.

The general description of the evolution from PIE to TB has been presented neatly by Mallory & Adams (1997:592), here further shortened (and $h_4$ left out):

| PIE | | TB |
|---|---|---|
| *p,b,$b^h$ | > | p |
| *t | > | t~c |
| *d | > | t~ts~ø |
| *$d^h$ | > | t~ts |
| *ḱ,ǵ,ǵ$^h$,k,g,$g^h$ | > | k~ś |
| *$k^w$,$g^w$,$g^{wh}$ | > | k~ś~$k^w$ |
| *s | > | s~ṣ |
| *j/i | > | y/(y)a~(y)ä |
| *w/u | > | w, TB w~y / a~ä |
| *m/m̥ | > | m/am~äm |
| *n/n̥ | > | n~ñ/an~än |
| *l/l̥ | > | l/al~äl |
| *r/r̥ | > | r/ar~är |
| *e/ē | > | (y)a~(y)ä/(y)e |
| *a,ā,o | > | ā |
| *o | > | e |
| *ū | > | o |
| *$h_1$ -$_3$ | > | ø |

For a more complete account refer, please, to Ringe (1996).

The relative chronology may give us some evidence on timing of borrowing, for reasons stated in the next section, it is not part of the automatic processing.

Some laws have been postulated to affect Tocharian. See e.g. Collinge (1985) for a partial discussion.

## 1.4. Chinese

Admittedly, the tradition of modern Sinologist reconstructions is somewhat shorter than that of Indo-Europeanists. Most reconstructions lack precise shape and the more we go into past, the more undetermined features of the system show up, in much similar, yet much more prominent manner, than in the Proto-Indo-European (PIE).

### 1.4.1. Old and Middle Chinese (morpho)phonology

The most prominent, referenced and widely used reconstructions of Old Chinese (OC) system in the West today were made by, in rough chronological order, Bernhard Karlgren, Edwin Pulleyblank, Fang-Kuei Li (李方桂), Sergei Starostin, William H. Baxter, Laurent Sagart, Axel Schuessler, Baxter & Sagart and possibly Shangfang Zhengzhang (郑张尚芳).

Karlgren (1957) was the first western systematic reconstruction, which was later in the 1970s revised by Li. Both are today considered outdated. Starostin's (1989) reconstruction later transformed into a part of his "Starling" online database while being enlarged and amended (For the reasons of APA compliance referred to here as Starostin 2006). It does not seem to be as widely referenced as his work on Sino-Tibetan (Starostin & Peiros 1991)[24], which has also been included in the "Starling" (to comply with APA referred as Starostin 2005). Baxter (1992) is still a partial standard, as it needs to be consulted for details along with Sagart (1999a) where the latest Baxter & Sagart (2014a) fail to comment. Schuessler (2007) reconstructs a system mostly compatible with those previously mentioned and is sometimes more complete at others less complete while taking into account only data that seem legitimate[25]. When doing research on Old and Middle Chinese, all of these have to be mentioned as none can be complete and none offers an explicit discussion on consensus.

Starostin (1989) is explicitly referenced already by Baxter (1992) and Zhengzhang (2003) by Baxter & Sagart (e.g. 2014a:115,115,213) effectively linking all works together. Comparing various reconstructions both needs to be present and needs to be brief. Therefore, tables depicting the systems described in two complementary up-to-date works Baxter & Sagart (2014a) and Schuessler (2007) follow.

#### 1.4.1.1. Old Chinese Phonology

No concise table of consonants or vowels as they are reconstructed s present in the two most-referenced works. What follows is an abstraction by the author, *none* of the sources explicitly list their system in a systematic manner.

[24] Interestingly enough, it seems to be used mainly by researchers interested in lexicostatistics, long-range comparison and macro-families.

[25] Explicitly stated is not adapting the forms to fit the Sino-Tibetan reconstruction (Schuessler 2006:122). That is not exactly correct since etymologizing is not done on forms, while information on semantics from cognates is part of the input. Personally, I find this approach to be most uncontroversial, however, as reconstructs could also be thought of as an algebraic system, sometimes results of this approach are lacking in usefulness where the near-homonyms-near-synonyms are not clearly distinguished.

| | | labial | | alveolar | | palatal | velar | | | uvular | | | glottal | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | plain | pharyngeal | plain | pharyngeal | plain | plain | pharyngealized | labialised | plain | pharyngealized | labialised | plain | pharyngealized | labialised |
| plosive | aspirated | pʰ | pˤʰ | tʰ | tˤʰ | | kʰ | kˤʰ | kʷʰ | qʰ | qˤʰ | qʷʰ | | | |
| | unvoiced | p | pˤ | t | tˤ | | k | kˤ | kʷ | q | qˤ | qʷ | ʔ | ʔˤ | ʔˤ |
| | voiced | b | bˤ | d | dˤ | | g | gˤ | gʷ | ɢ | ɢˤ | ɢʷ | | | |
| affricate | aspirated | | | tsʰ | tsˤʰ | | | | | | | | | | |
| | unvoiced | | | ts | tsˤ | | | | | | | | | | |
| | voiced | | | dz | dzˤ | | | | | | | | | | |
| fricative | unvoiced | | | s | sˤ | | | | | | | | | | |
| nasal | unvoiced | m̥ | m̥ˤ | n̥ˤ | | ɲ̊ | ŋ̊ˤ | ŋ̊ʷˤ | ŋ̊ʷ | | | | | | |
| | voiced | m | mˤ | n | nˤ | ɲ | ŋˤ | ŋʷˤ | | | | | | | |
| approximant | voiced | | | | | (j) | | | (w) | | | | | | |
| lateral | unvoiced | | | l̥ | l̥ˤ | | | | | | | | | | |
| | voiced | | | l | lˤ | | | | | | | | | | |
| trill | unvoiced | | | r̥ | r̥ˤ | | | | | | | | | | |
| | voiced | | | r | rˤ | | | | | | | | | | |

Table #-# - OCB – abstracted consonants Baxter-Sagart OC phonology[26]

[26] I would argue that a simple visualization shows that this is not a possible system if taken to represent a real language. The fact that atomization is probably impossible shows, there is something inherently wrong with it.

The Baxter-Sagart system involves voiced-unvoiced-aspirated opposition combined with labialisation feature for back consonants, distinction between velars and uvulars, and unvoiced sonorants. The system is non-defective, every position in a natural class is filled.

| | | labial | alveolar | palatal | velar | | glottal |
|---|---|---|---|---|---|---|---|
| | | plain | plain | plain | plain | labialised | plain |
| plosive | aspirated | (pʰ) | (tʰ) | | (kʰ) | (kʰʷ) | |
| | unvoiced | p | t | | k | (kʷ) | (ʔ) |
| | voiced | b | d | | g | (gʷ) | |
| affricate | aspirated | | (tsʰ) | | | | |
| | unvoiced | | ts | | | | |
| | voiced | | dz | | | | |
| fricative | unvoiced | | s | | | | |
| nasal | unvoiced | m̥ | | | ŋ̊ | (ŋ̊ʷ) | |
| | voiced | m | n | | ŋ | | |
| approximant | voiced | | | j | | w | |
| lateral | unvoiced | | l̥ | | | | |
| | voiced | | l | | | | |
| trill | unvoiced | | | | | | |
| | voiced | | r | | | | |

Table #-# - OCM – abstracted consonants as part of Schuessler (2007; 2009) system

Notably simpler, the Schuessler system does not involve cross-linguistically unattested consonants. The labiality and aspiration is not necessarily part of the oldest system, originally may only be a sequence of C+w/h (2009:xix).[27]

| | Front | Center | Back |
|---|---|---|---|
| High | i | | u |
| Mid | e | ə | o |
| Low | | a | |

Table #-# - OCB – abstracted vowels as part of Baxter-Sagart OC phonology

| | Front | Center | Back |
|---|---|---|---|
| High | i | | u |
| Mid | e | ə | o |
| Low | | a | |

Table #-# - OCM – abstracted vowels as part of Schuessler (2007; 2009) system

The Schuessler vowel system seems identical to the Baxter-Sagart. That may or may not be true, since Baxter & Sagart postulate additional consonantal features for the same reason Schuessler postulates circumflexed vowels series, which is stated to not reflect a vowel quality (directly) without explicitly stating what it stands synchronically (Schuessler 2009:xx).[28]

[27] Comparing grammars of languages like Czech and Thai with the proposed system, I have also decided to consider the glottal stop not to be a full phoneme, rather considering the final stop a feature of syllable connected to its composition (dead/live). This analysis has no overreach on the interpretations ensuing.

[28] Diachronically, the different sets stand for different reflexes (in LH and MC) of what seems to be attested as identical onset/vowel combination in earlier texts.

A note should be made concerning the phonemic length of vowels. Zhengzhang (2003) as some others before him believes in long vowels[29]. For the purpose of this study, this is considered a phonetic detail, although their reconstruction is based on the same data as reconstructions of others and so it has occasionally a certain influence (i.e. having mutual influence with consonant features) on the interpretation of the forms cited here.

Schuessler (2015) offers a critique of the Baxter & Sagart (2014a) system for being overly specific in features that no-one can be absolutely sure of while projecting them into the whole of PC-OC combination and opting for alternative theories, namely the pharyngealisation (Schuessler 2015:574-5). Quite correct is the critique of pharyngealised+aspirated ˁʰ/ ʰˁ series from a phonetic and typological standpoint where this combination is not only rare, it should not be possible at all. When a language with a pharyngealisation has triple voicing contrast, it is voiced, unvoiced and ejective (e.g. Ubykh). Aspiration is typically a phonetic detail emerging from tense-lax opposition/scale, like in English or Korean or it arises as secondary aspiration of a segment in contact with glottal fricative[30], often in systems that already have an aspiration (supposedly) e.g. Korean and Aryan languages. Baxter & Sagart (2014a:73) state that it is "quite rare" and the alternative is aspirated consonant in sequence with pharyngeal segment [ʕ][31]. That is not a very satisfactory alternative – still requiring a combination of the same phonetic features. The tense-lax opposition as a solution is not satisfactory either, as Baxter & Sagart (2014a:70-72) show, on account of comparative evidence. Since opposition velar-postvelar is not relevant to the compared material, the choice has been made here to preserve the aspiration and voicing while disregarding the information presented by pharyngealisation and/or special East Asian tense-lax which is not a real equivalent of voicing as in Indo-European language. Schuessler 2015:575 believes the system does not differentiate enough between different OC phases, this would be relevant here if the only consulted work were Baxter & Sagart (2014a), it shouldn't therefore pose a problem. One very important point for interpretation of Baxter & Sagart (2014a) is their reconstruction of different nasals according to Sinoxenic pronunciation which is sometimes quite problematic (Schuessler 2015:575-576). Forms commented on in the work presented do not suffer from this, with velar nasal realisations being undisputed.

In his review of Schuessler (2007), G. Starostin (2009:157) compares Baxter 1992, Starostin (1989) and Schuessler (2007), surmising that finals are compatible, while S. Starostin has different initials than Baxter and that Schuessler is informed by Baxter and Sagart (1999a) while choosing his own solution. Indeed, the initials Baxter's OCB, as Schuessler (2007) abbreviates it, and his own OCM are quite different at times as you can see from the tables. One thing they have in common is a large number of preinitials with unsure theoretical basis, which e.g. G. Starostin (2009) disputes. They are used here as part of argumentation, but they are not solely relied upon, therefore they should not pose a serious problem.

[29] These are present in some modern varieties also, it would therefore not be a counter-"sinoversal".

[30] Not a general laryngeal, though, which is rather part of the repair when adapting a loan – e.g. Czech uses unvoiced (post-)velar fricative [x] to mark aspiration having no aspiration on its own (*Thajsko* [txajsko] "Thailand").

[31] Although from context in reconstructions, what was possibly meant was [ħ] when using IPA, with preceding consonants being either unvoiced, sonorant (which is more often than not undefined for voicing, or transparent), or unsure and in only very few cases reconstructed as ⁺ bʕ/dʕ/gʕ sequence. Also of note is that a prevalent number of cases with pharyngealisation seem to have either a sonorant or a semi-vowel in them, leading back to the idea that it is a feature of the whole syllable. As for unexplained difference in treatment of *l- initial, I do not feel that there is any need to postulate any specific phonetic feature for a disappearing consonant, cp. e.g. ModK, Thai, etc. The pressure to somehow preserve a sound may be simply of a pragmatic source – it may represent a layer in the language, a register; this would explain why only in certain lexemes the phonemes are "fortified" while in others their pronunciation becomes non-phonemically lax in line with their general tendency, until they eventually disappear.

G. Starostin (ibid:157-8) criticises today's widespread use of *word families* which includes Sagart (1999a)[32] and Schuessler (2007), even though they are present in moderate amounts in the latter; the *word families* are comparable to the infamous phonesthemes[33] – words with *similar* pronunciation with *similar* meaning are treated as somehow belonging together without further evidence for its motivation.

Older systems generally differ in number of vowels and specifications of onset. One relatively new system which uses a larger number of vowels is Zhengzhang (2003), where an opposition of long-short and rounded-unrounded is present with a full set. The shift from a large number of vowels is general, where e.g. shift is visible in Baxter's dropping of high central vowel from his Baxter (1992) to Baxter & Sagart (2014a).

### Phonotactics

In Baxter & Sagart (2014a) system, every OC had an initial consonant (ibid:42), based on information from other languages[34] (e.g ibid:42), *preinitials* are thought to exist, of OC consonants only *j and *w cannot fill this position (ibid:51). They are either "loosely-attached (long variant prefixes)" or "short/tightly-attached (short variant prefixes)" (cf. e.g ibid:46-7, 54). The distinction may give an idea of being phonological, however, since it comes from morphological information, it is in essence morphonological, which is why the terms they distinguish are equated here. Abstracted into phonotactics, they either become part of a consonant cluster, become a minor syllable as they propose, or should form a real, full, syllable in case of long sonorants. While tightly attached preinitials were "simplified in different ways" in MC, loosely attached ones mostly disappearing "at times influencing the major syllable's initial" (ibid:52).[35]

From the standpoint of a phonetician, the idea that there are minor syllables with consonants that are farther from vowels than sonorants can exist while minor syllables with semi-vowels cannot exist is no less than strange. The problem may be rather in the definition of syllable than in the system, with some minor syllables actually not being sesquisyllabic at all. This would, of course, violate the principle of monosyllabicity of roots and problematize the reconciliation with one sign – (no more than) one syllable principle, nevertheless, it has been chosen at the solution how to adapt the system for the purpose of this work.

Maximal syllable has been postulated by Baxter & Sagart (2014a:53) as $C_1$ ə$C_2$ r$VC_3$ ʔ. Prefixes are expected to come before the preinitial (ibid:53-4). Others are much more conservative with solutions that could be surmised as $C_{prefix}$-CRVC, while looking more permissive at a first glance, all positions actually are far more restricted by rules generating the components, than in Baxter & Sagart (2014a) system. Of note is that in Baxter-Sagart system, pharyngealisation is postulated for the reason of preventing palatalization, non-pharyngealised consonant would therefore equate to *Cj in others (Baxter & Sagart 2014a:43).

Coda in Schuessler (2007:68-79) system and what is here presented as Baxter-Sagart's $C_3$ seems to be identical: p/t/k/m/n/ng/ʔ. Schuessler (ibid) notes that from PC to OC final -r was probably metathesized to a medial position.

For the Eastern Han timeframe, Schuessler (2007:120) believes there to already be no consonant clusters.

[32] And by extension Baxter & Sagart (2014a) which did not exist at the time when he wrote his review.
[33] A controversial sub-morphemic carrier of meaning.
[34] Loans and cognates.
[35] This could simply be interpreted as tightly attached being reconstructed with more certainty than the loosely attached ones, if we do not believe in separable prefix unattested as separate in OC or a reanalysis of often-preceding suffix as in Czech *ní* < *-n jí* (which could actually make some sense). Fusion of a word with a particle is considered to exist, e.g. Norman (1988:85).

### MC transcriptions

MC phonotactics as used by Schuessler (2007), Baxter & Sagart (2014a) and others allow only for a syllable CVC plus tone.

Since the language that is actually attested is considered to be *koiné* (as mentioned before) or at least a combination of various scribe's dialects, the exact shape of a general sound system would require a paper devoted only to this topic, which would also explain every dialect.

While the original purpose of this work would call for at least a partial treatment, the actual shape the work took based on truly unexpected findings makes this subchapter redundant.

The transcription of MC segments differs from author to author completely with Baxter & Sagart (2014a) being complicated by using only ASCII:

Baxter & Sagart (mostly 2014a:12-20) use <‘> for [ʔ] initial; nasals and dental plosives are also written in standard manner, other sounds are considered to be different across dialects and to account for that, they transcribe them in what they hope to be the easiest way (ibid:13). In short – the system is completely unreadable for the uninitiated and these forms should only be considered by those who understand the the reconstruction well. In short: -r- stands for palatalization of a kind, semivowels and vowel breaking (diphtongisation) are indicated by a sequence of a vowel sign plus semivowel sign where e.g. ju is different from yu not by general pronunciation, rather, by its treatment in respective dialects. Where the need arises, the pronunciation expected is commented upon in the text.

Schuessler (2007; 2009) uses transcription adapted from Baxter (1992), which is far more intuitive and should not pose a major problem for a linguist; for reasons of brevity: omitted, see Baxter (1992:27-32).

For treatment of tones, see 1.4.1.5.

### 1.4.1.2. Prefixes

Unlike modern varieties of Chinese, Old Chinese had a distinctive set of productive grammatic morphemes. Some of those proposed are listed in this and the next two subchapters in an **abridged**[36] version, details of those relevant are discussed in the dictionary part where needed, for others, please refer to the literature. The fact that OC probably possessed a morphology in the European sense does not mean that conversion was not possible and common.[37]

In OC, prefixes have been proposed by most authors to exist. In Baxter & Sagart (2014a), multiple (stacking) prefixation is possible, e.g. (ibid:54) “懶*[N-kə.]rˁanʔ > lanX> lǎn ‘lazy’; cf. pHmong *ŋglæn B ‘lazy’”[38].

Some of these prefixes are already unproductive in OC, some may even be petrified already in PC or PST. (see further).

Baxter & Sagart (2014a:53-57) postulate these prefixes (with details in Sagart & Baxter 2012):

OC *N- causes onset voicing[39], *Nə- disappears; typically V (verb)>V derivation.

[36] The simplification may cause slight differences in details with the original proposition.
[37] The fact that many characters can be used in almost any position in a sentence has led some to think that there are no word classes in Old Chinese. Zádrapa (2011) dispels that, also the simple fact that in different positions, the characters have different readings should convince even the completely uninitiated that this widely-held idea is a complete nonsense, since the readings are the actual words, not the characters.
[38] This shows what Schuessler and other have criticized, postulating improbable reconstructions based on comparative data where there is no need to presuppose the common origin of the full form in both/all languages.
[39] Effects are postulated for MC as part of their theoretical framework. The voicing part is often more important here than the nasality and there is no reason to rule it out as a coarticulation already in OC, if these prefixes existed.

OC *m$_{1}$ ,$_{2}$ - onset voicing, *mə-disappears; 1: V>V/N>V/V>N derivation adding volition, 2: (redundant) S marker with some classes.

OC *s$_{1}$ ,$_{2}$ - ; 1: increases V valency, 2: V>N.

OC *t$_{1}$ ,$_{2}$ - ; 1: intransitive V marker, 2: inalienable N makrer.

OC *k- ; sometimes V>N, other times unknown.

More generally, they mention these preinitials (ibid:46-49): *b(ə)-, *m(ə)-, *N-, *t- and general *C-. The first two seem to be needed for compatibility inside ST. The *t- preinitial has been proposed to account for some otherwise aberrant cases of palatalization.

As stated earlier, virtually any consonant may be preinitial in this system. The fact that the range of possibilities to fill the slot is so wide with no analysable semantics postulated should make this position in fact simply the initial in line with the saliency principles[40]. The saliency stemming from the position would then compete with the difference in saliency of the natural classes of $C_1$ and $C_2$ . It is hard to imagine how, e.g. [mt$^{h}$r-] combination could become something like voiced retroflex palatalised stop in combination with systematic changes that are at the base of other proposed contextual changes.

Schuessler (2007:16-19,24) proposes a simpler system, which does not differentiate between a preinitial and a prefix[41] but which abides the phonotactics:

ST>OC *m- introversion marker; ST>OC *s- extroversion marker (CAUS) + intensive/iterative, explaning MC<OC: s>ʑ/_l,j,w, sr>ʂ.

Voicing of initial consonants is supposed to have morphological role.[42]

Of note is that “Most OC morphemes are ST because they also occur in TB languages.” (ibid:16)[43]

#### 1.4.1.3. Suffixes

Baxter & Sagart (2014a:58-59) list three *-s suffixes: 1: most common, V>N; 2: N>V; 3: V>V (in Schuessler terms) endopassive to exoactive.

Schuessler (2007:16) ST>OC *-s/-*h PST/PSV and transitivisation; ST>OC *-k “of unknown function”. In contrast to Baxter & Sagart who propose complex prefixation, Schuessler (2007:17-18) proposes a rather complex suffixation – but moves it into PT with OC having these no longer productive and derives it from internal Chinese data with other branches serving as evidence, as stated before[44]: *-n$_{1}$ ,$_{2}$ 1: (redundant) N marker, 2: 3pers. pron.; *-ŋ TERM; *-t (redundant) N marker; *-k distributive marker. (ibid:40) he speaks of MC tones as morphemes where tone B (*shǎngshēng*) has an endoactive meaning and should go back to <OC> *ʔ.

#### 1.4.1.4. Old Chinese Infixation

Most transcriptions of OC and/or MC work with *-*r*- infix. As stated, e.g. by Schuessler (2007:19), it is not clear, whether this was truly an infix or prefix in OC and by the time of MC, it has blended with the initial consonant. While it is called *an* infix, Baxter & Sagart (2014a:57-58) identify at least three functions, in action verbs it marks distributiveness, in stative verbs it marks intensiveness, and in nouns marks distributed structure.

[40] I.e. the closer to the beginning the more salient the sound is.
[41] On the basis that preinitial is unidentified prefix.
[42] Thereby omitting the need for complicated nasal prefixes. In fact a traditionalist view.
[43] Which Baxter and Sagart take further, projecting *everything* of this kind from Tibeto-Burman into PST > PC > OC, hence their extensive prefixation scheme.
[44] Not all etyma that separate them are taken as proven here.

Other infixes are postulated, among them most important *-*n*- by Schuessler (2007:22-23), supposedly coming from an Austroasiatic source, it being a dialectal substrate (ibid:4-5).[45]

### Morphosyntax

The basic word order was Subject-Verb-Object, however, there were many constructions that violated that.[46]

Old Chinese had a productive reduplication mechanism[47].

From modern sources, at least Schuessler (2007:25-25) considers the possibility of the existence of productive re-analysis, backformation and re-cutting,[48] metathesis and convergence[49].

#### 1.4.1.5. Tones and togenesis

Since the possible contact between the languages has not been lined down to an exact time frame, tonogenesis, the emergence of supraphonemes[50], may play a role in how the words were being borrowed. Specifically, some may have been borrowed with the adaptation of tones while others without it.

Tonogenesis is traditionally supposed to arise from simplification of syllable when trying to preserve a meaning distinction while an unstable phonological system is reducing the number of phonemes by the way of reducing distinctive features. When a primary feature is being marginalised, the phonetic detail helps and secondary feature(s) takes over (a revision of the traditional view, for Vietnamese, is presented by Thurgood 2002)[51].

As described by Sagart (1999b), tone as part of morphology is not expected to have existed during the Old Chinese period until its latest stage and is first described in the Early Middle Chinese, reportedly by *Shen Yue* and *Zhou Yong*. The four tones (sìshēng 四聲) of MC are level tone (*píngshēng* 平聲), rising (*shǎngshēng* 上聲), departing (*qùshēng* 去聲) and entering (*rushēng* 入聲). The phonological status of entering tone in Middle Chinese as a whole is questionable – it only occurs with a plosive coda. When following the literal interpretation of rime tables, the tones are supraphonemes while final nasal-stop alternation is seen as allophonic.

The modern varieties' tones are not derived directly from these tones (Baxter & Sagart 2014a). The author's obvious conjecture is that they cannot therefore be, in effect, derived from the original consonants of a possible loanword. Unlike with IE cognates, there is therefore no simple set of rules that can be postulated to account for every modern phone – loanword phone correspondence algorithmically.

