Giancarlo Tomezzoli, Joseph Kreutz

THE LINGUISTIC POSITION OF THE TOCHARIAN

Povzetek

JEZIKOVNI POLOŽAJ TOHARCEV

Študije toharščine so se pričele koncem 19. stoletja, ko so bila arheološka raziskovanja Kitajskega Turkestana ali Xinjianga predstavljena na 12. mednarodnem kongresu orientalistov v Rimu (1899) in so jih postopno nadaljevali Rusi (I – V, 1899 – 1915), Britanci (I – III, 1900 – 1916), Japonci (I – III, 1902 – 1909), Nemci (I – IV, 1902 – 1914) in Francozi (I, 1906 – 1909). Toharski dokumenti so shranjeni v navedenih državah in iz njih sta spoznani dve različici: vzhodna toharščina ali toharščina A (TocA) in zahoda toharščina ali toharščina B (TocB). TocA in TocB sta bila spoznana kot Indo-Europska jezika tipa Kentum.

TocA ali tudi turfanščina, se je govorila v območju Turfana in najdbe so pretežno verske narave. TocB ali tudi Kučanščina se je govorila do 9. stoletja po Kr. v območju Kuha. Napisi v TocB so iz verskega, trgovskega in vsakodnevnega življenja. Napisi TocA in TocB so omogočili hipotezo, da so v tisočletju pr. Kr. v Kitajskem Turkestanu govorili Proto-Toharščino. Kasnejša raziskovanje dobro ohranjenih mumij (1800 BC) v Tarimskem bazenu, ki imajo kavkaške značilnosti, pa so dale hipotezo, da pripadajo Toharcem.

Vektorska analiza v glasovnostatističnem prostoru, razvita v tem prispevku, kaže, da imata TocA in TocB ureditev in razvoj verjetno neodvisno od drugih jezikov, opazne pa so skupne značilnosti med TocA, TocB in slovenščino, venetščino, kakor verjetno tudi z drugimi, tukaj ne upoštevanimi slovanskimi jeziki, pa tudi z oskijščino in latinščino. Za potrditev te hipoteze in boljši vektorski test v bodočnosti predvidevamo primerjavo več slovanskih, keltskih in uralsko-altajskih jezikov.

Introduction

The period between the end of the 19th and the beginning of the 20th century was particularly fruitful for archaeological discoveries in the Chinese Turkestan or Xinjiang [1] (cf. Fig. 1). It is at the beginning of this period the first discoveries of ancient cultures remains, of the underground canal system of the Turfan oasis, of the Bower's manuscript – the oldest known Sanskrit manuscript, of desiccated bodies and of the first Tocharian manuscripts. The archaeological discoveries in the Xinjiang or Tarim Basin were presented at the 12th International Congress of Orientalists in Rome on 1899 through the interventions of Hoernele, Klementz, Radloff and Sénart. This Congress can be considered the starting point of systematic cultural studies and further explorations of the Xinjiang.

Five Russian expeditions (I – V) took place in the period 1899 – 1915 under the direction respectively of Klementz (I), Kozlov (II), the brothers Berezovsky (III) and Oldenburg (IV, V) in several sites including: Kara-shahr, Turfan, Kucha. The discovered Tocharian manuscripts, only in part published, are preserved at the Oriental Institute of Saint Petersburg and at the Russian Academy of Arts and Science. Three British expeditions (I – III) took place in the period 1900 – 1916 under the direction of Stein in several sites including: Kara-shahr, Loulan, Miran, Niyan, Turfan, Yar-khoto. The manuscripts in different ancient languages, the artifacts discovered, the drawings and the photographs of the expeditions are preserved at the British Library in London, at the National Museum of India in Delhi and at the Hungarian Academy of Sciences in Budapest.

Three Japanese expeditions (I – III) took place in the period 1902 – 1909 under the direction respectively of Watanabe and Hori (I), Tachibana and Nomura (II), Tachibana, Yoshikawa, Watanabe (III) in various sites including: Dunhuang, Kizil, Kucha, Kumtura, Turfan. Only few Tocharian manuscripts were discovered. Four German expeditions (I – IV) took place in the period 1902 – 1914 under the direction respectively of Grünwedel, Huth, Bartus (I), Le Coq, Bartus (II), Grünwedel, Le Coq, Pohrt, Bartus (III), Le Coq, Bartus (IV) in various sites including: Bezelik, Khocho, Kucha, Sengim, Turfan, Yar-khoto. The Tocharian manuscripts discovered, almost entirely published, is preserved at the Staatsbibliothek zu Berlin. One French expedition (I) took place in the period 1906 – 1909 under the direction of Pelliot, Vaillant, Nouette (I) in various sites including: Dunhuang, Kara-shahr, Kucha, Subashi, Turfan. The discovered Tocharian manuscripts, only in part published, are preserved at the Bibliothéque National de France.

Where to find published Tocharian manuscripts in internet has been indicated by Malzahn [2–4].

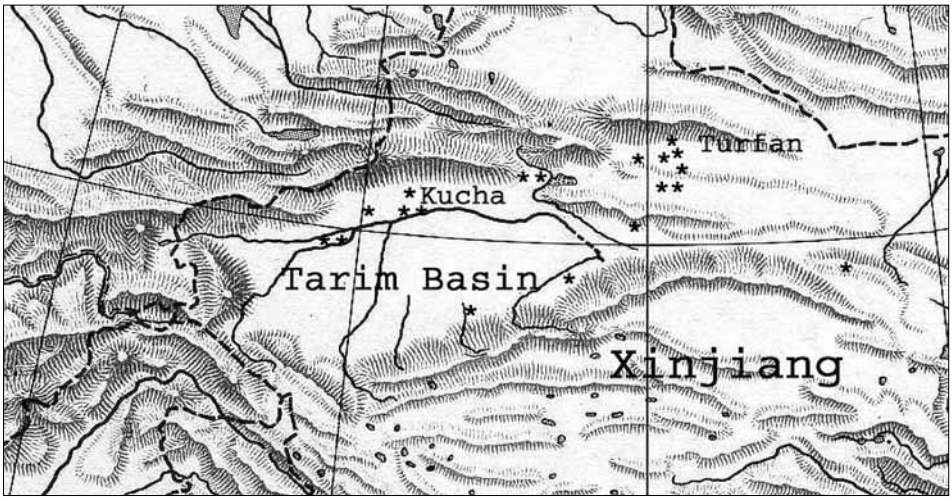

Fig. 1: Tarim Basin – excavation sites [1]

The Tocharian Question

The Tocharian question concerns mainly the epoch in which the Tocharian's arrived in the Xinjiang and from where they came. But the reply is largely unknown [5], pp. 88-98. The Tocharian surely arrived in the Xinjiang before the beginning of our era because of the presence of Tocharian lexemes in Indian Prakrits texts preceding our era. From the manuscripts collected is now clear that there existed at least two Tocharian languages: the Tocharian A (TocA) named also Turfanian, Arsi or East Tocharian and the Tocharian B (TocB) named also Kuchean or West Tocharian, both belonging to the Indo-European family and both originated from a Proto-Tocharian language (P-Toc or Common Tocharian). TocA became soon in our era a dead language used in religious manuscripts while TocB was a spoken language at least up to the 8th cen. AD The presence in the Tocharian manuscripts of words and linguistic elements belonging to Finno-Ugric languages and possible linguistic contacts of the Tocharian with East Iranian languages suggests that the Tocharians moved from their original homeland probably in the Pontic steppe, first to North and then to East.

The discovery in the Tarim Basin at the beginning of the 20th century of desiccated bodies or mummies and their study provided new information about the Tocharian history. The recent excavations of ancient cemeteries provided further rests of mummies and artifacts which give evidences of connections among the ancient Xinjiang cultures, the Pit-Grave cultures and the Afanasievo and Andronovo cultures of Central Asia, of the presence in the Xinjiang of Caucasoid populations in the period 1800 BC – 300 AD and of the arrival in the late Bronze Age and the early Iron Age of a second population of East Mediterranean type similar to the Saka of Pamir. The first mummies of Caucasoid somatic type, dating 1800 BC were discovered at Qäwrighul. The excavation of the cemetery of Yanbulaq provided other 29 mummies dating 1100 – 500 BC: 21 of Mongoloid somatic type and 8 of Caucasoid somatic type like those of Qäwrighul. The common features of the Caucasoid mummies are: elongated bodies, angular faces, recessed eyes, blond, red to deep brown hairs. Genetic studies ascertained that the Caucasoid mummies have a Haplogroup R1a (Y DNA) and an mtDNA haplotype characteristic of Western Eurasian populations. Other genetic studies on mummies of Xiaohe showed Y-DNA and mtDNA markers characteristic of an admixed population.

This scenario could be explained either by the hypothesis of Jandáček and Perdih [6-7] of "the expansion of nomadic stockbreeder proto-Slav from near East and (South-) Eastern Europe into Central Asia and about 4000 BC reaching as far as China mixing with indigenous peoples", or by the hypothesis of Mair [8-10] that the earliest mummies in the Tarim Basin correspond to Caucasoid or Europoid individuals arriving in the Tarim Basim around 3000 BC through the Pamir Mountains that mixed with Asian migrants arrived around 1000 BC while the arrival of the Uyghurs has to be placed around the year 842 AD after the collapse of the Orkon Uighur Kingdom in Mongolia under the push of the Kyrgyzs.

