UNIVERZITA KARLOVA V PRAZE
Filozofická fakulta

DIPLOMOVÁ PRÁCE

UNIVERZITA KARLOVA V PRAZE
Filozofická fakulta
Ústav Blízkého východu a Afriky

DIPLOMOVÁ PRÁCE

Korpus orchonských runových textů
Corpus of Orkhon runic inscriptions

Filip Kaas

Vedoucí práce
Doc. Petr Zemánek 2017

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

Podpis:

Rád bych na tomto místě poděkoval za velkou pomoc, usměrnění i trpělivost při tvorbě korpusu doc. Petru Zemánkovi. Mé nemalé díky patří také Hande Terzi, Adamu Pospíšilovi, Jiřímu Dyndovi, Janu Křivanovi a Martinu Gálisovi za podnětné postřehy a náměty vzešlé z diskuzí s nimi. Děkuji také speciálně Veronice Zikmundové a Jonáši Vlasákovi za četné semestry strávené čtením starotureckých textů, bez nichž by tato práce jinak nevznikla. Díky si zaslouží také Béla Brogyanyi, Béla Kempf a Bayarma Khabgataeva, kteří mi umožnili dostat se k potřebné literatuře v Čechách jinak nesehnatelné. Jinak děkuji všem kolegům a členům Ústavu obecné lingvistiky a Ústavu Blízkého Východu a Afriky, především Petru Kučerovi za výuku tureckého jazyka, bez níž by tato práce byla poloviční.

F. K.

Anotace

Cílem diplomové práce bude vytvořit elektronický korpus staroturecких orchonských runových textů. Diplomant(ka) provede a zdůvodní výběr textů; objem textů dosáhne nejméně 30 tis. run. Dále navrhne model datové struktury, která zahrne propojení nápisů s jejich elektronickou podobou (včetně kódování run, otázek transliterace a transkripce), a také další roviny popisu. Navrhne řešení základních problémů segmentace, a to jak na větné, tak slovní a případně i morfotaktické úrovni. Funkční prototyp korpusu vhodnou formou zpřístupní, celou proceduru popíše v textu práce.

Klíčová slova

staroturečtina, orchonské runové nápisy, elektronický korpus

Abstract

The goal of the submitted thesis is creating an electronic corpus of Old Turkic Orkhon runiform inscriptions. Author will argue the choice of texts he made; the minimum volume of textual material will be at least 30 000 characters. Author will propose a model of data structure that will connect inscriptions with their electronic counterpart (including discussion of the following problems: encoding of runes, transliteration and transcription) and also various other levels of description. Author will propose solution for basic segmentation problems (on both sentence, word and morphosyntactic level). Pilot version of corpus will be made accessible and the whole procedure will be described in the text of the thesis.

Keywords

Old Turkic, Orkhon inscriptions, electronic corpus

CONTENTS

1 INTRODUCTION

This thesis is a part of my work on developing a corpus of Old Turkic which I have been pursuing during my studies of linguistics and Turkish studies at the Faculty of Arts of the Charles University in Prague in the years 2015-2017. In the following paragraphs I will summarize the goals and the contents of the thesis.

Although the title of the thesis suggests, that the primary focus of this work is to describe the process of building an Orkhon runiform corpus, this thesis has a much broader range of problems to address. They range from touching upon the data structure of the proposed corpus, through technical difficulties tied to creating a corpus of a rather exotic script in the modern digital era, to commenting on some of the current practices in the study of Old Turkic texts.

The study of languages in general is to a certain extent shaped by the limits of manipulation with texts. With the advent of digital era the ideal candidate to be the instrument-of-choice for linguistic and philological research seems to be electronic searchable corpus. As McEnery & Wilson (1996: 123) note already two decades earlier: *"...computerised resources and tools used to analyse them have become part of most research on historical linguistics today"*. The reasons are obviously the possibility of advanced annotation and the incomparable speed and efficiency of searching through data. All of this makes the corpus a tool which is beyond compare if set against any other classical instruments of text linguistics. For many languages the work with electronic corpus has become standard for the analysis of texts. Unfortunately, this cannot be said about Old Turkic language. Although a few electronic corpora indeed exist, a complete corpus of Old Turkic texts, and especially a corpus that would cover also the oldest available documents, is still missing. In this work I would like to pave a road towards building such a corpus and propose solutions to problems that emerge alongside this enterprise. These include among others the problem of encoding of the Old Turkic runiform script, transcription, segmentation and glossing of Old Turkic text, designing of the data structure for a multi-level corpus, and in the end creating a searchable electronic corpus. Another important part of the work, that must be accounted for, is processing of Old Turkic data from five inscriptions into this new corpus (operationalization of damage annotation, transcription, segmentation, and glossing).

The vision of this project is to produce a comprehensive corpus of Orkhon runiform text documents with rich annotation and strong tools to work with. Eventually, what is now an aspiration might evolve into a platform that would not be a mere instrument, but a single place where various ideas could meet and be compared. Machine-readable texts and modern day online access are powerful instruments, that would be shame not to harness to its full potential.

In the following chapter I would like to introduce the reader to the historical and social context of Turkic and Uyghur Kaghanates, Orkhon inscriptions, their language, runiform script, and various transcriptions of Old Turkic. In Chapter 3 I review other projects that digitalized Old Turkic texts and comment on the technical solutions - especially encoding, fonts, and keyboard layouts - in order to be able to work with the runiform script on a computer. In Chapter 4 that constitutes the main body of the thesis I describe the process of building the Orkhon Runiform Corpus. The sections included in this chapter focus on the choice of the initial set of inscriptions, design of the spreadsheet data structure, alignment, marking of damage, metadata, and the overall operationalization of the language data into the corpus. In the last chapter I propose, how should the end-product stage of the corpus work.

In many places throughout this thesis, there are examples of Orkhon texts in transcription as well as in runiform script. More details about the pronunciation, transcription, or the script itself are in sections 2.3.1-2.3.4.

2 A SKETCH OF THE HISTORY OF THE TURKIC KAGHANATES, ORKHON RUNIFORM INSCRIPTIONS AND OLD TURKIC LANGUAGE

𐰇𐰕𐰀𐰚𐰇𐰚 : 𐱅𐰭𐰼𐰃 : 𐰽𐰺𐰀 : 𐰖𐰍𐰔 : 𐰘𐰼 : 𐰶𐰞𐰣𐱃𐰸𐰟𐰀 : 𐰚𐰃𐰤𐰀𐰺𐰀 : 𐰚𐰃𐱁𐰃 : 𐰆𐰍𐰞𐰃 : 𐰶𐰞𐰣𐰢𐱁

“When the blue heavens above and the brown earth below were created, humankind was made between the two...” (KT E 1)

2.1 The Early history of the Turkic people

Writing the history of the Turkic people before 6th century CE can rely only on indirect evidence. According to Chinese sources (Liu Mai-Tsai 1958: 5) Turks were one of the tribes that were part of the Xiong-nu federation. They were allegedly occupying pastoral lands on the Chinese frontier and based upon lexical analysis, it has been proposed that the original *urheimat* of the Turkic people lied in the Manchurian region (Golden 2011: 35).

The first record of the name *Turks* can be found in Chinese chronicles (6th century CE) as 突厥 (rendered in modern Mandarin as *tūjué*). This word is reconstructed as **duətkuat* for Early Middle Chinese (600 CE) and it is thought to represent the word *türküt* (plural of *türk*) (Golden 2011: 20).

The first appearance of the Turkic people on the stage of history dates back to the middle of the 6th century CE. They inhabited a region near the Altay mountains and were in a vassal relationship to Rourans. The dissolution of Tuoba state and the emergence of Western and Eastern Wei lead to an alliance between Western Wei and the tribe of the Turks. In 552 Bumin Kaghan from the Ashina clan rebelled against the Rouran Kaghanate, assumed leadership amongst the other local tribes and founded the First Turkic Kaghanate centered in today’s Mongolia.

The empire grew by dominating neighbouring tribes and controlling the Silk Road trade. From Ötüken, considered the holy land and centre of the kaghanate, laying in proximity to the Orkhon river (the place where capitals of many other successor states were, e.g. Karakorum), Turkic kaghans ruled a vast empire reaching west as far as the Sassanid Empire in Transoxania, the Byzantine Empire in Crimea and neighbouring with the Chinese Empire in the south.

The Turkic people were mostly herdsmen seeking new pastures. The prevailing religion was Tengrism, in which 𐱅𐰭𐰼𐰃 *teŋri* ‘the heaven’ was worshipped as the main deity. The religion incorporated many shamanistic practices as an important part of the cult, while being under long-term influence of Buddhism. In 584 two pretenders started a civil war, that lead to a split of the Turkic Kaghanate in two parts. The western part had its capital in Suyab, in today’s Kyrgyzstan and eastern part kept its capital in Ötüken. Due to the skilled horse archers who according to

Grousset (1970: X) gave the Turks the “*technical arm, that gave [them] almost as great an advantage over sedentary man as artillery gave modern Europe over the rest of the world”*, the situation in the northern steppes was one of the defining factors for the Chinese politics. In 630 the weakened Eastern Kaghanate was vassalized after a successful military campaign by the Tang dynasty.

After 50 years of subjugation a leader named Ilterish Kaghan revolted against Chinese sovereignty and established the Second Turkic Kaghanate in the year 682. Soon the Kaghanate gained control over the steppes and clashed with the expansion of Umayyad Caliphate in Transoxania. The deeds of Bilge Kagan and Kül Tegin, sons of Ilterish, were carved into stelae and comprise one of the most important Old Turkic texts, as will be seen in section 4.1.1.

Some scholars (Kljaštornyj 1994) assume that the ruling dynasty of Ashina were originally of Indo-European (Sogdian) origin and connect the name to Khotan-Saka *āşşena* ‘blue’, Sogdian *’γs’n’k* ‘green’ or Tokharian *âśna* ‘blue’ and some have also pointed towards the name *kök türk* ‘Blue Türk’, that appears in Bilge Kagan and Kül Tegin inscriptions (BK E 4, KT E 3) and suggested, that it is an Old Turkic translation of the name Ashina Türk (Golden 2011). The status of the name Göktürk (Celestial Turks) is still hotly debated (cf. Tezcan 1990) and the majority of scholars now doubt that it comprised an ethnonym.