For the purpose of this work, the phonetic detail in realisation is of no matter as is phonematic status of the tones. Whatever the case, the loanword adaptation must have respected the segmental properties.

There are historically three main notation standards in the Western scholarship for MC tones:

---

[45] If true, it should probably not be understood as productive in OC as a whole and as Austroasiatic speakers are geographically removed from the early IE speakers, it should be therefore ruled out from being part of morphology in contact dialects in question here. For this reason, it is not part of the argumentation presented here, unlike the *-*r*- infix(es).

[46] For details refer to one of the standard grammars of Classical Chinese, von der Gabelentz (1881).

[47] The topic is complex and need not be treated here above the level of the statement that it did exist, both partial and complete and developed over time, for details see e.g. Sun (1999).

[48] I.e. rebracketing.

[49] Two words' meaning influencing each other because of their forms being similar. I do not believe this to be a widespread phenomenon.

[50] Superphonemes, suprasegmental phonemes.

[51] Models postulated for various languages take into account loss of distinctive vowel length, onset voicing, loss and simplification of coda and various other reasons. For most languages, the specific mechanism is disputed. What can be seen from a spectrogram is that between any two speech sounds in realization of any spoken language, there are slight movements in pitch.

Karlgren (1957) and Li (1971) use colon for rising tone and hyphen for departing tone, no notation of the other two. As the entering tone cooccurs with final stop consonants, marking it further would be redundant.

Schuessler (2007:xi) uses ABCD for level, rising, departing and entering tone resp.

Baxter & Sagart (2014a; 2014b) use X for rising tone and H for departing and nothing for the other two tones in their notation but in tables they also indicate ABCD in fanqie transcription.

| | Karlgren, Li | Schuessler | Barter & Sagart |
|---|---|---|---|
| Level 平 | | A | A |
| Rising 上 | : | B | X B |
| Departing 去 | - | C | H C |
| Entering 入 | | D | D |

Table #.# Tone notation

### 1.4.2. Writing system

The Chinese writing system, or *Hanzi* (漢字/汉字), while logographic in essence, has some very useful properties for reconstruction of earlier stages of its existence and thereby of the archaic language features it preserves.

There are in parallel to the primary[52] (onomatopoeic) nouns and verbs, also primary characters – their shape should somehow reflect the concept they depict, they are ideographic[53]. Next, there are diagrams which may or may not be primary. Next, there are indexes, true logograms whose shape is derived from another character. Historically, in many instances, the "coining" of such characters did not include modification of the base character (for an abstract concept usually a similar sounding character), so in reality, these were cases of allography, only later were they modified to include a radical (a part derived from a semantically connected word) or a phonetic (a part derived from a similar sounding character). The original near-homophonic allographs are now widely called phonetic borrowings. The last type of character by the means of its creation we may set apart is a phono-semantic compound of a radical and a *phonetic*. On the surface, many times, they fell together with the previous type nowadays.[54,55]

#### Rationale for using the characters for etymologization

It is the author's belief that when sieving through the etymological dictionaries, we can usually identify them, since, unlike with the spoken form of a language, when the writing is concerned, as long as the corpus of older texts is extensive enough, we can be fairly sure of whether an attestation of a certain grapheme stage can be found if it is hypothesised to exist.

[52] Non-derived.

[53] For the purpose of this introductory paragraph, I have left out the pictograms as a separate category, since pictograms should have no connection to the language and therefore, strictly speaking, do not constitute a (part of) writing system. While it would be useful to set them apart from the rest if this was a palaeographic research, in a linguistic research, having them included in ideograms should prove an adequate simplification.

[54] Since these facts are widely known to people who deal with the Chinese writing in any way (e.g. every literate Chinese person), I felt no obligation to cite any particular source. For a good, far more extensive explanation, see e.g. Norman (1988:58-82), Slaměníková (2013), Pejčochová & Zádrapa (2009) (the last two in Czech).

[55] The traditional classification simplified here actually goes back to (Sturgeon 2011) the *Shuowen Jiezi* 說文解字 (late Han dynasty character dictionary) and is in full: 象形, 指事, 會意, 形聲, 假借 and 轉注, that is, pictogram, ideogram, combined ideogram, ideogram plus phonetic, loan and transfer, resp. Together, they form 六書, or the Six methods (of Hanzi forming).

Potentially every derived character which contains a phonetic component should be very useful to us since it contains a bit of the information on the then-current state of the spoken language in the area[56], i.e. where could the character's coinage be put in relation to the *lect*. In extension, we may therefore, by identifying certain character's history, possibly identify a certain word's history. The Chinese themselves have done that with varying success from the ancient times and so did everyone who undertook the effort of attempting an OC reconstruction.

As stated by Baxter & Sagart (2014a:2-4) the main source of reconstruction of MC[57] are traditionally the native so-called Rime (or Rhyme) dictionaries and tables; they also include explicitly Sinoxenic, i.e. the Chinese part of lexicon in non-Chinese languages (primarily Japanese, Korean, Vietnamese, to a lesser extent Thai, Khmer and other languages of East and South-East Asia).[58]

Rime dictionaries were an evolution of earlier dictionaries which included a systematic description of pronunciation. Rime tables were a reference material for explanation of Hanzi pronunciation and were being created in the Late Middle Chinese period.

Rhyme dictionaries 韻書 (Baxter-Sagart "rhyme books") used the *fǎnqiè* 反切 method of explaining reading, it gave a character for initial (onset) and final (rime and tone) for every homonym group of characters. The most important among them is *Qieyun* 切韻 from 604 CE and *Guangyun* 廣韻 from from 1008 CE (Baxter & Sagart 2014a:9-12).

The oldest two variants of Chinese writing system attested are Oracle bone script 甲骨文 *jiaguwen* (from here onward shortened JGW) and bamboo and silk manuscripts 簡帛 *jainbo* (from here onward shortened JB; also 簡牘 *jiandu*).

The modern system of standard characters is well coded in the Unicode in the CJK[59] block even in older versions. No problems with computer-assisted analysis were expected.

### 1.4.3. Sino-Tibetan evolution

The evolution of Sino-Tibetan has not yet been fully explored and agreed upon. If there were minor syllables present in PST, they were either preserved into PC or even OC in the likes of Baxter & Sagart (2014a), or they were already simplified with only a few possibilities remaining, in line with most of the other reconstructions. Exact **mapping** of both vowels and consonants in any transitional phase is problematic, since for reconstruction of a contact, even phonetic detail is important and while the reconstructions used today are largely compatible, since their differences are often not systematic, they cannot be accounted for in an elegant form that would show the benefits of proposed methodology. For use in automatic processing here, the information that is under consensus is largely unusable.

[56] Baxter & Sagart (e.g. 206-207) also expect, no doubt correctly, when those creating Rime dictionaries used their own dialects when creating them, adding information on *their* dialects. This is in fact one of basic parts of their reworking of methodology – to account for aberrants.
[57] And by extension OC.
[58] Schuessler, on the other hand, believes in searching for substrate *in* the OC.
[59] Chinese-Japanese-Korean, i.e. languages that use these characters. Sometimes written CJKV to include Vietnamese, the *Chữ Nôm*, however, were never standardised and therefore are only mapped to an extent and only in rather recent versions of Unicode (since 8.0, with full fonts support still lacking), language-specific character encodings exist with good font support, though.
None of the non-standard characters are to be used in computer processing.

# 2. Method

A semi-automatic processing of wordlist data has been used in combination with the study of literature and check against general constraints and actual texts.

## 2.1. Borrowing as a principle

There are many reasons for borrowing. Before modern language contact models were created, borrowing was thought to occur only when a language is missing a term for a novel concept. While that may be true in many cases, there are other motivations to be considered:

Bilingualism, language shift[60], diglossia[61] – extensive language contact leads to lexical transfer. Some social configurations and situational contexts are obviously more prone to support. Also some language-internal factors play a role.

Winford (2003:11-24) identifies three basic contact situations: language maintenance (borrowing serves as interference), language shift, creation of contact languages while problematising this division himself. The degree of borrowing ranges from a few lexemes to the incorporation of structural features. On his scale (ibid:23-24), the expected contact between Tocharian and Chinese would be casual borrowing situation under the language maintenance, with later convergence where an intense pressure on a minority groups (Tocharians in Tang China) would be expected – the expected outcome of "heavy structural diffusion" however does not happen. Also, we would expect from material evidence an intense inter-community contact already in earlier times, again, no heavy lexical diffusion is to be seen. Language shift has been postulated from Tocharian to Old Uyghur, this is, however, mostly irrelevant for this study.

The social aspect of the contact has not been well studied in the sense of general attitudes. The probability of a large number of multilingual speakers is high inside Tocharian space. The same is probably not true for the other side.[62]

Winford (2003) follows Van Coetsem in distinguishing between borrowing and imposition, where borrowing happens from L2 to L1 while imposition[63] is the other direction. For lexical borrowings, he proposes a classification (ibid:384, modified) following Haugen: 1. loan words: Direct loans, loan blends (morphological adaptation, etc.); 2. loan shifts: semantic extension (of an L1 word), loan translation (calque); 3. creations: (hybrid) creations (newly created words for foreign concepts), creations using only foreign morphemes (reverse of the previous).

### 2.1.1. Borrowability scale

As cited by Field (2002:35), the original idea of hierarchy of borrowability goes back to W.D Whitney and it was van Hout & Muysken (1994:41) who proposed that a basic hierarchy is thus: "Nouns > other parts of speech > suffixes > inflections > sounds."[64]

Based on other works, Field (ibid:36) summarizes that furthermore "Nouns > adjectives, verbs" and that (ibid:38) "content item > function word > agglutinating affix > fusional affix".

If we are to believe this scale, then the complete set would be:

[60] The gradual shift of speakers from one language to another spreading in domains until one language is not used anymore.
[61] The situation where one language has higher social status than the other, leading to the speakers of lower variety using the higher variety in some contexts.
[62] As attested by the number of borrowings *from* rather than *to* Chinese, see further.
[63] Traditionally *transfer*.
[64] By sounds, the author surely means phonemes, speech sounds are obviously being borrowed and adapted as part of the loan on any account.

`[Nouns > verbs, adjectives] (content words) > other POSs > agglutinating affix > fusional affix > phonemes.`

While this set looks intuitively correct, Field (ibid:41) goes further and postulates the Principle of System Compatibility (PSC) in dichotomy with the Principle of System Incompatibility (PSI):

PSC "Any form or form-meaning set is borrowable from a donor language if it conforms to the morphological possibilities of the recipient language with regard to morphological structure."

Versus PSI "No form or form-meaning set is borrowable from a donor language if it does not conform to the morphological possibilities of the recipient language with regard to morpheme types."

In other words, if the borrowing language has more complex morphology, it should readily borrow grammatical items, if it is simpler, it should not be able to borrow any part of grammar that is incompatible with its morphology. Obviously, that is too strong a statement as no language has a morphology simplifiable to the absolute macro-types[65]. Also, it should not be understood as meaning that grammatical morphemes are not borrowable into isolating language as part of a prosodic word, only that the retention of their grammatical function is hampered. Also, when the contact is extensive, the morphology of languages could arguably change – for a non-controversial example see e.g. the so-called *xenoclisis*[66] of Romani. This is, by no means, an attempt to imply that Chinese should have any kind of Tocharian-derived grammatical words.[67]

As the words are usually borrowed not in isolation but in a context, it is not only the dictionary forms that are transferred from one language to another. Special case is creolisation and pidginization, where in the process of lexicalisation, the L1 words are more often than not completely misunderstood not only on a semantic level, but also on a syntactic level[68] - since spoken language prosody is not strictly bound to the syntax (cp. e.g. the definition of a word in morphology vs in the phonetics).

The borrowability scale is not used as a rule here for the reasons stated, it is used as a motivation device.

### Language universals

Before we delve into the topic, it is important to say that the author does not consider himself a proponent of the strong idea of language universals. There seems to be actually a strong negative correlation between surety of findings and their usability, in other words, absolute universals are absolutely useless while frequency universals are usually no more than mere descriptors, they have a minimal predictive power. That being said, some may be used as a hint on restraints and motivation for repairs (i.e. adapting to the phonotactics and morphology) where no other indicators are present.

### Semantic similarity

Semantic similarity of similar forms in different languages, even for those in contact, should by itself not constitute a basis for a decisive statement on whether the words in question are cognates. There is a large number of words that are by pure coincidence quite similar. It is not by chance that the work's findings support this notion.

---

[65] Analytical, agglutinative, flexive.

[66] Term used by specialists and some general linguists (furthered in Prague by V. Elšík) to refer to the peculiar system of declination/conjugation in Romani, which has been borrowed as a whole with a large number of lexical items from Greek. While it works as part of the system, it also marks (relatively) recent loanwords. Its counterpart is *oikoclisis*. Not used as such, it could be extended e.g. to the ModJ honorific prefixes written 御：お/ご (read as o-/go-, resp., in front of Japanese / Sino-Japanese words).

[67] I would argue, however, than more often than not, when an isolating language borrows a word from a flectional one, it will borrow it with morphemes attached (e.g. infinitive, 3sg, ergative, absolutive markers, which are obligatory in "neutral" contexts in languages possessing them).

[68] For example, in some French-lexicalized creoles (e.g. the Antillean Creole) *dlo* is thought to be ultimately from French *de l'eau* "of water". In this case, a prosodic word has become a word in a syntactic sense.

### 2.1.2. Segments adaptation

When a word is being borrowed, from a synchronic view, it has to be adapted, that is, the phonological system of the donor is expected to have some incompatibilities with the recipient and these are to be dealt with so that it can be pronounced, so that the word can be integrated into the lexicon.

The adaptation happens as a repair either at the time of borrowing, or online for not fully integrated items.

Among the basic strategies are:

- Place of articulation approximation
- Manner of articulation approximation
- (De)voicing
- Unpacking (epenthesis)
- Elision
- Blending

Why speaker choose unpacking has been studied by Vendelin & Peperkamp (2004). In the paper, their premise can be surmised as (ibid:1-2): 1. the indicators that lead to the speaker's choice whether to perform unpacking are not necessarily on the phoneme that is being unpacked, the reason may be stemming from another phoneme with a secondary feature that defines the properties of the phoneme in question (vowel tenseness for English final stops). 2: That the adaptation lies in decoding, rather than coding.

When speaking of borrowing in diachrony, one usually reconstructs a word to a form that approximates some timeframe and compares it with another approximation in another language, searching for what seem like exact matches. The concept of phonetic adaptation then is removed from the process. When searching for cognates in contact with Chinese, working with this illusion is common, as shown in 3.3.

Kenstowicz & Suchato (2006:4) suggest that Frisch, Pierrehumbert & Broe's (2004) formula

```
(i) similarity = shared natural classes / (shared natural classes + unshared natural classes)
```

could be applied to loanword adaptation. The original idea was to incorporate it into a script for automatic assessment. In reality, we may never know enough information about features of reconstructed languages in contact – while similarities can be typically analysed, some features are by definition undefined, underspecified.[69]

Matras (2007:37) speaks of three types of sound change related to borrowing: incorporation of L2 phonemes, adjustment of phoneme to fit L1 system, incorporation of borrowed phoneme into inherited words.[70] He goes on to postulate several implicational hierarchies: (ibid:37) C borrowed > V; prosodic features > segmental features; phonological features in a loanword > independent features (ibid:38).

#### On suppletion

If a regular unattested word-form that contradicts the information provided by analysis showing a suppletion corresponds to the supposed cognate in the recipient language, it is taken a proving this form existed at a certain time. If the supposed cognate supports a weak suppletion in the donor language (i.e.

[69] For OC, that is the fricative – trill status of *r, definition of vowels for frontness (as attested by the need of some authors to postulate pharyngealization), a place of articulation of certain onset phones, and even many times their presence. Not every language is as underspecified as OC of course. The Baxter-Sagart overspecification much criticized by other authors actually shows promise, being an encoding of features, whichever they may be, may lead to its analyzability by some of the heuristic methods for cognate search.

[70] The first and third do not apply here.

the correspondence is not perfect and not reconcilable with adaptation principles), it is an argument against these words being cognate, not a basis for postulating a suppleting form.

### 2.1.3. Tone adaptation

Every language has its own phonological system. Tone languages have a system that requires a syllable to be defined for tone, therefore when borrowing from other languages, they need to adapt not only segmental but also the suprasegmental part of a word.

Theoretically, the adaptation could take into account the original tones in case of a tonal language or a language which has tonal accenting. In those cases, the tones should correspondent somehow. When borrowing from L2 where there is no lexical tone, there may be indicator of correspondences which emerged as secondary characteristics of a dynamic accent in the source language. There may also be no inherent tone in the word being borrowed in which case the language must generate it by its own rules. There are, however, languages, where this is done in all cases, like Thai.

Modern Mandarin loanword tone adaptation has been studied[71], yet tone adaptation for older stages is generally neglected to the point that tonogenesis in some languages is or has been thought of as arising from contact with Chinese (e.g. Sagart 1999:11 holds this view), e.g. in Vietnamese[72] so the correlation of tones would not be considered part of adaptation in Chinese but rather the spread of its characteristics to the other language.

As stated earlier, tone proper should not play a direct role in adaptation of loanwords in this study since the stages of interest still had segmental correlates which in contact with non-tonal languages would be exactly the markers of what tone should result.

## 2.2. Source material

Sources available to the public for studying both languages have in the recent years begun to become relatively abundant considering that all of these languages are reconstructs. For this reason, the term “primary source” used in the next subsection should not be understood as meaning raw data, it is not the purpose of this study to revise reconstructions of respective languages as such.

### 2.2.1. Primary

Primary sources used are dictionaries of Tocharian and Chinese. For additional information, text databases have been consulted.

Most important source of computer-processable data for Tocharian is the CEToM. CEToM, which is short for a Comprehensive Edition of Tocharian Manuscripts (referred to here in the APA format as Malzahn 2017), is an attempt at creating a digital database of all currently discovered manuscripts in Tocharian. The digitization of text is done in Unicode transcription/transliteration mix. It includes a large set of manuscripts, most with photographs of varying quality, all with metadata describing the content (inventory number, place of discovery, language, etc.) and most with transliteration and possibly edited transcription and translation.

This source has been used as an input data into scripts (see 2.4.2).

Most important source of lexemes for Chinese are the Baxter & Sagart’s addenda to their 2014 magnum opus (Baxter & Sagart 2014b) which are conveniently published on-line in XLSX format for use in computer-aided research. The Middle Chinese wordforms are explicitly stated to be renderings of features representing the fanqie readings, an intermediate stage between the reconstruction of Old

[71] E.g. Miao (2005)

[72] The phonological tone in Austroasiatic languages is quite rare leading to some early hypotheses of Vietnamese tone being borrowed along with a large number of words from Chinese. While this has since been refuted, the language contact is still seen as enforcing the pre-existing cline by many.

Chinese and the Modern varieties without strictly being bound to them. This notation should suffice for the comparison itself, it should prove to be problematic for the diachronic alignment still.

Primary literature in this case means dictionaries of both Chinese and Tocharian.

#### Discussion

Other possible sources are:

TITUS, or *Thesaurus Indogermanischer Text- und Sprachmaterialien* (roughly "Thesaurus of PIE text and speech material") – is basically a database of partially annotated texts from old Indo-European languages. The outputs of the project's Tocharian branch[73] have been integrated into CEToM (Malzahn 2017 "about…"). It has therefore not been consulted.

Schuessler (2007), which is used also as a tertiary and to an extent as a secondary source, could not be used for technical reasons in automatic processing. For reasons cited at other places, it is also considered to very reliable when used in conjunction with other sources. It is used in combination with B&S2014 and Baxter (1992) and sometimes other for manual data analysis.

Chinese Text Project – one of Chinese classics databases, shortened usually CText (referred to here as Sturgeon 2011). This was used for two purposes: as a text database, it was used to search for early attestations without considering it complete, i.e. considering unattested words in major works as such without extending this to a statement that they did not exist at an earlier time (the dating of works is taken over from there). Second, as a source of OC and MC views on etymology of characters.

Starling, an exhaustive and easily computer-processable etymological database that includes data from most well-studied world languages as well as reconstructed languages nearly to the point of Proto-World has been considered for an inclusion as a primary source and rejected, for justification see comments on dictionaries and consensus.

These sources have been used only to facilitate and supplement the framework for data manipulation.

### 2.2.2. Secondary

There are no, strictly speaking, comprehensive descriptive grammars of the kind of modern general linguist's grammars for either of the languages in question. For Chinese, the closest to it is Sagart (1999). Learner grammars have been written for Classical Chinese (i.e. the written form), among them famous von der Gabelenz (1881). Norman (1988) is to be considered an encyclopaedia or a handbook. Sketches of grammar have been presented by various authors as part of their presentation of the reconstruction. A lot has been written on the topic of Tocharian grammar, some of those works are referenced here. A learner's grammar is Pinault (2008) and of course dictionaries as part of their system describe some part of grammar.

The following subchapters surmise findings scattered in literature on the general topic.

#### 2.2.2.1. Indo-European loanwords in Chinese

WOLD (Wiebusch 2009) lists only 15 entries with absolute certainty of a loanword status in Chinese, of those: *shīzi* 獅子 "lion" also discussed here < Persian *šer*, *níngméng* 檸檬 "citrus" < either Persian or English or Arabic, *bōli* 玻璃 "glass" < Sanskrit (SKR), *héshang* 和尚 "priest" < SKR, *sēng* 僧 "priest" < SKR, *bāshi* 巴士 "bus" < English, *mǎdá* 马达 "motor" and 咖啡 *kāfēi* "coffee" from a "European Colonial Language". As *probably borrowed* (of 11) are classified *tǎ* 塔 "tower" < SKR, and *mǎ* 馬 "horse" from an "Unidentifiable Indo-European" or from an unidentified source, which is also discussed here.

[73] It contains manuscripts from the Berlin Turfan collection, London collection and Paris collection (Gippert & Martínez & Korn 2016).

In OC and MC times, loanwords have been postulated from Indo-Iranian languages, Indo-Aryan (as seen above, pertaining to religious topics), and possibly Tocharian were sources in literature presented in this work.

### Borrowing into Chinese from other languages

From Schuessler (2006), Baxter (1992), Baxter & Sagart (2014a), Dybo (2007)[74], it would seem that the primary sources of borrowings from non-IE languages in earlier times were Turkic, Tungusic and Mongolic languages; Austroasiatic might be considered a substrate language. Other languages of the area did have contact.

#### 2.2.2.2. Previous studies on Tocharian-Chinese language contact

Most of the studies on the contact between Tocharian and Chinese deal with borrowing from the latter to the former. They are usually in the semantic fields of administration, titles, measures, calendar, crop, crop produce and cultural items; and are more numerous in TB. What follows is a non-exhaustive list.

Lubotsky & Starostin (2003:262-265) surmise in, to this date the most complete list, based on previous Adams' first edition of (2013), these borrowings from Chinese in Tocharian: TA TB *klu* "rice", TB rapaññe "last month of the year" < "winter sacrifice", TB *cāk, tau* "(dry measures)", TB *cāne* "money", TB *śakuse* "brandy", TB *ṣaṅk* "(measure of volume)", TA *yāmutsi* TB *yāmuttsi* "waterfowl" < "parrot" (with a note that this is either through Iranians, or not true); following Grenet and Pinault, they list also TB *ṣitsok* "millet alcohol" – probably contaminated by TB verb for "to drink", TB *ṣipāṅkiñc* "abacus; to which they add their own TA TB *cok* "lamp", TA *truṅk* TB *troṅk* "cave", TA *ri* TB *rīye* "town" (also see an entry in wordlist), TA *lyäk* TB *lyak* "thief" (refusing the IE etymologies previously postulated), TA < TB *tseṃ* "blue", possibly TA *nkiñc* TB *ñkante* "silver" and very improbable TB *kapci* "authentication". Of note is Toch *-a-* in *cāk* vs MC palatalisation – this may mean the borrowing is earlier. If against Lubostky & Starostin's proposition (which is the most probable one) this was a MC loan, there would be an information crucial to the definition of OC phonology.

On many of these, there is a consensus, as shown, e.g. by Schuessler, who cites Mallory apud Mair, citing OCM **g-luʔ* to be the source of Tocharian *klu* "rice".