Up to now, Tocharian manuscripts and fragments of manuscripts have never been found in direct relation with the Caucasian mummies, however, the identical geographical

location and the common non-Chinese origin suggest that said mummies are related with the Tocharians. Because no Tocharian document have yet been found later than the 8th cen. AD it is admitted [11] that the Tocharian language disappeared because of the mixing after the 840 AD of the last Tocharians with the new arrived Uyghurs.

The Tocharian Language

Transliterations of Tocharian manuscripts [12-14] have been published in the recent past, and because many Tocharian manuscripts of religious content have corresponding texts written in Brāhmī or Sanskrit, translations of Tocharian texts were possible and various Tocharian grammars [15-35] appeared. The Tocharian manuscripts [36] are written mainly in North Indian Brāhmī script (aksaras and anusvaras) and in a small number in Manichean script. Both TocA and TocB have a common inventory of vowels and the same consonant inventory. In both languages all the vowels are short and there is no certain evidence of vocalic length. The diphthongs which existed in P-Toc were monophthongized in TocA, diphthongs remained only in TocB. In TocB the accent is normally on the second syllable in words with more than two syllables and on the first syllable in words of two syllables, however, it could not be the case for TocA.

Both TocA and TocB maintain the three grammatical genders masculine, feminine, neuter distinct from the notion of biological gender, unless the noun represents something of animate in which case grammatical and biological gender normally coincides. Both TocA and TocB maintain the categories of singular, dual and plural as the PIE but there are suggestions also of **paral** – indicating natural pairs: hands, eyes – and **plurative** – indicating an individual in a plurality [37]. Both TocA and TocB distinguish human and

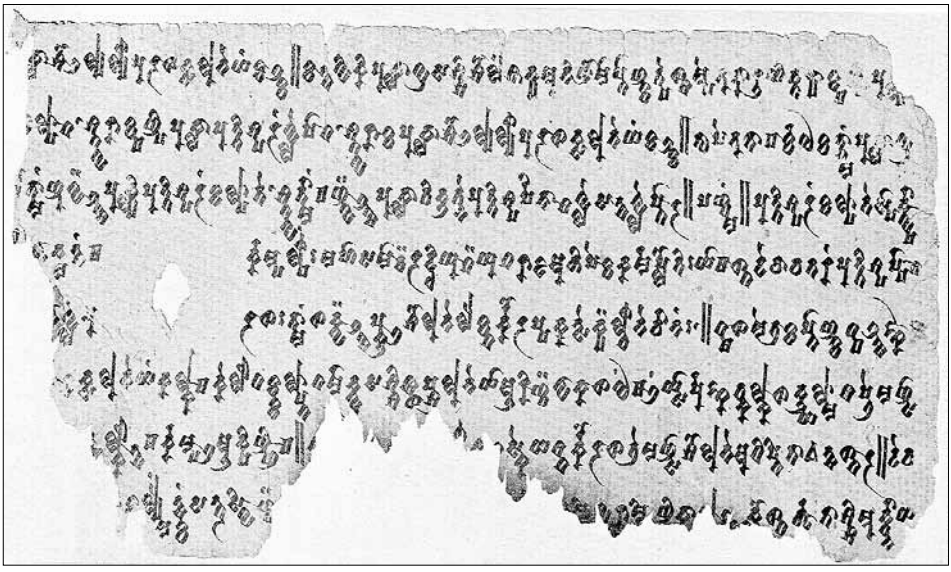

Fig. 2: Tocharian manuscript fragment THT 888 from the Berlin Collection [3]

non-human substantives. The Tocharian verbs distinguish three singular persons and three plural persons and three numbers: singular, dual and plural. The Tocharian use **active sentences** in which the subject is also the agent and **middle or medio-passive sentences** in which the subject is the agent and also the object of the action. Some Tocharian verbs occur only in the **middle** or **deponent** form. The Tocharian verbs distinguishes the **indicative**, the mood of simple fact, composed by present and past tense (imperfect, preterit), the **subjunctive**, the mood of hypothesis or supposition, the **optative**, the mood of whish or desire and the **imperative**, the mood of direct command. **Future** can be expressed either by the present, or by the subjunctive or by verbal nouns. The Tocharian verbal system distinguishes, by changing the verbal stem: the **causative** or **factitive** verbs i.e. verbs causing another action – and the non-causative or **base** – i.e. verbs non-causing another action. The Tocharian prose texts exhibit the basic structure of **subject – object – verb** (SOV). Fig. 2 shows an example of Tocharian manuscript fragment.

Linguistic Comparison Studies

Many linguistic comparison studies, sometime by contradicting, sometime by rectifying previous studies [37], tried to determine a possible linguistic position of the Tocharian with respect to other languages.

Van Windekens [38-41] proposes, for example, the TocB substantive *ñerwe = today* as Uralic in origin, while for its semantic equivalent TocA *ārco* an origin through the IE adverb **art = exactly, precisely* and the IE verb **keu- = to shine*; for the TocB substantive *trau = a capacity measure* an origin linked to the IE substantive **drou = barrel, bath tub*; for the TocB verb *yaukk- = to use* an agreement with the Sanskrit verb *yuj = put in activity, to employ*; for the TocAB *pānto = tutor, patron* a link with the IE verb **bhendh = to tie, to bind*; for the TocB verb *mik- = to close the eyes* an agreement with the Slavic verb *směžiti = to close the eyes*, with the Slovenian dial. verb *mežáti = to have closed the eyes.*

Pinault [42], in an accurate analysis of the fragment TocB PK LC I (Pelliot Koutchéen Lettres Commerciales I), by using information derived from the Res Rusticae of Varro (Livre II, ch. 2) and of the Encyclopedia of Agriculture (Lagri) identifies the terms: *alyi yrim = gray lambs* or *"agneaux élévés"* or "agneaux grandis", *orotsana awi = big sheeps, aiyyāna śānta = small ovine cattle* "*petit bétail ovine*". He concludes that the Tocharian places itself at the same level of the great languages of the IE family attested before our era: Hittite, Indo-Iranian, Greek, and Latin and for many terms the Tocharian preserves forms that are as archaic as, or more archaic than, the other languages.

Lubotsky [43], after having acknowledged that the Chinese (C) substantive *mi = honey* is Indo-European, probably of Tocharian origin through the Old Chinese (OC) *mjit/**mit* TocB *mit = honey* < Proto-Toc **m'ət* < PIE **med^{h}u-*, proposes as possible Chinese loan words from Tocharian concerning charriots technology:

(1) C *shéng = chariot* (with four horses) < … < OC **Ljings / *Ləngs* – TocB *klenke*, TocA *klank = vehicle*, TocAB *klānk = to ride, to travel* (by vehicle);
(2) C *gŭ = nave of a wheel* < … < OC **kok / *kōk* – TocB *kokale*, TocA *kukäl* = chariot;

(3) C *fú = spokes of a wheel* < … < OC **pjik / *pək* – TocB *pwenta* (pl.) < P-Toc **pəw*;
(4) C *guǐ = wheel-axle ends* < … < OC **k^{w}rju? / *k^{w}r?* – TocB *kwarsär*, TocA *kursär = league, mile, vehicle, means of salvation*, P-Toc *k^{w}ärsär*;
(5) C *zhöu = carriage pole* < … < OC **trju / *tru* – TocA *tursko = draft-ox;*
(6) C *kuò = leather* < … < OC **k^{w}hak / k^{w}hăk* – TocA *hâc = skin, lude* < P-Toc **k^{w}ac-*;
(7) and as possible Chinese loan words from Tocharian concerning town building:
(8) C *jì = masonry* < … < OC **tsjit* < **tsjik / *tsik* – TocAB *tsik- = to build, to form* < P-Toc **ts'əik*;
(9) C *lǐ = village, hamlet* < … < OC **C-rji? / *C-rə?* – TocB *riye* TocA *ri = town* ;
(10) C *yuán = wall, garden park* < … < OC **wjan / wan* – TocAB *want = to envelop, to surround*;
(11) C *zhēn = post in a wall, support* < … < OC **trjeng / *treng* – TocB *trenk-*, TocA *trank- = to be fixed;*
(12) C *bi = wall* < … < OC **pek / *pēk* – TocB *pkante*, TocA *pkant = hindering, obstacle;*
(13) C *chéng = city, wall, fortified wall* < … < OC **djeng / *deng* – TocAB t*ank = to hinder, to impede.*

Concluding, he observes that only the OC words in the examples (1), (2), (5), (6), (8), (12) and possibly (9) can be positively identified as borrowings from Tocharian.