The Sogdians played important role in the state as the kaghanate was dependent on skilled administrators with background in sedentary cultures. The trade connections and multilingualism of Sogdians allowed them to serve as Chinese-Turkic interpreters in the Tang Empire (Bahry 2016: 15). This influence is manifested well enough, if we look at the inscriptions from the First Kaghanate, which are in fact written predominantly in the Sogdian language and script (the most famous Bugut inscription is dated to 581 CE, cf. Kljaštornyj & Livšic 1972, Yoshida & Moriyasu 1999, Alyılmaz 2003).

In 744 the balance of power changed and an alliance of three Turkic peoples, Uyghurs, Basmyls and Karluks, seized power and the Uyghur leader Kutlug Bilge Kaghan eventually became founder of the Uyghur Kaghanate, which can in many respects be considered as a successor state of the Second Turkic Kaghanate. In 763 Tengri Bögu Kaghan changed the state religion of the Uyghur Kaghanate to Manicheism.

2.2 Orkhon inscriptions

The Orkhon inscriptions are the earliest-known texts written in any Turkic language whatsoever. The oldest text(s) can be dated back to the late 7th century CE, but the most important originated in the 8th century CE. They are named after the river Orkhon, where the first sizable inscriptions (*Kül Tegin* and *Bilge Kaghan*) were discovered by Nikolaj Jadrintsev's expedition in 1889. They are written in the so called runiform script. The distribution of these inscriptions is not limited to the Orkhon river basin, but covers also basins of other Mongolian rivers (Selenge, Tuul, and others) and areas without surface water (Gobi desert).

Besides the inscriptions found in the Orkhon area, textual artifacts written in nearly the same language and script are found all over southern Siberia up to the Tien Shan mountains and Ferghana valley. These artifacts they are usually called after the area in which they are found - Yenisei inscriptions, Altay inscriptions or Talas inscriptions. The first of the Yenisei inscriptions were discovered for Europe actually already during the years 1721-22 by Strahlenberg and Messerschmidt (E 31 Uybat III). In 1907 sir Aurel Stein made a discovery at the Mogao Caves near Dunhuang (Gansu province, China), where he found among other things manuscripts written in Old Turkic language and the runiform script.

New inscriptions are still being discovered, among the more important ones being for example multiple inscriptions found in the Republic of Altay, the inscription of Bombogor (Mongolia) discovered in 2004, and the latest flashnews - the inscriptions of Chang'an (Xi'an, China) and Sükhbaatar (Sükhbaatar aimag, Mongolia) found in 2013, which still do not have any published edition.

For complete list of inscriptions cf. Kempf (2004), Sertkaya (2008). For detailed information about Altay inscriptions see Tybykova, Nevskaya, Erdal (2012). Latest edition of Yenisei inscriptions is Aydın (2011).

The term Orkhon inscriptions is used in multiple ways. Sometimes it denotes only the three most important inscriptions (Bilge Kaghan, Kül Tegin, Tonyukuk), eventhough Tonyukuk inscription lies more than 300 km away from Orkhon river. I prefer to think of the Orkhon river basin as the symbolic center of the textual production and I use the name Orkhon inscription as a synekdoche referring to all the Old Turkic textual artifacts found in the whole of today's Mongolia.

The uniqueness of the inscriptions from Orkhon resides in the fact, that the texts are carved in large stone stelae, which usually constitute a part of a larger memorial complex. These texts are comparatively longer than inscriptions from Yenisei or Altay and are especially valuable thanks to the information that they provide about the history, culture and language of the Turkic society at that time. From the linguistic point of view the significance of these memorials lies in the fact, that these are preserved as the oldest documents written in a Turkic language and give an insight into many aspects of the history of the Turkic language family.

Besides fragmentary grafitti and texts on stones, coins, or tamgas (sealers), the largest volume of the Orkhon inscriptions are epitaphs carved in memory of political elites of the Turkic and Uyghur Kaghanates (kaghans, generals and other officials).

2.3 Old Turkic language

The Old Turkic language is the oldest Turkic language for which there exist preserved linguistic data. Aside from one short sentence found in the Chinese sources from the 4th century CE, continuous written tradition of Old Turkic starts in the 8th century at the latest. Turkic languages were during these centuries already in their prime supplanting Indo-European languages in the area of Central Asia. Contacts with Chinese and Indo-European languages can be demonstrated on multiple personal names and titles borrowed from Chinese and Sogdian, as well as other loanwords (e.g. Old Turkic *tümen* '10 000' from Tocharian B *tumane*, Old Turkic *öküz* 'ox' from Tocharian B *okso, or* Old Turkic *kunçuy* 'spouse' from Chinese).

Once in a while the idea referred to as the Altaic hypothesis (claiming that Turkic, Mongolic and Tungusic, and sometimes also Korean and Japanese languages are genetically related) reappears in academic discussions, but since the end of 20th century, the hypothesis has been heavily criticised and sees less and less acceptance (for summary of discussions on the Altaic macrofamily see Vovin 2005).

Eventhough Old Turkic is the oldest Turkic language and possibly also does not fundamentally differ from Common Turkic (the common ancestor of Turkic languages), Old Turkic is considered a dead end in the Turkic dendrogram. According to some scholars the genetically closest living branch are the Oguz languages (Turkish, Azeri and Turkmen), the speakers of which migrated to the Transoxanian region during the 5th and 6th centuries.

Erdal proposes to call Old Turkic the language which is constituted by material underlined by three following corpora (Erdal 1998, 138; 2004, 6-10):

1) Old Turkic runiform texts (since 8th century CE): era of the Second Turkic Kaghanate, the Uyghur Kaghanate and the Kyrgyz Kaghanate. They comprise over 200 texts of largely fragmentary texts from Central Asia and South Siberia (part of them are difficult to decipher).

2) Old Uyghur texts (since 9th century CE), mostly discovered in Xinjiang and Gansu provinces in China. Approximately 75% of these texts are comprised of Buddhist literature, rest is Manichean, Nestorian and non-religious literature. Large part of this corpus are translations from other languages.

3) Karakhanid texts (11th century CE). Two most important texts are *Kutadgu Bilig* by Yusuf Balasaguni and *Dīwān Lughāt at-Turk* by Mahmud Al-Kashgari.

It should be noted that not a pair of these corpora are linguistically homogeneous. On the other hand the variability in grammar and phonology among these three corpora is not necessarily bigger than variability within each single corpus (Erdal 2004, 11). As Erdal further notes:

“The three corpuses mentioned above represent a coherent group of fuzzy dialects differing most in the lexicon (as they belong to different cultural domains), certainly also in morphology and in some ways also in phonology. Syntactic differences may in part be due to the fact that the corpuses contain different textual types, but also reflect the gradual Turkification of much of the population using Uygur, and historical development. Translations, which constitute most of our corpus 2 (though by no means all of it), were, in particular, carried out by bilingual committees.”

In this work I will use two names for the language of Orkhon inscriptions. Old Turkic will refer to the general variety, for which there is enough data to produce a reasonable grammar, while the term Orkhon Turkic will be used when the language variety will be considered as a counterpart to other more specific language varieties especially Yenisei Kyrgyz, and Old Uyghur.

2.3.1 Sketch description of Old Turkic language and phonology

This chapter includes only a general outline of the grammatical properties of Old Turkic. Old Turkic grammar will be discussed in more detail in sectoins 4.3.4 and 4.3.5, where I will focus on the glossing of the texts. The bulk of this chapter will be Old Turkic phonology, while I will slowly drift towards how the sounds were represented by the runiform script.

Typologically, Old Turkic corresponds to what a type of language that is traditionally called agglutinative. The language does generally not cumulate morphemes, has no inflectional classes, only a few morpheme alternations and shows almost no suppletion (the exception being negated forms of some participles). The language has vowel harmony, although this feature is not as developed as it is the case in most of the modern Turkic languages. The prevailing syntactic order is head-final (the language has postpositions, suffixing, SOV word order). For a detailed description of its grammar see Tekin (1997), Erdal (2004).

The language had 8 vowels, that can be put in the classical three-fold vowel symmetry: back (a, ı, o, u) vs. front (e, i, ö, ü), unrounded (a, ı, e, i) vs. labialized (o, u, ö, ü), and high (i, ü, ı, u) vs. low (e, ö, a, o).

	Front		Back	
	lab-	lab+	lab-	lab+
High	i [i]	ü [y]	ı [ɯ]	u [u]
Low	e [ɛ]	ö [ø]	a [ɑ]	o [o]

Table 1: Vowel Harmony.

Besides the eight vowels in Table 1 there is some evidence for vowel /ė/ representing close-mid front vowel [e]. The evidence stems from the Yenisei insciptions, which have a distinct grapheme for this sound. In Orkhon inscriptions this vowel is usually written with the same grapheme as /i/ and Erdal (2004, 45) considers it an innovation that appeared at some stage of Old Turkic (probably still not during 8th century). The distinction between /e/ and /ė/ is highly contested as no modern Turkic language expresses this opposition in script. Full set of 8 long vowels (a:, ı:, e:, i:, o:, ö:, u:, ü:) is reconstructed for Proto-Turkic, but in the case of Orkhon inscriptions, there is only a handful of examples of words with long vowels.

The reconstructed consonant inventory of Old Turkic (Table 2) is straightforward. Some of the phonemes might have had front and back alophones, especially velar and uvular /k/, and /g/ (Erdal 1998: 139-140, 2004: 62). There is b - v alternation with [b] realization at the word onset and [v] in the rest of the positions with the exception of some words (cf. Ölmez 2015b: 683-685).

	labial	dental	post-alveolar	palatal	velar
nasal	m	n		ɲ	ŋ
stop	p b	t d			k g
trill		r			
fricative	v	s z	ʃ		
affricate			t͡ʃ		
approximant		l		j	

Table 2: Consonant inventory.

2.3.2 Old Turkic runiform script

In this chapter I will show, how Old Turkic sounds were represented in script, while technical difficulties tied to using Old Turkic script will be discussed in section 3.2.