Adams (2013) notes yet more possible borrowings in TB : *poylā* from unknown <MC> word (ibid:434), names of months, e.g. *meñe-rapañ* and other *rāp* "month"(e.g. ibid:503,573-574), *yāywyeṃ* "convoy" (ibid:532), (probably a) personal name *Śiṅke* (ibid:689), *ṣau* "receipt" (ibid:727), *simā* "adjutant, marshal" (ibid:758)[75], *hwuṣṣi* "vice commissioner" (ibid:797), *tsuṃ* "inch" (ibid:810), *tsyāṅk* either "soy sauce" or "wild rice", *tsyāṅkune* "general" (ibid:814).

Ching (2011) wrote a convincing paper on TB *kaum* "silk".

Schwarz & Blažek (2015:26-30) discuss *lwāke* "a (ceramic) vessel".[76]

Since Shaughnessy's (1989) treatise about chariots, scholars concentrate on the semantic areas for borrowing from Tocharian to Chinese to cultural artefacts pertaining to areas where the Chinese had supposedly lacked in inventions before the advent of Indo-Europeans in Central Asia.

For some lexemes, the correct "sidedness" of borrowing has not been clearly established. For those, see corresponding entries in 3.3.

[74] And all the other texts referenced here.
[75] It is interesting that a word derived from a word supposedly borrowed from the language should end up being borrowed back. While not unheard of, it does point to a fact that this might not be what really happened.
[76] The proposition seems quite interesting, its evaluation is not part of this work, as is of other words in this section. They are taken to be correctly analysed while their analysis is not further used.

#### Discussion

Most sources cited here should be reliable, some others are taken as potentially biased due to their author having widely known controversial beliefs that are in direct (or indirect) relation to the topic, and/or are published in a medium dedicated to publishing controversial topics and solutions which are nevertheless deemed by the author so interesting that they cannot be ignored.[77]

The topic of this work is not widely researched which should lead to the ability to create an exhaustive list of relevant texts.

The most controversial figure is with little doubt S. A. Starostin, a (co-)author and/or reformulator of many hypothetical macrofamilies[78]. Among them the Altaic family, Dené-Causasian family and Borean family. While by themselves, these ideas could simply be ignored, when the reconstruction of possible intermediate stages is concerned, they are sure to influence the outcome and change some shapes and meanings to accommodate for what is considered by Starostin to be an input from bound data. His reconstructions of Old Chinese lexemes have therefore been only included on an individual basis after consideration.

Some chose to partially or completely disregard[79] Starostin's work, but the volume, spread and accessibility of his work makes it necessary to process it in some way.

Another source to be taken with some reservations is the journal Sino-Platonic papers, which encourages "unconventional and controversial" texts (the warning is present in every paper – e.g Shaughnessy 1989). While meant as controversial, it does occasionally provide an insight which could be considered in line with mainline thought while being unconventional in some other way. At places, other sources refer to Sino-Platonic papers, often taking the theories and hypotheses presented there as fully in line with the consensus.

TLS – Thesaurus Linguae Sericae, a semantic mapping of Chinese words, was considered to be used, but since the word families (discussed elsewhere) are not taken to be proven, its use was not deemed practical.

The site chineseetymologyonline.com (Sears 2013) which gets some critique is used as a supplementary source due to its easily accessible listing of probable character forms.

### 2.2.3. Tertiary

Theory sources supply theoretical input for creating a framework.

Previously mentioned Schuessler (2006) is used to bind the Baxter & Sagart system to realistic phonological systems.

Etymological dictionaries of PIE forms are considered part of the framework, since the theory behind them always influences reconstructed forms in substantial ways. Used is LIV (Rix 2001) and NIL (Wodtko & Irslinger & Schneider 2008) with some consultation from Mallory & Adams (1997) and others compatible with tri-laryngeal theory. The work has been informed by the Leiden IEED series.

#### 2.2.3.1. Universals

The Universals Archive (Plank et al. 2009) lists 2029, often duplicate, items including those, that were refuted by evidence. Those with the slight possibility of being relevant here[80] are (referred to only by

[77] Obviously, nothing I would consider a pseudo-science has been included.
[78] I would differentiate between macrofamily and phylum – on the grounds of whether the supporters view the correlations as resulting from genetic relationship or from a possible prehistoric *Sprachbund*, resp.
[79] G. Starostin (2009) mentions this observation about Schuessler's (2006) treatment of some forms.
[80] As stated earlier, true and useful absolute universals are a matter of belief; I have therefore left them out of the theoretical framework. To illustrate the point - some, that may apply, would be (simplified): UA 926-927 (C[syllabic]<*CV: not true), 1328 (historical tendency towards phonological symmetry: "tendency" renders this not absolute), 1764 is an exception, 1768 (compensatory effects of nasal neutralization affect heavy syllables

UA number): 614, 686, 725, 1005, 1087, 1097, 1243, 1252, 1760, 1762, 1763, 1786, 1850, 1854, 1855, 1856, 1860, 1861; 671, 1764, 1787, 1801, 1932, 1988. They are in fact mostly useless, since many of them have been refuted, others are very general or very specific. One that may be useful is 671 – "some lexical items are more stable than others. … tend to be the same for all languages." For more on this idea see note on Leipzig-Jakarta list in the next section.

More useful are phonological universals when taken as *statistical*, not absolute, they help evaluate the probability of presented ideas. For this, Greenberg & Ferguson & Moravcsik (1978) is far more valuable[81].

#### 2.2.3.2. Loanword databases

Loanword databases are usually made by and for typologists. In this work, the decision has been made not to use them. Justification follows.

WOLD – The World Loanword Database. Possibly very useful when in its final form, contains Tocharian A and B, Old Chinese and Middle Chinese. Closer look reveals that etymologies are not properly sourced in-place and that both Tocharian A and B have only one entry – "wheel" as a donor to Mandarin Chinese, which is time-wise impossible to be done directly. Entries for Chinese as a donor are more numerous but due to the objective and format of the database similarly unusable as a reference for etymological work.

WALS – The World Atlas of Language Structures. It is a database of languages transposed over a single model[82]. It is used here as a source of quantitative-data input on probability of certain features, it does not include data on dead languages and is therefore not used as a source of direct evidence.[83]

Leipzig-Jakarta List – not a loanword database as such, is an evolution of Swadesh list[84] based on actual scientific method. The words on a list should pose terms most resistant to borrowing. This list is used here as a supplement to the borrowability framework. No explicit reference to it is being made.

## 2.3. Basic rules of theoretical framework

Based on data from generally agreed upon sources in up-to-date versions, literature on the topic has been reviewed and put together to form general rules. Most of the framework has been discussed either in 1 and preceding subchapters so as not to repeat the same information only to evaluate it, a smaller part is discussed in the next subchapters. What follows is a set of basic rules that are strictly adhered to. These rules are to be taken as something that is required by the scientific method, where other authors diverge from it, it is commented on openly.

Lubotsky (1998:381) states his simple but necessary requirements for cognates, these are taken over with certain modification:

1. The form and semantics have to similar in both languages.
2. OC word has to be isolated in ST.
3. A good etymology has been found for Tocharian side in PIE.

---

before light: part of compensatory rules in general), 1770 (coda N weakening before onset nasals: not universal, e.g. Korean initial de-nasalisation, though still not fully part of phonotactics), 1953 (tonal languages have rich V system, stress accent impoverishes it: probably true but useless) ; 1792-4, 1797, 1890-6 (not absolute, part of borrowability theory), 1928-37 (either refuted or statistical).

[81] For some reason not one of UA sources.

[82] The work is based on universal generalisations informed by comparative concepts (ed. by M. Haspelmath). Its sources however also include traditional descriptive grammars. Comparative concept is a structure postulated for comparing concrete languages as opposed to language-specific descriptive categories. For details, see Haspelmath (2010).

[83] There are many gaps (not all structures are concepts are analysed in every language) for this reason it is not used as an authoritative source here (compare e.g. Czech and Mandarin).

[84] A list of basic concepts shared by every human culture and their corresponding *signifiants*.

4. “The word must belong to a semantic field liable to borrowing.”

The first rule has been reformulated to: any form that is postulated to be a cognate with another form must be adaptable by a set of explicit rules to conform to the phonology, phonotactics and morphology of the other language at the time of borrowing. The semantics must be fully compatible in the same timeframe at least in line with what could be expected on the basis of universals and on the knowledge of those languages’ semantics and *pragmatics*.

The second rule has been reformulated to: the recipient language must not have cognates that are attested at the same time or earlier than in the recipient language, if such exist, at least one must be explainable as an intermediary to prove the connection, postulating unattested intermediary is not allowed.

The third rule is suspended in those cases where all other rules obtain, in which case the donor and recipient role is either reversed based on secondary evidence, the donor changes status to intermediary, where evidence points to it, or the word is considered to have an unclear and unexplainable etymology given the evidence at the time.

The fourth rule is relegated to a supporting rule based on the evidence that borrowing of a word in any semantic field is possible, however improbable.

A rule has been added that requires a situation where a borrowing might have occurred. If there is no thinkable situation, even in the case of compatibility of both sound and meaning, these are to be ruled out as impossible.[85]

## 2.4. Used software

Most of the software known widely in the historical linguist community is focused on very specific topic and is tailored for it, e.g. the so-called *electronic Neogrammarian*[86] or *CARP*. No doubt many Swiss-army knives have been made but the problem is that their authors have often neglected to release them to the public or they were too cumbersome to use without extensive data preparation. To overcome this problem, a set of simple tools was programmed with the hope they may help others as they are mostly devoid of theory and are thought by the author to be extremely easy to use even with no prior knowledge of computational linguistics, only basic knowledge of historical linguistic principles and notation is required.

The author’s original idea was to create scripts that would automatically create paradigm tables for every dictionary entry of the fusional language type using conjugation and declination information provided. That, however, proved very early to be an impossible task as no real-world language has an entirely regular morphology and as stated in the introductory chapter, while Tocharian is in this regard exceptional, it is certainly not so in a way that would enhance its auto-processing possibilities.

Luckily, the CEToM project is already developed into a stage where it can safely by used, to an extent. A script has been made by the author for *OpenRefine* which in combination with *wget*[87] downloads and pre-processes the wordlist offered by the site. The resulting data is in CSV/TSV format[88], which is suitable for processing of large data[89] because of lack of unnecessary features and dependencies. While CEToM offers an API[90] to access its content, the simplicity of HTML presentation of web did not substantiate its use.

[85] Consider borrowings of family terminology, numbers, invectives, even grammatic constructions.
[86] The Proto-Algonquian reconstruction attempt.
[87] A standard *nix (Unix and Linux) tool to recover web pages. If set incorrectly, it may be seen as a malicious bot (computer program with a certain restricted functionality in certain application supplanting a human menial work) both by moral and technical standards. The script to download the pages in this case has been set so as not to overload the server to a 6 seconds’ wait between requests.
[88] CSV – comma separated values, TSV – tab separated values. Tabular data in simple text.
[89] No overhead, no special RAM/CPU/OS requirements for processing software – tools for simple text will suffice.
[90] Application programming interface – a way to connect to an application’s functions from other applications.

### 2.4.1. Computer assisted approach

Previous studies have been done by both linguists and programmers on the viability of automatic search for cognates in the world's languages. Most use either heuristics based on probability scale, Hall & Klein (2010) to cite a recent example, while others have worked with language-informed models. As Hall & Klein (ibid:1030) point out, the second kind actually works on already proposed cognates. As they note, methods for automatic cognate detection have been proposed, most of them working on language pairs (2010:1030-1031), their method works on the whole family.

None of the frameworks available seem to be specifically oriented towards finding cognates in non-related languages, where identification of a pre-form should come from different sets of rules (for one family, the transformational rules are two sided, when borrowing and adapting, the rules change the form in such way that little no information on the original is necessarily preserved). The volume of data for languages where only extremely limited amount of contact is postulated is different, the rules may be ad-hoc, no meaningful statistical information may possibly be extracted to teach HMM, etc. For reconstructed languages from different families where the contact seems to have been limited, a semi-automated approach seems to be advisable.

### 2.4.2. The scripts

For the purpose of this work, a previously-existing set of scripts[91] made by the author has been extended. The set is made in combination of standard Linux CLI[92] tools.[93]

The original set was a simple "historical linguist's calculator", that is, the input was a text, a set of chronologically ordered sound laws in a generativist notation and the output was the same text after going through the changes with accompanying list of the output phonemes and their evolution. Substitute characters[94] were allowed to a certain point and the change could be conditioned. No module for phonetic feature was created[95].

These scripts were used as a single module in a set of scripts made for this work specifically.

While the used set of scripts may seem like a mass lexical comparison, their purpose is not to prove genetic relationship, only to hint at possible shared lexicon. No heuristics[96] are to be used for drawing definite conclusions. The script is used for the purpose of finding all possible – though not probable – cognates, or more precisely, words with similar form at certain point in time, to be checked against a literature by a human operator. This should rule out human error from repetitive tasks a simple "calculator" can do and also the need to specially prepare data for compatibility with the semantic module of the software used for comparison which may oversimplify to the extent the database itself becomes unusable to a non-specialist who cannot repair the incongruences.

First script explicates Baxter & Sagart (2014b), that is, disambiguates all possible reconstructions hinted in the notation. The second "transforms" one language into another. Third compares differences between forms reconstructed in dictionary and those resulting from transformation. Next, human operator has to

[91] Made for, and available in original form at, the Department of comparative linguistics, Faculty of Arts, Charles University.
[92] Command line interface, as opposed to GUI, Graphical user interface.
[93] For more "tech-savvy": interpreter is set to Bash, only AWK and sed are used, which means the scripts should be compatible with Cygwin and/or any other implementation of *nix compatibility layer for MS Windows.
[94] The most common ones: *C* for generic consonant, *V* for generic vowel, # for word boundary, _ for focus (in conditioning). Further substitution characters could be added to the list by editing a file but were deemed unnecessary at that point.
[95] The calculator was to be superseded by a more advanced variant with phonetic features, teacher's/student's modules and an easy-to-use GUI made by a colleague in Java. I am unaware of the outcome of the effort as our mutual cooperation ended at early stages.
[96] The word is being used in the sense attributed to it in computer engineering, that is, algorithm that approximates results based on previous results, usually without a help of underlying theory.

manually check all data regarding the semantic correspondences. Finally, all possible candidates exhibiting phonetic and semantic similarities are gathered and compared with previous research.

# 3. Data analysis

## 3.1. Script input

Based on adaptation rules abstracted from literature, a list of word-forms has been inputted into the series of scripts.

What follows is a set of correspondences between words that were deemed to be undoubtedly borrowed or at least compatible in all respects. The sources cited are 1. those that proposed the connection, 2. those that include an up-to-date

Based on[97] Lubotsky (1998), who proposed or at least mentioned most of the possible cognates in his relatively small yet dense article, no rules have been abstracted – see respective entries in 3.3.

Based on Ching (2011): Synchronic adaptation rules (personal names, place names…)

(p. 66) TB *Kumpantiśke / Kumpāntiśke* (PN) 白俱滿失雞, Early MC (bai) *kyə-man'-ɕit-kɛj*

TB *Kumpānte/Kumpanti* (<- SKR *kumbhaṇḍī*) (PN) -> 俱滿提 Early MC *kyə-man'-dɛj*

*Kumpantile (PN) 俱潘地黎/白俱滿地黎 Early MC (bai) *kuə-pʰan-diʰ-lɛj/kuə-man'-diʰ-lɛj*[98]

Other authors mentioned these words:

TB kuśiññe "kuchean" from *kućі(ye) etymology uncertain[99] -> Quizi (龜茲) (Adams 2013:198). Baxter & Sagart (2014b) reconstruct OC *[k]ʷə, MC **kjuwA for the first character. The second is reconstructed by Schessler (2009:102) as LH tsɨ<tsiə, explicitly referring to Kuch: Middle Han *ku-tsə.

Adams (2013:354) mentions TB *Nāri** (PN), in Chinese sources as *Nali*. The stage probably meant is MC, OC would have no reason to adapt *r as *l, MC reflex of an older borrowing would be also different from *l. No characters are given and none are derivable by common knowledge, no tones written. Only taken as a support. He also mentions (ibid:544) TB Yurpāṣka taken over from Lévi, adapted as MC **jia-ba-shi-kɛj* (PN), no characters are given, but the last two are supposed to reflect *-śke*. This would seem like a later adaptation considering different resyllabification strategy, the final consonant is no longer available. More information on MC tones in this word would be needed to properly judge it, the transcription is too broad.

Pulleyblank (1966) list place names and personal names. Among them 焉耆 (p. 20) which he derives from TB *ārkwi* TA *ārki*. While many other of his propositions are still not certain, this one is without a doubt, with TA being the original word[100]. Interesting is that the form seems to fit OC better than MC, also of note is a possible indication of intersonoric[101] voicing. The voicing has been postulated for late TB on the basis of Manichean bilingual already by Peyrot (2008:88-90) as part of lenition which also fricativises the consonant (in his case for /k/ and /p/). Chinese does not give any indication of earlier fricativisation.

The terms *Yuezhi* (connected with *ywati*) and *Dayuan*, *tuxuolo*, etc. (connected with *taxwar*) are not considered to be without a doubt of Tocharian origin and therefore not used as part of input.

---

[97] Or rather, being informed by.

[98] The aspiration seems to reflect the SKR form mentioned earlier, without the retroflex features. This form is taken with a some reservation.

[99] Although, as he notes, it might be connected to "shining, white" which would explain the dynastic name in Chinese (白 seen in examples from Ching 2011).

[100] This also agrees with where the speakers have been localized.

[101] I use this term for intervocalic voicing to emphasize that never does this happen solely in between vowels, it happens also when one of the phones in contact is a sonorant.

Pulleyblank (1966:20-21) mentions 祁羅(漫), 析羅(漫), as does Schwarz (2006), according to Adams (2013:250) these are TB klyomont, TA klyomänt. The first character in both words does not seem to be reconcilable with our knowledge of MC: gijA and sekD (Baxter & Sagart 2016). The first one then may go back to (late) OC. None are useful for comparison which is attempted here. Some more words have been proposed, being similarly considered unproblematic by some while non-reconcilable by others (Arśe[102], …). Those worth mentioning due to their importance, or interesting character, or due to them being discussed often with some outcome, are part of Wordlist, some would need a separate publication, these are not listed. The question of whether to use words that supposedly fit perfectly in their forms as a base of adaptation rules is of importance. For reason of minimal controversy, I have left these out.

These rules for adapting TB to Early MC have been abstracted (for MC used Baxter-Sagart notation):

(p -> ph), k -> k, ti -> tshi, r/l -> l, t -> d[103], m -> m, n -> n (final n may necessitate X, probably not), ć -> tsy, final -r -> -n, final -ś -> -t; ā>a, i>ej/ijH, u>u/ju (palatalising on non-palatalising MC context).

Clusters are simplified or unpacked. Open syllables are taken to be live syllables (i.e. not ending in ʔ reflex). While other explanations have been proposed, it seems that homorganic plosive may blend with homorganic nasal. The first character in Kucha should probably reflect an intermediate stage between Baxter-Sagart OC and MC where labialisation was already unpacked and the palatalization of non-pharyngealised consonants did not yet happen, either way, the reflex is indeed juw.

All the adaptations would seem to have happened in the Later Han, i.e. early time of TA TB clear diversification. If so, then earlier extensive cultural contact would not be reasonable to consider, unless Tocharians were migrating through Chinese territories to the Tarim Basin.

Using place names and personal names might be somewhat problematic as the adaptation could be ad-hoc, but it still should reflect the native speaker's perception of the other language. Another possible controversial decision is not using the other way for deciding the mutual sound correspondences. The reason is that borrowing when it occurs is one way process that involves the processes of repair in the recipient language, not the donor language and therefore adaptation from one language to another does not necessarily create the same allophones and/or phonemes as does the other direction.[104]

From general phonetic principles and specific Sino-Tibetan evolution, one would need to account for adapting of PT forms, as those mentioned above should roughly correspond to early MC, there is no need to simulate an evolution, only general principles are needed to be applied.

On the basis of universals and other parts of framework, a probable expansion of the adaptation scheme: T -> T, T -> M /[sonorant+]_[sonorant+], T -> T[palatal+] /_i, T -> t /_#, vowels preserved – broken where would be expected from an OC reflex.[105] Some context adaptation which result from

[102] TB ārśe as cited by Adams (2013:57) mean either "Agnean" in connection to TA ārśi, "Aryan" or "monk". He believes there to be a problem since he does not allow for homonymy. My problem is in the form: Yanqi (焉耆). Baxter & Sagart (2014b) reconstruct OC *ʔa[n], MC **ˈjen + OC *[g]rij, MC gij. This would look more like the SKR name, than the Tocharian. Pulleyblank's (1966) etymology seems better. Interesting is a possibility that the OC final of the first character was -n with rhotic element being reflected by the second character. This would be only a speculation, though.

[103] The context is specific, this may in fact indicate that in Tocharian, there was indeed an intersonoric voicing.

[104] Consider e.g. Austrian German *Powidl*, a Damson plum marmalade, from Czech *povidla* (I don't consider Polish *powidła/o* a probable alternative as other parts of Austrian cuisine also derive from strong Czech presence in Vienna, e.g. *Buchteln*, a kind of sweet pastry often filled with *Powidl*). Correlation /p/->/p/, yet [p]->[pʰ]. Note also the morphological adaptation of Cz. *-dla* NACT.FEM.PL suffix to native -l DIM with rebracketing. In contrast, Cz. *plech* "sheet metal" < German *Blech* /b/->/p/, [b̥]->[p]. The complete set of correspondences for labials in loanwords would /p-p, p-b, b-b/. Obviously, for a complete set, the correlation chart would be enormous, even for a completely synchronous analysis where we do know the realizations including phonetic detail.

[105] Either because borrowings were in the intermediate stage, or because they would for some mystic reason fall into division III/IV. Why would ku be analysed as kyu is out of my reach, possibly some kind of regressive assimilation of place in the case of the word in question.

suprasegmental features of the whole word either in MC or in Tocharian seem to play a role. These would need a separate study with better identified cognates.

## 3.2. Script output

Since the general rules of evolution of pre-forms in ST has not been fully established and disagreement is even on many forms, the author opts for a hopefully non-problematic direct contact between potentially synchronically co-existing phases of respective languages.

### 3.2.1. TB transposed onto Chinese

After comparing the script series' output with Baxter & Sagart (2014b), these forms seem to be compatible with some TB words:

pang

謗,pangH,*pʶaŋ-s ,slander

舫,pangH,*pʶaŋ-s ,boat

tang

當,tang,*tʶaŋ ,"match (v.); have the value of, rank with"

黨,tangX,*tʶaŋʔ ,500 families; relatives

tong

登,tong,*k-tʶəŋ ,a kind of sacrificial vessel

鐙,tong,*k-tʶəŋ ,ritual vessel; lamp

燈,tong,*k-tʶəŋ ,lamp

登,tong,*tʶəŋ ,step up

登,tong,*tʶəŋ ,ascend

twan

耑,twan,*tʶor ,tip (n.)

端,twan,*tʶor ,tip (n.)

短,twanX,*tʶorʔ ,short

斷,twanX,*tʶo[n]ʔ ,cut in two

段,twanH,*tʶo[n]-s ,hammer

斷,twanH,*tʶo[n]ʔ-s ,cut off; decide

twat

掇,twat,*tʶot ,"pick, gather"

pjop

法,pjop,*[p.k]ap ,"model, law"

灋,pjop,*[p.k]ap ,"model, law"

kang

鋼,kang,*C.kʶaŋ ,cast iron; steel

亢,kang,*k-ŋʶaŋ ,lift high

剛,kang,*kʶaŋ ,strong; hard

綱,kang,*kʶaŋ ,guiding rope of net

mang

芒,mang,*mˁaŋ ,"awn, beard of grain"

忙,mang,*mˁaŋ ,flurried

nang

囊,nang,*nˁaŋ ,"sack, bag"

曩,nangX,*nˁaŋʔ ,in past times

nyem

髯,nyem,*nam ,whiskers

染,nyemX,*C.n[a]mʔ ,to dye

nyit

日,nyit,*C.nik ,sun; day

衵,nyit,*nik ,a lady's clothes nearest to the body

wang

汪,'wang,*qʷˁaŋ ,vast; pool

尪,'wang,*qʷˁaŋ ,emaciated

jang

央,'jang,*ʔaŋ ,end (v.)