Lubotsky and Starostin [44] propose the words: TocA *kom*, TocB *kaum = sun, day;* TocA *āle*, TocB *alyiye* = palm* (of the hand); TocA *tor*, TocB *taur = dust*; TocB *ām* = silence*; TocA *kanak*, TocB *kanek = cotton cloth*; TocB *olya = more*; TocA *tmām*, TocB *t(ɯ) māne = ten thousand, a myriad*; TocB *pärśeri* = (head)louse*; TocB *yase* = shame*; TocAB *kärk- = rob, steal*; as Turk words borrowed by the Tocharian. They observes that "Some of them must already been borrowed during the Common Tocharian period and some representing the stage anterior of the Proto-Turk sound changes *lj > j and *ʹr > z. The latter would date the Turk – Tocharian contacts by a period prior to the separation of the Bulgar (Chuvash) branch, most probably around the beginning of our era". Concluding, they propose as:

– Early loans: the words: TocAB *klu = rice;* TocB *rapaññe = of the last month of the year*; TocA *trunk*, TocB *tronk* = hollow, cave*; TocA *ri*, TocB *rīye = town*, showing pre-Han or Early-Han phonetic peculiarities, entered Tocharian not later than the 2nd century BC;
– Middle Chinese loans: the words: TocB *cāk = hundred quarts* (dry measure); TocB *cāne* a unit of money*; TocB *tau = ten quarts* (dry measure); TocB *śak(ɯ)se* = brandy*; TocB s*ank* a wet or dry measure of volume; TocA *yāmutsi*, TocB *yāmuttsi* a kind of water flow; TocB s*itsok = millet-alcohol;* TocB s*ipānkiñc = abacus*; TocAB *cok = lamp*; TocA *lyäk*, TocB *lyak = thief*; TocAB *tsem = blue,* showing Middle Chinese phonetic features, entered Tocharian in the period 2nd – 7th century AD

Concerning in particular the TocB s*itsok*, we have to observe that in Slovenian *sok = juice* and that in general *žit-sok = a juice from grain.* For beer-making several grains can be used, so that TocB *sitsok* could also mean *beer.*

Ivanov [45] pointed out surprising similarities between Tocharian and Balto-Slavic languages:

TocA *āle*, TocB *alyina* = *palm* (of hand), Lit. *délna*, Let. *delna*, Old Slav. *ДЛАНЬ*, Old Rus. *ДОЛНЬ*;

TocB *wrauña* = *raven, crow,* Lit. *várna*, Russ. *ворона;* Slovenian. lit. *vrana*, dial. *urana*

TocB *wrauśke* = *little raven*, Russ. *ворон-енок*;

TocA *wrsār*, TocB *ysāre* = *fruit-stone*, Russ. *овес* = *oats*;

TocB *reki* = *language*, Old Slav. *РЬЧЬ*, Russ. *речь*;

TocB *pruk-* = *to jump*, Russ. *прыгать*;

TocB *tāpär(k)* = *now*, Russ. *теперь*;

TocA *pik-*, TocB *pik-* = *to write, to draw*, Lit. *piẽšti*, Old Slav. *ПИСАТИ*.

All these linguistic comparison studies, as well many other [46-50] give valuable information concerning the possible linguistic position of the Tocharian with respect to other languages. However, in our opinion they consider only limited sets of Tocharian words, without considering the Tocharian languages and the other languages as a whole. Moreover, the concept of linguistic position of a language in these studies remains largely undefined.

Mathematical Studies

Methods, based on different definitions of linguistic distance, have been developed in the past for measuring the distance between languages, dialects and variants in the same or different family languages. For example, the linguistic distances between Dutch and Irish dialects were measured by Nerbonne [51-53], Kessler [54], Heeringa [55], Kruskal [56] presented the Levenstein distance technique and Viaregge [57] presented hybrid techniques applied respectively by them and other authors for determining the linguistic distance when applied to Corpuses of well-known words having well known phonetics and grammar rules. However, also in these works only a limited set of words is considered in determining the linguistic distance.

More recently, Silvestri and Tomezzoli [58-59], by determining the frequencies of all the vowel and consonants in selected Languages Databases (LDs) formed by including texts and inscriptions, measured the Euclidian (or Pythagorean) Linguistic Distance between Latin, Slovene, Venetic and Rhaetian languages. The determination of the frequencies of the vowel and consonants was necessary because the Venetic and Rhaetian inscriptions considered for building the corresponding LDs are written in continuo, i.e. without subdivision in words. The result was that Venetic and Rhaetian have Euclidian (or Pythagorean) linguistic distances closer to the Slovene than to the Latin.

In an accurate study, Perdih, Tomezzoli and Vodopivec [60] applied the method of the Principal Components Analysis (PCA) to sound frequencies in LDs of the languages: Basque (Bq), Estonian (Es), Etruscan (EtB, EtT), Finnic (Fi), Greek (Gr), Hittite (Ht), Latin (LaC, LaS), Luvian (Lu), Mycenaean (My), Old Church Slavonic (Cs), Oscan (Os), Old Phrygian (PhA, PhT, PhV2), Rhaetic (RtB, RtT, RtV, RtV2), Old Slovenian (Sl),

Umbrian (Um), Venetic (VeB, VeT, VeV, VeV2) and Venetian (Vz). The PCA represents a rotation of the alphabetic or sound coordinate system such that the largest variations or agreements in the LDs are displayed by a limited set of variables (10) called Latent Variables or Principal Components.

The study confirmed the results by Silvestri, Tomezzoli [58-59] and provided various other relevant results, for example, Old Phrygian is close to Venetic and Rhaetic; Etruscan is also close to Venetic and Rhaetic but not close to Hittite and Luvian from which it might have derived; Latin is closest to Greek as well as Oscan, Umbrian, Mycenaean and Estonian; Estonian is close to Finnish, Old Phrygian, Venetian, Finnish; Venetian a Romanic language spoken in the former Venetic territory is closer to Venetic and Old Slovenian than to Latin, of which it retains many other characteristics.

Vectorial Formalism in the Alphabetic Statistic Distribution Space

To improve our knowledge about the linguistic position of languages, let us define an Alphabetic Statistic Distribution Space (ASDS) as a multi-dimensional space having coordinates a, …, u, b, …, z. We define an Alphabetic Frequencies Distribution (AFD) vector V as a multidimensional vector whose components v_1 ... v_n are the relative frequencies of the alphabetic characters a, …, u, b, …, z of a language calculated by using its respective LD. Such an AFD vector V:

$$V = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \quad \text{(Eq. 1)}$$

Having the relative frequencies v_1 ... v_n corresponding to said language represents this language in the ASDS. The norm of an AFD vector V is defined by:

$$norm(V) = \sqrt{\sum_{i=1}^{n} v_i^2} \quad \text{(Eq. 2)}$$

and the scalar product of two AFD vectors V and W by:

$$V \cdot W = \sum_{i=1}^{n} v_i \cdot w_i \quad \text{(Eq. 3)}$$

Having two AFD vectors V and W representing two different languages “V” and “W” in the ASDS, their cosine function cos(V,W) is defined by:

$$\cos(V,W) = \frac{V \cdot W}{norm(V) \cdot norm(W)} \quad \text{Eq. 4)}$$

A value of cos(V,W) close to 1 indicates that the two AFD vectors V and W are practically aligned and thus the statistical properties of language “V” and language “W” have strong linear dependence. A value of cos(V,W) close to 0 indicate that the two AFD vectors are practically orthogonal and that their statistical properties are linearly independent. Of course, the more similar two languages “V” and “W” are, the strongest their linear dependency is and this is reflected by the value of cos(V,W) of their AFD vectors V, W.

The distance dist(V,W) of the two AFD vectors V and W is defined by:

$$dist(V,W) = \sqrt{\sum_{i=1}^{n}(v_i - w_i)^2} \quad \text{(Eq. 5)}$$

and represents the Euclidean distance of the languages "V" and "W" in the ASDS. In the ASDS the closer to 0 is the dist(V,W) of the AFD vectors V, W, the more alphabetically similar are the corresponding languages "V" and "W". The dist(V,W) corresponds to the main diagonal SuS = $(\Sigma\ (y_i\text{-}x_i)^2)^{1/2}$ used by Perdih [61–62] for establishing a sequence of increasing distances between the languages considered in [60].

Cosine and distance of the two AFD vectors V and W representing the languages "V" and "W" are a good representation of their mutual linguistic position in the ASDS.

AFD vectors can give good insights about the mutual positions of the corresponding languages and if AFD vectors are calculated for representing a single language in different times, they give good insight about its evolution. Etymology and phonetics can add of course further language positional information. In the following, we will refer to the linguistic position of different languages as the values of the cosine and distance of their respective AFD vectors.