The label *runiform* script goes back to the 19th century, when the script in which Old Turkic was written, was still not deciphered and some believed, that because of its superficial resemblance with the Germanic runic alphabet, there was some sort of genetical relationship between the two. This has been proven to be the untrue as the resemblance can be easily explained by the writing technique used for this script - carving into stones - that encourages the tendency towards using certain shapes of letters.

The Old Turkic runiform script was deciphered on the basis of Kül Tegin and Bilge Kagan inscriptions by Vilhelm Thomsen in 1893. He correctly guessed the language as Turkic and used the Chinese inscription on the western side of Kül Tegin stele to identify first words.

The origin of Old Turkic runiform script is still uncertain (Clauson 1970, Róna-Tas 1998b). There are currently two dominant hypotheses. The first one, proposed already by Thomsen himself, assumes that runiform script is a derivation from a script with Aramaic origin. Some scholars propose Sogdian (e.g. Coulmas 1999: 512) as a plausible source, while yet another viable option is Kharosthi, discussed especially in connection with the Issyk inscription (Harmatta 1999: 521). If the script indeed was transmitted from other language, it was well adapted to the phonology of Old Turkic language. The second hypothesis accounts for the origin of the script calling it an autonomous innovation (there is surprising difference between the structure of the runiform script and any Aramaic-based script), Mallitskij (1897) proposes, that the script is developed from Turkic *tamgas* (seals). Some of the characters were proposed to have iconic meaning, e.g. 𐰸 *ok* “arrow”, 𐰋 *eb* “house (tent)”, 𐱃 *at* “horse”. These might hypothetically have belonged to the base set of single syllabic logographs, from which the script might have evolved similarly to Arabic alphabet.

This section will concentrate on the script itself. In Tables 3-6 there is a list of runiform characters, together with their transliteration (used by Tekin (1995), notice that some characters are transliterated as ligatures and are underscored), my transcription (that I will comment on in more detail later in section 2.3.4) and their reconstructed pronunciation in IPA. There are four graphemes for vowels (𐰇 , 𐰆 , 𐰃 , 𐰀), and they do not follow the same lines of vowel harmony. The rounded vowels are divided to high (𐰃) and low (𐰀), while the unrounded vowels are divided to front (𐰇) and back (𐰆).

The most interesting feature of the runiform script is deploying two sets of characters (Table 4) to mark the same consonant phonemes, but in combination with different vowels. One set of characters is used with back vowels, second set with front vowels. The script is more-or-less alphabetic meaning that there are means to encode every single sound by itself, but the consonant characters have intrinsic vowel associated to it in majority of their occurrences, thus the script shows some features of abugida. There are thus multiple ways of reading a single consonant character (with vowel preceding, superseding, or absent). Usage of the feature of two consonant character rows corresponds functionally to the absence of marking of vowels in some positions (eventhough the application of this rule is not stable).

character	transliteration (Tekin 1995)	transcription	sound (IPA)
𐰀	A	a / e	[ɑ] / [ɛ]
𐰃	I	ı / i / ė	[ɯ] / [i] / [e]
𐰆	U	u / o	[u] / [o]
𐰇	Ü	ü / ö	[y] / [ø]

Table 3: Vowel letters.

𐰉	B	ab, b, ba	[b]	𐰋	b	eb, b, be	[b]
𐰑	D	ad, d, da	[d]	𐰓	d	ed, d, de	[d]
𐰍	G	ag, g, ga	[g]	𐰏	g	eg, g, ge	[g]
𐰞	L	al, l, la	[l]	𐰠	l	el, l, le	[l]
𐰣	N	an, n, na	[n]	𐰤	n	en, n, ne	[n]
𐰺	R	ar, r, ra	[r]	𐰼	r	er, r, re	[r]
𐰽	S	as, s, sa	[s]	𐰾	s	es, s, se	[s] / [ʃ]
𐱃	T	at, t, ta	[t]	𐱅	t	et, t, te	[t]
𐰖	Y	ay, y, ya	[j]	𐰘	y	ey, y, ye	[j]
𐰴	K	ak, k, ka	[k]	𐰚	k	ek, k, ke	[k]

Table 4: Back and front consonants.

𐰲	ç	aç, eç, ç, ça, çe	[t͡ʃ]
𐰢	m	am, em, m	[m]
𐰯	p	ap, ep ,p	[p]
𐱁	ş	aş, eş, ş	[ʃ]
𐰤	z	az, ez, z	[z]
𐰭	ng	aŋ, eŋ, ŋ	[ŋ]
𐰪	ny	añ, eñ, ñ	[ɲ]

Table 5: Equivocal consonants.

𐰸	oK	ok, uk, k, ko, ku	[k]
𐰜	ök	ök, ük, k, kö, kü	[k]
𐰶	ıK	ık, kı, k	[k]
𐰱	iç	iç	[t͡ʃ]
𐰨	NÇ	anç, enç, nç	[nt͡ʃ]
𐰦	NT	ant, ent, nt	[nt]
𐰡	LT	alt, lt	[lt]

Table 6: Consonants with intrinsic vowel and double consonants.

2.3.3 Rules for writing Old Turkic runiform

In this section I will formulate a system accounting for the represantation of vowels in Old Turkic language. It is based on observation of others (Tekin 1995) and my experience. I propose 3 rules concerning writing of vowels. These rules are not in any way absolute, or predictive, but they work in majority of cases. For example one of the most frequent word *bodun* ‘people, tribe’ is written mostly as 𐰉𐰆𐰑𐰣 BUDN (KT E 14), but we find also 𐰉𐰆𐰑𐰆𐰣 BUDUN (KT E 14) and accusative version 𐰉𐰑𐰣𐰍 BDNG (KT N 7) without any vowel letters whatsoever. The word *üküş* ‘lot, majority’ is once written as 𐰇𐰚𐱁 Ükş (BK N 7), once as 𐰜𐰇𐱁 ökÜş (KT E 10). The following three rules work as an algorithm, if one wants to ‘predict’, how a word was written, it is necessary to take transcription of the given word, look at the first syllable, then look at all the following syllables, and finally check the last syllable.

Rule about the first syllable: The vowel letter 𐰀 (/a/, /e/) is not written in the first syllable 𐰴𐰺𐰀 KRA *kara* ‘black’, 𐱅𐰢𐰼 tmr *temir* ‘iron’. If there is any of the other vowels (/i/, /ı/, /ö/, /o/, /ü/, /u/), they are represented by their respective letters 𐰇𐰔𐰀 ÜzA *üze* ‘above’, 𐱅𐰇𐰼𐰚 tÜrk *türk* ‘Turkish’. If there is letter 𐰀 (/a/, /e/) written in the first syllable, it means that the vowel is long 𐰀𐱃 AT *āt* ‘name’, compare to 𐱃 T *at* ‘horse’.

Rule about the following syllable: If there are two consequent vowels of the same labialization, rounded after rounded, unrounded after unrounded (analogy to vowel harmony), the second vowel is not marked 𐰴𐰃𐰞𐰣𐰢𐱁 KILNmş *kılınmış* ‘created’, 𐰴𐰍𐰣 KGN *kagan* ‘kaghan’, 𐰆𐰞𐰺𐰯 ULRp *olurup* ‘sitting’. Otherwise the second vowel is marked 𐰚𐰃𐰚𐱁𐰇𐰼 kIkşÜr *kikşür* ‘incite’, 𐰲𐰇𐰢 çÜm *eçüm* ‘my ancestor’.

Rule about the last syllable: If the word ends in vowel, it is marked 𐰴𐰺𐰀 KRA *kara* ‘black’, 𐰾𐰇 sÜ sü ‘army’.

The system of consonant writing is more complex. There are words that have multiple ways of being represented in script (in Tables 4, 6 we can count five characters that represent phoneme /k/). For example the name *Kül Tegin* is written in three different ways 𐰜𐰇𐰠𐱅𐰃𐰏𐰤 ökÜltIgn (KT E 26), 𐰚𐰇𐰠𐱅𐰃𐰏𐰤 kÜltIgn (KT E 27), and 𐰚𐰇𐰠𐱅𐰏𐰤 kÜltgn (KT N 8).

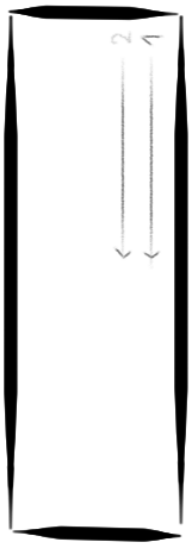
The back consonants are used to mark consonants in syllables that have back vowels and front consonants are used to mark consonants in syllables that have front vowels 𐰉𐰺𐰍𐰢𐰀 BRGmA *barıgma* ‘going’ (KT E 23), 𐰋𐰼𐰏𐰢𐰀 brgmA *bėrigme* ‘giving’ (BK E 21). The letters from table XXX are used irrespective of the front and back distinction. The front 𐰾 s consonant letter is used to mark both /s/ and /ş/ sounds 𐰾𐰃𐰤 sIn *sen* ‘you’ (KT S 8), 𐰚𐰃𐰾𐰃 kIsI *kişi* ‘person’ (KT S 7).

The phoneme groups /ok/, and /uk/ are usually represented by character 𐰸 oK and the phoneme groups /ök/, /ük/ by character 𐰜 ök. For example 𐰇𐱅𐰜𐰤 Ütökn *ötüken* ‘Ötüken (placename)’ (KT S 3). If the letter 𐰸 oK is used after the letter 𐰆 U, it means that the vowel /o/, or /u/ is long 𐰖𐰆𐰸 YUoK *yōk* ‘not existing’ (KT E 39). The letter 𐰶 ıK can be used only for group of phonemes /ık/ in the middle or at the end of the word 𐰔𐰶𐰪𐰀 zıKnya *azkıña* ‘a little bit’ (KT E 34).