央,'jang,*ʔaŋ ,center (n.)

殃,'jang,*ʔaŋ ,calamity

鴦,'jang,*ʔaŋ ,female mandarin duck

After comparing this list with CEToM's output, no form has been deemed suitable for inclusion into Wordlist.

## 3.3. Wordlist (monosyllabics)

This chapter lists all the lexical entries relevant to the research, both those that discussed elsewhere before and used for abstracting rules for the method and those that emerged from its application.

Structure is such:

`Entry_number Han_character` *`pinyin`* `"oldest_Chinese_meaning"` *`Chinese_reconstruct`* `sidedness`[106] *`Tocharian _candidate`* `rough_timeframe ■ first_proposed ■ explanation_and_comments (● separates arguments) ♦ oft-confused_word`[107].

Entry number is followed by a mark designating an outcome of the analysis. Entry marked with contradiction sign (⊥) is not considered a probable or even possible loan by the author but is mentioned here for the reason of a discussion having occurred in earlier works. Unsure connections where cognate status (possibly by other means than direct borrowing) is suspected are marked with lozenge (◊) and Tocharian etymologies considered proven are marked with white square (□).[108] That is, in order of trustworthiness: □ > ◊ > ⊥.

[106] ← Tocharian presupposed as source ↔ borrowing possible both ways, ↜ Tocharian source through other language.

[107] E.g. allograph, homograph, near-homonym cited in literature without proper reference to character, etc.

[108] A combination of different algebraic notations has been used knowingly, to anyone considering that an insult, I hereby apologize.

The Chinese and Tocharian candidates given before an explanation are based on the original proposition amended by inputs from the discussion and emended to show the reason for postulating the connection; quite often, no source cites them as such.

Put apart is the work(s) that proposed it first or others refer to ultimately, sometimes through some intermediary.

The list is alphabetical in Latin script order of Putonghua readings. While this is not ideal, as the modern word does not necessarily reflect (directly) the word borrowed and/or reconstructed, it is a compromise used also by Baxter & Sagart (2014b) which was deemed best being approachable to everyone without knowing the languages involved. From Baxter & Sagart (2014a) has also been taken over their amalgamated criticised "not-MC" without further explanation for transcribing MC where the precise information on sound value at that period is not necessary.[109]

## A

### 3.3.1. B

01 ⊥ 壁 *bì* "wall" **pek* ← **pək*° OC-PT? ■ Lubotsky (1998) ■ Lubotsky (1998:387) sees "TB *pkante* TA *pkant* 'hindering, obstacle'" coming from PT **pəkənte* < PIE "*bʰeg- 'to break'" as clear cognates with 壁 where the positive identification stems from Tocharian loss of phonological voicing (ibid:388).

● There is really very little reason to believe the hypothesis of borrowing to be true, the Chinese word seems like a primary noun[110] – and changing word class from a major one to another in the process of borrowing is uncommon[111]. If we thought that the word was borrowed as a verb, there would have to be either an OC verbal reading, or at least we should be able to find a cognate that would also fit the reconstruction, which we don't.[112] The pragmatic side of the proposal seems rather peculiar, there is probably no meaningful situation where the meaning postulated for entering OC could be in L2 incorrectly inferred as such (and if inferred correctly, it would not have this meaning). The only realistic situation would be where an attacker would refer to a wall of his enemy as an obstacle, in which case, it is highly probable the defender wouldn't really need to borrow such a word.[113] ● Baxter & Sagart (2014b) reconstruct MC **pek < OC *C.pˤek "house wall"[114]. Schuessler does not mention this word but in (2007) but in (2009:133) he has OCM *pêk LH pek. Zhengzhang (2003) reconstructs *peeg.[115] PST form is not reconstructed by Starostin (2006). ● Adams (2013:438-439) discusses TB *pkante*, reconstructs PT **p(ä)känte* mentioning that it "as if" reflects PIE **bʰ(e)gn̥to-* from **bʰeg-* 'to break'. Morphologically, he supposes it to be similar to TB epiṅkte 'within'; although (ibid:94-95) he states this word has an unknown etymology, with Winter's (1941,1976) derivation from PIE **bʰeg-* being unconvincing. ● Wodtko (2008:6) reconstructs PIE **bʰ(e)gn̥to-* only referring to the Tocharian words in reference to Adams (1999:407), which is identical to the aforementioned entry. ● From what has been stated above, the Tocharian word does not seem to possess a satisfactory reconstruction semantically

[109] What they did was that they introduced archigraphemes in the likes of ones proposed for unified Romani writing. Somewhat hindering the immediate understanding without a study may be that nearly all sounds are represented by di- or trigraphs for easier processing on American keyboard. For complete explanation see Baxter & Sagart (2014a:12-20).

[110] At least its character's usage suggests that, overt marking is not required for derivation, therefore conversion from a different word is a possibility, cp. 聖.

[111] I am unaware of a case where a verb was borrowed as a noun.

[112] Which is quite rare, actually.

[113] To be just, the TB *poṣiya* TA *poṣi* "wall" has a peculiar etymology, if we believe it, indeed: A2013:435-6 states it goes back to PIE "*pusiyehₐ 'that, which divides'".

[114] If we took word families seriously, 邊 *biān* "side" would work quite well: Baxter & Sagart (2014b) OC *pˤe[n] where *C- could be a prefix and both *-*k* and *-*n* would be suffixes allowing us to postulate an earlier verb with form not similar enough to the Tocharian side of equation.

[115] Voiced finals are now considered outdated, they were part of now outdated hypotheses of tonogenesis.

and doubtful reconstruction morphologically[116]. No Chinese word has been found to connect with the form-meaning combination.[117]

### 3.3.2. C

02 ◊ 車 *chē* "chariot" ***(C)K**[118]***(r/j)a(C)** <- *kVkVl PC-PT/postPIE ■ Suggested Pulleyblank (1962), implied by Schuessler (2007). ■ Already Pulleyblank (1966:30) cites the connection with Tocharian as a possibility. ● OCM ***k-hla*** reconstructed by Schuessler (2007:182), is postulated as following Bauer (1994) to go back to an Indo-European word for wheel, citing only TB *kokale* TA *kukäl* and Greek "*kýkla* or *kýkloi*" for comparison. ● Bauer (1994) expands on Mair (1990:45) who believes that PT form is not compatible with the form reconstructed for OC. ● Lubotsky (1998:384-385) supposes a meaning more akin to "cart" and rejects the connection with the Tocharian, deriving 車 *chē*/*jū* from inherited word(s) for "to dwell" – *jū* 居 and *chǔ* 處 "This fact seems to indicate that Chin. *jū* and *chē* orginally referred to a cart where the nomads put all their belongings and where they lived." (ibid:385). However, this is completely implausible due to the physical nature of old Chinese chariots[119], furthermore no parallel of the semantic side exists[120] in languages of the world. ● Schuessler (2007:62) comments on a rare OCM ***k-hl-*** mentioning that all words with this initial but 車 are inherited, that the written records show that the *k- was still there in the beginning of their writing that the MC reflex was not from OC voiceless *lh-.[121] ♦ Bauer (1994:4) lists OC words for "wheel" and "chariot": 車, 輪, 轂, 輅, 輾 with various, now outdated reconstructions. While he does connect these words to a common IE source, it does not seem to be necessary. More details on Bauer and Mair's ideas: 軥轆.

03 ⊥ 乘 *chéng* "to ride" ***kə.ləŋ*** ← ***klānk*** OC-PT ■ Lubotsky (1998) ■ ♦ For discussion see 乘 *sheng*.

04 ⊥ 城 *chéng* "city wall" ***deng*** ← ****tɑnk°** OC-PT? ■ Lubotsky (1998) as a possibility ■ Lubotsky (1998:387) OC *djeng ← TA TB "tank- 'to hinder, impede' < PIE *teng$^{h}$- Although the semantic side of the equation is quite attractive. … and words for 'city wall' are frequently borrowed. … is not without problems. … it can reflect OC *gjeng/*geng.", PT source is postulated as unattested deverbal noun (ibid:386) ● The optimistic view that the semantics could be connected is not to be agreed with, also the OC reconstructions point to a voiced initial, which is in direct conflict with both Tocharian and PIE initial. ● The updated reconstruction in Baxter & Sagart (2014b) "城 chéng dzyeng (dzy- + -jeng A) *[d]eŋ city wall" keeps the problematic underdefined OC initial. If we believe Schuessler (2007:7), the difference in vowel could be of dialectal origin in case of a borrowing. ● Adams (2013:306) states that TA TB "täṅk- reflect PTch. *täṅk-, probably from PIE *teng$^{h}$- 'pull back'.", this clarification of Tocharian forms seems to further mitigate the possibility of connection on phonetic basis. ● Rix (2001:657) has a very peculiar PIE form "*t$^{h}$eng$^{h}$- 'ziehen'" (pull), "Durch Verlust von *s* mobile aus

[116] The -n stem in the PIE extension of 'break' is unexplained, probably being postulated to derive the meaning already there. A nasal presens marker would be placed *inside* the root.
[117] Could this be an Indo-Aryan loan with a shift in meaning? Metathesis does happen as part of borrowing sometimes.
[118] Velar or uvular plosive.
[119] First: earliest depictions of this character show spoked wheels, meaning it was an advanced technology which would probably not be used by primitive nomads, second, their size would not allow for such use, third, this character is attested in contexts of battle, e.g. *XianWen* 憲問 "子曰：「桓公九合諸侯，不以兵車，管仲之力也。如其仁！如其仁！」" (Sturgeon 2011).
[120] To my knowledge.
[121] The writing style makes it difficult to understand whether the only difference in OC was the *k-. Earlier in the work, Schuessler (2007:7) discusses the voiceless sonorants, stating that what is transcribed as *hl- and *lh has different reflexes in MC due to *hl being coming from 'Rural', or non-literary dialect of OC. Rural and *foreign* (in his meaning substrate) words seem to be understood by Schuessler to behave similarly in most respects.

*$st^heng^h$- < *$sd^heng^h$-? … gegen *th₂ spricht z.B. *e in germ. *$þinhslō$- “[122]. ● Rix’s (2001) reconstruction could possibly allow phonetically for a borrowing from an improbable pre-PT/postPIE form with voiced initial without the presupposed s-mobile[123] with a more compatible meaning. If we were to suppose this, we would probably still need a way to cope with Grassmann’s Law,[124] which would cause prePIE **$d^heng^h$- > PIE *$deng^h$- > PT †*tseṅk*- rendering the form yet again incompatible.[125] ● Lubotsky (1998:387) himself mentions an alternative analysis to OC dental by Bodman (1980:160) with OC initial *gj/g-, connecting the OC word with Tib. *ḥjengs* ‘to fill, fulfill’. For reasons stated in other parts of this work, this is not very probable, also the semantic side does not fit well, Lubotsky’s refusal should then be correct; other aspects of the analysis only seem to fit when explicitly trying to account for a postulated connection. ♦ Same semantic field: 塁.

### 3.3.3. D

05 ⊥ 兑 duì “passage, opening” *$lot$ <- *$lot$- OC-PT? ■ none, discussed by Lubotsky (1998) ■ Lubotsky (1998:381-382) finds connection with “Toch. AB *lut*- 'to remove, drive away', B *lyauto* 'opening', A *lot* 'hole', cf. also A *lyautaṃ* 'ravine, chasm', B *laute* 'moment, period'” to be “tempting” but refuses to speculate based on too little evidence to suggest that there is a possibility of a borrowing in this semantic field. ● The phonetic correspondences seem to hold if we allow for some leniency of the kind Schuessler (2007) shows for substrate; the semantic side also seems to be possible for some. The motivation for borrowing these words seems lacking. Of the proposed allofams, only 奪 is reconstructed by Starostin (2006) with phonetically unconvincing Kachin *khrut* as a cognate. Therefore, the ST etymology of OC words is not considered to be proven. The PIE etymologies for Tocharian listed by Adams seem unconvincing, however: TB *laute*, TA *lot* (2013:612) were previously connected with *leudH ‘that, which is cut off’ by Winter and *loud°[126] connected with Old Norse *laut* ‘depression in the ground’ and *leyti* ‘moment’ by Hilmarsson. TB *lyauto* ‘hole, opening’ is considered to be unrelated. ♦ As connected to 兑, Lubotsky (ibid:381-2) sees *duó* 奪 ’take away, deprive’, *yuè* 閱 ’opening, hole’, possibly also *tuō* 脫 ‘take off, let loose’[127] and tuì 脫 ‘easy, leisurely’[128]. These would not even constitute allofams, being far too removed semantically.

06 ⊥ 奪 duó ■ none? Mentioned by Lubotsky (1998) ■ For discussion see 兑.

[122] “By the loss of s-mobile from *$st^heng^h$- < *$sd^heng^h$-?” (*$st/d^heng^h$- do not have a separate entry). “Against th₂ <as source of $t^h$> speaks e.g. *e in Proto-Germanic *$þinhslō$.” (Please note that primary T+MA root is disallowed and secondary ones are mostly controversial; to me it seems that an artefact of method has been levelled up to be a rule, personally, I find no problem with the root used by Lubotsky).

[123] PIE word-initial *s which, based on some word-forms, should be reconstructed, while based on others it should not. Rix (2001) mentions s-mobile preform only in the footnote quoted above.

[124] Deaspiration of the first of two mediæ aspiratæ into plain voiced plosive in sequence. Whether it occurred only in Greek and Indic is disputed. It has been postulated for some etymologies by Ringe (1996), more in 1.3.4.

[125] With this word, anyway – but maybe more compatible with other words with similar meaning, see 塁 (not to suggest that there is actually a connection that can be proven at this point).

[126] My abbreviation.

[127] Probably meant 挩? For this entry updated Baxter & Sagart (2014b) agrees in meaning, reading and reconstructions. If the character meant to be taken literally, either we have to change OC form to *$mə\text{-}l^{ʂ}ot$, MC to *dwat*, and/or change meaning. Error in Baxter & Sagart (2014b) is improbable due to consistent agreement with Baxter & Sagart (2014a).

[128] Baxter & Sagart (2014b) do not create a reconstruct for this reading/meaning combination. Neither does Schuesser, although her does seem to refer to it (2007:504).

### 3.3.4. E

07 ⊥ 厄 *è* “part of a yoke”[129] **ʔrek* <- **h₃ reĝ-* OC-prePT/postPIE ■ Lubotsky (1998) ■ Lubotsky (1998:384) may have made some invalid conjectures: first, he considers this to be a ring of unknown purpose through which, he speculates, reigns went to “horse bits”. First, traditional Chinese yoke meant here was used not only for horses, but also for oxen and was quite primitive, comparable to bovine-oriented yokes from other parts of world, being only a piece of wood through which two rings that held the heads in place went and an optional central ring to pull them, meaning it is not the purpose of the ring that is lacking explanation.[130] Second, he compares a reconstructed initial glottal stop of OC with the reconstructed third laryngeal of PIE. This is extremely risky as we have no idea of how it was actually pronounced, with opinions ranging from $x^w$ or $\chi(^w)$ to a voiced segment (the other solution is postulating $h_a$).[131] Lubotsky sees the PIE root “**h₃ reg-* ‘to make straight, steer’” in TA TB räk- ‘to stretch’ with possibly preserved meaning of ruling or steering in personal name *Kleṅkarako*. ● Ching & Ogihara (2013:112) apud Malzahn (2017) see the aforementioned personal name as containing TB *akau* which is either a personal name or a title.[132] ● Updated reconstructions: Baxter & Sagart (2014b) as MC **'eak OC *qˤ<r>[i]k where Baxter & Sagart (2014a:58) “Infix *<r3> in nouns marks distributed structure (double or multiple objects)” analysis shows there is no connection to the PIE word; the word is seen as deverbal distributive noun from “縊 *q[i]k-s (dial. > *qek-s) > 'jieH> yì ‘strangle’”. Other words with this infix are mentioned further substantiating the claim. ● Reversing the donor and recipient languages will not give any more sensible information than the one originally proposed. ♦ No doubt connected are characters 戹, 軶, not discussed here.[133]

### 3.3.5. F

08 ⊥ 烽 *fēng* “beacon fire” **pʰuar* ← †*pu(h₂ )ār* PC/OC-prePT ■ Adams (2013)? ■ Adams (2013:421) TB pūwar “(a) ’fire’; (b) ‘digestion’; (c) ‘beacon fire’” < PT **pŭwār*, suggests the PT word may be an ancestor to *fēng* he reconstructs in OC as *pʰuaN using Pulleyblanks MC reconstruction **pʰuawŋ*. ● Baxter & Sagart don’t mention 烽 (2014a; 2014b), neither does Schuessler (2007), Zhengzhang (2003:319) reconstructs OC **phoŋ* which would rule out possible borrowing – already problematic would be accommodating for the initial OC *pʰ where either: one would need to say the initial PT /p/ was clearly aspirated or that the PIE laryngeal was preserved and caused secondary aspiration when adapting, even more hypothetically, in line with Schuessler (2007; 2009), the *p-h sequence would have to be borrowed into PC with metathesis of *h with *u. A second problem would be the final velar nasal. Velarity could be arrived at by adding DIM -śke and nasality by coming from LOC:SG, cp. TB *pwarne*. No diminutive form of this word is attested, however and it seems rather strange that neighbours and occasional enemies would speak to the other of their beacon fires using diminutives, or in any way to begin with. ● Also note that many languages and language families have similar sounding words for fire, e.g. ModK 불 [pul], ModJ ひ *hi*, [134] Khmer ភ្លើង *pləəŋ* and Thai เพลิง *plɤɤŋ*, also Thai *fai* ไฟ and Lao *fai* ໄຟ, Ket бо’к *boʔk*[135], Miskito *pata*, etc. PIE has more terms related somehow to fire, except for *péh₂ wer/n- (Wodtko & Irslinger & Schneider 2008:540), cf. also the dissimilar *hı ̯ngʷnis (Mallory & Adams 1997:202), *pel- (Rix 2001:469), bʰeh₂ - (Wodtko & Irslinger & Schneider 2008:7)[136] and others. This may very well be one of the concepts having prevalently iconic representations. The

[129] An interesting evolution “yoke ring” ► ModM “predicament”, similar semantic expansion/shift is present in other languages (e.g. Czech *jho*).
[130] Also – in connection to his hypothesis that shéng
[131] And to pretty much anything except for, probably, only and exactly [ʔ].
[132] Personally, I find Lubotsky’s explanation quite believable.
[133] Baxter & Sagart (2014b) reconstructs same meaning, same form.
[134] Altaic cognates are not considered to be valid by the author.
[135] Ket lacks phonemic /p/ (and /f/), therefore I consider the /b/ similar enough.
[136] My personal idea of PIE phonological system supposes secondarily aspirated devoiced consonant in postPIE (possibly *pʰa<**bha or more extreme **bạɻ) with MA series an artifact of method.

semantic field does not support an idea of borrowing, the phonetics could only be reconciled by relative chronology unsupported by material evidence.

09 ◊ n 輻 *fú* “spokes of wheel” postPST-postPIE/prePT ■ Lubotsky (1998)? ■ Lubotsky (1998:383) believes that OC “*pjɨk/pək” is connected to TB “*pwenta* (pl.) < PToch. *pəw- < *puH- ‘spokes of wheel’” which he connects to SKR *paví* also mentioning Bodman’s theory that later phases of OC merged *-ʔ and *-k which would allow for the PIE laryngeal to show up this way. ● The idea of final *-ʔ/k alternation relates to the concept of word families, under which Baxter & Sagart (2014a:61) discuss them, mentioning that in OC they are already not productive.[137] ● Baxter & Sagart (2014b) reconstruction for this character is OC *pək. ● Adams (2013:422) states for TB *puwe* “presumably reflects a putative PIE *pewes- (nt.) whose only suggested relative is the isolated SKR paví- (m.) ‘wheel-band, metallic point of spear’” while being sceptical. ● Neither Rix (2001) nor Wodtko & Irslinger & Schneider (2008) mention the PIE stem. ● Monier-Williams (1899) explains pavī पवि aside from afore mentioned meanings also as “an arrow, a thunderbolt, speech, iron band on सोम-stone”[138] with possible original meaning “brightness, sheen”. ● The meaning of “light” is at the base of many words for explanation, the imagery of other words as stemming from it is even more evident, yet, there does not seem to be an evident way to connect this word to other IE words for light[139]; however, purifying properties of light are cross-culturally common, and similar stem Rix (2001:480) reconstructs as “1.*peu̯H- ‘reinigen, läutern’” from which is cited as descending a number of words with similar meaning in Sanskrit, e.g. *pávate*[140]. The idea of a sun-wheel’s purifying rays being at the basis of other meanings would be imaginable.[141] Adams (2013:421-422) mentions that the words for “fire” TB *pūwar* TA *por* are hard to reconcile with common etymology from PIE “fire” *$peh_2$ *wr*, it may be that they descended from similar (and possibly related) **pewH* + adjective **r* by elision, making it “purifying (fire)”[142]. Still, any etymology of the IE words seems to be highly speculative and proposing a descendance from a loosely connected stem reconstructed from a few words in a single language is not to be taken as more than a suggestion. The IE etymology is therefore considered *not* proven. If the “Tocharian route” proved to be possible, it could arguably mean that the OC word either directly or indirectly descended from it. ● A search in literature does not reveal any connection of the OC word to other ST words that has been postulated to this date. There is therefore no reason to rule out the borrowing, even though the general idea of borrowing words for chariotry does not hold due to other words’ borrowing being disproven.

### 3.3.6. G

10 ⊥ 狗 gǒu ♦ See 犬.

11 ⊥ 牯 *gǔ* “ox” **K(V)u* ← †*ko*? ■ Gamkrelidze & Ivanov (1984)? ■ Gamkrelidze & Ivanov (1984:935) mention what seems to be this word in their list of Tocharian borrowings. This hypothesis is ● Baxter & Sagart (2014b) reconstruct OC *Cə.kʷˤaʔ for *gǔ* 牯 “male (bovine) and for *gǔ* “ram” 羖. Both are

[137] Here must be noted that Schuessler believes Baxter and Sagart don’t distinguish OC a PC well, meaning this may well change the timeframe such that it would not allow for a contact, both families’ speakers being geographically unarguably distant at that period. Of the oldest Neolithic cultures, only *Peiligang* (makers of the famous *Jiahu* symbols) seems to be closer to western territories, however, we have no information on their language and to my knowledge there was no chariot found there.

[138] My guess is this refers to the ceremonial use of the *soma* (Avestan *haoma*) plant for intoxication.

[139] The PIE stem *$b^heh_2$ - does not seem to be possible to connect to the indian word and there is no reason to connect it to the Tocharian word.

[140] Or rather, this word is solely reconstructed based on Vedic Sanskrit.

[141] E.g. *soma* stone should have a part of in a purification process. For wheel-rim and also wheel-spokes, the meaning of “light” would be ideal, explaining why both IE languages have different parts of wheel connected (image of corona: wheel vs center: rays).

[142] It would seem that “fire” may be a descendant of “to purify” already in PIE with the supposed laryngeal metathesis being reversed in order.

males of large domestic grass-eating animals with similar “uses”, this would lead to obvious conclusion that both could originally be one word. ● While 牯 seems to be unattested in Pre-Qin and Han texts, 羖 is attested in Western Han, e.g. *Shuoyuan* 說苑 (Sturgeon 2011), *Shuowen* defines it as “夏羊牡曰羖”[143] (Sturgeon 2011) which does not seem to be completely in line with its actual use but does represent a certain contemporary view. The fact that one character is attested earlier would not be a problem if we found not only combinations 羖羊 but also 羖牛 which we don’t. This probably makes the hypothesis of a connection nil. ♦ See also 牛.