Linguistic Position

For determining the linguistic position of TocA and TocB with respect to the languages already considered in the study of Perdih, Tomezzoli and Vodopivec [60], as first step, we prepared:

- the Tocharian A Language Database (Toc-A-LD) comprising: in the order the transliterations of the TocA manuscripts: YQ 1.30, YQ 1.29, YQ 1.32, YQ 1.28, YQ 1.17, YQ 1.16, YQ 1.15, YQ 1.3, YQ 1.9, YQ 1.1., YQ 1.2, YQ 1.4, YQ 1.42, YQ 1.8, YQ 1.14, YQ 1.13, YQ 1.5, YQ 1.6, YQ 1.7, YQ 1.12, YQ 1.11, YQ 1.10, YQ 1.31, YQ 1.33, YQ 1.43, YQ 1.21, YQ 1.22, YQ 1.44, YQ 1.44, YQ 1.23, YQ 1.24, YQ 1.25, YQ 1.26, YQ 1.41, YQ 1.20, YQ 1.18, YQ 1.19, YQ 1.27, YQ 1.39, YQ 1.34, YQ 1.35, YQ 1.36, YQ 1.37, YQ 1.38, YQ 1.40 from the Maitreyasamiti-Nāṭaka text of the Xinjang Museum, China [63], of the A255 = THT 888 manuscript from Maitreyasamiti-Nāṭaka [63], pp. 89-108 (cf. Fig. 2), of the A5a5-10a2 manuscript: Histoire du peintre, du mécanicien et de la fille mécanique [16], pp. 251-268, of the No. 60 = T III Š 85 6 fragment [14], p. 34, of the No. 336 = T III S 92.33 + 64.4 fragment [14], p. 184;
- the Tocharian B Language Database (Toc-B-LD) comprising the transliterations of the following manuscripts: THT 1540 [64], pp. 321-339, Tok. B 298 – Poème épigraphique adressé à la mort [16], pp. 15-18, Tokh. B 496 – Poème d'amour en tokharien B [16], pp. 19-36, Commentaire des règles du Vinaya (Discipline monastique) Tokh. B Texte 1: PK AS 18A [16], pp. 61-88, Buddhastotra en Tocharien B de la collection Petrovskij (St. Petersburg) SI P/1a1 [16], pp. 294-311, Confession d'une femme pour ses écart de langage B 241 = THT 241 [16], pp. 329-350, Laissez-passer de caravanes en Tokharien B – LP1, LP2, LP11 [13], pp. 351-358, Comptabilité du monastère PK D.A.M.507(8)

[16], pp. 359-374, Lettres commerciales B 492, PK L C X, Nr. 20 = T III. Š 95. 13 + Š 98 [16], pp. 34-35, No. 367 = T III. M 146. 8 [12], p. 242, No. 552 = T III. MQ 73. 5 [13], p. 348.

As second step, the alphabetic characters in the Toc-A-LD and Toc-B-LD have been aggregated according to the pronunciation rules used in [59] and the frequencies of the alphabetic characters a, …, u, b, …, z determined. As third step, the cos (V,W) and the dist (V,W) according to equations (3), (4), of the AFD vectors V and W for each pair of languages in the set formed by the TocA, TocB and the languages considered in [60] have been determined (cf. Tables 1-3 Cosine; Tables 4-6: Distances). As final step, in order to represent the linguistic position of the TocA and TocB with respect to the languages of [60], the values of the cos(V, W) and dist (V,W) in which W represents the AFD vector for the TocA and V represents the AFD vector of each one of the TocB and the other languages of [60] have been put in a diagram (cf. Fig. 3).

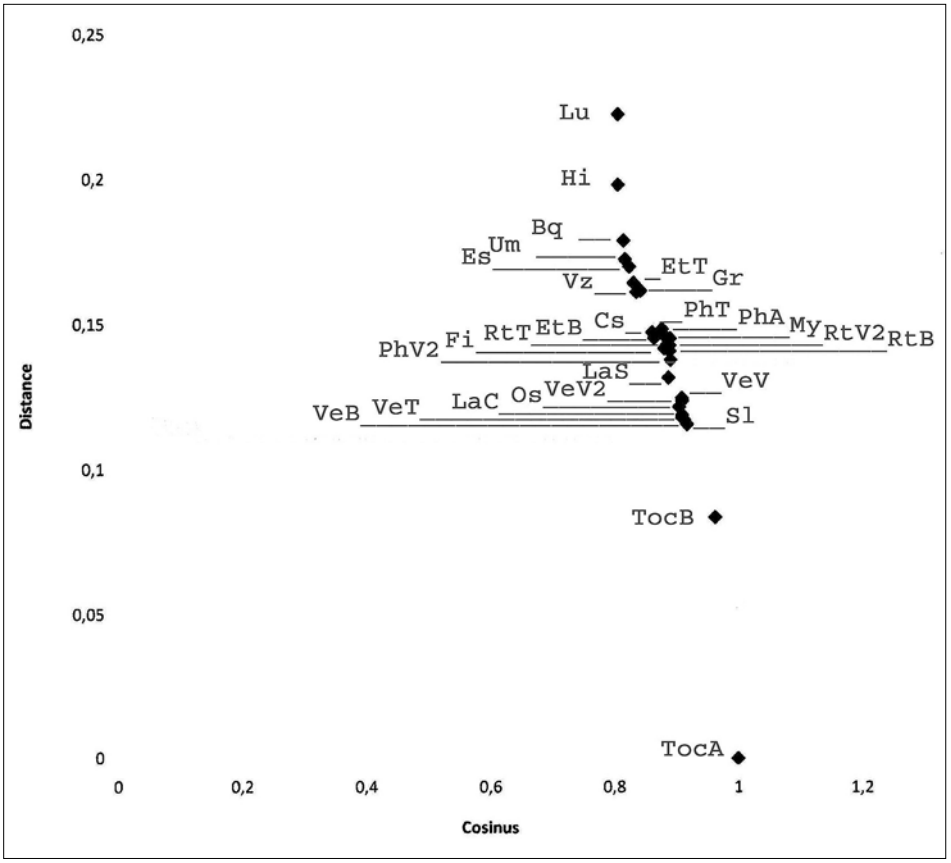

Fig. 3: linguistic position (Cosine and Distance of the respective AFB vectors) of the languages with respect to TocA and TocB

Table 1: Cosine of the AFD vectors of the languages Bq … LaS

	Bq	**Cs**	**Es**	**EtB**	**EtT**	**Fi**	**Gr**	**Hi**	**LaC**	**LaS**
Bq	**1,0000**	0,8434	0,9349	0,9169	0,8962	0,9167	0,9299	0,8346	0,9138	0,9176
Cs	0,8434	**1,0000**	0,8562	0,8547	0,8083	0,9081	0,9144	0,6420	0,8837	0,8835
Es	0,9349	0,8562	**1,0000**	0,9182	0,9105	0,9483	0,9086	0,8109	0,9288	0,9262
EtB	0,9169	0,8547	0,9182	**1,0000**	0,9731	0,9212	0,8788	0,8541	0,8882	0,8990
EtT	0,8962	0,8083	0,9105	0,9731	**1,0000**	0,9090	0,8339	0,8509	0,8799	0,8898
Fi	0,9167	0,9081	0,9483	0,9212	0,9090	**1,0000**	0,9400	0,7875	0,9318	0,9269
Gr	0,9299	0,9144	0,9086	0,8788	0,8339	0,9400	**1,0000**	0,7225	0,9314	0,9370
Hi	0,8346	0,6420	0,8109	0,8541	0,8509	0,7875	0,7225	**1,0000**	0,7955	0,7814
LaC	0,9138	0,8837	0,9288	0,8882	0,8799	0,9318	0,9314	0,7955	**1,0000**	0,9955
LaS	0,9176	0,8835	0,9262	0,8990	0,8898	0,9269	0,9370	0,7814	0,9955	**1,0000**
Lu	0,7934	0,6513	0,7828	0,8331	0,8461	0,7702	0,6853	0,9675	0,7572	0,7344
My	0,8828	0,8717	0,8487	0,8374	0,7772	0,8649	0,9187	0,7459	0,9134	0,9081
Os	0,8677	0,8828	0,9229	0,8765	0,8754	0,9205	0,9009	0,8008	0,9836	0,9769
PhA	0,9300	0,9073	0,9306	0,9004	0,8686	0,9340	0,9265	0,8229	0,9064	0,8933
PhT	0,9339	0,9007	0,9308	0,8985	0,8668	0,9310	0,9280	0,8250	0,9066	0,8937
RtB	0,8865	0,8659	0,9210	0,9201	0,9301	0,9259	0,8295	0,8568	0,9114	0,8981
RtT	0,8984	0,8588	0,9304	0,9263	0,9361	0,9281	0,8364	0,8682	0,9145	0,9020
RtV	0,8712	0,8708	0,9101	0,9110	0,9267	0,9223	0,8222	0,8409	0,9104	0,8964
Sl	0,9197	0,9626	0,9310	0,9243	0,8915	0,9420	0,9464	0,7916	0,9604	0,9604
Um	0,8923	0,8345	0,8793	0,8602	0,8658	0,8749	0,9053	0,7303	0,9568	0,9646
VeB	0,8735	0,9634	0,8569	0,8736	0,8269	0,9282	0,9361	0,7061	0,9147	0,9064
VeT	0,8793	0,9436	0,8494	0,8777	0,8332	0,9282	0,9443	0,7116	0,9139	0,9084
VeV	0,8497	0,9687	0,8366	0,8625	0,8181	0,9168	0,9198	0,6960	0,9028	0,8944
Vz	0,9411	0,8903	0,9182	0,9039	0,8444	0,9139	0,9654	0,7545	0,8984	0,9031
PhV2	0,9271	0,9219	0,9323	0,9004	0,8729	0,9389	0,9314	0,8138	0,9201	0,9070
RtV2	0,8620	0,8751	0,9029	0,9045	0,9219	0,9203	0,8194	0,8337	0,9093	0,8951
VeV2	0,8499	0,9674	0,8359	0,8639	0,8217	0,9173	0,9207	0,6972	0,9052	0,8977
TocA	0,8144	0,8609	0,8240	0,8639	0,8305	0,8803	0,8420	0,8054	0,9097	0,8876
TocB	0,8676	0,9110	0,8774	0,8943	0,8794	0,9262	0,8666	0,8148	0,9198	0,9025