Letter iç is used only for the phoneme group /iç/ at the beginning of the word 𐰱𐰚𐰓𐰃 içkdI *içikdi* ‘was related to’ (BK E 37). The letter 𐰡 LT is used only in syllables with back vowels ın order to mark group of phonemes /lt/ 𐰉𐰆𐰡𐰃 BOLTI *boltı* ‘he/she was/became’ (BK E 37). Geminates are usually represented by only one letter 𐰋𐰇𐰚𐰠𐰃 bÜklI *bökküli* ‘Korean’ (BK E 8), but compare 𐰸𐰆𐰉𐰺𐱃𐰢 oKUBRTm *kuvratdım* ‘I gathered’ (BK N 7) and 𐰸𐰆𐰉𐰺𐱃𐰑𐰢 oKUBRTDm *kuvratdım* ‘I gathered’ (KT E 10).

The only punctuation character are two dots on top of each other (the character is similar to a colon). It is used to mark a boundary between words or syntactical phrases, but it is difficult to find any exact consistent patterns. As Rybatzki (1999: 220) notes: “*...one cannot avoid the impression that, for some details, every inscription has its own rules of punctuation.*” There is no other marking present in the script, that would be used for marking the boundaries of longer syntactical units (for example sentences).

The letters of the runiform script were written from right to left in rows running from bottom to top (see Fig.1). This is basically the same way that Chinese letter were written, except the letters being rotated by 90 degrees (Fig.2). The Chinese cultural influence in Central Asia was immense, influencing writing direction of more literary cultures like the Sogdian (Novák 2016: 48).



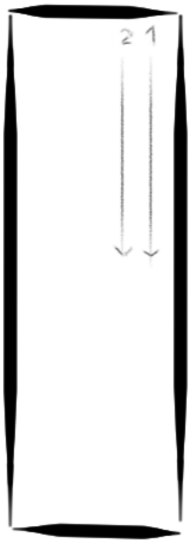


Figure 1: Old Turkic writing direction Figure 2: Chinese writing direction

In 20th century two damaged manuscript pages with partial alphabet listings were discovered in Xinjiang. They are named Toyok and Ryukoku and they indicate that Old Turkic might have had a “standard” alphabetic order.

Apart from the runiform script, the Old Turkic language has been written in number of other scripts. Analysis of the orthography of Old Turkic documents written in Old Uyghur, Arabic, Manichaean, Syriac, Sogdian, Brahmi, Tibetan, or Phagspa together with scant evidence from Chinese and Greek sources, offer some clarity on the phonology of the language.

The Yeniseian variety of the runiform script (for complete table of characters see von Gabain 1941: 12,) was used from 9th to 11th century by Yeniseian Kyrgyz. The script has different graphemes for a majority of phonemes (cf. Everson 2008: 20), but otherwise the script stays structurally similar to the Orkhon runiform. The only exception is the special letter (𐰅) used to represent the phoneme /ė/ in Yeniseian inscriptions. In Orkhon inscriptions this phoneme is mostly represented by the letter 𐰃 I.

2.3.4 Transcription of Old Turkic language

“...our knowledge of the phonetics, particularly, of early Turkish, is so imperfect that it would be foolish to use anything more scientific than a very simple transcription alphabet, sufficiently refined to ensure that each letter represents a sound or sounds distinct from those represented by any other letter, but not so refined as to provide separate representation for sounds so close to one another that there is really no means for determining which of them should be used in particular case.” (Clauson 1962, 34)

More than 120 years passed already since the first editions of Orkhon Turkic inscriptions have been published. Surprisingly enough the number of different transliterations and transcriptions, that are used in Old Turkic studies, is high. After I tried to look up all the various transcription systems to be able to decide, what features are marked and why, I daresay that the number of systems might be actually higher than the number of scholars working on Old Turkic language (since different transcriptions are used in different publications sometimes). It is only natural, that a transcription system changes as deeper understanding of the language is acquired. But I believe that high amount of various transcription systems is not necessary.

To tackle the problem more closely I excerpted transcription of the first line of the Ongi inscription from three different contemporary editions (Erdal 2010, Berta 2010, Ölmez 2015):

(ä)çüm(ü)z : (a)pam(ı)z Y(a)ma : q(a)γ(a)n : tört b^{2}ul(u)ŋ(u)γ : (e)tm(i)ş : yıγm(ı)ş : y(a)y(ı)m(ı)ş : b(a)s^{2}m(ı)ş : ol q(a)n yo°q : bolt^{u}qda : k(e)srä : (e)l yitm(i)şi : ç(ı)γ(a)ñ ... q(a)z[γ](a)nm(a)d^2(ı) [m(ı)z] : (e)l(lä)d^{ü}k (e)l(i)n :

eçẅmẅz apamız yamı qaγan tört bulwñwγ ėt°miş yıγmış yaymış bas°mış ol qan yoq bolɒwq°δa kės°re ėl yit°miş ıç°γınmış q... r...

ėçümiz : apamız : yamı : kagan : tört : buluŋug : ėtmiş : yıgmiş : basmiş : ol kan yo°k : boltokda : kėsre : ėl yitmiş : ıçgınmış : k^1... r^2....

Before I comment on the three transcription styles in particular, I want to mention, what function I believe the transcription level should fulfill, either in a classical edition or the Orkhon runiform corpus. The transcription should represent reading of a particular word to the best of our knowledge at the current point in time. There are of course no means to retrieve direct informations about how the language was pronounced. That is why transcription is based naturally on the reading of the characters of the inscription, but the transcription itself should not attempt to give information about how the word was written, but it should confine itself only to giving information about the phonological shape of the word.

First point about Erdal’s transcription system is, that it fulfills more roles, than it needs to. It marks every sound, that is ‘missing’ in the script by enclosing it in brackets. This practice escalates visually up to the point, that the transcription is difficult to read. The upper index b^2 in the word b^2*ul(u)ŋ(u)γ* again marks use of the front consonant character in a word, that consists of predominantly back consonant characters. This does not gives us any information, about the phonology of the word, but about the way the word was represented in the script. And as such should be part of transliteration, but not transcription.

The second transcription system (by Arpád Berta) distinguishes between front and back non-labial plosives (t - ɒ, d - δ, k - q, g - γ) and postulates two other vowels marked as *w* and *ẅ* (mid labialized back and front vowel). The system is more careful, but it would be difficult to decide, if there would be enough evidence for these distinctions in modern Turkic languages.

The last transcription system used by Mehmet Ölmez is certainly a step ahead in being more simple and reader-friendly, and does not leave the interpretation upon the reader. Similarly to the other three transcription systems the characters are based on modern Turkish alphabet. I follow Clauson in his preference to use transcription based on modern Turkish alphabet (1972: vii). But besides his argument, that Old Turkic phonetic system was most likely very similar to modern Turkish, I would argue, that it is a pragmatic decision to base this transcription system on the script, that is used by the most numerous group of Turkic speakers, and is also the alphabet of a country that has multiple research institutions working in Old Turkic studies.

3 OLD TURKIC AND COMPUTERS?

In this section I will address at first the previous attempts (successful) to create electronic corpora of Old Turkic texts. Secondly I will comment on what technical problems might one encounter with rendering Orkhon runiform script on computer screens and how to solve them.

3.1 Earlier Old Turkic corpora

There are currently three corpora of Old Turkic texts, that are intended for academia. In this section I present basic information about the three projects.

3.1.1 VATEC (Vorislamische Alttürkische Texte: Elektronisches Corpus; Erdal, Gippert, Röhrborn & Zieme 2003)

The first digitalization project of Old Turkic texts is the project VATEC. It was created during the years 1999-2003 under the leadership of Marcel Erdal (Frankfurt), Jost Gippert (Frankfurt), Klaus Röhrborn (Göttingen) and Peter Zieme (Berlin). The corpus consists of many Buddhistic (e.g. Altun Yarok, Maitrisimit, Xuanzang biography), Manichaean (Chuastuanift), Nestorian texts and also includes Book of Omens (Irk Bitig). All the texts are translated into German or English, they are morphologically segmented and glossed. The corpus does not make use of runiform characters, all texts are transliterated instead. This is understandable as the situation around historical fonts and encoding was very different in 2003 compared to current situation. The whole corpus is online accessible and the data can be investigated also through a search engine interface. It enables user to search directly through the linguistic material, as well as the metadata.

Another two corpora are corpora in more-or-less philological sense. They evoke the classical edition of texts with additional features enabled by computer. Both of these two projects are online accessible, but lack any query interface (without which the data are "trapped" on individual pages).

3.1.2 Altai Corpus (Tybykova & Nevskaya 2013)

Altai Corpus is project that was developed in the years 2003-2013. The leaders of the project are Irina Nevskaya and Larisa Tybykova. The focus of this corpus was to collect and document Old Turkic runiform inscriptions from the Republic of Altay. It contains around 100 localities. The texts are usually short (couple words on average), they are transliterated, transcribed, translated and commented, and contain also edited runiform text (in form of a picture). Every inscription includes informations about the locality, history of research, readings of different researchers and high-resolution photographs.

3.1.3 DTRI (A Database of Turkic Runiform Inscriptions; Károly & Rentzsch 2017)

The latest addition in the family of Old Turkic corpora is DTRI. The project started in 2015 and it is developed by László Károly and Julian Rentzsch. The database aims to provide an edition of all runiform inscriptions. Until now 7 inscriptions are accessible on the website (short inscriptions from Tuva). Aside from the basic information about the inscription and locality, there are transliteration, transcription, and translation levels, as well as comments and photographs.

I will make use of multiple solutions from the VATEC corpus in the section 4.3.4-4.3.5 about segmentation and glossing of the data. All three corpora were an inspiration for creation of the list of metadata (section 4.3.7).

3.2 Handling runiform script

One of the important levels of every inscription are the original runiform characters. In earlier publications the common way to represent Old Turkic characters was by means of transliteration (rendering inscriptions from left-to-right and top-to-bottom). Since the Unicode initiative proposal for encoding Old Turkic characters in 2008, and the following implementation of the Old Turkic runiform script in the version 5.2., the need to represent runiform characters in transliteration decreased, as one important obstacle ceased to exist. Old Turkic has now own dedicated Unicode block located in range from U+10C00 to U+10C4F. This block consists of 73 characters designed to represent both Orkhon and Yeniseian character sets of the runiform script. The punctuation sign is encoded as U+205A (named *two dot punctuation*) and strictly speaking it is different character than a colon.