12 ⊥ 轂 *gǔ* “nave of wheel” **k$^{ʕ}$ok* ← ***kok*° OC-CT? ■ Blažek (1997) ■ Blažek (1997:235) used Karlgren OC reconstruction **kuk* to partially restore TA *ku////* “nave, hub” to *kuk*° and proposes two possible sources for TA word: either “a derivative or a compound of A *kukäl*, acc.pl. *kuklas*, B *kokale* ‘wagon, chariot’” or “a metaphorical use of A *kukäṃ*, B *kukene* (du.) usually translated as ‘heels’”. The OC word would then be a direct borrowing from Tocharian. ● Lubotsky (1998:383) states that PT "wheel" ► "chariot" in TB *kokale*, TA *kukäl*; where OC *-o- “clearly points to Tocharian provenance”, sic. ● Adams (2013:214) states ‘cart, wagon, chariot’ TA kukäl and TB “kokale reflect PTch *käuk(ä)le”; Adams (2013:191) states that TB *kuke** “‘heel’ (?)” from PT ***kukäne*/***kukene* without delving further into etymology. ● Baxter & Sagart (2014b) reconstruct “nave of a wheel” OC **[k]$^{ʕ}$ok* > MC **kuwk ● CText does not show bone script or bamboo for 轂, however, it is already attested in Warring States period (*Dao De Jing* and others). ● Based on refined periodisation and reconstructions, it is improbable that the source of OC word would be a PT wordform of “wheel ► chariot” (or “heel”) directly. Blažek’s proposal for TA(?) source does not work when comparing timeframes[144]. To reconcile this, one would have to place the derivation before TA time with TB losing/not attesting this meaning[145]. Otherwise, it does seem to fit perfectly.[146] The problem with supposing that the word must have been borrowed is that it is based on an idea that lexemes from this semantic field could have been and *were* borrowed from Tocharian into Chinese. This is supported supposedly by archaeological evidence, which, however, is circumstantial – that around the time of a *possible* first contacts, the Chinese *suddenly* invented this branch of technology. Shaughnessy (1989) speaks of other cultural items being probably borrowed along the with chariot around the same time period: weapons, 7-days week, god/heaven, he makes no mention of Tocharians directly. The case of this borrowing is not considered proven here while also not considered disproven.[147]

13 ⊥ 軌 *guǐ* “wheel ruts” ■ Lubotsky (1998) ■ Lubotsky (1998:383) connects this word and 逵 with TB *kwarsär*, TA *kursär* ‘league, mile; vehicle, means of salvation’, with PT as calque from SKR *(pra)yoǰana* without stating that there is any certainty. ● Lubotsky (ibid:383) cites the meaning as ‘wheel-axle ends’ with OC form **k$^{w}$rjuʔ/*k$^{w}$ruʔ* probably from Baxter (1992). This meaning is also reconstructed by Starostin (2005). Schuessler (2007) does not reconstruct this character, however, he mentions tài, dì “OCM *dês, *dâs” ‘wheel-axle cap’ (ibid:614) 軑, to which we could add 轟, 轄, 軸

[143] Big/great sheep’s male is called *gǔ*. (First character translated using secondary meaning according to context, “summer” would probably not fit well).
[144] In the periodization accepted here, PT or CT still in 1st century AD (following Carling 2005). Warring states period ending no later than 221BC. An opposing view holds Blažek & Schwarz (2008; 2011:127), who seems to believe that the divergence took place around 400BC which would make this analysis a correct one with highest degree of probability. The validity of the used method of counting lexical replacements over a certain period of time to arrive at a formula for timing changes in an open system, which language certainly is, however, is to be evaluated by each reader.
[145] Parallels to this evolution are known in other languages, e.g.
[146] It was made to fit perfectly, after all. In a way, it is definition by circle – etymology of one word is proven by supplying missing parts in the other based on the assumption that they are indeed connected. Still, it cannot be ruled out as possible PIE ancestor does support this analysis. For the vowel correspondence between OC *o/u and foreign *o/u see Schuessler (2007:112-115).
[147] As such, it is not a valid candidate for deriving adaptation rules.

and 軔 from Baxter & Sagart (2014b) as characters somehow relating to wheel-axle. ● Baxter & Sagart (2014b) did not change the reconstruction, having OC *kʷruʔ – for the other identifiable meaning 'wheel-ruts'. Checking in old texts, OC character dictionaries *Shuowen* and *Guangyun* (Sturgeon 2011) seem to prefer Baxter & Sagart (2014b) meaning over the one preferred by Lubotsky and Starostin. ● As an obvious phono-semantic compound[148], the etymology of the character will not give us any meaningful hint as to the original meaning of the word, it is possible also that it was used to write homonymous cognates in which case, there is no telling whether it could be possible to connect the etymology outside the language. ● Baxter & Sagart (2014b) reconstruct 軑, which would seem to fit the meaning originally proposed best, as OC *[l]ˁe[t]-s, which cannot be connected to the Tocharian words. 轂 with allograph 轄 'wheel-axle cap with linch-pin' is reconstructed as OC *[g]ˁrat, 軸 as OC *lruk 'wheel-axle' and 軔 as 'ornate band on axle-cap of wheel' OC *[l̥]ru[n]. None of these words seem very similar to Tocharian kVrsVr. ● The semantic side of the proposed connection between OC and Tocharian words seems rather strange, if the Tocharians were in a position to show and/or explain the inner workings of their vehicles to the Chinese, why would they not explain that they are speaking about a road, not a part of a vehicle? The connection with "wheel-ruts" 逵 seems much more reasonable with meaning "axle-ends" coming from a different source. One could hypothesize that this character was used primarily for axle-ends with near-homonymy and ambiguous contexts having it in time come to be used as a substitution for the other, although there is absolutely no indication of this in texts. ♦ Since the connection to this character specifically seems hypothetic at best, the rest of discussion follows in the entry for *kuí* 逵.

H

### 3.3.7. J

14 ⼯ 塅 *jí* "masonry" ***tsjik*- <- **tsik*- (pre)OC-PT ■ Lubotsky (1998) ■ Lubotsky (1998:385-386) considers the word to come from unattested noun derived from PT ts'əik- "'to build, form' < PIE dʰeigʰ 'to knead clay, make walls'" and mentions Bodman, Coblin and Baxter as connecting the Chinese word to Tibetan *rtsig* "so that this word has been borrowed not only in Chinese, but also in Tibetan." (ibid:386) ● Baxter (1992:301) indeed mentions Chinese "*jí* < *tsit* < **tsjit* < **tsik* 'masonry'" in connection with "Tibetan *rtsig-pa* 'to build, to wall up; a wall, masonry'" although the reason he reconstructs **-ik* as the OC pre-form is exactly to accommodate for the TB form, to reconcile a set of regular correspondences to prove shared etymology.[149] ● Schuessler reconstructs[150] two complete homophones, in (2007:294) OC *tsit 'coaled part of a burning torch, to burn or scorch earth' < *tsik PST 'to smolder' and (ibid:295) 'to wall up, wall, masonry' with same forms in OC/PST. This reconstruction is quite strange in that regard that the first word to have negative connotations as proved by denotation of TB cognates he derives from PTB m-(t)sik 'burn, angry', connection to *jì* 瘠 "sick" is consider improbable. ● The solution to on the one hand reconstruct two words on the basis of seemingly incompatible meanings and also not to accept a probable convergence is unexpected. It is, however, probably correct, for two words to converge they should probably be both either associated with positive or negative feelings if their semantics is not directly related. Alternative would be that the character is simply used for homophones. Why would a homophone of a word potentially connected to taboo not disappear is even harder to answer. Yet another version is that the Chinese at the time already knew how to make fired bricks and these words are ultimately from the same root. This would seem the most probable, if so, no space for a borrowing would be possible. ● Adams (2013:807) for TB tsik- reconstructs PT *tsäik- from PIE dʰeiĝʰ-.

[148] The modern right side meaning nine is usually used only as a phonetic component, rarely being a simplification of a part of an original diagram. The sound value fits here quite well.
[149] In other words, Baxter does not speak about loanwords, only about sharing cognates with certain sequences of segments and bases it on this very example.
[150] By "reconstructs", I mean that he cites the entry with his form and the meaning he chose from (MC) dictinoaries.

● Starostin (2006) PST *[c]ik citing as a possibility the last theory mentioned earlier. ●[151] Rix (2001:140-141) cites PIE "*$d^heie\hat{g}^h$- 'bestreichen, kneten'[152] Präsens *$d^hei\hat{g}^h/d^hi\hat{g}^h$-" as the ancestor to word forms of Young Avestan *uz-dišta*, Armenian *edēz*, Gothic *digan*, etc. with TB "tsikale 'zu formen'[153]" as possible descendant form with a note on Winter's emendation to *tsiṅkalle* which was refuted by Hackstein. If this form were correct, it would be possible to phonetically connect it to OC forms for "building" other than one proposed by Lubotsky. ♦ Could MC 砌 be connected directly?[154]

15 ◊ 車 *jū* "chariot" ■ None (by extension). ■ Schuessler (2007:182) mentions two meanings for this reading: "a piece in chess" and "(literary) carriage", he reconstructs OC **ka*. ♦ For discussion see primary reading of 車, *chē*

### 3.3.8. K

16 ㄴ 逵 kuí "thoroughfare" ■ Lubotsky (1998) ■ Lubotsky (1998:383) connects 逵 and 軌 with PT **$k^w$ärsär* while inferring that the [r] went through metathesis as part of adaptation because OC "probably had no final *-r*." ● Against that: first, while final **r is not reconstructed by all experts[155], the medial **r may very well be prefixed in many or even all cases[156], which would rule out the metathesis on the grounds of simply being counter-universal[157]. Second, OC had unarguably resonant coda, places where some would reconstruct **-r are reconstructed by others with conservative **-n. If the word was being adapted into the language, it would be very unnatural for it to change the place rather than articulation.[158] ● Baxter & Sagart (2014b) reconstruct **[g]$^w$ru* 'thoroughfare', Schuessler (2007) doesn't mention this character. Both Zhengzhang (2003) and Starostin (2005) reconstruct voiced initial. Voiced initial is incompatible with any stage in the evolution of the Tocharian word imaginable when in isolation. ● Adams (2013:253) seems to offer a satisfactory resolution of TB *$k^w$arsär* etymology as probably being

---

[151] Very hypothetically, I would connect the OC word 塓 to 城 *chéng* *[d]eŋ "city wall" (reconstruction by Baxter & Sagart 2014b) if anything. Metathesis of prefixed *s- with cooccurring blending – "*$s_2$ -" from Baxter & Sagart (2014a:56) ? – would then be an explanation for ts-t initial correspondence, this process has been mentioned already by Schuessler (2007:58) with large skepticism. The rest seems to be non-reconcilable for the moment, so this hypothesis is probably invalid, unless we postulate the meaning of "masonry" to be distributive and of "wall" to be "terminative" from some verb *[t]V "to build" (Schuessler 2006:18 mentions such suffixes *-ŋ and *-k). This way, we could even connect these words to 蒸 "twigs" and pretty much anything we wish. Even so, for semantic parallel see e.g. Czech *zedník* (wall:ACT) "mason" -> *zednictvní* (*zedník* + nominalising suffix) "masonry", also Slovak *murár* (wall:ACT) "mason", etc. Connection with 城 should be equalled to connection with 成 and 盛 (same reconstruction, e.g. Schuessler 2006:185, however, he reconstructs "OCM *geŋ ?"). Another term for brick-laying exists: 砌 qì, no reconstruction is done in Baxter & Sagart (2014b), and although Zhengzhang (2003) has *s$^h$iids, the series' initial s- corresponds consistently to Baxter & Sagart (2014a) ts-like sound (*[ts$^h$]i[t] for 七, *[ts$^h$]$^s$i[t] for 切, etc.). It may be a descendant as it is attested much later (cf. Sturgeon 2011), therefore it will not provide us with information on etymology of the word in question.

[152] "Spread, knead."

[153] "To form, mold."

[154] Mentioned Baxter (1992:327).

[155] Baxter & Sagart (2014b) do list e.g. *chún* 鶉 as OC *[d]ur.

[156] Cp. Tibetan.

[157] The sonority hierarchy universal, which is nearly absolute due to its physiological basis, would not allow for a final -r to move into an initial minor syllable position; I am unaware of a word in any language where a coda consonant moved to a second position in onset, save for a few dubious examples, though it is probably not impossible.

[158] This could happen in more ways: either the consonant has been weakened in the donor language to a semi-vowel and there might not be a trace in the recipient, or the manner of articulation may change to suit phonotactic rules of the language, or it may be unpacked. E.g. Thai adapts foreign loans with final *-r* as having final *-n*, even though it does have initial C*r* clusters, name Thatcher from non-rhotic English variety is rendered as แทตเชอร์ *tɛtsɤɤ* versus old Pali borrowing มนเทียร<มณเฑียร moont$^h$iæn from มนฺทิร "monthir" (the second example cited from "thai-language.com…").

derived from PIE "*$\hat{k}_{\text{e}}$rs-r-u- 'a [distance of] running'", deverbal noun from *$\hat{k}$ers- with cognates in various languages. Therefore, reversing the way of borrowing does not seem to be needed, either. ♦ For more on the original thesis see 軌 *guǐ.*

17 ⊥ 鞹 *kuò* "leather" OC-PT? ■ Lubotsky (1998) ■ Lubotsky (1998:384) 鞹 OC *k$^{w}$hak/*k$^{w}$hāk connects with TA kāc 'skin, hyde' < PT *k$^{w}$ac- < PIE *kuH-ti adding IE cognates. Lubotsky supposes the original meaning to be connected exclusively to chariot based on secondary meanings (which could be explainable as a simple semantic extension). He refuses the connection to Tibetan (s)kog-pa 'rind, shell', Burmese ə-khok 'tree bark', adding that uH>wa is attested only in Tocharian. ● The sound change would only be relevant if we were to suppose the etymology is of IE origin without any doubt. As for OC word's cognates, there is no reason to not see them as such. On the phonetic side, adapting PT *-c as OC *-k would seem reasonable, semantic side is also compatible. ● Baxter & Sagart don't discuss the character in question, they do reconstruct (2014b) 革 OC *k$^{ʕ}$rək 'hyde, skin' and 郭 OC *k$^{wʕ}$ak 'outer wall'. Both words seem to serve as a phonetic component while the first one also as a semantic. Schuessler (2007:341) reconstructs 鞹 OC *khwâk, 郭 OC *kwâk as possibly connected to it with *zhóu* 糓 'husk' OC *kûk as "somewhat similar". He adds TB cognates – Jiarong *werk$^{h}$wak* and Kiranti *kwak/kok-te* 'skin'. The word seems to have therefore a good ST explanation. There seems to be no need for an external source explanation, while it is not ruled out.

### 3.3.9. L

18 ⊥ 里 *lǐ* "village"[159] *C-rəʔ <-> *wriH OC-prePT/postPIE ■ Lubotsky (1998) ■ Lubotsky (1998:386,388) considers an OC form "*C-rjɨʔ/*C-rəʔ" to be "probably" descended from Tocharian, connected to TB *riye* TA *ri* 'town' with <hypothetical> "PIE *$\underset{\sim}{u}$riH-eH$_2$ , cf. Thracian βρία" as a support. ● A few years later, Lubotsky & Starostin (2003:264) reverse the direction "The Indo-European etymology of Toch. A *ri*, Toch. B *rīye* is thus rather questionable. On the other hand, Peiros and Starostin (1996,2: 77) reconstruct Sino-Tibetan *riəH, adducing Jingpo *mare$^{1}$* `town'." in reference to the original argument. ● Baxter & Sagart (2014b) refine the reconstruction of 里 to MC **liX < OC *(mə.)rəʔ*, in Baxter & Sagart (2014a) no further detail is given to explain the m- minor syllable in this word, although from general discussion by the authors (ibid:48, 53-56), if we believe the form to be analysable by productive means in OC, then only the long version (in contrast to *m-rəʔ) should be reconstructed, otherwise the outcome would not yield the *l-* initial. None of the prefixes offered by Baxter & Sagart (2014a:53-56) fit the semantic side well[160]. Schuessler (2007:19) mentions PST prefix m- already unproductive in PC which marks intransitive[161] verbs, which seems to be even less useful. Either way, the *mərəʔ form seems to fit the probable cognate in Kachin (Jingpo)[162]. ● The word is not considered for the above stated reasons to have come to ST from IE sources.

### 3.3.10. M

19 ⊥ 馬 *mǎ* "horse" m° <- m° PC/PST-? ■ PIE supposed by many, Tocharian origin sometimes supposed to be suggested, probably by misunderstanding where forms are not given. ■ The word has been discussed by numerous sources, no additional information is to be given here. ● Gamkrelidze & Ivanov

[159] The character was also used to write a homophonic unit of distance, which is possibly related (rather) to 理 (roughly "divide into regular sections"), see e.g. Schuessler (2007:349-350).
[160] Possibly m$_{ɪ\ c}$ for agentive/instrumental nouns with note that no available verb seems to fit both form and expected meaning (*rVʔ), closest would be shēng 生 "to live", Baxter & Sagart (2014b) OC "*sreŋ (or *s.reŋʔ)". I suppose the initial s° could be dealt with as PST *s-, for the difference in final, I see no obvious hypothesis.
[161] Actual term used was *introvert* meaning inward-oriented with explication connecting it to (in)transitivity, more detail on this nomenclature Schuessler (2007:38-9).
[162] Jingpo is a language of a group which is part of larger Jingpho/Kachin ethnic, they speak a separate branch of Sino-Tibetan (cf. SIL International 2017b).

(1995:472-473)[163] speak of Celto-Germanic term projected as PIE *mark$^{h}$o by Pokorny, which they believe to be a Wanderwort from unknown Asian source, which is a reversal of the typical hypothesis of the origin. They connect the word to Mongolian *morin*, Tungusic *murin*,[164] Korean *mal*, ST *mraŋ. They (ibid:479) discuss that the similar mythologies regarding horses must have travelled from Europe to Asia, (ibid:828) they hypothesize supposedly following Pulleyblank (1966:31-32) that the idea of a horse-drawn sun in OC was of IE origin. ● The idea that a culture should have a need to borrow a concept of sun being dragged by a horse carriage is the same as borrowing the concept of flat Earth or a god of rain, the idea is so generic, it could easily arise in an ancient context quite separately. ● Baxter & Sagart (2014b) reconstruct OC *m$^{ʕ}$raʔ, Adams (2013:796) records TB *haye*, yakwe (ibid:518). There is no reason to connect the OC word to any of the two mentioned. Of note is TB simā "marshal" borrowed from Chinese (ibid:758).

20 □ 蜜 *mì* "honey" *mit/mət <- *mᶨətə OC-(pre)PT ■ Usually projected to Polivanov (1916)[165] ■ Lubotsky (1998:379) says "We have known for 80 years (since Polivanov 1916) that the Chinese word for honey is likely to be of Indo-European, probably Tocharian, origin". Polivanov (1916) makes no mention of any Indo-European branch, though. ● The Chinese word's etymology has been discussed countless times, often with the conclusion that Tocharian is indeed the original language. The reasons are considered convincing, no complete discussion will be present. Most recently, Meier & Peyrot (2017) studied it, coming to the conclusion that Polivanov (1916) simply did not know the Tocharian word yet. As they note (ibid:8), some objections have been made to the specific form of an etymon in PT, for details see their article. They reiterate that the traditional reconstruction of OC *m(j)it is correct and that TB mit$^{ə}$ and TA †mät from PT *mᶨətə are cognates. They search for a perfect match for both consonants and vowels, postulating that either the borrowing is rather late, or from a prePT stage with intermediate *mᶨitə (ibid:18).

### 3.3.11. N

21 ⊥ 牛 *niú* "ox" ŋ$^{w}$ə <- *$g^{w}$ou- OC/PC-postPIE ■ Gamkrelidze & Ivanov (1984)? ■ Cited and also ruled out in Lubotsky (1998:381) as "often proposed" borrowing from PIE with "only one phoneme *$g^{w}$ in common". He cites TB *kau* TA *ko*. ● The connection with Tocharian seems unlikely as not even that single OC/PIE phone seems to correspond in respect to voicedness in Tocharian proper. ● Gamkrelidze & Ivanov (1984:935) cite OC mi̯ĕt 'мед', OC k`iwen 'собока', OC *.ngi̯ə̯u and *`kuo 'бык' < Toch. 'корова', ModM *chu* 'свиня'[166] using Karlgren reconstructions as coming from Tocharian without citing source or reason, obviously considering them proven. Characters are not cited but are probably: 蜜, 犬, 牛 with 牯, and 豬 resp. ● The Tocharian words are impossible to reconcile with the voiced onset in Chinese. Baxter & Sagart (2014b) reconstruct OC "*[ŋ]$^{w}$ə(< uvular?)". This would open up a possibility of borrowing from a form N/ywɑ. Neither Tocharian nor any other geographically close IE language seems to show this form. ♦ Word connected by some authors is 牯.

### P

### 3.3.12. Q

22 ⊥ 犬 *quǎn* "dog" *K$^{w}$en ← *k$^{w}$en OC-PT ■ ? ■ Lubotsky (1998:381) refutes on the basis of having an established Sino-Tibetan etymology[167] with reference to Benedict (1972:44). ● A quick look at Baxter & Sagart (2014b) wordlist with check in CText whether it is attested in OC texts shows these

[163] A reworking of the original Gamkrelidze & Ivanov (1984) in English, taking much less radical stance than the original in many formulations.
[164] To my knowledge, the Tungusic form is usually identical to Mongolian.
[165] A text rarely to be found in libraries, fortunately kindly digitized by the Orenburg Regional Library.
[166] All discussed in this work, in order: *"honey, dog, bull < cow, pig"*.
[167] Actual formulation is Tibeto-Burman, this would not, however, rule out borrowing ultimately coming from IE languages, therefore it is interpreted here as a mistake.

(near-)synonyms: 狗, 獀, 豺[168], (獑, 狸)[169], *Shuowen* shows additional [1] and 獟 with circular definition, although to the first one "一曰逐虎犬也" is added; since they still lack modern reconstructions[170], let us not delve into speculation. ● Schuessler (2007:18) adds derivational suffix for (denominal) substantives *-n to a supposed PST **kwi* "dog" to make up the OC word. This would rule out borrowing from IE. ● Baxter & Sagart (2014b) reconstruct OC *[k]$^{wh\varsigma}$[e][n]ʔ which would require the borrowed word to be *K$^{wh}$αR. The labiovelar fits, the vowel would have to be adapted, which is also not a problem (grave onset), the coda seems to fit also. The problem lies in the fact that a very large set of forms fits into this frame, giving no surety. ● Schuessler (2007:257-258) reconstructs 狗 OCM *kôʔ<*klôʔ which fits the possible cognate 犬 while ruling out the Tochrian side. Baxter & Sagart (2014b) reconstruct 狗 *Cə.k$^{\varsigma}$roʔ, 獀 OC *ke[t]-s, 豺 *m-[ɢ]$^{\varsigma}$a[r]-s. They seem similar enough to consider PC **k$^{\varsigma}$we. The unvoiced initial would then be a problem for PIE word, which would have to be already unvoiced. Alternatively, the words were borrowed at separate times from/through various languages.

R

### 3.3.13. S

23 ⿱乘 *shèng* "chariot" **kə.ləŋ+s* ← **klāṅk-* PC-PT ■ Lubotsky (1998) ■ Lubotsky (1998:382) uses OC reconstruction from Baxter (1992) MC *zyingH* < OC **Ljɨŋs*/**Ljəŋs*[171] to support idea that OC < TA TB klānk 'to ride' < "PIE **kleng*" with "obvious derivation from *chéng*" < **ləŋ* with further borrowing into Tibetan. ● *Chéng* is a verb, that would mean that "chariot" was not borrowed, rather the verb "to travel using a vehicle" was borrowed or that somehow nominal form came to be borrowed as a verb ● Baxter & Sagart (2014b) update the etymology: MC ***zyingH* < OC **Cə.ləŋ-s*. ● A note regarding N→V: borrowing as **Cə.ləŋ* + OC **-s* (V>N function, Baxter & Sagart (2014a:58-59) *-s$_1$ > MC *-H) would make sense to preserve nominal class[172] with backformation[173] leaving out the *-s, possibly occurring at the same time as borrowing. As backformation of the kind suggested as a way to preserve the PT source is not considered to be proven by every scholar to exist in OC[174], it is highly unlikely there is any way to sustain the original claim if we suppose N→V. If the claim was that the athematic root verb without suffix was borrowed, then apart from the problem of borrowing a verb for a kind of travel on a vehicle for which the language supposedly lacks a noun, possible segmental discrepancies arise. ● Schuessler (2007:185; 2009:116) reconstructs *chéng* OC **m-ləŋ* and *shèng* **m-ləŋh*[175], "exopassive[176]

[168] *Shuowen* states "胡地野狗", roughly "hound from lands of *Hu*", where Hu means western/northern barbarians.

[169] Projections?

[170] They do not appear in Zhengzhang (2003), Schuessler (2007) or Baxter & Sagart (2014b).