Discussion

The frequencies of the alphabetic characters derived from the Toc-A-LD are certainly influenced by the religious nature of the texts considered, i.e. by an higher presence of alphabetic characters corresponding to Buddhist terms. The Toc-B-LD, containing text referring to various aspects, of the day life, is less influenced by that effect. Anyway, the Toc-A-LD and the Toc-B-LD fulfil the criteria set out by Perdih [65] because they contain

Table 2: Cosine of the AFD vectors of the languages Lu … Um

	Lu	My	Os	PhA	PhT	RtB	RtT	RtV	Sl	Um
Bq	0,7934	0,8828	0,8677	0,9300	0,9339	0,8865	0,8984	0,8712	0,9197	0,8923
Cs	0,6513	0,8717	0,8828	0,9073	0,9007	0,8659	0,8588	0,8708	0,9626	0,8345
Es	0,7828	0,8487	0,9229	0,9306	0,9308	0,9210	0,9304	0,9101	0,9310	0,8793
EtB	0,8331	0,8374	0,8765	0,9004	0,8985	0,9201	0,9263	0,9110	0,9243	0,8602
EtT	0,8461	0,7772	0,8754	0,8686	0,8668	0,9301	0,9361	0,9267	0,8915	0,8658
Fi	0,7702	0,8649	0,9205	0,9340	0,9310	0,9259	0,9281	0,9223	0,9420	0,8749
Gr	0,6853	0,9187	0,9009	0,9265	0,9280	0,8295	0,8364	0,8222	0,9464	0,9053
Hi	0,9675	0,7459	0,8008	0,8229	0,8250	0,8568	0,8682	0,8409	0,7916	0,7303
LaC	0,7572	0,9134	0,9836	0,9064	0,9066	0,9114	0,9145	0,9104	0,9604	0,9568
LaS	0,7344	0,9081	0,9769	0,8933	0,8937	0,8981	0,9020	0,8964	0,9604	0,9646
Lu	**1,0000**	0,7324	0,7713	0,8336	0,8343	0,8688	0,8763	0,8612	0,7757	0,6901
My	0,7324	**1,0000**	0,8694	0,9010	0,9028	0,8261	0,8309	0,8215	0,9179	0,8847
Os	0,7713	0,8694	**1,0000**	0,8949	0,8929	0,9219	0,9221	0,9234	0,9562	0,9267
PhA	0,8336	0,9010	0,8949	**1,0000**	0,9996	0,9014	0,9092	0,8958	0,9410	0,8472
PhT	0,8343	0,9028	0,8929	0,9996	**1,0000**	0,8971	0,9057	0,8906	0,9382	0,8481
RtB	0,8688	0,8261	0,9219	0,9014	0,8971	**1,0000**	0,9988	0,9977	0,9308	0,8573
RtT	0,8763	0,8309	0,9221	0,9092	0,9057	0,9988	**1,0000**	0,9946	0,9301	0,8663
RtV	0,8612	0,8215	0,9234	0,8958	0,8906	0,9977	0,9946	**1,0000**	0,9289	0,8564
Sl	0,7757	0,9179	0,9562	0,9410	0,9382	0,9308	0,9301	0,9289	**1,0000**	0,9085
Um	0,6901	0,8847	0,9267	0,8472	0,8481	0,8573	0,8663	0,8564	0,9085	**1,0000**
VeB	0,7136	0,9226	0,8921	0,9338	0,9292	0,8670	0,8629	0,8731	0,9528	0,8514
VeT	0,7124	0,9207	0,8833	0,9270	0,9240	0,8512	0,8497	0,8553	0,9404	0,8623
VeV	0,7157	0,9058	0,8897	0,9259	0,9201	0,8708	0,8643	0,8801	0,9516	0,8377
Vz	0,7179	0,9259	0,8586	0,9388	0,9407	0,8181	0,8286	0,8055	0,9264	0,8703
PhV2	0,8273	0,9064	0,9104	0,9982	0,9974	0,9095	0,9153	0,9065	0,9531	0,8573
RtV2	0,8577	0,8202	0,9244	0,8912	0,8855	0,9950	0,9905	0,9992	0,9299	0,8526
VeV2	0,7156	0,9054	0,8923	0,9239	0,9181	0,8702	0,8637	0,8799	0,9519	0,8409
TocA	0,8054	0,8899	0,9046	0,8789	0,8767	0,8890	0,8846	0,8883	0,9091	0,8169
TocB	0,8301	0,8779	0,9234	0,9154	0,9121	0,9557	0,9501	0,9552	0,9481	0,8406

more than 700 alphabetic characters. Consequently, they are of sufficient size for avoiding possible selection effects on the frequencies of the alphabetic characters due to their size. However, the frequencies of the alphabetic characters in the Toc-A-LD and Toc-B-LD, as well as in the LD of the other languages, reflect not only the phonetic characteristics of the corresponding language but also semantic features linked to the kind of life of the corresponding people.

Table 3: Cosine of the AFD vectors of the languages VeB … TocB

	VeB	VeT	VeV	Vz	PhV2	RtV2	VeV2	TocA	TocB
Bq	0,8735	0,8793	0,8497	0,9411	0,9271	0,8620	0,8499	0,8144	0,8676
Cs	0,9634	0,9436	0,9687	0,8903	0,9219	0,8751	0,9674	0,8609	0,9110
Es	0,8569	0,8494	0,8366	0,9182	0,9323	0,9029	0,8359	0,8240	0,8774
EtB	0,8736	0,8777	0,8625	0,9039	0,9004	0,9045	0,8639	0,8639	0,8943
EtT	0,8269	0,8332	0,8181	0,8444	0,8729	0,9219	0,8217	0,8305	0,8794
Fi	0,9282	0,9282	0,9168	0,9139	0,9389	0,9203	0,9173	0,8803	0,9262
Gr	0,9361	0,9443	0,9198	0,9654	0,9314	0,8194	0,9207	0,8420	0,8666
Hi	0,7061	0,7116	0,6960	0,7545	0,8138	0,8337	0,6972	0,8054	0,8148
LaC	0,9147	0,9139	0,9028	0,8984	0,9201	0,9093	0,9052	0,9097	0,9198
LaS	0,9064	0,9084	0,8944	0,9031	0,9070	0,8951	0,8977	0,8876	0,9025
Lu	0,7136	0,7124	0,7157	0,7179	0,8273	0,8577	0,7156	0,8054	0,8301
My	0,9226	0,9207	0,9058	0,9259	0,9064	0,8202	0,9054	0,8899	0,8779
Os	0,8921	0,8833	0,8897	0,8586	0,9104	0,9244	0,8923	0,9046	0,9234
PhA	0,9338	0,9270	0,9259	0,9388	0,9982	0,8912	0,9239	0,8789	0,9154
PhT	0,9292	0,9240	0,9201	0,9407	0,9974	0,8855	0,9181	0,8767	0,9121
RtB	0,8670	0,8512	0,8708	0,8181	0,9095	0,9950	0,8702	0,8890	0,9557
RtT	0,8629	0,8497	0,8643	0,8286	0,9153	0,9905	0,8637	0,8846	0,9501
RtV	0,8731	0,8553	0,8801	0,8055	0,9065	0,9992	0,8799	0,8883	0,9552
Sl	0,9528	0,9404	0,9516	0,9264	0,9531	0,9299	0,9519	0,9091	0,9481
Um	0,8514	0,8623	0,8377	0,8703	0,8573	0,8526	0,8409	0,8169	0,8406
VeB	**1,0000**	0,9920	0,9948	0,9232	0,9463	0,8759	0,9945	0,9174	0,9235
VeT	0,9920	**1,0000**	0,9827	0,9314	0,9375	0,8562	0,9828	0,9129	0,9094
VeV	0,9948	0,9827	**1,0000**	0,8933	0,9403	0,8851	0,9998	0,9091	0,9275
Vz	0,9232	0,9314	0,8933	**1,0000**	0,9372	0,7983	0,8922	0,8353	0,8423
PhV2	0,9463	0,9375	0,9403	0,9372	**1,0000**	0,9034	0,9387	0,8907	0,9267
RtV2	0,8759	0,8562	0,8851	0,7983	0,9034	**1,0000**	0,8851	0,8895	0,9567
VeV2	0,9945	0,9828	0,9998	0,8922	0,9387	0,8851	**1,0000**	0,9097	0,9270
TocA	0,9174	0,9129	0,9091	0,8353	0,8907	0,8895	0,9097	**1,0000**	0,9629
TocB	0,9235	0,9094	0,9275	0,8423	0,9267	0,9567	0,9270	0,9629	**1,0000**

As shown in Fig. 3, there is a close linear correlation between calculated distances and cosine values of the AFD vectors, showing that these two evaluation tools appear to present the same sequence of increasing linguistic distance. The diagram of Fig. 3, in which TocA has been assumed as reference for the calculations, shows the group TocA, TocB on the lower right side. Closer to TocA and TocB group is a first group formed by Slovenian (Sl), Venetic (VeV, VeV2, VeT, VeB), Latin (LaC) and Oscan (Os). More distant