Having Old Turkic characters as unique codepoints opened many options for using the Old Turkic runiform on computers. With proper rendering of the characters on the screen secured by installed fonts, there is no reason not to use Old Turkic runiform in the same way as any other script.

3.2.1 Fonts

User has to install one of the available fonts, that includes characters for Old Turkic runiform. As of 2017 I have found only a few fonts with filled Old Turkic code points. The fonts available may differ for particular distributions of operation system. Distributions of Windows 7 - 8.1 have a pre-installed font named *Segoe UI Symbol*. Since version Windows 10, the support for Old Turkic has been moved from *Segue UI Symbol* to *Segoe UI Historic*.

None of the standard fonts that are part of Linux distributions is able to render Old Turkic characters. The solution is either downloading a shareware font *EversonMono* (at: http://www.evertype.com/emono/). Or downloading an opensource font *Quivira*. *Quivira* has but one cosmetic disadvantage. It is a serif font, and its serif runiform characters have very ahistoric appeal, and the inaccuracy feels deceptive in some cases, compare the rendering in Example 1 with the version presented at the beginning of the Chapter 2.

(1) 𐰭𐰤𐰀𐱅𐰭𐱅 : 𐰚𐰇𐰠𐰞 : 𐰽𐰇𐰀 : 𐰖𐰃𐰓𐰤 : 𐰆𐰠 : 𐰍𐰞𐰀𐰋𐰀𐰑𐰀 : 𐱅𐰞𐰤𐰀𐰇𐰀 : 𐱅𐰞𐰃𐰞 : 𐰸𐰃𐰓𐰀𐰞 : 𐰍𐰞𐰀𐰣𐰢𐰃

3.2.2 Keyboard layouts

Once researcher is able to properly render characters on screen, the question, how to write in Old Turkic runiform presents itself. Without any instruments, there is only a clumsy option of “adding special character” in text editors, or ‘copy paste’ runiform characters from some other source. In order to be able to work with Old Turkic texts, I created two keyboard layouts for encoding runiform characters, one for Windows operating system and one for Linux distributions. Both versions have the same key mapping. They are devised to encode runiform characters and edit runiform texts. They are based on *Turkish Q keyboard layout* and I paid special focus to map keys in a mnemotechnic manner to ensure user-friendliness. The layout exploits the possibility of encoding back and front consonant characters by upper case and lower case letters (using the Shift key), while other basic text operations (copy/paste, undo action) are not hindered. More detailed information, keyboard layout files, map list, and instructions for installation are to be included as an addendum to this paper and will be part of the online electronic corpus.

I have created also a keyboard layout for transcription of Old Turkic. The basic characters stay the same as in Turkish Q keyboard layout (that is capable of writing all Turkish as well as English characters) and as an addition I included characters, that are used to transcribe Old Turkic. This ensures, that user does not have to switch between different keyboard layouts, when he writes in a set of these languages at once. I aimed to provide a keyboard layout, that would be able to code all the characters from different transcription systems, and some characters used for transliteration (e.g. upper index numbers).

4 BUILDING ORKHON TURKIC RUNIFORM CORPUS

4.1 Choosing Inscriptions

There are more than 500 discovered documents written in the Old Turkic runiform (Sertkaya 2008: 26). From this number only a fraction are inscription found in the area of today's Mongolia. Numbers from recent listings of inscription from Mongolia are the following: Alyılmaz (2003) lists 79 inscriptions, Kempf (2004) lists 43 inscriptions, Sertkaya (2008) lists 88 inscriptions. The exact number of Orkhon inscriptions is difficult to obtain, as some of the inscription, that are listed by some do not have any text, are lost, or the data are just not available. Sometimes multiple inscriptions from one locality are counted in various manners (e.g. where Kempf (2004: 43) counts one inscription of Açit Nuur, Alyılmaz and Sertkaya (2008: 2246) counts two different, simply labeled as Açit Nuur I, and II). The situation is further complicated, because some inscriptions are often known under several different names (the most famous being probably Şine-Usu / Moyun Çor / Selenge inscription). The inscriptions are usually named 1) after the person, that the inscription was erected for (Kül Tegin, Tonyukuk), 2) the place, where the inscription was found (Bombogor, Sükhbaatar), or 3) a close by river (Ongi, Selenge).

The longest and most famous inscriptions are the following trio Kül Tegin (KT), Bilge Kagan (BK) and Tonyukuk (T). Before the latest discoveries in 2013 were made (especially reports about Sükhbaatar stelae seem very promising), these three inscriptions accounted for 2/3 of the data from Mongolia. Together they consist from approximately 26 000 characters. All of these three inscriptions are usually dated to the decade form 725 to 735 CE. They are relatively well conserved and certainly most-studied Orkhon inscriptions.

Other group of inscriptions that presents itself are various moderately long inscriptions from times of either Turkic (682-744) or Uyghur Kaghanates (744-840). In this group I would include the following inscriptions: Küli Çor, Ongi, Tez, Tariat, Şine-Usu and Süci. They were erected for the same reason as KT, BK and T, being part of larger funerary memorials, or relating the history of the Turkic people.

The third group is the rest of the inscriptions - be it a short funerary inscription or one-word long epigraph on a metal coin. The common denominator of this group is the fragmentary character and shortness, that does not allow for more specific characteristics.

The aim of the Orkhon runiform corpus project is to eventually include all the inscriptions from the Orkhon area. But for purposes of the thesis the number of inscriptions, that will be processed, is limited. I have decided to include the KT, BK, and T inscriptions for the simple reason, that they are the best conserved and most studied inscriptions, so I expect less problems with the segmentation and glossing. Besides that, I have chosen one inscription from each other group (Ongi inscription, and Bombogor). In the next paragraphes I submit a short summary about each of these five inscriptions. Kül Tegin and Bilge Kagan inscriptions will be treated together, because they were found at the same place and share some characteristics.

4.1.1 Inscription of Kül Tegin and Bilge Kagan

Kül Tegin and Bilge Kagan inscriptions were erected in years 732 and 734/735 CE as part of a memorial complex next to the Orkhon river forty kilometers north from today's town of Kharkhorin (Khöshöö Tsaidam, Arkhangai aimag). They were discovered for the Western world by Nikolaj Jadrincev in 1889. Both inscriptions are part of the Finnish (Heikel 1892) and Russian atlases (Radloff 1892-99). Another important editions are Orkun (1936), Gabain (1950), Malov (1951, 1959), Tekin (1968), Berta (2004), Alyılmaz (2005), Ölmez (2015).

Kül Tegin was younger brother of Bilge Kaghan and according to the inscriptions, he was leading armies of the kaghanate. The stele has four sides. Three sides are written in Old Turkic, one side has both Chinese and Old Turkic text. There is also a short text on a turtle piedestal, that the inscription was originally placed on. Total number of lines (inscription + piedestal) is 76. The size of the stele is 331 x 122-128 x 41 cm (Alyılmaz 2005: 9).

Bilge Kagan was the ruler of Turkic Kaghanate in the years 716-734/735. The inscription was found in the distance of one kilometer away from the Kül Tegin's monument. The distribution of text is similar to the inscription of Kül Tegin (three sides of Old Turkic, one side of Chinese). Total number of lines is 77. The inscription is slightly bigger 369 x 122-126 x 78 cm (Alyılmaz 2005: 103), but the text have seen more damage and the stele is broken in two parts.

The inscriptions of KT and BK share extensive part of the text (KT S 1-11 = BK N 1-8, KT E 1-30 = BK E 1-24). This fact gave us a lot of information about, how the Old Turkic inscriptions were written, because the fragments are not exact copies of each other. The two versions differ mostly in ortography and punctuation, but sometimes the wording is not identical as well. Compare Examples 2, and 3 of the sentence *bolmış teŋri küç bėrtök* "...became. (Because) Heaven gave (them) strength...".

(2) 𐰉𐰆𐰞𐰢𐰽 : 𐱅𐰭𐰼𐰃 : 𐰜𐰇𐰲 : 𐰋𐰃𐰼𐱅𐰜 (KT E 12)
BULMs : tngrI : ökÜç : bIrtök

(3) 𐰉𐰆𐰞𐰢𐱁 : 𐱅𐰭𐰼𐰃 : 𐰚𐰇𐰲𐰋𐰃𐰼𐱅𐰜 (BK E 11)
BULMş : tngrI : kÜçbIrtök

In Examples 4, and 5 one word was omitted by the scribe.

(4) 𐰋𐰃𐰠𐰏𐰾𐰔𐰼𐰨 (KT E 5)
bIlgszrnç
biligsiz erinç

(5) 𐰋𐰃𐰠𐰏𐰾𐰔𐰼𐰢𐰾𐰼𐰨 (BK E 6)
bIlgszrmsrnç
biligsiz ermiş erinç

These repetitions enable us to supply correct interpretation in many places even though the text is very damaged or even missing.

Various inscriptions have often idiosyncracies in the style, or in the appearance of some characters. Where KT uses letter 𐰾 s (especially in past perfect suffix *-miş*), BK more often than not makes use of letter 𐱁 ş. Compare the following two Examples 6, and 7 roughly translated as "...(kaghan) fed (you) because you were rebellious, (you) were with your kaghan and (you) went with him...".

(6) 𐰚𐰇𐰼𐰏𐰇𐰭𐰤 : 𐰇𐰲𐰇𐰤 : 𐰃𐰏𐰓𐰢𐰾 : 𐰋𐰃𐰠𐰏𐰀 : 𐰴𐰍𐰣𐰭𐰣 : 𐰼𐰢𐰾 : 𐰉𐰺𐰢𐰾 (KT E 23)
kÜrgÜngn : ÜçÜn : Igdms : bIlge : KGNngN : rms : BRMs
küregüŋin üçün igidmiş bilge kaganıŋın ermiş barmış

(7) 𐰚𐰇𐰼𐰏𐰇𐰭𐰤 : 𐰇𐰲𐰤 : 𐰃𐰏𐰓𐰢𐱁 : 𐰴𐰍𐰣𐰭𐰀 : 𐰼𐰢𐱁 : 𐰉𐰺𐰢𐱁 (BK E 19)
kÜrgÜngn : Üçn : Igdmş : KGNngA : rmş : BRMş
küregüŋin üçün igidmiş kaganıŋa ermiş barmış

Another peculiarity of KT and BK is the often use of letter 𐰏 g in final position of the word instead of the letter 𐰭 ng. This concerns especially 2nd person verb conjugation and 2nd person possessive suffixes. Compare the following Examples 8, 9, 10, and 11.