[171] In his notation L stands in opposition to l in clusters, following Bodman through Baxter 1992. The matter of OC/PC two *l*s is a bit complicated, as noted by Lubotsky; here, he considers the clusters simplified already at the time of borrowing with the chosen reflex being closer to PT cluster. Interestingly enough, Baxter (1992:232-234) does not differentiate these clusters in such way. As seen further, Baxter & Sagart opt for a different solution. Note the Schuessler (2007) two *l*s different from these with cluster – preserved in OC, as seen in 車.

[172] Elšík (2009:284) shows what to me seems like an example: Selice Romani *kóbás-kiň-a* "[kind of] sausage" ← Hungarian *kolbász* + (borrowed South Slavic) FEM noun marker + (borrowed Greek) FEM.NOM.SG

[173] The process of reanalysis of morphemic boundaries, e.g. English *hamburg-er* ([dish] from Hamburg) ► ham-burger (ham-hamburger), Jap. アルバイト *arubaito* (Germ. *Arbeit* "work") ► *aru-baito* > *baito* バイト "part-time job".

[174] L. Zádrapa (Pers. comm., in Czech, translation) "Desuffixation, though not impossible, is not considered <by sinologists> to occur in OC at this moment." Nevertheless, at least some consider backformation of some kind a possibility (see 1.4.1)

[175] Schuessler (2007:16) PST *-s > OC *-s/h PSV.

[176] Schuessler (2007:38-39) explains it as a combination of outward direction of action (S->O) and passive voice. The complete list is endoactive, exoactive, endopassive, exopassive. Approximation could be made by reading "'middle' for 'endoactive', 'active' or 'causative' for 'exoactive', 'passive' for 'exopassive'."

derivation of *chéng*” in line with *communis opinio*[177] (i.e. deverbal noun derived from verb undisputed, only needed by the newly invented controversial way to substantiate a possible N→V) without further specifying the origin of OC form.[178] Probable, unproblematic, semantic evolution of inherited verb: “to climb/mount”[179] > ModM “to ride” versus other ST “ascend” by narrowing and shift. ● Starostin (2006) reconstructs two different words for a yet earlier stage: PST *lǎŋ (?) “rise, ascend” PST *liŋ “mount, a k. of vehicle”, with a note that the latter contaminated the former in OC. Cognates mentioned are: Tibetan *laŋ* “rise”, Kachin *luŋ*$^{2}$ “ascend”, Lepcha *tă-ljaŋ* “the high place” for the former; Burmese *hlańh* “vehicle”, Kachin *leŋ*$^{2}$ “vehicle, wheel” for the former. ● While using the etymology of a character to substantiate a claim about the etymology of a word is debatable, the fact stands that no trace of a vehicle seems to be present in its oldest forms.[180] The form reconstructed for PTB by Starostin (2006) does not have a connection to a character which could be semantically connected: other than the one in question, 畋 “to hunt” and 田 “a field” (meaning “round, rotate” is also reconstructed, with no character mentioned”). If it was contaminated, it should therefore have already occurred in PC. For this reason, it is imaginable that some borrowing occurred inside the ST family in the direction proposed by Lubotsky, although this may have been from a deverbal noun from a verb semantically expanded, no IE contact is necessary. ♦ Not to be confused with 車

24 ⊥ 獅 *shī* “lion” ● Pulleyblank (1962:226) states that “There is no reason to regard 子 here as the noun forming suffix of Modern Mandarin. In earlier passages it is always treated as an inseparable part of the word and it is only much later that *shih* alone comes to be used for ‘lion’.” ♦ For more information see 獅子

T

W

X

### 3.3.14. Y

25 ◊ 鴈 *yàn* “wild goose” **ŋrans* ↔ **Ken(t)s* PC-PT ■ Adams (2011:39) ■ Adams (2011:39) suggests that correspondence between OC **ŋ(r)a-n-s* and CT *kents** looks promising – however, he also notes two problems: first is the non-correspondence of the onset, second that OC in his view is also very hypothetically derivable from **ŋa* ‘domestic goose’. Indeed, the process of secondary noun derivation by marker -n in *PC* has been postulated, as stated in 1.4.1.3, this would, however, probably mean double suffixing[181], otherwise the *inherited word* hypothesis seems like a good explanation. ● Later, Adams (2013:207) postulates for item TB *kents** meaning ‘goose/bird?’ and loosens up the semantic side a bit. ● Baxter & Sagart (2014b) reconstruction of MC ***ngaenH* < OC **C.[ŋ]ˁrar-s* ‘wild goose’ does not agree with coda of the Tocharian root, and also the initial already mentioned by Adams himself as not

[177] Note that G. Starostin (2009) does not agree with the widespread reconstruction of **m-* minor syllable simply on the basis of TB cognate with *bC-* initial cluster, pointing out the inconsistency, shown e.g. in 乘 (the word is discussed here) missing it.

[178] The original form of the character in Oracle bone script seems to be a diagram that does not indicate any connection to vehicles. If we exclude the shift of the word’s primary meaning before the character’s creation, the derivation from chéng would then be the correct one, though not from a loan, rather from the meaning “to mount, climb” extended to “ride”.

[179] Compare “to mount stairs”, also “to mount a horse”, also “the bills mounted” and more. Words for “going up” usually develop very wide meanings with the pass of time. Originally “to go up the mountain”. Semantically widened, shifted, narrowed, varied. Similar shifts in Manchu (ᡠᡥᡝᠰᡳᠮᠪᡳ *wesimbi* with CAUS suffixes may stand for “to ascend, go up, raise, promote, lift, submit to present (emperor)…” as listed in Gorelova 2002:249), Japanese (*agaru* 上がる, *noboru* 上る), French (*descendre*), Czech (*vystoupit*), others.

[180] I would see a man climbing up something, possibly a plant. In any case, *graphic indication* of a connection to riding shows only after the Bronze script evolved into Seal script. cf. Sears (2013).

[181] Strictly speaking, the PC suffix which was already unproductive in OC may have also been already unanalyzable. This would mean it is a simple case of suffixation.

agreeing. ● While both Tocharian and Chinese words do look similar and do have similar meaning, various birds have similar sounding names in the area. This might indicate a chain of borrowing and gradual shift in meaning or a simple case of onomatopoeic base.[182] ● Note that words for animals make strange semantic twists when borrowed and even when inherited – cp. e.g. Turkish *deve* "camel" and Yakut *taba* deer[183]. Some words may seem like a clearly motivated borrowing when in fact, they are newly coined and their semantics may therefore not be compatible.[184] ● Schuessler (2007:222) connects 雁 (seemingly an allogram of 鴈 – Baxter & Sagart reconstruct identical forms and meanings) with 鵝 *é*, in his system reconstructed as **ŋâi*, in Baxter & Sagart (2014b) this is *ŋˁa[r]. As for *how* these two words could be connected in the Schuessler system – he is not specific. From the general information he provides, nothing can account for that. He accounts for that in Schuessler (2009:254) where he reconstructs **ŋrâns* in reality mirroring Baxter & Sagart. ● As for the ability to reconcile the final of Baxter-Sagart and PT, as shown, not every system reconstructs OC **-r*, we can therefore opt to believe in conservative *-n. The initial is a bigger problem: borrowing of an undisputed nasal voiced onset from an undisputed unvoiced non-nasal onset is possible under special circumstances like those presupposed by Pulleyblank (1962)[185], against that, however: 1. no widespread recognition seems to manifest in modern reconstructions, 2. we would have to speculate on a dialectal dropping of the onset in Tocharian, then the potentially palatalising vowel could be understood as ʔyə and possibly be adapted as *ŋa*, then again we would have to postulate a dialectal change between division III sign, and the chain continues… The words for goose is therefore not considered a probable or even candidate for a loan into Chinese.[186] ● Yet another kind of wild goose, *hóng* 鴻 OC *[g]ˁoŋ (Baxter & Sagart 2014b), exists. Not connected by anyone, if we hypothesized that the borrowing occurred in the PC-postPIE times, this could more easily be of IE origin, its connection to other ST words being not obvious.[187] ♦ Also written 雁.

26 ⊥ 營 yíng "lay out, plan" *wʲVŋ <- *wäńk OC-PT ■ Lubotsky (1998)? ■ Lubotsky (1998:381) notes OC *w(j)eng in connection with TB wäńk- 'to prepare' in his rules stating the Tocharian side does not have a good IE etymology. It is unsure what made him choose this word combination. ● Baxter & Sagart (2014b) *[ɢ]ʷeŋ "demarcate, encamp", as Schuessler (2015:590-593) comments, *ɢ comes from a hypothesis that forbids initial **y* for which it basically stands. If we accept this, the phonetic side does indeed fit. The semantic side also is not problematic. Schuessler (2007:576) reconstructs OCM *weŋ 'to lay out, plan, build, encamp, surround' and sees this as part of Austro-Asiatic substrate, giving e.g. Old Mon wiṅ 'surrounding', Khmer viaṅa 'enclosed, encirceled' and others as cognates or allofams. This is not exactly convincing; however, Adams (2013) does not comment on the TB word, which means this may be a ghost word.[188] Malzahn (2017) does list the word with the meaning "to prepare, offer (food)".

27 ⊥ 垣 *yuán* "wall" †wjɑN <- †wɑnD° preOC?-postPIE? ■ Lubotsky (1998) ■ Lubotsky (1998:386) sees this along with 園 and 圓 as possibly coming from an unattested deverbal noun of Toch. provenience, connected with TA TB want- "envelop, surround" from PIE *wendʰ-. He sees all three

[182] "Chicken" (the semantic actual semantic range is quite wide, similar to the English word – a bird to be eaten) Thai *gaiˈlɔ̀*, ModV *gà*, others… In animal emulation speech, duck, goose and many other birds have similar sounds in langauges of the world (cf. Czech *kač:*RED, *gá:*RED, *kvok*:RED for sounds of three different domestic birds etc.)
[183] For more on Proto-Turkic *debe see Dybo (2007:58-9).
[184] Cp. *Canary Islands* and *canary* (bird). (Harper 2017)
[185] Mentioned elsewhere in this study.
[186] For reasons of brevity, I will not discuss here the possibility of reversing the direction. The proposed PIE etymon seems to me like a good candidate for the TB word, even in the case that we find that semantics prove to be somewhat loosely connecting those two.
[187] There is still no need to see any connection with Tocharian. Also, this should not be understood as a hypothesis of the actual origin of the word, it only illustrates that unless required by genetic reasons or by direct evidence, one should not postulate loanwords.
[188] A word that is not attested but is cited by linguists.

characters as being originally for the same word. He goes on to cite Pulleyblank who sees words related to roundness as a word family, which he doesn't agree with. ● Baxter & Sagart (2014b) reconstruct 垣 *[ɢ]ʷar, 園 *C.ɢʷa[n], 圓 *ɢʷ<r>en. If we accept the idea of word families, the reconstructions would support that. If we take ɢ to stand for /j/ in other systems (mentioned elsewhere in this work), the common etymon would be **jwɑ(R)[189]. While the vowel could be reconciled with PC/OC form, there is no reason for a palatalised onset. Even if we suppose the final to be -n, there is little reason not to borrow the form as †wɑnt. While the semantics seem to be connectable for some or even all of these words, on the basis of forms being non-reconcilable, this is not supposed to be a valid hypothesis.

28 ⊥ 園 *yuán* "garden, park" ■ Lubotsky (1998) ■ ♦ See 垣.

29 ⊥ 圓 *yuán* "circle, cicumference" ■ Lubotsky (1998) ■ ♦ See 垣.

30 ⊥ n 閱 yuè■ None? Mentioned by Lubotsky (1998) ■ For discussion see 兑.

### 3.3.15. Z

31 ⊥ 楨 *zhēn*[190] ■ Lubotsky (1998) ■ Lubotsky (1998:386-387) uses Karlgren's reconstruction unamended by Li[191] "'post in the wall, support' < EC *trjeng* < OC **trjeng*/**treng*" in connection with "Toch. B *trenk*-, A *trank*- 'to be fixed to', PIE **dʰergʰ*-"; if true, from unattested deverbial noun (ibid:386) ● Neither Baxter & Sagart (2014b) nor Schuessler (2007) mention this character, Zhengzhang (2003) reconstructs *teŋ in the series with 貞 with onsets which could be abstracted as *(r)D-, of which only 貞 and 赬 are reconstructed also by Baxter & Sagart (2014b): *treŋ and *[t.kʰ]reŋ corresponding to Zhengzhang (2003) *teŋ and *tʰeŋ. Probable updated OC forms therefore do not conflict with the one used by Lubotsky. ● Shuowen stresses the "wood" in meaning "剛木也。从木貞聲。上郡有楨林縣" but Karlgren's analysis seems to be uncontested so any direct connection to a word with a meaning of "wood" would not be necessary have to be true[192]; still, Schuessler (2007:612) reconstruction of zhēng OC 蒸 *təŋ "'brushwood' (as firewood)" seems like a good candidate for a cognate. This word has a PTB form with Written Tibetan and Written Burmese cognates[193]. Baxter & Sagart (2014b) reconstruction agrees in form with Schuessler, while Zhengzhang (2003) reconstructs incompatible (?) *kljɯŋ. ● For the TB words treṅk, entreṅkätte, treṅkäl and treṅke Adams (2013:338) states "perhaps from PIE **dʰrengʰ-/dʰrenĝ-*", a nasal-infixed[194] 'hold fast to' as "*élargissement* of **dʰer*- 'id.'". Rix (2001) does not reconstruct the expanded stems, although (ibid:145) the root **dʰer*- 'befestigen, fixieren'[195] is present.[196] While the exact form of PIE stem is in doubt, this does not seem to be a valid reason to not consider the word inherited. ● Adams (2013:338) also mentions possible connection of PT *träṅk to TB traṅko 'sin', ibid:332 he states *traṅko* should be probably connected with träṅk- 'lament' as "*'that which is lamented'" and his alternative as 'that, which clings'. A semantic interpretation "that,

[189] Again, following Schuessler in considering the labialization unsure, possibly from a sequence of segments.
[190] ModM -n is a result of dialectal variation. The velar coda for OC is supported by comparative evidence. While some OC *-ŋ > ModM -n may be disputed – against Baxter & Sagart (2014a) inconsistent way of marking these see e.g. Schuessler (2015:575-576) – there seems to be a general consensus about this one.
[191] Karlgren (1957:221), identified rather by number: 834L.
[192] The motivation for connecting the word to "wood" is not only the character etymology, but also the idea of palisades and Chinese borders traditionally marked by trees. Also my native intuition in Czech, where words *roští* and *klestí*, both meaning roughly "brushwood" while also having a separate meaning could have had a role.
[193] Consider especially WT *tʰaŋ* 'pine, fir, evergreen tree' Coblin (1986:79) apud Schuessler (2007:612) compared with ModM 桢 'evergreen shrub' ("CC-CEDICT").
[194] Originally "nasalized".
[195] Attach, fixate.
[196] The stem as it would look after a Grassmann's law with the correct meaning is discussed (ibid:126) but wouldn't one expect PIE **d* to become TB *ts-/ś-* if coming from full grade? Also, the ablaut vowel is in wrong place for that. Connection with PIE initial **dr*- is already refused by A2013:38 on the basis of regular outcame in TB *r*-.

to which one clings" would seem more probable, cp. etymologically (intralinguistically) more clear TB *treṅkäl** and *treṅke** "clinging, worldly attachment" (ibid:338-9), also cp. English translation of Buddhist terms: SKR *upādāna* उपादान and *rāga* राग with Chinese equivalent of the first one being cóu/qǔ 取, in classical Chinese rougly "take, acquire"[197]. ● One should be hesitant to consider something connected to Tocharian religious terminology to come from Chinese and as there seem to be possible cognates with better semantic connection in other ST branches than with unattested verbal noun for the Chinese word; Lubotsky's hypothesis seems *improbable*.

32 丄 輈 *zhōu* "carriage pole" ■ Lubotsky (1998) ■ Lubotsky (1998:384) believes that 輈 'carriage pole' OC *tr(j)u is connected to TA *tursko* with tentative meaning "draft ox" from PIE *$d^hur(h_1)$- (no meaning stated), he points out the "metathesis" of -r-. ● First, if metathesis is supposed to happen, it would have to occur in accordance with Schuessler (2007:69) in between the PST and OC stage, other variants do not seem to be reasonable, as noted elsewhere in this work. Second, Lubotsky means to show that the general fuzziness of the semantic reconstruction serves his purpose while this is in fact a reason to doubt the hypothesis. Third a PIE stem with a meaning possible to connect is *$d^her$- 'befestigen, fixieren' (reconstruction by Rix 2001:145), which would have an unexplainable u-vocalism; Rix (2001:159-160) reconstructs PIE stem with the correct form with a meaning of 'beschädigen, verletzen', which does not obtain. If we believe the LIV to be an authoritative source, then the reconstruction offered by Lubotsky is a projection. ● Adams (2013:338) mentions TB truskäññа 'binding, bond, harness' in connection to the TA word, going back to PT tursk 'bind, harness' from PIE *$d^hwrH$-sḱe/o- with supposed cognates in Hittite tūriye- 'harness' and Sanskrit dhū́r 'yoke'. This seems convincing. ● Schuessler (2007:623) connects hypothetically the OC word, he reconstructs *tru, with OC *tu 舟 "boat" with semantic evolution from *"trunk". Baxter & Sagart don't reconstruct the mentioned word, they, however, reconstruct 轅 OC *[ɢ]$^w$a[n] with the same meaning, which could not be connected based on the onset. ● If we are to connect PT and PC hypothetical stems, it would have to be at a time, when the voicing was no longer distinguished in Tocharian, which is not a problem, and we would have to move the TB metathesis to a dialect of PT already, which is a big problem. CT would not work for reasons of chronology. TA would not fit at all. A slight possibility would be of a borrowing from Chinese to TB – which was relatively a common occurrence, with TA form then being either a morphological adaptation of an Indic form or an inherited word as postulated before and suggested by the supposed Hittite cognate. The OC etymologies do not seem convincing enough to reverse the direction. The borrowing as originally postulated seems possible, however, it is based on a premise that the technology was definitely borrowed from Tocharians, on an unattested verb, on unparalleled sound changes/adaptations and a PIE stem reconstructed based on words from three branches where two could be connected. For this reason, this borrowing is considered to be highly improbable.

## 3.4. Compounds

Phrasal meanings (idioms) and compounds in Old and Middle Chinese are somewhat problematic. What ensues is a discussion of those that are considered to be so in Chinese by other sources.

33 ◊ 阿魏 *āwèi* "ferula asafoetida" ← TB aṅkwaṣ(ṭ) ■ misunderstood as either Bailey (1946) or Pulleyblank (1962) ■ Lubotsky (1998:379) believes this to be uncertain, as the words are "Wanderworte, of unknown etymology", sic. ● Schuessler (2009:211,291) corresponding entries show OC forms: *ʔâi and *ŋwəi/ŋwəih/ŋwəs, "asafoetida" is explicitly stated to come from TB *ankwaṣ* -> OC ʔâi-ŋwəis ● Zhengzhang (2003) 阿 "*qaal", Zhengzhang (2003) 魏 "*ŋguls"; this version of reconstruction would seem to rule out borrowing from any source in the vicinity in OC time. ● Baxter (1992:313) cites Pulleyblank (1962) as claiming TB origin.[198] ● Pulleyblank (1962:217) mentions the Tocharian word

[197] Originally, the character depicted a right hand taking an ear.
[198] Also, probably for the purpose of consistency of the text, he modifies the reconstructions originally proposed to fit in his transcription rules. No obvious change in the reconstructions themselves seems to have been done.

among cognates citing Bailey (prob. 1946), but makes no explicit claim as to its origin "Bailey gave a number of examples in which Chinese diphthongs in -i appeared to represent a foreign sibilant or dental fricative. … 阿魏 M. ʔɑ-ŋi̯wəi\, 央匱 M. ʔi̯ɑŋgi̯wi\ = Khotanese aṃguṣḍä, Tokh. B. ankwaṣ, Uighur 'nk`pwš'".[199] *Earlier* in the text (ibid:99), he is even less clear – leading to some other author's obvious confusion: "Chinese syllables with M. -i̯- are found representing foreign words with vocalic initials where there is no reason to expect y-. … 央匱 M. ʔi̯ɑŋ-gi̯wi\ (besides 阿魏 M. ʔɑ-ŋi̯wəi\) = Tokharian B. ankwaṣ 'asafoetida' (Bailey 1946, p.786)". ● Baxter & Sagart (2014a:121) cite Bailey (1946) apud Pulleyblank (1962:99) as postulating a loanword from TB and consider their reconstruction of the second character to support the connection: "阿*qˤa[j] > *ʔaj > *ʔa = MC 'a> ē ('slope, river bank')" "魏*N-qʰuj-s > *N-qʰwəj-s > *ɴχwəj-s > *ŋwəj-s > ngjw+jH> wèi 'high'" ● *In fact*, Bailey (1946:786) is quite clear and considers the source of both to be KS *aṃguṣḍä* with 阿魏 <MC> "*â-ngjʷei*", 央匱 <MC> "*i̯ang-g`jʷi*"[200], where "The Chinese can be interpreted as **anguž* with *ž* expressed by final *-i̯*."[201] ● A reconstruction based on Baxter & Sagart (2014a: 121)/Baxter & Sagart (2014b:25,114) "*ʔˤa (< *qˤaj)" + "N-qʰuj"<(-s)> would allow for the interpretation of original form of borrowed item as OC /(C)a(j)(.)N(.)Kʰuj(s)/; if we believe this version, it is highly improbable that a borrowed item would add an aspiration unless there were some semantic reasons (folk etymology) or general rules[202]. A form twisted by reanalysis[203], however, allows for a far greater phonetic variation than of a single feature, making the actual source even more uncertain. This plant's centre of origin is Central Asia (Mahendra & Bisht 2012), which does support the idea that the term could have been an early borrowing from a language of the area and the extreme phonetic similarity seems impossible to ignore, in no version of any reconstruction seems to be a place for synchronic way of adaptation of a potential PT word. ● The word does not appear in pre-Han or Han major texts; the earliest mention seems to be *Tongdian* 通典 encyclopaedia (Sturgeon 2011) from the late eighth century CE , the alternative word seems to be attested even later (in major text only as late as KangXi). This means both words are relatively recent (also supported by the fact that their OC reconstructions mutually incompatible, obviously the characters were used for phonetic value). That means the word is indeed borrowed and the MC timeframe should be the correct one with the writing discussed in this entry being the original one, hopefully closer to the word from which it has been adapted. When looking at KS form, it would seem obvious the TB word is descended from KS (ṣd->ṣt). The Tocharian version, which lost in some versions the dental, seems to be the more probable source. ♦ Synonymous with 央匱

34 ⊥ 軱轆 *gūlū*[204] "wheel" ■ None, discussed by Bauer (1994) in extension of Mair (1990). ■ Mair (1990:45) mentions the ModM *chē* "chariot" in connection to PIE **kʷékʷlo-*, connects it to TB *kokale*, TA *kukäl* and Proto-Iranian **čaxra*, which he prefers on the presumption that the IE language would have had to lost the labiality distinction for velars by the time of borrowing. Bauer (1994:6) surmises that all words in IE and ST languages he lists are connected with source in some IE language. Partial list of his Chinese words is (ibid:8-9): ModM *kūlu*[205], *zhēkūlu, kūlur, kūliúliur, kulu, kóulou* all connected in meaning to roundness and rotation. PST form is taken from an unreleased Starostin work as **kʷ(r)eł*. Starostin (2006) no longer reconstructs that form, opting rather for **r[ua]ɬ* connected directly

[199] Obviously, this formulation is quite easy to misunderstand.
[200] In <> brackets are emendations.
[201] The idea that voiced coda retroflex fricative was de-fricativised into a semivowel seems to be phonetically possible but highly speculative. The Baxter & Sagart (2014a) solution with *-s* suffix (and "de-retroflexion") seems more appealing.
[202] E.g. in Thai, voiced plosives of Pāli are realized counter-universally as unvoiced aspirates, e.g. คุณ *khun(a)* "virtue", also "23:SG:DEF" from *guṇá* गुण "virtue" (P. Youyen, pers. comm.). This has probably historical reason (interference of intermediate language, probably Old Khmer).
[203] In this case, it would be of the kind of *bridegroom*, not e.g. *hamburger*.
[204] Also transcribed as kūlū.
[205] Dialectal variant reading for the entry.

to Chinese *lún* 輪 “wheel”. ● The characters appear in oldest sources separately (e.g. *Liu Tao* 六韜 “著轉關轆轤八具”), which means either full integration of the loanword in the lexicon already in the PC (PST) stage, or that it is not a loanword at all. ● The word may, again, like many words that seem to be either *Wonderwörter* or coming from a single proto-language,[206] be of iconic origin, cp. onomatopoeic “rumbling sound” ModM *gūlu* 咕嚕 (“CC-CEDICT”), also see previously cited passage[207]. Schuessler (2007:353) cites LH *lek-lok* “spinning (wheel)”. Various (partial) reduplications of characters for wheel exist. ♦ See also 車.