Table 4: Distance of the AFD vectors of the languages Bq … LaS

	Bq	**Cs**	**Es**	**EtB**	**EtT**	**Fi**	**Gr**	**Hi**	**LaC**	**LaS**
Bq	**0,0000**	0,1635	0,1076	0,1206	0,1347	0,1221	0,1118	0,1854	0,1227	0,1202
Cs	0,1635	**0,0000**	0,1527	0,1490	0,1739	0,1231	0,1179	0,2639	0,1336	0,1327
Es	0,1076	0,1527	**0,0000**	0,1155	0,1217	0,0943	0,1248	0,1964	0,1081	0,1097
EtB	0,1206	0,1490	0,1155	**0,0000**	0,0653	0,1149	0,1408	0,1739	0,1307	0,1231
EtT	0,1347	0,1739	0,1217	0,0653	**0,0000**	0,1240	0,1662	0,1753	0,1375	0,1309
Fi	0,1221	0,1231	0,0943	0,1149	0,1240	**0,0000**	0,1018	0,2083	0,1072	0,1108
Gr	0,1118	0,1179	0,1248	0,1408	0,1662	0,1018	**0,0000**	0,2364	0,1066	0,1023
Hi	0,1854	0,2639	0,1964	0,1739	0,1753	0,2083	0,2364	**0,0000**	0,2026	0,2086
LaC	0,1227	0,1336	0,1081	0,1307	0,1375	0,1072	0,1066	0,2026	**0,0000**	0,0267
LaS	0,1202	0,1327	0,1097	0,1231	0,1309	0,1108	0,1023	0,2086	0,0267	**0,0000**
Lu	0,2278	0,2859	0,2329	0,2102	0,2022	0,2388	0,2749	0,0987	0,2445	0,2540
My	0,1511	0,1551	0,1697	0,1743	0,2036	0,1611	0,1258	0,2329	0,1303	0,1344
Os	0,1507	0,1334	0,1122	0,1365	0,1394	0,1154	0,1274	0,2000	0,0500	0,0587
PhA	0,1139	0,1275	0,1123	0,1332	0,1527	0,1098	0,1156	0,1922	0,1293	0,1376
PhT	0,1111	0,1326	0,1126	0,1352	0,1544	0,1127	0,1149	0,1914	0,1299	0,1381
RtB	0,1452	0,1533	0,1199	0,1205	0,1128	0,1165	0,1756	0,1734	0,1264	0,1351
RtT	0,1371	0,1567	0,1122	0,1155	0,1075	0,1144	0,1716	0,1664	0,1236	0,1320
RtV	0,1560	0,1524	0,1293	0,1287	0,1171	0,1206	0,1810	0,1832	0,1289	0,1382
Sl	0,1185	0,0758	0,1057	0,1059	0,1293	0,0989	0,0942	0,2038	0,0768	0,0759
Um	0,1377	0,1629	0,1423	0,1494	0,1481	0,1460	0,1264	0,2323	0,0838	0,0761
VeB	0,1489	0,0749	0,1546	0,1416	0,1677	0,1106	0,1037	0,2418	0,1168	0,1217
VeT	0,1446	0,0925	0,1569	0,1372	0,1625	0,1100	0,0964	0,2382	0,1156	0,1183
VeV	0,1646	0,0722	0,1685	0,1520	0,1759	0,1211	0,1184	0,2485	0,1285	0,1335
Vz	0,1022	0,1308	0,1160	0,1219	0,1573	0,1203	0,0762	0,2209	0,1258	0,1219
PhV2	0,1153	0,1156	0,1096	0,1313	0,1486	0,1045	0,1105	0,1963	0,1182	0,1271
RtV2	0,1619	0,1506	0,1347	0,1335	0,1212	0,1226	0,1829	0,1875	0,1303	0,1396
VeV2	0,1641	0,0732	0,1683	0,1506	0,1736	0,1203	0,1174	0,2477	0,1263	0,1308
TocA	0,1789	0,1474	0,1701	0,1454	0,1645	0,1418	0,1617	0,1983	0,1189	0,1316
TocB	0,1565	0,1250	0,1485	0,1368	0,1463	0,1159	0,1551	0,1964	0,1201	0,1317

is a second group formed by Phrygian (PhT, PhA, PhV2), Mycenaean (My), Rhaetic (RtT, RtV2, RtB), Etruscan (EtB) and Finnish. Additionally distant is a third group formed by Venetian (Vz), Greek (Gr), Etruscan (EtT), Estonian (Es), Umbrian (Um) and Basque (Bq). The fourth group formed by Luvian (Lu) and Hittite (Hi) is the most distant group from the TocA and TocB group.

Table 5: Distance of the AFD vectors of the languages Lu … Um

	Lu	My	Os	PhA	PhT	RtB	RtT	RtV	Sl	Um
Bq	0,2278	0,1511	0,1507	0,1139	0,1111	0,1452	0,1371	0,1560	0,1185	0,1377
Cs	0,2859	0,1551	0,1334	0,1275	0,1326	0,1533	0,1567	0,1524	0,0758	0,1629
Es	0,2329	0,1697	0,1122	0,1123	0,1126	0,1199	0,1122	0,1293	0,1057	0,1423
EtB	0,2102	0,1743	0,1365	0,1332	0,1352	0,1205	0,1155	0,1287	0,1059	0,1494
EtT	0,2022	0,2036	0,1394	0,1527	0,1544	0,1128	0,1075	0,1171	0,1293	0,1481
Fi	0,2388	0,1611	0,1154	0,1098	0,1127	0,1165	0,1144	0,1206	0,0989	0,1460
Gr	0,2749	0,1258	0,1274	0,1156	0,1149	0,1756	0,1716	0,1810	0,0942	0,1264
Hi	0,0987	0,2329	0,2000	0,1922	0,1914	0,1734	0,1664	0,1832	0,2038	0,2323
LaC	0,2445	0,1303	0,0500	0,1293	0,1299	0,1264	0,1236	0,1289	0,0768	0,0838
LaS	0,2540	0,1344	0,0587	0,1376	0,1381	0,1351	0,1320	0,1382	0,0759	0,0761
Lu	**0,0000**	0,2586	0,2386	0,2068	0,2064	0,1862	0,1816	0,1905	0,2367	0,2724
My	0,2586	**0,0000**	0,1574	0,1395	0,1385	0,1848	0,1819	0,1883	0,1277	0,1486
Os	0,2386	0,1574	**0,0000**	0,1367	0,1387	0,1193	0,1186	0,1203	0,0801	0,1085
PhA	0,2068	0,1395	0,1367	**0,0000**	0,0085	0,1362	0,1304	0,1410	0,1045	0,1652
PhT	0,2064	0,1385	0,1387	0,0085	**0,0000**	0,1395	0,1332	0,1449	0,1075	0,1653
RtB	0,1862	0,1848	0,1193	0,1362	0,1395	**0,0000**	0,0151	0,0215	0,1128	0,1600
RtT	0,1816	0,1819	0,1186	0,1304	0,1332	0,0151	**0,0000**	0,0327	0,1126	0,1544
RtV	0,1905	0,1883	0,1203	0,1410	0,1449	0,0215	0,0327	**0,0000**	0,1164	0,1622
Sl	0,2367	0,1277	0,0801	0,1045	0,1075	0,1128	0,1126	0,1164	**0,0000**	0,1203
Um	0,2724	0,1486	0,1085	0,1652	0,1653	0,1600	0,1544	0,1622	0,1203	**0,0000**
VeB	0,2629	0,1230	0,1308	0,1096	0,1138	0,1544	0,1562	0,1525	0,0864	0,1567
VeT	0,2629	0,1251	0,1337	0,1149	0,1179	0,1622	0,1624	0,1619	0,0948	0,1491
VeV	0,2631	0,1353	0,1364	0,1166	0,1215	0,1541	0,1574	0,1498	0,0920	0,1673
Vz	0,2607	0,1211	0,1476	0,1056	0,1047	0,1790	0,1733	0,1869	0,1060	0,1450
PhV2	0,2103	0,1352	0,1249	0,0186	0,0229	0,1296	0,1251	0,1329	0,0923	0,1581
RtV2	0,1925	0,1893	0,1202	0,1445	0,1486	0,0314	0,0431	0,0124	0,1164	0,1648
VeV2	0,2629	0,1355	0,1342	0,1180	0,1228	0,1541	0,1574	0,1497	0,0910	0,1651
TocA	0,2227	0,1453	0,1216	0,1466	0,1485	0,1409	0,1430	0,1431	0,1179	0,1725
TocB	0,2087	0,1547	0,1177	0,1259	0,1287	0,0912	0,0966	0,0926	0,0986	0,1686

Conclusion

The fact that the linguistic positions of the TocA and TocB appears quite isolated from the linguistic position of other languages suggests that TocA and TocB would have had a formation and evolution rather independent. However, the closer presence of the first group formed by Slovenian (Sl), Venetic (VeV, VeV2, VeT, VeB), Latin (LaC) and Oscan (Os) would indicate more important mutual influences between TocA, TocB, Slovenian,