(8) 𐰾𐰇𐰭𐰜𐰏 (BK E 20)
sÜngökg
süŋük-üŋ
bone-POSS.2SG

(9) 𐰾𐰇𐰭𐰜𐰇𐰭 (KT E 24)
sÜngökÜng
süŋük-üŋ
bone-POSS.2SG

(10) 𐰉𐰺𐰑𐰍 (BK N 7)
BRDG
bar-d-ıŋ
go-PST-2SG

(11) 𐰃𐰲𐰚𐰓𐰭 (T W 3)

Içkdng
içik-d-iŋ
depend-PST-2SG

4.1.2 Inscriptions of Tonyukuk

Tonyukuk inscription consists of two separate stelae. The first stone is preserved in better condition than the second. They were discovered in 1897 by Dmitrij Klements close to river Tuul in Bayanzürx sum, Töv aimag (about 60 km southwest from Ulaanbaatar). The exact date the stelae were erected, is still debated, but most proposals range between 720-730 (cf. Tekin 1995: 13, Alyılmaz 2005: 184). The inscription of Tonyukuk is smaller than KT/BK, the measurements of first stone and second stone are 243 x 64 x 32, and 217 x 45 x 28 cm. Total number of lines is 62.

The inscription was erected to commemorate death of Tonyukuk, advisor and military leader of the Second Turkic Kaghanate. Tonyukuk was born in China, during the subjugation of Turks by Tang dynasty. He played important role during the Second Turkic Kaghanate, as he served four different kaghans from 682 until his death. The most important editions of Tonyukuk inscription are Radloff (1899), Orkun (1936), Malov (1951), Tekin (1968), Rybatzki (1997), Erdal (2004), Alyılmaz (2005), and Ölmez (2015).

The literary style of T is different from KT and BK. While relating life story of Tonyukuk, it is full of parables and proverbs. The langauge of T insctiption also exposes dialectical variation. The clitic *ben* used for marking 1st singular on verbs takes the form *men* (example XXX) in T inscription. The inscription has seen a lot of damage. It has been left for a long time without any shelter. The damage by exposure to severe weather conditions, researchers (that for example strived to make the text more visible by painting), and birds made multiple sections non-readable.

(12) 𐰇𐰭𐰼𐰀𐰶𐰃𐱃𐰪𐰍 : 𐰇𐰠𐰼𐱅𐰲𐰃 : 𐱅𐰃𐰼𐰢𐰤 (T W 10)
ÜngrAıKITnyG : ÜlrtçI : tIrmn
öŋre kıtañıg ölörteçi tėr men

öŋre	kıtañ-ıg	ölör-teçi	tė-r=men
east	Kitan-ACC	to.kill-FUT.PTCP	to.say-AOR=1SG

‘(you) will kill Kitans in the east I say’

4.1.3 Ongi inscription

The Ongi inscription was discovered in 1891 by Radloff and Jadrincev. It is situated close to the source of Ongi river in Uyanga sum, Övörkhangai aimag (about 200 km south from KT and

BK, 450 km southwest from Ulaanbaatar). The measurments are 154 x 41 x 12-15 cm. The inscription relates history of Second Turkic Kaghanate. The most important editions are Radloff (1895), Orkun (1936), Clauson (1957), Malov (1959), Tekin (1968), Ōsawa (1999), Berta (2004), Aydın (2008), Erdal (2010), Ōsawa (2011), and Ölmez (2015).

Multiple parts of the inscription are damaged and the reading is unclear. Some parts of the inscription are parallel to T, KT and BK. Ong 1-3 is short summary of KT E 1-11. Characters 𐰏 g, 𐰋 b, 𐱅 T sometimes acquire slightly different forms.

4.1.4 Inscription of Bombogor

The Bombogor inscription is located in Shiveeny Kherem district, northwest of Bombogor sum, Bayankhongor aimag. The complex was discovered by expedition organized by Archeology Institute of Mongolian Academy of Sciences in 2004. The size of the stele is 133 x 20-47 x 16-20 cm. There are altogether 5 lines of text. Part of the stele are 32 tamgas representing sub-tribes and families living under Kaghanate (User 2015: 2). Important editions of the text are Battulga (2005), Suzuki (2010), and User (2015).

4.2 Data structure

4.2.1 Data structure and XML markup

The corpus, if it is built right, might serve as a useful tool for analysing various aspects of language, that may otherwise escape researcher’s attention. The corpus may provide statistical data about freqeuncy, that are beyond capabilities of traditional manual approach, or just facilitate researcher to find a desired example. In this chapter I will try to construct data structure, into which various Orkhon runiform texts may be converted in order to be used as part of the Orkhon runiform corpus. The focus will be to provide such a structure, that will be able to store all important informations, and at the same time the amount of information and marking will not handicap the possibility to search through the data.

What information should be component of the corpus? If we look at most of the editions of Old Turkic texts, three levels usually appear - 1) transliteration (or original text in runiform), 2) transcription, and 3) translation. These three levels can be paraphrased as following: how the text looks like, how do we read it, and what does it mean.

As I already indicated earlier, since there are no obstacles to use runiform script on computers (the only thinkable obstacle is the writing direction, which causes some programs to malfunction), I prefer to use the runiform script over transliteration. They both fulfill the same functional niche - giving us information about how the text looked like. But the representation of text by the original runiform has one simple advantage - it represents the script more truly.

What format should one choose, when one has three corresponding levels of text? The three levels need to be aligned in certain way, in order to be properly rendered in the corpus interface. There are multiple ways to do this. One way is to make use of XML marking. There are multiple projects that aim to provide standardized encoding for digitalized texts, for example Text Encoding Initiative (TEI), or Corpus Encoding Standard (CES). The Text Encoding Initiative (TEI) develops standard set of guidelines used to represent text in digital form and is active more than 20 years. It has established reputable standard widely used by many institutions and researchers. The Guidelines are very inspirative reading for anyone, who plans to format text for electronic corpus, as it benefited from an input from many researchers and can draw from experience with all different kinds of texts.

This is how raw TEI annotated sample text from Inscription of Aphrodisias (Reynolds, Roueché, Bodard 2007) looks like:

```
<div type="edition" lang="grc">
      <head lang="en">Edition</head>
```

```
<ab>
     <lb n="1"/>
     <w lemma="οὗτος">
          <supplied reason="lost" cert="low">οὗτος</supplied>
     </w>
     <w lemma="ὁ">
          <supplied reason="lost">ὁ</supplied>
     </w>
     <w lemma="τόπος">
          <supplied reason="lost">τόπο</supplied>
          <unclear reason="damage">ς</unclear>
     </w>
     <w lemma="ἱερός">ἱερὸς</w>
```

This is TEI encoding of four words, three of which are lost, and part of one is damaged. These raw data are to some extent encoded manually. They are created with help of XML editor, that makes it a bit more user-friendly by prompting set of allowed tags, and validating the syntax. XXX - It is not a format, that everybody is used to work with.

And more importantly the orientation in a longer text, that is encoded this way, is in my opinion very demanding. On the other hand TEI format has indisputable advantages, aside from being recognized standard, it is encoded in HTML-like code and can easily be transformed into electronic edition, once the work is done. After careful and long consideration, I decided not to use TEI encoding for the data of the Orkhon runiform corpus. In the next section I propose my own way to represent data for the corpus.

4.2.2 Structure of the spreadsheet

What I preferred instead as the means to encode all the data, is a classical spreadsheet. Compared to XML-marking. It embodies a format, that every researcher is familiar with. I used different columns for different levels of annotation, and rows representing syntactic units of text (words). I will comment shortly on the overall structure and alignment, and in the following sections (4.3.1-4.3.7) I will cover the individual levels of annotation one by one. The Structure level in section 4.3.7, the Original runiform level in section 4.3.1, and the Transcription level in section 4.3.2. The following Table 7 illustrates the first part of the format. The first row are the names of what I call *levels*, and they are used to subsume columns, that mark related information in one set. Names of the columns are commented in more detail in individual sections and below.

Structure			Original runiform			Transcription	
name	side	# of line	full reading	damage	comm.	transcription	comm.
Bom	F	1	Front				
Bom	F	1	1				
Bom	F	1	𐰴𐰆𐱃𐰞𐰍	??𐰆𐱃?		kutlug	
Bom	F	1	𐰸𐰨𐰖𐰭	𐰭?--		kunçuyuŋ	
Bom	F	2	2				
Bom	F	2	𐰃𐰠𐰋𐰃𐰠𐰏	𐰃𐰠(𐰋)?(𐰠𐰏)		ėlbilge	
Bom	F	2	𐰸𐰆𐰨𐰖𐰭	𐰸𐰆𐰨𐰖𐰭		kunçuyuŋ	
Bom	F	3	3				
Bom	F	3	𐱃𐰆𐰡𐰣𐰃	𐱃𐰆𐰡𐰣𐰃		tultunı	
Bom	F	3	:			:	
Bom	F	3	𐰞𐰆	𐰞𐰆		alu	
Bom	F	3	𐰴𐰺𐰞𐰸	𐰴𐰺𐰞𐰸		karluk	
Bom	F	4	4				
Bom	F	4	𐰸𐰉𐰺𐰯	𐰸𐰉𐰺𐰯		kubrap	
Bom	F	4	:	:		:	
Bom	F	4	𐱃𐰆𐰡𐰣𐰞𐰑𐰃	𐱃𐰆𐰡𐰣𐰞𐰑𐰃		tultunladı	
Bom	S	1	Side				
Bom	S	1	1				
Bom	S	1	𐰇𐰔𐰀	𐰇𐰔𐰀		üze	
Bom	S	1	𐱅𐰭𐰼𐰚𐰀	𐰀----		teŋrike	
Bom	S	1	𐰽𐰺𐰀	𐰽𐰺𐰀		asra	
Bom	S	1	𐰘𐰼𐰚𐰀	𐰘𐰼𐰚𐰀		yėrke	
Bom	S	1	:	:		:	
Bom	S	1	𐰘𐰇𐰚𐰦𐰜𐰢	𐰘𐰇𐰚𐰦𐰜𐰢		yüküntüküm	
Bom	S	1	𐰉𐰺	𐰉𐰺		bar	
Bom	S	1	𐰼𐱅𐰃	𐰼𐱅𐰃		erti	
Bom	S	1	𐰖𐰭𐰡𐰸𐰢	𐰖𐰭𐰡𐰸𐰢		yaŋıltokum	
Bom	S	1	𐰖𐰸	𐰖𐰸		yok	
Bom	S	1	:	:		:	
Bom	S	1	𐰉𐰽𐰢𐰠𐰍	𐰉𐰽𐰢𐰠𐰍		basmıllıg	
Bom	S	1	:	:		:	
Bom	S	1	𐰉𐰆𐰑𐰣𐰍	?𐰉𐰆𐰑𐰣		bodunug	

Table 7: The Bombogor inscription (Structure, runiform and transcription).