35 ⊥ 駃騠 *juétí* “(a superior kind of) horse” *kuet-dei ← *y(V)kwe postOC-PT? ■ prob. Schuessler (2007) ■ The word seems to be connected by all authors as to the Xiongnu[208] horse which in itself could work as an argument against it being of Tocharian provenience. Pulleyblank (1962:245-246) is probably the first to dispute its meaning as being identical to the modern *hinny* and would like to see it as a Yenisenia *kuti* > *küti*. Dybo (2007:87-88) discusses its Turkic connections coming to the conclusion it comes from Proto-Turkic “herd” **güdü-t-üg* > **güd-t-üg*. She also metions (p.88) Bailey’s hypothesis that it indeed is a “mule”, of Iranian origin. ● Schuessler (2007:326) translates it as "a superior type of horse of the north barbarians" mentioning that it shows similarity to the TB *yakwe*, not stating a definitive etymology. He reconstructs as the oldest form LH **kuet-dei*. This form does not seem to be reconcilable directly, with Tocharian being a recipient from the same source. ● The word positively appears in OC sources, e.g. Yi Zhou Shu 逸周書 (Sturgeon 2011). ● Baxter & Sagart don’t reconstruct the characters. ♦ The modern meaning of mule is already attested in Shuowen.

36 ⊥ 月氏 *ròuzhī* ■ None – added for clarification. ■ Alternative reading of 月氏 referring an alternative[209] usage of 月 for *ròu* “meat” (肉) instead of *yuè* “moon”. As the reading is supposed to refer to the sound value of a borrowed self-designation, it is now considered incorrect and is becoming obsolete. Dai (2006) the pronunciation variation. ● Baxter & Sagart (2014b) do not reconstruct reading ròu for 月, but it should be the same as one reconstructed for “meat, flesh” OC *k.nuk which would rule out any connection to Tocharian †ñäkät° if it were the correct reading preserving the original sound (rather than a derogative exonym or such). Either way, this word probably cannot be connected to the Tocharian languages: Baxter & Sagart (2014b) reconstruct OC *ke for 支 making the complete OC version *k.nuk-ke ♦ For discussion see 月支.

37 ⊥ 獅子 *shīzǐ* “lion” *sri-tsVʔ ← †*ṣ*VćVk PC/MC-PT? ■ (Pelliot 1931), Pulleyblank (1962) ■ TB secake TA śiśak as a source of LH **ṣi-tsiəʔ* e.g. Schuessler (2007:461) “獅子” following Pulleyblank, later (2009:23) he even uses the word to support a theory on the relative chronology of Chinese tonogenesis.[210] ● Lubotsky (1998:379) considers the words to be *Wanderworter*, refusing the Tocharian origin. ● Not mentioned by Baxter & Sagart (2014a). ● Behr (2005) lists words used in the area: Turkic, Mongolic, Tungusic *arslan*, which was also borrowed into some languages in Europe and is in form incompatible with the Chinese word; SKR *siṁha* for which some postulate a PIE root (Meillet, Dolgopolsky) with the meaning “leopard”, others (Thieme) postulate a separate Indic evolution as a taboo replacement with meaning “dangerous”, yet others (Mayrhofer) see it as a loanword from

[206] *Mothers* nearly universally at some reconstructed stage contain [m] (ModM *mǔ* 母, Zulu °*mama*, Egyptian *mwt*, etc.), *fathers* p- or t-, vocatives a, etc.
[207] Possibly reduplication?
[208] Connected to various non-IE peoples, most notably Huns, which e.g. De la Vassière (2005) disputes.
[209] Cp. Classical Chinese usage (full homographs) with ModM usage as “meat” in radical only; probably source of the confusion.
[210] Final ʔ at LH stage is supposed to stand for foreign final consonants at LH stage showing that tone C was still not present. The theory itself is not criticized here.

unknown source; Behr (ibid:5-6) proposes that Iranian *šrV* has been a source of the Tibetan word, from which other Tibeto-Burman languages would borrow it. Tocharian words' Iranian origin is refuted on the basis of incompatibility of chronology; mentioned is Pelliot (1931) who first proposed the origin of OC as being Tocharian. Also mentioned is Lüders, who rather believed the word came as part of Chinese zodiac. Behr (ibid:9-10) lists various hypotheses for the source of Tocharian words and how they do not obtain (e.g. derivation from "mane"). Following Ringe, Behr (ibid:10) considers no etymology to be convincing. This view is followed here. He reiterates (ibid:11) the incompatibility of TA TB forms in case of common etymon, showing the words were probably borrowed after the two separated, which he considers to be the time of first attestation in Chinese.[211] As many other authors do, he (ibid:12) postulates a source of OC and TB to be an unknown language. ● Adams (2013:723) reiterates and reformulates his own older hypothesis that TB *ṣecake* and *TA śiśäk* in combination with OC reconstruction *srjij-tsiʔ (Baxter 1992:323) would give PT **ṣīcäke* with some discrepancies which he would attribute to some irregular influences. He goes on to support Blažek's (2005:89-90) idea that MC word was borrowed from Kashgarian.[212] ● Since no convincing etymology has been proposed to this date for either of the words, this borrowing is not considered proven. ● Although Pulleyblank (1962:226) considers the 子 to not be a derivational suffix of ModM, in OC it was used as a sign of respect[213], speculatively, it could therefore have a similar function to *raca* in Thai *raca-sinto*. This would allow for an analysis yet again of a single-syllable word as the original one. ♦ Descendant by backformation (Pulleyblank 1962:226) is 獅.

38 ⊥ *歙侯 *xìhóu* "(Yuezhi) ruler" **CyɑpKu* ← **yɑpku* OC/MC-PT/? ■ Adams (2013) and Pulleyblank (1966) separately ■ Pulleyblank (1966:28) cites the form as 歙候 with modern reading *hsi-hou* (no tone indication)[214] making the understanding whether he actually meant to use these characters a bit complicated. The reconstructions used are either MC **hi̯əp-ɦu* or MC **śi̯əp-ɦu*[215]. The word is cited to mean *Da Yuezhi* 大月氏 rulers, one of which founded the Kushan empire. Cited is Bailey's[216] hypothesis of Iranian origin, "yam- 'to lead', with addition of a suffix -uka" while at the same time refuted on the basis of the word being unattested and contact not being proven for the timeframe. Suggested is connection to Tocharian words discussed further. ● Adams (2013:528-529) uses OC reconstruction **hjep-ɣu* and does not mention characters in TB entry for "yāpko* (n.) '± duke, count palatine, sub-king' and considers the connection with, as he himself states, extremely hypothetical PT "*yāp(ä)ku- and it would be possible to see in it an agent noun related to TchB *yapoy*/TchA *ype* 'land, country'" as a possible source while admitting the term is a *wanderwort*. ● Baxter & Sagart (2014b:118) 歙 MC ***xip* OC **q*ʰ*(r)[ə]p* "contract (v.)", Baxter & Sagart (2014b): 侯 huw *[g]ˁ(r)o "feudal lord"; this version leaves no place for connection with Tocharian both semantically and phonetically, with *C(r)əp-K(r)o not being a possible adaptation phonetically, having initial consonant an aspirated plosive and semantically – the meaning would be analysable as "contracted lord", i.e. something in the line of *protector* or *vassal*. The first OC **r* is compatible with Tocharian when taking into account that in other reconstructions (older ones) it often stands for [j] system-wide, the second one is, again, incompatible, while *o as per Schuessler (2007) often stands for foreign *u and vice versa. Since the phonetic side is problematic in the oldest attestation on PT, not OC side and the word is clearly analysable in OC rather than PT, it seems that the direction has to be reversed, even though the word is rather obscure in Chinese

[211] The timeframe, while following other sources and methods, does agree with the one presented here.
[212] I.e. Tocharian C, a language attested only through sources in other languages. This does not hold any persuasiveness, since there are too many variables.
[213] Uses in Classical Chinese range from son, child, master, part of names.
[214] Which is probably a mistake, texts seem to indicate 侯 as the second character, more details further. 候 would mean "wait upon" according to Baxter & Sagart (2014b), reconstructed OC form seems identical to the character used hereon, except for final (nominalizing?) *-s.
[215] Pulleyblank has a nasty tendency to write MC forms while speaking of OC words. This might be one of those.
[216] Not specified, although, with near certainty Bailey (1946) or Bailey (1951) Asia Major.

(歙 often stands for personal names with corresponding pronunciation and composites in the era are rather sparse, instead it is better to speak of phrasal meanings). ● If the word is taken to be a borrowing from the OC time, it is not a very probable hypothesis, the written attestation is, however, from the intermediate period (it is present in Han Shu 漢書 "與歙侯戰")

39 ㄴ 央匱 *yāngkuì* "asafoetida" ← aṅkwas ■ Misunderstood as either (Bailey 1946) or Pulleyblank (1962). ■ Baxter & Sagart (2014b:132,38) "*ʔaŋ" + "*[g]ruj-s" The word is thus reconstructed here based on the presupposed existence of the characters in OC timeframe needed to derive it from indigenous sources. Old texts with this character combination, however, do not seem to exist and the resulting form is incompatible with reconstruction of the synonym. The OC form is probably a projection. ♦ For more information see synonym 阿魏.

40 ㄴ 禺知 yúzhī ■ Indirectly Pulleyblank (1966), possibly others ■ Pulleblank (1966:19) mentions as being an earlier form of 月氏. ● For the second character, Baxter & Sagart reconstruct OC *tre, the first is not reconstructed and cannot be easily abstracted in the full version of their system, it could range from *la to *[ɢ]ʷ(r)a, if we take extremely hypothetically 隅 as being the same based on information from Zhengzhang (2003) who reconstructs both as *ŋo, it should be *ŋ(r)o; together, that would make *ŋ(r)o-tre; it is reasonable to consider the *r as adapter of foreign palatalization or retroflexion due to its MC reflexes. This would make the original in Tocharian n/ñ(ä)k.ṭ/ty° which does seem to fit ♦ Thierry (2005:4) lists all the supposed synonyms: 月氏, 月支, 禺知, 禺氏, 牛氏.

41 ㄴ 月支 *yuèzhī* "(historical) Tocharians" ■ Pulleyblank (1966)? ■ Pulleyblank (1966:17) believes that Chinese initial *ŋ reflects foreign *y- or *ø- before Tang period[217], with *yw- commonly occurring in Tocharian. The form in donor language would be "something like" *ywati. He goes on to give examples of personal names, titles and names of nations that seem phonetically similar in his reconstruction (ibid:18-22) and connects the words he sees as similar to different foreign words at different places. In the summary (ibid:36), he again, only reconstructs *ywati. ● Baxter & Sagart (2014b) reconstruct OC *[ŋ]ʷat > MC **ngjwot for the first character and for the second part: OC *ke > MC **tsye for 支, and nothing for the zhī reading of 氏, but the same can be expected. The OC versions do not seem compatible with ● The word has to be of OC timeframe, e.g. *Shan Hai Jing*[218] 山海經 mentions "月支之國", Yi Zhoushu 逸周書 (same period) uses "月氏" in a list of Northern countries in direct speech, *Mutianzi Zhuan* 穆天子傳 (again same period) mentions "禺知", in Guangzi 管子 there is both "禺氏" and "牛氏" (Sturgeon 2011). If these all indeed are referring to the same tribe, this shows a large variation in adaptation, meaning probably that none of them is ideal. ♦ Probably more common form with second character semantically more regular, reading-wise more confusing 月氏[219]; see also 禺知.

## 3.5. Ad-hoc adaptations

There is a number of words that were not (fully) integrated into lexicon, only phonetically adapted for (ad-hoc) use by Chinese speakers (readers)[220]. The variation in used characters for the same sound is therefore expected to be present more so than for other discussed words. Listed here are proposed place names and personal names deemed uncertain, those considered to be unquestionably valid are mentioned in 3.1.

[217] While confusingly using forms from his reconstruction of Middle Chinese.
[218] Warring states era.
[219] The second character stands for "tribe/family".
[220] I.e. citations of foreign words.

42 ⊥十撐利[221] *chēnglì*? “Tocharian word for heaven” *tsengli ∽ *kilyomo(nt) OC-PT ■ Schwarz (2007) ■ Schwarz (2007:20) proposes based on a supposed Xiongnu *tsengli* (stated to be attested in Chinese records as) “撐利” that the Tocharian form is ancestral to the Xiongnu term, “s největší pravděpodobností převzali místní výraz jazyka některého z kmenů Yuezhi, který je příbuzný s toch. B *klyomo*/ A *klyom* ‘posvátný’. Jeho rekonstrukce **kilyomont*/ **kilyomo*, resp. **kaelum* dobře odpovídá lat. *caelum* ‘nebe’.”[222] ● There is absolutely no reason to connect the word[223] to Indo-European sources, the easiest way it to propose a Turkic origin from Proto-Turkic **teŋri* (as Dybo does, in a way), even though the word itself is sometimes, rather unconvincingly, regarded as borrowed into Turkic from Indo-European. ♦ More widely known term with identical modern pronunciation and supposedly the same meaning in historical records is 撐犁 in the Xiongnu title 撐犁孤塗單于, “the son of heavens”[224]. ● *Hanshu* 漢書 states: 匈奴謂天爲「撐犁」[225] (Sturgeon 2011). ● Neither Baxter & Sagart (2014a; 2014b) nor Schuessler (2007) offer reconstructions of the first character, Zhengzhang (2003) reconstructs **rt$^{h}$aaŋ*[226], Baxter & Sagart (2014b) reconstruct the second character as **[r][i]j*[227] arguably supporting a Turkic origin hypothesis if connected and definitely discarding the Tocharic hypothesis. ● The word 天 *tiān* itself is sometimes connected with Turkic *tengri* e.g. Shaughnessy (1989) apud Schuessler (2007:495), although those Sinologists that explicitly address this usually reject it, e.g. Schuessler (2007:495) and connect it with “TB cognates: WT, OTib. Steŋ ‘above, upper part, that which is above’”, etc. ● The TB word *klyomo* is not considered to mean “sacred” or be connected to a meaning “heaven” by Adams (2013:119): “*klyomo* ‘noble’ (< *’having fame’)”, further (ibid:250) convincingly connects the root to PIE **k̂leumon-* “and TB “klyaus- (vt.) ‘hear, listen to’”. For comparison see Rix (2001) entry. ● Dybo (2007:82-83) discusses *chengli* written 撐黎 as coming from Turkic, which was originally disputed by Pulleyblank (1962:241)[228]. This form does not seem to appear in OC texts, from MC (Northern Song) *Taiping Yulan* 太平御覽 does include it (Sturgeon 2011). The word is from the timeframe of Turkic contacts

43 ⊥ 崑- *Kūn*° “Kil° (placename)” **kun*° <- **kil*° ■ Lin (1998), extension of Pulleyblank (1966) ■ Schwarz (2007:20) states that in place-names “*Kunshan* 昆山”(<崑山/崐山>)[229] and “*Kunlun* 昆侖 (<崑崙/崐崘>) supposedly synonymous with *Qilian* 祁連山[230]“Podle Lina (Lin Meicun 1998:482) může být stčín. Kun velmi dobře přepisem první slabiky rekonstruovaného tocharského **kilyomo*: stará čínština

[221] To my knowledge, no attestation of this combination of Chinese characters is present in texts. Schwarz (2006) has used simplified characters, the first character is presented here in traditional form, the second one has both forms identical. Either the second character is a mistake on his part, or it is a word that is simply not present in materials available to me. Since the word is obviously a phonetic rendering, I suppose the latter here.

[222] Translation: “Most certainly, they took over the local term from a language of one of the Yuezhi tribes, one that is a relative of the TB klyomo / TA klyom ‘sacred’. His reconstruction *kilyomont/*kilyomo, resp. *kaelum is a good match for Latin caelum ‘heaven’”. From the context, it is unsure, whether “him” refers to Pinault (1998) or Pulleyblank <1962>.

[223] Of which I found no attestation and none is given in the source.

[224] ‘chēnglí gūtú shànyú 撐犁孤塗單于 „Velký syn Nebes“, což je na první pohled přibližná podoba čínského císařského titulu tiānzǐ 天子 ”Syn Nebes“. Poprvé je titul shànyú zaznamenán u vůdce kmene Xiongnu Toumana. (Lattimore 1951: 450)’ (Hejdová 2012:18)

[225] Translation: ‘The Xiongnu call heaven “撐犁”’.

[226] The Baxter & Sagart (2014a) equivalent would probably be something like **t$^{hʕ}$raŋ*.

[227] With different tone slso **[r]$^{ʕ}$[i]j*.

[228] Others disputed that the word is inherited in Turkic languages believing in its loan status, for discussion see Dybo (ibid).

[229] Not to be confused with the city located in the Jiangsu 江蘇 province of Eastern China. This word and the next are actually shorter variants of the full name of the Kunlun mountains 崑崙(之)山.

[230] In the text itself, the form is 祁连. The statement is incorrect as Qilian is part of Kunlun.

totiž nerozlišovala hlásky *l* a *n*."[231] And continues to quote Cen Zhongmian[232] apud Lin (1998:479) stating "*Tianshan* 天山 v době dynastie Han pojmenováno variantně jako 祁羅漫 *Qiluoman*/ 析羅漫 *Xiluoman*, tj. nejspíš rovněž přepisem tocharského slova **kilyomont*/ **kilyomo* 'posvátný, nebeský'" ● Why these were not accepted by author of this work into placenames and discussed in 3.1 should be obvious when looking at reconstructions of respective characters: 崑 is not reconstructed by either Baxter & Sagart (2014a; 2014b) or Schuessler (2007) but the character with which it merged is reconstructed as 昆 *[k]ˤu[n] "elder brother" making the only sure phone the one that does not agree with the proposed source word; further, when put together, reconstructs for 祁羅漫 and 析羅漫 become MC gij-la-manH < OC [g]rij-rʕaj-ma[n]-s and MC sek- la-manH < OC [s]ʕek-rʕaj-ma[n]-s, resp. completely ruling out the second word and making the first one very dubious. ● Zhengzhang (2003:393) reconstructs the 昆 series as **kuun(ʔ)* with 崐, a variant of 崑 as **kuun*[233]. Of note is that in his system, there is final *-*l* so any preservation of possible labiality of the coda by the means of compensatory rounding of vowel is out of question. This version does not therefore support the view that there is any connection with the Tocharian word. ● It would seem no reconstruction would favour borrowing from PT, the MC Qilian on the other hand seems to be without any doubt. ● It should be noted that 祁羅 and 析羅 were proposed as coming from Tocharian already by Pulleyblank (1966:20), where he also stated his belief that Turking *tenrgi* comes from the same source.

[231] Translates as: "According to Lin (Lin Meicun 1998:482) may the OC Kun very well be a transcription of the first syllable of a reconstructed Tocharian *kilyomo since the Old Chinese did not differentiate the phones *l* and *n*."

[232] Work/year and page not mentioned.

[233] Looking at CText, the variant reconstructed by Zhengzhang (2003) seems to be attested earlier (*Shangshu* 尚書 of Spring and Autumn period) than the other.

# 4. Discussion

Regarding *Wanderwörter*[234], while the ultimate source of the word may be of question, where the anthropological and linguistic data seem to suggest borrowing in a certain way, I do not believe it should be a priori discarded.

Regarding *semantic fields*. In SPP, there has been a discussion on further semantic fields where Indo-Europeans may have had influence on Chinese (for a rather old, but quite typical and somewhat influential example see Chang 1988, where he goes as far as to consider OC a mixed language). One is mythology, where there are parallels in evolution of deities. For these parallels, the correlation is seen by the author of this study as hypothetical at best and is not discussed here since they rarely if ever are postulated to have influence on word forms.

Others spoke about cultural contacts without specifying that language exchange occurred. Schafer 1963:85 speaks of Tocharians as intermediaries between East and West, although he seems to confuse the Iranian ones with the ones of concern here at times. Iranian Tocharians supposedly had an influence on Chinese music (Kishibe 1952:76-86 apud Schafer 1963), quite controversially even on Japanese (*Toragaku* by corruption from Bactrian endonym *Tūkhara* + Japanese *gaku* in the seventh century CE, mentioned by Ariyoshi (1940: 233) apud Waterhouse (1991:75)). Hitch (1993) speaks of Manichean[235] hymns supposedly composed in Tocharian B, (ibid:96), he refers to Schafer 1963:52 stating that Kucheans[236] had a large influence on Tang music. If there is a semantic field where borrowing is possible due to the situational context, it would be sharing ways of artistic expression.

A topic completely left out is purely phatic communication (derogatives and euphemisms). The problem lies in sparse attestations of taboo-related concepts, much less forbidden words, in early sources of any written language. A thorough study of spoken Chinese of this day would have to be done to better reconstruct lexemes that may come from foreign sources. Since the desired outcome of using a coarse language at someone is to calm oneself by angering the other person, it is easily conceivable that one would learn a foreign swearword in order to better insult them. Some such words could easily survive for long periods of time for their iconic value.

When the research in Chinese historical linguistics shifts from reconstructing words written in characters and from analysing langue to features preserved only in parole, evidence of many loanwords may yet arise. Some attempts have already been made – as seen in much criticised Baxter & Sagart (2014a) inclusion of irregularities. Still more has to be done to shift from written sources as primary input to secondary, to one framing the analysis rather than forming it.

Notes regarding calquing and borrowing in general: while there certainly is a number of calques when there is a large-scale contact, these are very hard to identify and most of the times speculative at best. Since this work has in effect tried to prove that the borrowing in the case of Chinese-Tocharian contact has been for the most part in a single direction from the Chinese with only very few, if any, items going the other way at different times from various semantic fields. This should in itself disprove the validity of an attempt to locate any concrete calques, since where there are no identifiable cultural items to borrow, there is very little to base your claims on. A rough semantic similarity of components in compounds with the same meaning across languages is for the endocentric ones basically a must.

Considering how problematic Tocharian loanwords have proven to be, I suspect similar problems for borrowing from other Indo-European branches, even though they are far better understood. This study would seem to indicate that when dealing with contact linguistics of pre- and proto-forms, non-critical citing of conclusions of others is something to be wary of, since even a great scholar can base his

[234] Internationalisms of unknown provenience.
[235] Followers of the prophet Mani.
[236] The usage is confusing as he speaks in the text both about Kuchean Tocharians, and Iranian and Turkic speaking Kucheans.

assumptions only on knowledge at the time of the writing, which in this field can become obsolete in some respects with every new finding.

A note should be made on making uninformed exact comparisons using semantics as reconstructed by Baxter & Sagart (2014b), who only approximate the OC meaning. As careful as the author could be, some mistakes are unavoidable with rarely occurring words, in those cases, other dictionaries have been consulted, those are cited in-place, preferring OC sources rather than modern interpretations, still, even *Shuowen* is famous for being incorrect at times and should not be understood as explanatory dictionary which it is not.

The Tocharian *self-designation* is a matter of long debate in Indo-European studies. At present, we are not aware of any above the local, city level. Even though it should be relatively easily identifiable context-wise in Tocharian manuscripts themselves, I believe that in Chinese sources, one will be found once we step away from the idea that it has to be directly analysable in their own language[237].

Many words are incompatible when in isolation, if we were to postulate an intervocalic lenition causing voicing of plosives in Tocharian, it would be possible to phonetically equate the words with their Chinese counterparts, since a lexical word is rarely uttered in isolation[238], the only thinkable situational context would be explanation and repeating in order to teach the word. In languages where voicing contrast is not present as such, intervocalic sonorisation is quite common, the Tocharian script, however, does not seem to indicate such possibility.[239] At least Kim 1998:159 seems to suggest that at some stage of development from PIE to PT at least one sound could have been voiced intervocally, with large scepticism.