Table 6: Distance of the AFD vectors of the languages VeB … TocB

	VeB	**VeT**	**VeV**	**Vz**	**PhV2**	**RtV2**	**VeV2**	**Toc A**	**Toc B**
Bq	0,1489	0,1446	0,1646	0,1022	0,1153	0,1619	0,1641	0,1789	0,1565
Cs	0,0749	0,0925	0,0722	0,1308	0,1156	0,1506	0,0732	0,1474	0,1250
Es	0,1546	0,1569	0,1685	0,1160	0,1096	0,1347	0,1683	0,1701	0,1485
EtB	0,1416	0,1372	0,1520	0,1219	0,1313	0,1335	0,1506	0,1454	0,1368
EtT	0,1677	0,1625	0,1759	0,1573	0,1486	0,1212	0,1736	0,1645	0,1463
Fi	0,1106	0,1100	0,1211	0,1203	0,1045	0,1226	0,1203	0,1418	0,1159
Gr	0,1037	0,0964	0,1184	0,0762	0,1105	0,1829	0,1174	0,1617	0,1551
Hi	0,2418	0,2382	0,2485	0,2209	0,1963	0,1875	0,2477	0,1983	0,1964
LaC	0,1168	0,1156	0,1285	0,1258	0,1182	0,1303	0,1263	0,1189	0,1201
LaS	0,1217	0,1183	0,1335	0,1219	0,1271	0,1396	0,1308	0,1316	0,1317
Lu	0,2629	0,2629	0,2631	0,2607	0,2103	0,1925	0,2629	0,2227	0,2087
My	0,1230	0,1251	0,1353	0,1211	0,1352	0,1893	0,1355	0,1453	0,1547
Os	0,1308	0,1337	0,1364	0,1476	0,1249	0,1202	0,1342	0,1216	0,1177
PhA	0,1096	0,1149	0,1166	0,1056	0,0186	0,1445	0,1180	0,1466	0,1259
PhT	0,1138	0,1179	0,1215	0,1047	0,0229	0,1486	0,1228	0,1485	0,1287
RtB	0,1544	0,1622	0,1541	0,1790	0,1296	0,0314	0,1541	0,1409	0,0912
RtT	0,1562	0,1624	0,1574	0,1733	0,1251	0,0431	0,1574	0,1430	0,0966
RtV	0,1525	0,1619	0,1498	0,1869	0,1329	0,0124	0,1497	0,1431	0,0926
Sl	0,0864	0,0948	0,0920	0,1060	0,0923	0,1164	0,0910	0,1179	0,0986
Um	0,1567	0,1491	0,1673	0,1450	0,1581	0,1648	0,1651	0,1725	0,1686
VeB	**0,0000**	0,0366	0,0319	0,1112	0,0976	0,1514	0,0320	0,1156	0,1174
VeT	0,0366	**0,0000**	0,0568	0,1036	0,1051	0,1619	0,0561	0,1171	0,1272
VeV	0,0319	0,0568	**0,0000**	0,1347	0,1037	0,1471	0,0065	0,1247	0,1153
Vz	0,1112	0,1036	0,1347	**0,0000**	0,1052	0,1908	0,1349	0,1613	0,1664
PhV2	0,0976	0,1051	0,1037	0,1052	**0,0000**	0,1354	0,1049	0,1378	0,1164
RtV2	0,1514	0,1619	0,1471	0,1908	0,1354	**0,0000**	0,1468	0,1429	0,0912
VeV2	0,0320	0,0561	0,0065	0,1349	0,1049	0,1468	**0,0000**	0,1237	0,1154
TocA	0,1156	0,1171	0,1247	0,1613	0,1378	0,1429	0,1237	**0,0000**	0,0836
TocB	0,1174	0,1272	0,1153	0,1664	0,1164	0,0912	0,1154	0,0836	**0,0000**

Venetic, Latin (LaC) and Oscan (Os), and less important mutual influences between TocA, TocB and the languages in the other groups. However, in order to confirm this hypothesis and to better test our vectorial formalism we plan to consider in the future more Slavic, Celtic and Uralo-Altaic languages.

Acknowledgement

We thank prof. A. Perdih for having provided us with the final frequency characters counting data of the linguistic databases considered in the study of Perdih, Tomezzoli and Vodopivec [60] as well as for the support in the preparation of the present paper.

References

1. H. A. Fellner, *The Expeditions to Tocharistan*, In Instrumenta Tocharica, M. Malzahn (ed.), Universität Verlag Winter, Heidelberg **2007**, 13-36.
2. M. Malzahn, *Tocharian Texts and Where to Find Them*, In Instrumenta Tocharica, M. Malzahn (ed.), Universität Verlag Winter, Heidelberg **2007**, 79-112.
3. *Berlin Collection:*
   http://titus.fkidg1.uni-frankfurt.de/texte/tocharic/tht1.htm,
   http://www.bbaw.de/forschung/turfanforschung/dta/mainz/dta_mainz_index.htm,
   http://www.bbaw.de/forschung/turfanforschung/dta/mainz/dta_mainz_Bemerkungen,
   htm#mainz0655(6),
   http://www.bbaw.de/forschung/turfanforschung/dta/u/dta_u_index.htm.
4. *London Collection:*
   http://idp.bl.uk.
5. E. Kuzmina, *The prehistory of the Silk Road*, Victor H. Hair (ed.), University of Pennsylvania Press, Philadelphia **2008**, part 2-3.
6. P. Jandáček, A. Perdih, *A novel view of the origins, development and differentiation of Indo-Europeans*, Proceedings of the Sixth International Topical Conference, Origin of European, Ljubljana 6-7 June 2008, A. Perdih (ed.), Založništvo Jutro, Ljubljana **2008**, pp. 88-98, available at:
7. http://www.korenine.si/zborniki/zbornik08/novel_ie_view.pdf.
8. J. P. Mallory, V. H. Mair, *The Tarim Mummies*: Ancient China and the Mystery of the Earliest Peoples from the West, Thames & Hudson, **2000**.
9. *Tarim mummies*, http://en.wikipedia.org/wiki/Tarim_mummies,
   http://en.wikipedia.org/wiki/Tarim_mummies
10. http://discovermagazine.com/1994/apr/themummiesofxinj359http://discovermagazine.com/1994/apr/themummiesofxinj359.
11. http://en.wikipedia.org/wiki/Tocharian_languages.
12. E. Sieg, W. Siegling, *Tocharische Sprachreste*, Sprache B, Heft I, Die Udānālankāra-Fragmente, Vandnhoeck & Ruprecht, Göttingen, **1949**.
13. E. Sieg, W. Siegling, *Tocharische Sprachreste*, Sprache B, Heft 2, Fragmente 71-633, Vandnhoeck & Ruprecht, Göttingen **1953**.
14. E. Sieg, W. Siegling, *Tocharische Sprachreste*, I. Band B, Die Texte, A Transcription, Vereinigung Wissenschaftlicher Verleger, Berlin und Liepzig **1921**.
15. E. Sieg, W. Siegling, W. Schulze, *Tocharische Grammatik*, Göttingen **1931**.
16. G. J. Pinault, *Chrestomathie Tokharienne Textes et Grammaire*, Peeters Leuven-Paris **2008**.
17. Tocharian Online – *Lesson 1,*
    http://www.utexas.edu/cola/centers/lrc/eieol/tokol-1-X.html.
18. Tocharian Online – *Selected Annotated Bibliography,*
    http://www.utexas.edu/cola/centers/lrc/eieol/tokol-E.html.
19. Tocharian Online – *Lesson 2,*
    http://www.utexas.edu/cola/centers/lrc/eieol/tokol-2-X.html.