The first and second rows of the spreadsheet are names of the corresponding columns. The first three columns are structural annotation (denoting the *name* of the inscription, *side* (or generally part) of the inscription, and *number of the line* where the text is located). The second three columns are level designated to store data about the original runiform text. The first of the three is called *full reading*, and it should ideally represent the proposed reading, that has the most consensus. The second column labeled *damage* is used for marking damaged characters. The third column is *commentary*. The last two columns in Table 7 are representing transcription, again they consist of two columns. The first is the most consented way of the transcribing the word, the second column is *commentary* to transcription.

As we can see, every line represents one word and by word I mean in this context the shortest syntactic unit. As a consequence, there are ortographic words (area between to punctuation marks), that are separated into multiple lines, compare with Table 8:

Structure			**Original runiform**			**Transcription**	
name	**side**	**# of line**	**full reading**	**damage**	**comm.**	**transcription**	**comm.**
KT	E	1	𐰲𐰇𐰢			eçüm	
KT	E	1	𐰀𐰯𐰢			apam	
KT	E	1	:			:	
KT	E	1	𐰉𐰆𐰢𐰣			bumın	
KT	E	1	𐰴𐰍𐰣			kagan	
KT	E	1	:			:	
KT	E	1	𐰃𐱁𐱅𐰢𐰃			iştemi	
KT	E	1	𐰴𐰍𐰣			kagan	

Table 8: Separation of ortographic word into rows.

Three orthographic words 𐰲𐰇𐰢𐰀𐰯𐰢 : 𐰉𐰆𐰢𐰣𐰴𐰍𐰣 : 𐰃𐱁𐱅𐰢𐰃𐰴𐰍𐰣 are thus separated into six syntactically independent words *eçüm apam bumın kagan iştemi kagan* ‘my ancestors (and) forefathers Bumin Kagan (and) Istemi Kagan’. As we will see later with the second part of the spreadsheet, the transcription level (syntactic words) is considered as a pivot level. It constitutes the cornerstone of the structure and all the annotation is aligned to it. The second part of the spreadsheet is illustrated in Table 9. Again it is divided into three levels: *segmentation, glossing* and *further annotation.* More informations about the individual columns are in sections 4.3.4-4.3.6 and below.

Segmentation				**Glossing**				**Further annotation**			
lex. root	**suffix 1**	**suffix 2**	**c.**	**root glos.**	**suffix 1 glossed**	**suffix 2 glossed**	**c.**	**POS**	**sp.sem.**	**sp.morph.**	**references**
kutlug				Kutlug				n	prop	-lug derivation	cf. Aydın (2011a, 18), User (2015, 3-4), Rybatzki (2000)
kunçuyuŋ				princess				n	title		cf. User 2011...
ėlbilge				El_Bilge				n	title	-ge derivation	about -ge derivation cf. Erdal...
kunçuyuŋ				princess				n	title		cf. User 2011...
tultun-	ı			grave-	poss.3			n			
:											
...											
karluk				Karluk				n	tribe	nomadic tribe...	
kubra-	p			assemble-	cvb1			v			
:											
tultunla-	dı			bury-	pst.3			v			
üze				above				pp			
teŋri-	ke			heaven-	dat			n			
asra				below				pp			
yėr-	ke			ground-	dat			n			
:											
yükün-	tük-	üm		worship-	obj.ptcp-	poss.1sg		ptcp		letter -t- is missing...	
bar				exist				a			
er-	ti			to.be-	pst.3			v			
yaŋıl-	tok-	um		to.err-	obj.ptcp-	poss.1sg		ptcp			
yok				exist.neg				a			
:											
basmıl-	lıg			Basmil-	adjvzr1			a	tribe	nomadic tribe...	
:											
bodun-	ug			people-	acc			n			

Table 9: Segmentation, glossing and further annotation.

Table 9 can be split to three parts again. The first part (first four columns) account for morphological segmentation of the word. First column represents the lexical root, second and third columns represent suffixes, and fourth column is a commentary to segmentation. The next four columns (glossing) are analogical to the first four columns (segmentation). Each of the first three columns of the two groups correspond to each other. Column 5 is glossing of column 1. Column 6 is glossing of column 2, and column 7 is glossing of column 3. The last column is again commentary.

The last four columns are dedicated to further annotation. First column represents *part-of-speech*, second is *special semantics*, third is *special morphology*, and fourth are *references* to special morphology or semantics. I will talk about these in more detail in section 4.3.6. There is also last part of the spreadsheet, but I will not present it here. It constists of columns, that represent data adopted from various editions. These data can be easily compared and they can be included as a commentary in the corpus interface.

Before proceeding to discuss editing of individual columns, I want to sum up the last couple of paragraphs again. There are groups of columns in the spreadsheet that are aligned to each other. The most important one is transcription column, that is filled as first and other columns are aligned to it later. After each section there is a commentary.

The advantage of spreadsheets in the current stage of the corpus development is (apart from being intuitive to work with) the fact, that it is easier to design query language to search through columns and rows of a table, than XML-marked text. The spreadsheet also makes it easy for editor to add further level of annotation by simply adding more columns to the spreadsheet (in order to for example mark the painted and greasy parts of Tonyukuk inscription).

Obviously the TEI marking language has the advantage of having elaborated system of text encoding. But I think that it is not necessary, when the Orkhon Runiform Corpus is just in its beginning. The amount of annotation, that is conceivable in this moment, depends heavily on the amount of work on Orkhon inscriptions done up to now. As an example - to make use of any more elaborated marking of damage (including gaps between words) would be only halfway work without proper photographs, or editions. All the possibilites of TEI structural mark-up (it can easily create various paragraphs, shapes, change writing directions) are not needed at this point of time as most of the longer Orkhon inscriptions the text is neatly structured in lines. And just right now the Orkhon runiform corpus does not aspire to account fo the chaotic structure of the rest of the inscriptions (e.g. inscriptions of Iche Achete, cf. Räjäbov & Mämmädov 1993: 156-157). The decision to design more-or-less own format of data structure is flexible and can change at later date. Tables and TEI encoded text are to a certain extend isomorphic structures, and it should not be difficult to convert the spreadsheets to XML-marking eventually, if the advantages of the latter prevail.

4.3 Formatting data for the corpus

In this section I will describe in more detail the functions of various columns into which the data is organized the corpus and the methods I used for preparing the data so that they can be fed into it.

4.3.1 Original runiform level - how the text looks like

What I call the *Original runiform* level subsumes three columns, labeled *full reading*, *damage*, and *commentary*.

The content of the *Full reading* column is the reading of the given part of the inscription to the best of our knowledge. I understand the qualifier “to the best of our knowledge” as implying the accord with the opinion, currently shared by most researchers. The data in the *Original runiform* level are encoded in Old Turkic runiform script (section XXX) and are formatted in right-to-left writing direction and right horizontal cell alignment. What constitutes an ‘orthographic word’ (simply defined here as the string of characters between two punctuation marks) is often split in multiple rows, based on the number of syntactic words the orthographic chunk contains.

Structure			**Original runiform**			**Transcription**	
name	**side**	**#**	**full reading**	**damage**	**c.**	**transcription**	**c.**
Bom	F	1	Front				
Bom	F	1	1				
Bom	F	1	𐰸𐰆𐱃𐰞𐰍	??𐰆𐱃?		kutlug	
Bom	F	1	𐰴𐰨𐰖𐰭	𐰭?--		kunçuyuŋ	
Bom	F	2	2				
Bom	F	2	𐰃𐰠𐰋𐰃𐰠𐰏	𐰃𐰠(𐰋)?(𐰠𐰏)		ėlbilge	
Bom	F	2	𐰴𐰆𐰨𐰖𐰭	𐰴𐰆𐰨𐰖𐰭		kunçuyuŋ	
Bom	F	3	3				
Bom	F	3	𐱃𐰆𐰡𐰣𐰃	𐱃𐰆𐰡𐰣𐰃		tultunı	

Table 10: Illustration of the damage marking.

As has been already mentioned in section 2.3.3, the ortography of Orkhon runiform inscriptions is highly volatile. It is surprising that often the same words, though repeated in almost immediate proximity can be written differently, even though, from a cognitive point of view, one would expect the word be primed after its first appearance and thus be written in the same way. Compare the word *ermiş* in Example 13:

(13) 𐰋𐰃𐰠𐰏𐰾𐰔𐰼𐰢𐰾𐰼𐰨 : 𐰖𐰉𐰞𐰴𐰼𐰢𐱁𐰼𐰨 (BK E6)
bIlgszrm**s**rnç : YBLKrm**ş**rnç
biligsiz ermiş erinç yavlak ermiş erinç
'(they) were not wise and they were wild'

In the next two examples 14, 15 the word *kıltı* 'he did' is subjected to change in orthography in two different ways. In Example 14 the change affects the phonemes /lt/. In the first occurence the phonemes are written with two separate graphemes, whereas in the second occurence ligature is used. In Example 15 the the phoneme /k/ is written by the letter 𐰴 K in the first occurrence, but the second time the letter 𐰶 ıK appears instead.