Most of the words listed here have been proposed as coming from unspecified IE language before the Tocharian origin was suggested. Where it was necessary and possible, I have tried to comment on that, however, the format of a diploma thesis restricts the topic and presentation, limiting the time and size allotted to the work. Therefore, some words discussed may still have an open possibility of being borrowed into Chinese from external source, a far more extensive research would be needed to explain them considering how large this text has become while only refuting invalid and outdated theories for one possible source. Words that are postulated as coming from an IE language without being mentioned or directly connected with a word mentioned as being of Tocharian origin specifically, have been left out for obvious reasons of size constraints.

Regarding *methodology and presentation*. In explaining method, the obvious listing of common methods in Chinese historical linguistics that has been taken over as a supplement for argumentation has been left out. Most of them are common to all comparative linguistics branches with some restrictions given the nature of the script. One – using character etymology has been commented upon in many places, one thing should be, however, repeated – it is highly problematic to use a graphic component of the word in arguments concerning the time *before* the word has been attested in writing. A more detailed study of variation in earliest forms where those are going back to a time where iconicity still played a big role may show additional information, still. Arguably, some characters may contain

[237] E.g. Czech self-designation does not have a conclusive etymology that is agreed upon, even though it is quite obviously not an exonym.

[238] Usually, a phonetic word has a content word with grammatical words "attached" to it. Unless those are encliticons, they are sure to cause some sort of coarticulation. Even for ModM where there is supposedly no morphology, in careful speech where coarticulation is supposed not to be prominent exists the *Erhua* 儿化 phenomenon of Northern dialects (suffixed *-r* becoming a coda and blending into the vowel).

[239] Very speculatively, among other reasons, various archaic scripts' scribes were aware of and able to indicate phonetic, rather than phonemic, differences and at least sometimes they did, I do not believe that to be true for TB and, more importantly, TA, traditionally thought of as a literary language where native speaker's intuition doesn't interfere, cp. Avestan script. The script does have the capability to express these differences. While in itself not an argument, with a large number of borrowings from languages that do have voicing contrast, one would expect them to indicate it in their own language also given the span of centuries.

elements incompatible with native scribe's intuition, which could point to a multimodal communication (drawing as an aid where communication in speech is complicated or impossible).

The dictionary/wordlist part has a specific notation originally devised to be usable in the same way as Baxter & Sagart (2014b) (uncomplicated computer processing) in case of finding a large number of cognates and was graphically inspired to an extent by Adams (2013). Since presentation could be thought of as part of method, a question could arise whether such a small number of entries (and most of them refuted) needs its own, relatively complicated, presentation style. As the work was in need of a commented explanatory dictionary aimed at non-specialists which includes yet clearly differentiates both correctly and incorrectly identified cognates, without any standard to use, this style was created. Problematic is citing in accordance with APA which does not allow for the most natural way of referencing dictionary entries by their respective numbers only.

The words discussed here were the ones deemed to be not religion-related. In Buddhist terminology, there is probably some influence (either lexical or in form of interference), considering Kuchean monks did serve as translators into Chinese.[240] Some words (*lion*) were not discussed in full considering the debate has been done by specialists on both fields for the last 100 years without a fruitful end, with no information to add, what could be considered a reference to an authoritative literature has been given with a short summary. Some more words not found by the method may still be of interest due to their form-meaning correspondence, yet the author is very sceptical as to their connection, considering the outcomes of this work's findings.

Some words from the compounds section would also fall under the ad-hoc adaptations, e.g. 歙侯, my justification for their placement is my analysis of their meaning. The border between a citation of a foreign word and its usage is a complicated matter which would need more attention from text linguists, my justification for such division even without a proper preceding study is that adaptation mechanisms should vary by register with ad-hoc adaptations always being the most turbulent (cp. instability of phonetic features of most learners' interlanguage[241]).

A note on used dictionaries: An argument may arise against the near-exclusion of S. Starostin (1989). Baxter & Sagart (2014a) and Schuessler (2007) differ profoundly in their approach, the former try to maximize information at the price of losing clarity while the latter tries to minimize the variation in presented information. Together, they weight out most of their shortcomings, leaving little need for a third one.

As every comparative linguist understands, reconstruction relies on a knowledge of material and previous works, on proven and/or provable and/or consensual method, and instinct/judgement. While the work presented should cause as little disagreement in the reader as possible, some introspection was needed and input from historical sociolinguistics and pragmatics would be in order for better explanations at times.

An argument against the used division of information between Introduction and Method could arise, possibly a valid one. The explanation of why the division was done this way lies in the author's abhorrence of redundancy, some comments on interpretation have to be made and yet explaining the OC reconstructions to those who chose to read from the Methodology chapter, believing their own understanding of the topic to be sufficient, would be indeed fruitless. Critique of chaotic approach to explaining methods have been previously made by various researchers against all of the used sources for Chinese (as discussed in 1.4).

*Final remark*: As can be seen from the general style, large number of abbreviations and referrals to other literature, the topic is so complex, it would require at least twice the size of this text to even begin to be

[240] See e.g. Hansen (2012:65-76) for one account.
[241] E.g. English written <th> can be for a Czech learner: [t, f, s, d, z, v], with or without context substitution, or, finally, the correct ϑ and δ.

complete and fully commented upon at the present state of knowledge. Even with the new technologies at my disposal most of those who originally worked on the topic didn't have and the time allotted by the authorities, this is little more than a sketch. A more detailed evaluation of loanword adaptation principles and language contact situations as applied to reconstructed languages with indirect cultural contact evidence would be in order. As new discoveries in all the fields needed to be taken into account present themselves on a nearly daily basis, the author expects this work to be outdated to an extent relatively soon.

# 5. Conclusion

This work has been proposed based on a knowledge the author believed true and self-evident.

Presented here was an attempt to summarize, evaluate and expand the findings in literature pertaining to the possible borrowing from the Tocharian languages into Chinese. While originally meant as a basis for further research, the evaluation had a devastating effect on the idea of an attested direct mutual influence between the two cultures. From the previously postulated 43[242] possible cognates, 2 have been found to be consistent with the current knowledge in regards to (Proto-)Tocharian as a source of a word that is not a personal/place name, of which only one is probable.

While possibly depressing to some[243], since the language evidence does not go in line with the evidence of a cultural influence, the work serves as a proof that lexeme exchange is not a necessary part of a large-scale cultural exchange, even where linguistic contact is inevitable, showing the historical linguistics, much more historical contact linguistics, the need to take into account socio-pragmatic and anthropological (not only archaeological) input; and dispels some ideas that may very well be widely held by those less initiated into this very specific topic.

The original author's vision of proving his own approach to a corpus based study of languages with partially prepared data proved to be ill-advised. The reason was that there needs to be a theoretical bias to work with, positive evidence to input. Where little or no positive evidence is present, the approach cannot bear additional results.

An important finding from the comparison of cognates has been postulated – intersonoric voicing in , possibly already in Common Tocharian or even Proto-Tocharian stages.

[242] The number is relatively arbitrary – the value includes different readings and suggested words not extensively discussed.
[243] Including the author.

# 6. Bibliography


"CC-CEDICT" Accessed 10.4.2017. Available at: https://www.mdbg.net/chinese/dictionary?page=cedict

"thai-language.com - มนเทียร" Available at: http://www.thai-language.com/id/145634

Adams, D. (1999). *A dictionary of Tocharian B.* Leiden studies in Indo-European 10. Amsterdam: Rodopi.

Adams, D. Q. (2011). Three additions to the Tocharian B aviary. *Tocharian and Indo-European Study*, *12*, 33-43.

Adams, D. Q. (2013). *A Dictionary of Tocharian B.: Revised and Greatly Enlarged*. Amsterdam: Rodopi.

Ariyoshi, S. (Ed.) (1940). *Zoho Rikkokushi* (12 Vols.). [Supplement to the six national histories]. Asahi Shinbunsha.

Bauer, R. S. (1994). Sino-Tibetan *kolo "Wheel". Mair, V.H. (Ed.) *Sino-platonic papers*. Available at: http://sino-platonic.org/complete/spp047_sino-tibetan_wheel.pdf

Baxter, W. H., & Sagart, L. (2014a). *Old Chinese: A new reconstruction*. Oxford University Press.

Baxter, W. H., & Sagart, L. (2014b). "*Baxter-Sagart Old Chinese reconstruction (version 1.1, 20 September 2014)*" Available at http://ocbaxtersagart.lsait.lsa.umich.edu/

Behr, W. (2005). Hinc sunt leones – two ancient Eurasian migratory terms in Chinese revisited (2). *International Journal of Central Asian Studies* (Vol. 10). 1-24. Seoul: Institute of Asian culture and development.

Blažek, V. (1997) Is it possible to restore Tocharian A ku//// "nave, hub"? *Tocharian and Indo-European Studies 7*, 234-235. (Reprinted Schwarz, M. (Ed.) (2011). *Blažek, Václav: Tocharian Studies. Works 1*, 30-31. Brno: Masarykova Univerzita).

Blazek, V. (2005). Hic erant leones: Indo-European" lion" et alii. *The Journal of Indo-European Studies*, *33*(1), 63-102.

Blažek, V., & Schwarz, M. (2008). Tocharians who they were, where they came from and where they lived. *Lingua Posnaniensis*, (50), 47-74. (Reprinted Schwarz, M. (Ed.) (2011). *Blažek, Václav: Tocharian Studies. Works 1*, 113-147. Brno: Masarykova Univerzita).

Carling, G. (2005). Proto-Tocharian, Common Tocharian, and Tocharian-on the value of linguistic connections in a reconstructed language. In *Proceedings of the Sixteenth Annual UCLA Indo-European Conference: Los Angeles, November 5-6, 2004* (pp. 47-71). Journal of Indo-European Studies, Monograph 52.

Chang, Ts. T. (1988). Indo-European Vocabulary in Old Chinese. Mair, V.H. (Ed.) *Sino-platonic papers*. Available at: http://sino-platonic.org/complete/spp007_old_chinese.pdf

Ching, Ch. J. (2011). Silk in ancient Kucha: on the Toch. B word kaum* found in documents of the Tang period. *Tocharian and Indo-European Studies*, *12*, 63-82.

Ching, Ch. J., & Ogihara, H. (2010). A Tocharian B Sale Contract on a Wooden Tablet. *Journal of Inner Asian Art and Archaeology*, *5*, 101-127.

Collinge, N. E. (1985). *The laws of Indo-european* (Vol. 35). John Benjamins Publishing.

Dai, X. L. (2006). Yuezhi (Yuèzhī) hu? Yuezhi (Ròuzhī) hu? – Qianlunyuyuande weiguan bianhua. *Hunan Daxue xuebao* (Sheihui kexue ban), 20(6), 101-108. [Rouzhi or Yuezhi: On the Micro-changes of the Language. *Journal of Hunan University (Social Sciences)*] Hunan: Hunan University.

De la Vaissière, É. (2005). Huns et Xiongnu. *Central Asiatic Journal*, *49*(1), 3-26.

Dybo, A. V. (2007). *Lingvisticheskiye kontakty rannikh tyurkov. Leksicheskiy fond. Pratuyurkskiy pyeriod*. Moscow: Vostochnaya lityeratura.

Elšík, V. (2009). Loanwords in Selice Romani, an Indo-Aryan language of Slovakia. *Loanwords in the world's languages. A comparative handbook*, 260-303.

Field, F. W. (2002). *Linguistic borrowing in bilingual contexts* (Vol. 62). John Benjamins Publishing.

Fortson, B.W. (2010). *Indo-European Language and Culture: An Introduction*. John Wiley & Sons.

Frisch, S. A., Pierrehumbert, J. B., & Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural Language & Linguistic Theory*, *22*(1), 179-228.

Gamkrelidze, T. V., & Ivanov, V. V. (1984). Indoevropeiskiy yazyk i indoevropeytsy. *Tbilisi: Tbilisi State University Publishing House*.

Gamkrelidze, T. V., & Ivanov, V. V. (1995). *Indo-European and the Indo-Europeans: A Reconstruction and Historical Analysis of a Proto-Language and Proto-Culture. Part I: The Text. Part II: Bibliography, Indexes* (Vol. 80). Walter de Gruyter.

Gippert, J. & Martínez, J., Korn, A. (2016) "TITUS" Available online at: http://titus.fkidg1.uni-frankfurt.de/framee.htm?/texte/texte2.htm#toch

Greenberg, J. H., Ferguson, C. A., & Moravcsik, E. A. (Eds.). (1978). *Universals of human language: phonology* (Vol. 2). Stanford University Press.

Hall, D., & Klein, D. (2010, July). Finding cognate groups using phylogenies. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 1030-1039). Association for Computational Linguistics.

Hansen, V. (2012). *The silk road: a new history*. Oxford University Press.

Harper, D. (2017). "Online Etymology Dictionary" Available at: http://www.etymonline.com/index.php?term=canary

Haspelmath, M. (2010). Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, *86*(3), 663-687.

Hejdová, J. (2012). Vztahy Číny a tureckého kaganátu. (Bachelor thesis) Praha: Univerzita Karlova v Praze. Available at: https://is.cuni.cz/webapps/zzp/detail/86244/

Hitch, D. (1993). The Kuchean Hymn in Manichean Script. *Tocharian and Indo-European Studies*, *6*, 95-132.

Juge, M. L. (1999, February). On the rise of suppletion in verbal paradigms. In *Berkeley Linguistics Society* (Vol. 25, pp. 183-94).

Karlgren, B. (1957). Grammata Serica Recensa. Stockholm: The Museum of Far Eastern Antiquities. *Bulletin*, *29*.

Kenstowicz, M., & Suchato, A. (2006). Issues in loanword adaptation: A case study from Thai. *Lingua*, *116*(7), 921-949.

Kim, R. (1999). The development of labiovelars in Tocharian: a closer look.

Kim, R. (2006). Tocharian. Keith Brown (Ed.) *Encyclopedia of Language & Linguistics, Second Edition*, (Vol. 12), 725-727. Oxford: Elsevier.

Kishibe, Sh. (1952). Seiikigaku touryuu ni okeru kogaku raichou no igi. Rekishi to bunka: Rekishigaku kenkyuu houkoku, 67-90. (Vol 1). Tokyo: Toukyou daigaku kyouikugakubu jinbun kagakka kiyou.

Lattimore, O. (1951). *Inner Asian Frontiers of China.* New York: American Geographical Society.

Li, F. K. (1971). Shangguyin yanjiu. (Studies on Archaic Chinese phonology). *Tsing Hua Journal of Chinese Studies*, *9*(1-2), 1-61.

Lin, M. C. (1998). Qilian and Kunlun–the earliest Tokharian loan-words in Ancient Chinese. *The Bronze Age and Early Iron Age Peoples of Eastern Central Asia*, *1*, 476-482.

Lubotsky, A. (1998). Tocharian loan words in Old Chinese: chariots, chariot gear, and town building. Available at: https://openaccess.leidenuniv.nl/bitstream/handle/1887/2683/299_040.pdf

Lubotsky, A. M. (2003). Turkic and Chinese loan words in Tocharian. Available at: https://openaccess.leidenuniv.nl/bitstream/handle/1887/16336/299_058.pdf?sequence=2

Mahendra, P., & Bisht, S. (2012). Ferula asafoetida: Traditional uses and pharmacological activity. *Pharmacognosy reviews*, *6*(12), 141. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3459456/

Mair, V. H. (1990). Old Sinitic* Myag, Old Persian Maguš, and English “Magician”. *Early China*, *15*, 27-47.

Mallory, J.P. & Adams, D. Q. (1997). *Encyclopedia of Indo-European Culture*. Taylor & Francis.

Malzahn, M. (2017) “About the project” https://www.univie.ac.at/tocharian/?About%20the%20project

Malzahn, M. (2017) “CEToM” https://www.univie.ac.at/tocharian/

Miao, R. (2005). *Loanword adaptation in Mandarin Chinese: Perceptual, phonological and sociolinguistic factors* (Doctoral dissertation, Stony Brook University).

Norman, J. (1988). *Chinese*. Cambridge University Press.

Peiros, I. & Starostin, S. (1996). *A comparative vocabulary of five Sino-Tibetan languages*. Melbourne: The University of Melbourne, Department of Linguistics and Applied Linguistics.

Pejčochová, M. & Zádrapa, L. (2009). *Čínské písmo*. Academia.

Peyrot, M. (2008). *Variation and change in Tocharian B* (Vol. 15). Rodopi.

Peyrot, M. (2015). "TOCHARIAN LANGUAGE," *Encyclopædia Iranica*, online edition. Accessed on 27 July 2015. Available at: http://www.iranicaonline.org/articles/tocharian-language

Pinault, G. J. (1998). Tocharian languages and pre-Buddhist culture. *The Bronze Age and Early Iron Age Peoples of Eastern Central Asia*, *2*, 358-371.

Plank, F. et al. (2009) “The Universals Archive”. Available at: https://typo.uni-konstanz.de/archive/intro/index.php

Polivanov, Y. (1916) Indoevropeyskoye *medhu~ obshchekitayskoye mit. Zapiski Vostnochnogo otdyeleniya Impyeratorskogo Russkogo Arkheologicheskogo obshchestva. (Vol 23, Pt. 3-4). Available at: http://elibrary.orenlib.ru/index.php?dn=down&to=open&id=1696

Ringe, D. A. (1996). *On the Chronology of Sound Changes in Tocharian: From Proto-Indo-European to Proto-Tocharian* (Vol. 1). Eisenbrauns.

Ringe, D. A. (1996). *On the Chronology of Sound Changes in Tocharian: From Proto-Indo-European to Proto-Tocharian* (Vol. 1). Eisenbrauns.

Rix, H. et al. (2001). *Lexikon der indogermanischen Verben: Die Wurzeln und ihre Primärstammbildungen* (2nd ed.). L. Reichert.

Sagart, L. (1999). *The roots of old Chinese* (Vol. 184). John Benjamins Publishing.

Sagart, L. (1999b). The origin of Chinese tones. In *Proceedings of the Symposium/Cross-Linguistic Studies of Tonal Phenomena/Tonogenesis, Typology and Related Topics.* (pp. 91-104). Institute for the Study of Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies. Available at: https://hal.archives-ouvertes.fr/docs/00/09/69/04/PDF/TOKYO_tone_published.pdf

Schafer, E. H. (1963). *The golden peaches of Samarkand: a study of T'ang exotics* (Vol. 742). Univ of California Press.

Schuessler, A. (2007). *ABC etymological dictionary of Old Chinese*. University of Hawaii Press.

Schuessler, A. (2009). *Minimal Old Chinese and later Han Chinese: a companion to Grammata serica recensa*. University of Hawaii Press.

Schuessler, A. (2015). New Old Chinese. *Diachronica*, *32*(4), 571-598.

Schwarz, M. (2009). K reáliím "tocharského" období - některé příspěvky z "Dunhuang and Turfan Studies (Bachelor thesis). Olomouc: Palacký University Olomouc. Available at: http://theses.cz/id/0t10gj/

Schwarz, M., & Blažek, V. (2015). Jména nádob v tocharských jazycích. *Linguistica Brunensia*, *63*(1).

Sears, R. (2013). "Chinese etymology". Available at: http://chineseetymology.org/CharacterEtymology.aspx?submitButton1=Etymology&characterInput=%E4%B9%98

Shaughnessy, E. (1989). Western Cultural Innovations in China, 1200 BC. Mair, V.H. (Ed.) *Sino-platonic papers*. Available at: http://sino-platonic.org/complete/spp011_shang_china.pdf

SIL International (2017a) "Chinese | Ethnologue" (*Ethnologue, 20th edition).* Available at: https://www.ethnologue.com/subgroups/chinese

SIL International (2017b) "Chinese | Ethnologue" (*Ethnologue, 20th edition).* Available at: https://www.ethnologue.com/language/kac/20

Slaměníková, T. (2013). *Ideogramy v moderní čínštině*. Olomouc: Univerzita Palackého.

Starostin, G. (2009) Review of Schuessler. *Journal of Language Relationship*, *1*, 155-162.

Starostin, S. A. (1989). *Rekonstruktsiya drevnekitayskoy fonologicheskoy sistemy*. Moscow: Nauka, Glav. red. vostochnoy lit-ry.

Starostin, S. A. (2005). "Chinese Characters". Available at: http://starling.rinet.ru/cgi-bin/query.cgi?root=config&morpho=0&basename=\data\china\bigchina

Starostin, S. A. (2006). "Sino-Tibetan Etymology". Available at: http://starling.rinet.ru/cgi-bin/query.cgi?root=config&morpho=0&basename=\data\sintib\stibet

Sturgeon, D. (Ed.) (2011). "Chinese Text Project". Accessed 10.4.2017 Available at: http://ctext.org

Sun, J. (1999). *Reduplication in old Chinese* (Doctoral dissertation). University of British Columbia.

Thierry, F. (2005). Yuezhi et Kouchans. Pièges et dangers des sources chinoises. *Afghanistan: Ancien carrefour entre l'est et l'ouest*, 421-539.

Thurgood, G. (2002). Vietnamese and tonogenesis: Revising the model and the analysis. *Diachronica*, *19*(2), 333-363.

Tremblay, X. (2007). The spread of buddhism in Serindia—Buddhism among iranians, tocharians and turks before the 13th century. In *The spread of Buddhism* (pp. 75-130). Brill.

Unicode Roadmap Committee & Unicode Consortium (2017). "Roadmap to the SMP. *Authored by Michael Everson, Rick McGowan, and Ken Whistler. Revision* 9.0.1." Available at: http://www.unicode.org/roadmaps/smp/smp-9-0-1.html

Van Hout, R., & Muysken, P. (1994). Modeling lexical borrowability. *Language variation and change*, *6*(01), 39-62.

Vendelin, I., & Peperkamp, S. (2004). Evidence for phonetic adaptation of loanwords: an experimental study. *Actes des Journées d'Etudes Linguistiques*, *2004*, 129-131.

von der Gabelentz, G. (1881). *Chinesische Grammatik mit Ausschluss des niederen Stiles und der heutigen Umgangssprache*. Leipzig: T. O. Weigel. https://archive.org/details/chinesischegram00gabegoog

Waterhouse, D. (1991). Where Did Toragaku Come From?. *Musica asiatica*, *6*, 73-94.

Wiebusch, T. (2009). Mandarin Chinese vocabulary. *Haspelmath, Martin & Tadmor, Uri (eds.) World Loanword Database.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at: http://wold.clld.org/vocabulary/22

Wilson, L. (2015). Preliminary Proposal to Encode the Tocharian Script. Available at: http://www.unicode.org/L2/L2015/15023-tocharian.pdf

Winternitz, M., Monier-Williams, M., Leumann, E., & Cappeller, C. (1899). A Sanskṛit-English Dictionary Etymologically and philologically arranged with special reference to Cognate Indo-European Languages. New Edition, greatly enlarged and improved with the Collaboration. Oxford: The Clarendon Press. Available at: http://www.sanskrit-lexicon.uni-koeln.de/scans/MWScan/2014/web/index.php

Wodtko, D. S., Irslinger, B. S., & Schneider, C. (2008). *Nomina im Indogermanischen Lexikon*. Universitätsverlag Winter.

Zádrapa, L. (2011). *Word-class Flexibility in Classical Chinese: Verbal and Adverbial Uses of Nouns* (Vol. 2). Brill.

Zhengzhang, F. Sh. (2003). *Shangu yinxi: Old Chinese Phonology*. Shanghai: Shanghai Educational Pubishing House.

## Adopted graphic material

None.

## A note on used fonts

All fonts used are under various free-and-open-source licenses, this document may therefore be reprinted and copied limited only by regulations of Charles University at time of reading.

Full list[244]: I.Ming (Traditional Chinese), FandolSong (Simplified Chinese), Koku Mincho (Japanese), Khmer OS Freehand (Khmer), Noto Serif Lao (Lao), Baekmuk Batang (Korean), Noto Sans Mongolian (Manchu), FreeSerif + FreeSans (scripts not supported by other fonts), Arimo, Tinos, Cormorant Serif, Inconsolata.

Due to the required PDF/A2-A compliance, some characters had to be deleted from final manuscript since they are mapped to the Unicode PUA (Private Use Area). The alternative of replacing all/only certain fonts with outlines was deemed an unsatisfactory solution. Please see attachment "omissions" for the list.

## List of attachments

prilohy.zip:

- Scripts (complete set)
- Omissions required by technical standard

[244] Due to an error in word processor, some other fonts' empty subsets may be embedded.