20. Tocharian Online – *Lesson 3,*
http://www.utexas.edu/cola/centers/lrc/eieol/tokol-3-X.html.
21. Tocharian Online – *Lesson 4,*
http://www.utexas.edu/cola/centers/lrc/eieol/tokol-4-X.html.
22. Tocharian Online – *Lesson 5,*
http://www.utexas.edu/cola/centers/lrc/eieol/tokol-5-X.html.
23. Tocharian Online – *Lesson 6,*
http://www.utexas.edu/cola/centers/lrc/eieol/tokol-6-X.html.
24. Tocharian Online – *Lesson 7,*
http://www.utexas.edu/cola/centers/lrc/eieol/tokol-7-X.html.
25. Tocharian Online – *Lesson 8,*
http://www.utexas.edu/cola/centers/lrc/eieol/tokol-8-X.html.
26. Tocharian Online – *Lesson 9,*
http://www.utexas.edu/cola/centers/lrc/eieol/tokol-9-X.html.
27. Tocharian Online – *Lesson 10,*
http://www.utexas.edu/cola/centers/lrc/eieol/tokol-10-X.html.
28. *Tocharian A texts,*
http://www.ling.upenn.edu/~kurisuto/germanic/tocharian_a_raw_notes.html.
29. Tocharian Online – *Tocharian A: English Meaning Index,*
http://www.utexas.edu/cola/centers/lrc/eieol/tokol-EI-X.html.
30. Tocharian Online – *Tocharian A: Master Glossary,*
http://www.utexas.edu/cola/centers/lrc/eieol/tokol-MG-X.html.
31. Tocharian Online – *Tocharian A: Base Form Dictionary,*
http://www.utexas.edu/cola/centers/lrc/eieol/tokol-BF-X.html,
http://www.utexas.edu/cola/centers/lrc/eieol/tokol-BF-R.html.
32. *Tocharian B texts,*
http://www.ling.upenn.edu/~kurisuto/germanic/tocharian_b_raw_notes.html.
33. Tocharian Online – *Tocharian B: English Meaning Index,*
http://www.utexas.edu/cola/centers/lrc/eieol/txbol-EI-X.html.
34. Tocharian Online – *Tocharian B: Master Glossary,*
http://www.utexas.edu/cola/centers/lrc/eieol/txbol-MG-X.html
35. Tocharian Online – *Tocharian B: Base Form Dictionary,*
http://www.utexas.edu/cola/centers/lrc/eieol/txbol-BF-X.html.
36. M. Malzahn, *The Tocharian Verbal System, in Brill's Studies in Indo-European Languages & Linguistics*, C Melchert, O Hackstein (eds), Vol. 3, BRILL, Leiden – Boston, **2010**.
37. W. Winter, *Nominal and Pronominal Dual in Tocharian, Language*, 38, **1982**, 111-134.
38. A. J. Van Windeckens, *Sur l'origine indo-européenne de quelque mot tokharien – I*, Orbis 13, **1964**.
39 A. J. Van Windeckens, *Sur l'origine indo-européenne de quelque mot tokharien – II*, Orbis 14, **1965**, 501-504.
40. A. J. Van Windeckens, *Sur l'origine indo-européenne de quelque mot tokharien – V*, Orbis 19, **1970**, 165-171.
41. A. J. Van Windeckens, *Sur l'origine indo-européenne de quelque mot tokharien – VI*, Orbis 19, **1970**, 526-528.
42. G. J. Pinault, *Terminologie de Petit Bétail en Tocharien*, Studia etymologica Cracoviensia, 2, **1997**, 175-218.
43. A. Lubotsky, *Tocharian Loan Words in Old Chinese*: Chariots, Chariot Gear, and Town Building. In: The Bronze Age and Early Iron Age Peoples of Central Asia, Victor A Mair (ed.), Institute for the Study of Man, Washington D.C. **1998**, 379-390.

44. A. Lubotsky, S. Starostin, *Turkic and Chinese Words in Tocharian, Language in time and space*. A Festschrift for Werner Winter on the occasion of his 80th birthday, B. L. M. Bauer, G. J. Pinault (edd.), Berlin – New York **2003**, 257-269.
45. V. V. Ivanov, *Balto-slaviano-toxarskie izoglossy*, Balto-slavianskie issledovanija, **1988**, Band 1986, 45-60.
46. A. J. Van Windekens, *Morphologie Comparée du Tocharien*, Louvain, **1944**.
47. A. J. Van Windekens, *Le Tocharien confronté avec les autres langues indo-européennes,* Vol. I, La phonétique et le vocabulaire, Louvain **1976**.
48. A. J. Van Windekens, *Le Tocharien confronté avec les autres langues indo-européennes*, Vol. II, La morphologie nominale, Louvain **1979**.
49. A. J. Van Windekens, *Le Tocharien confronté avec les autres langues indo-européennes*, Vol. I, La morphologie verbale, Louvain **1982**.
50. *Papers from the Workshop on Tocharian*, Leiden, September 5, 1987, Voll 1, 2, Ed J Hilmarsson, Reykjavik **1988**.
51. J. Nerbonne, W. Heeringa, *Measuring Dialect Distance Phonetically*, Workshop on Computational Phonology, in J. Coleman (ed.), Madrid **1997**, 12-15, available at: http://odur.let.rug.nl/~nerbonne/paper.htmlhttp://odur.let.rug.nl/~nerbonne/paper.html.
52. J. Nerbonne, W. Heeringa, *Computational Comparison and Classification of Dialects*, 2nd Congress of Dialectologists and Geolinguists, Amsterdam **2002**, 1-16.
53. J. Nerbonne, W. Heeringa, E. Van der Hout, S. Otten, W. Van de Vis, *Phonetic Distance between Dutch Dialects*. In: G Durieux, W Daeleman & Gillis (eds.), CLIN VI, Papers from the sixth meeting, University of Antwerpen, Center for Dutch Language and Speech, Antwerp **1996**, 185-202, available at: http://odur.let.rug.nl/~nerbonne/paper.htmlhttp://odur.let.rug.nl/~nerbonne/paper.html.
54. B. Kessler, *Computational Dialectology in Irish Gaelic*, Proceeding of the 6th Conference of European ACL, Dublin **1995**, 60-66.
55. W. Heeringa, C. Gooskens, *Norwegian Dialects examined Perceptually and Acoustically*, Computers and the Humanities, Groeningen **2003**, 295-297.
56. J. B. Kruskal, *An Overview of Sequence Comparison*. In: Time Warps, String Edits and Macro Molecules. The Theory and Practice of Sequence Comparison, 2nd Edition, CSLI, D Sankoff, J Kruskal (eds.), Stanford **1999**, 1-44.
57. W. Viaregge, A. Rietveld, C. Jansen, *A Distinctive Feature Based System for the Evaluation of Segmental Transcription in Dutch*. In: Proceedings of the 10th International Congress of Phonetic Sciences, Dordrecht **1984**, 654-659.
58. M. Silvestri, G. Tomezzoli, *Linguistic Computational Analysis to Measure the Distances between Ancient Venetic, Latin and Slovenian Languages*. In: Proceedings of the Third International Topical Conference: Ancient Settlers of Europe, Ljubljana 10-11 June 2005, A Perdih (ed.), Založništvo Jutro, Ljubljana **2005**, pp. 77-85.
59. M. Silvestri, G. Tomezzoli, *Linguistic Distances between Rhaetian, Venetic, Latin and Slovenian Languages*. In: Proceedings of the Fifth International Topical Conference: Origin of Europeans, Ljubljana 8-9 June 2007, A Perdih (ed.), Založništvo Jutro, Ljubljana **2007**, pp. 184-190.
60. A. Perdih, G. Tomezzoli, V. Vodopivec, *Comparison of Contemporary and Ancient Languages.* In: Proceedings of the Sixth International Topical Conference: Origin of Europeans, Ljubljana 6-7 June 2008, A Perdih (ed.), Založništvo Jutro, Ljubljana **2008**, pp. 40-87.
61. A. Perdih, *Comparison of Some Methods of Estimation of Linguistic Distances*. In: Proceedings of the Eighth International Topical Conference: Origin of Europeans, Ljubljana 4-5 June 2010, A Perdih (ed.), Založništvo Jutro, Ljubljana **2010**, pp. 78-86, available at:
62. http://www.korenine.si/zborniki/zbornik10/perdih_ling_comparisson.pdf.

63. Ji Xianlin, *Fragments of Tocharian A* Maitreyasamiti-Nāṭaka of the Xinjang Museum, China, Trends in Linguistics, Studies and Monographs 113 Ed. Werner Winter, Mouton de Gruyter, Berlin, New York, **1998**.
64. K. T. Schmidt, *THT 1540* in Instrumenta Tocharica, ed. M. Malzahn, Universität Verlag WINTER, Heidelberg, **2007**.
65. A. Perdih, *Linguistic Analysis based on the Frequency of Sound Pairs and Triplet*s, published in this publication.

Abstract

The study of the Tocharian language started at the end of the 19th century because of the archaeological discoveries in the Chinese Turkestan or Xinjiang presented at the 12th International Congress of Orientalists in Rome (1899) and successively made by the Russian (I – V, 1899 – 1915), British (I – III, 1900 – 1916), Japanese (I – III, 1902 – 1909), German (I – IV, 1902 – 1914) and French (I, 1906 – 1909) archaeological expeditions. Because of these expeditions now collections of Tocharian documents are preserved by institutions of the corresponding countries. From said documents two kinds of Tocharian has been identified, indicated as East Tocharian or Tocharian A (TocA) and West Tocharian or Tocharian B (TocB). Surprisingly, TocA and TocB have been recognized as Indo-European languages of type Kentum.

TocA, also indicated as Turfanian, was spoken in the region of Turfan. The documents in TocA are principally of religious nature. TocB, also indicated as Kuchean was spoken up to the 9th cen. AD in the region of Kucha. The documents in TocB are of religious, commercial and daily life nature. The existence of TocA and TocB brought to the hypothesis that in the millennium BC in the Chinese Turkestan a single Tocharian language or Proto-Tocharian was spoken. Because of the recent discovery in the Tarim Basin of well-preserved mummies (1800 BC) having Caucasian somatic characteristics have brought to the hypothesis that said mummies belonged to Tocharian peoples.

The vectorial formalism in the Alphabetic Statistic Distribution Space, developed in this paper, indicates that TocA and TocB had a formation and evolution rather independent from the other languages considered and that mutual influences existed mainly between TocA, TocB, Slovenian, Oscan and Latin. However, in order to confirm this hypothesis and to better test our vectorial formalism we plan to consider in the future more Slavic, Celtic and Uralic-Altaic languages.