(14) 𐰆𐰍𐰞𐰃𐰤 : 𐰸𐰆𐰞𐰴𐰃𐰞𐱃𐰃 : 𐰾𐰠𐰚𐰶𐰃𐰔𐰆𐰍𐰞𐰃𐰤 : 𐰚𐰇𐰭𐰴𐰃𐰡𐰃 (BK E7)

UGLIn : oKULKI**LT**I : slkıKIzUGLIn : kÜngKI**LT**I
ogılin kul kıltı ė̇şilik kız ogılin küŋ kıltı
'(they) made slaves of you sons and servants from your noble daughters'

(15) 𐰉𐰖𐰴𐰃𐰡𐰃 : 𐰔𐰍𐰇𐰚𐱁 : 𐰶𐰃𐰡𐰃 (BK E 14)
BY**K**ILTI : zGÜkş : **ıK**ILTI
bay kıltı : azıg üküş : kıltı
'he made (poor people) rich and he made few (people) a lot'

On a rare occassion one letter is shared by two words (Examples 16, 17). In some cases this could be explainable, if we accounted for this phenomenon as a kind of a geminate (Example 16) as geminates are usually written by only a single letter (cf. section 2.3.3).

(16) 𐰃𐰑𐰸𐱃 (BK E 25)
IDoKT
ıdok kut
'Idok Kut (name of a Basmil subtribe)'

But this is not the case, because there are examples where this omission occurs with vowels as well. In Example 17 the letter 𐰇 Ü is "shared" by the words *yençü* and *ügüzüg* (but we know that in the position at the beginning and the end of the word, the vowel 𐰇 Ü is never omitted; more about vowel omission in runiform script in section 2.3.3). Compare also with Example 18, where the vowel letter 𐰇 Ü is carved twice in the inscription.

(17) 𐰖𐰨𐰇𐰏𐰔𐰏 (T W 44)
ynçÜgzg
yençü ügüz-üg
pearl river-ACC
‘Syr Darya (river)’

(18) 𐰖𐰨𐰇𐰇𐰏𐰔 (BK N 3)
ynçÜÜgz
yençü ügüz
pearl river
‘Syr Darya (river)’

The second column labeled *Damage* is designed to store information about damaged parts of the text. Marking damaged and missing parts of a text is an important piece of any historical corpus. Marking damage can be only as accurate as allowed by either the existing editions of the text or direct visual evidence. Ideally, damage annotation should be drawn from the oldest reliable source available (rubbings of Heikel, Jadrincev, Radloff, Ramstedt, Malov, and other scholars). On the other hand as Clauson (1962: 43-44) notes, there is always a danger in relying on some sources:

“It should be added that at least one of the most recent photographs, that of Malov 1952, No. 49, is not a photograph of the stone itself but a photograph of the stone after the letters had been chalked in, and careful scrutiny of the photograph shows that some of these chalk marks do not exactly follow the original letters. Thus it is hardly too much to say that, with very few exceptions, none of the hand copies were made by people who could read the texts which they were copying, and none of the editors had actually seen the original inscriptions which they were editing. The results have in some cases been disastrous.”

To the best of my knowledge no edition elaborates on annotation of damage of KT, BK, and T inscriptions beyond actually providing the pictures of the inscriptions as published in the Russian Atlas (Radloff 1892), for comparison of retouched and “vanilla” photographs from Russian Atlas, compare Figures 3, and 4. In most of the classical editions the marking of damage is usually indicated rather vaguely and the length or character count of the missing parts of the inscription was usually not taken into consideration. Alyılmaz (2005) prepared an edition of KT, BK, and T, where he records the current state of the inscription. Moriyasu & Ochir (1999) accounted for the damage and missing parts of the texts of various inscriptions from First, Second and Uyghur kaghanates.

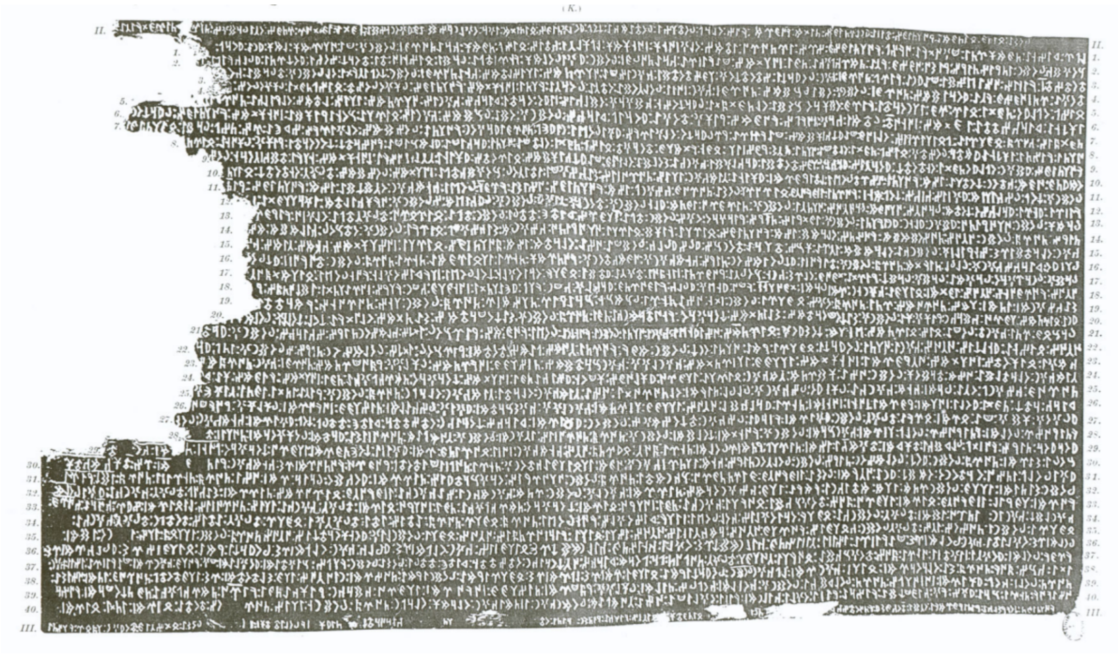


Figure 3: Kul Tegin’s inscription, eastern side, retouched (Radloff 1892: XVIII), the current standard numbering of lines is reversed.

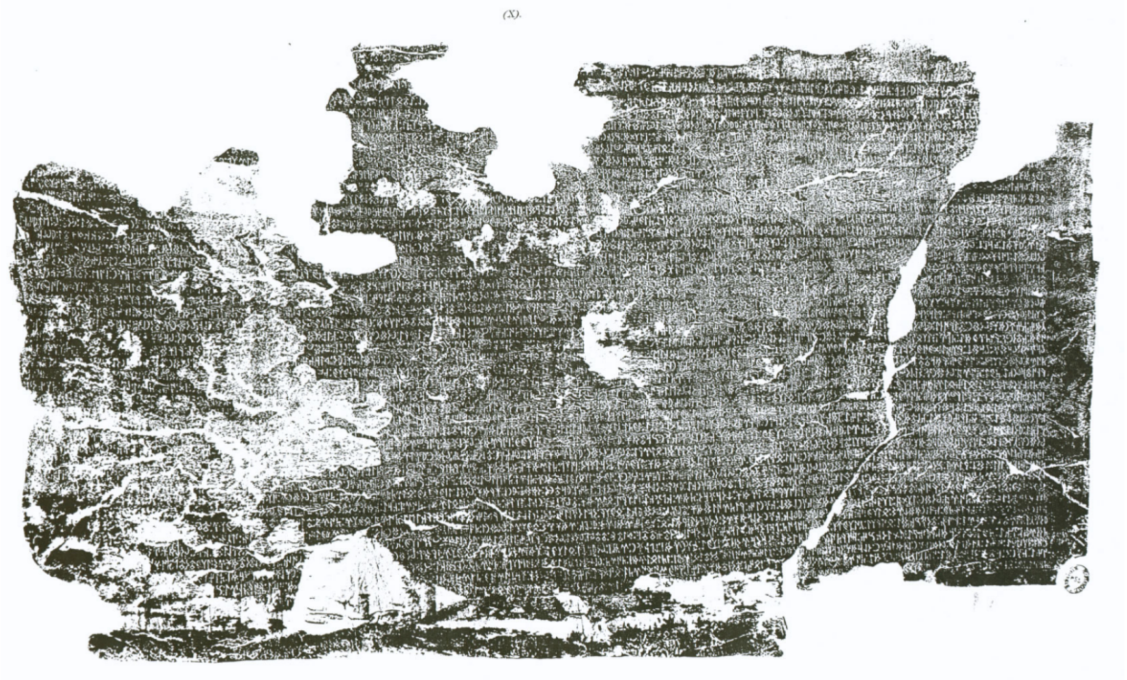


Figure 4: Bilge Kagan’s inscription, eastern side, no retouche (Radloff 1892: XXII).

In the case of the 5 inscriptions, that have been prepared for the Orkhun Corpus (section 4.1) I used three different sources. I used Alyılmaz (2005) edition for the damage annotation of the KT, BK, and T inscriptions. Osawa (2011) for the Ongi inscription and User (2015) for the inscription of Bombogor. Thus the damage annotation of the various inscription in the corpus is inconsistent. The annotation of damage in KT, BK, and T accounts for damage that is visible on the inscriptions now, but might have been clearly readable a century ago. The annotation of Ongi is based on rubbings acquired by Ramstedt in 1909. The damage annotation of the Bombogor inscription represents both the current and the oldest version at once, since only a few years passed since its discovery.

Another thing that should be accounted for is the intensity of damage. Marking the intensity of damage, and the damaged or missing areas is standard in classical philology (Reynolds, Roueché & Bodard, 2007). TEI Guidelines offer two different features related to damage, that can be accounted for. The first is the intensity of damage, and the second is the certainty of the reading that is supplied. In my experience, these two are to a large degree intertwined and cannot be easily separated. Often simply more damage simply means less certainty in supplying a reading.

For the needs of the Orkhon Corpus I propose a four-level distinction, based on both of these criteria combined (see Table 11). First level corresponds to a text, that is not damaged and that is clearly readable. I record this simply with repeating the same character in the damage column. So in case that a word is not damaged, the full reading and damage columns will contain the same data (the word *kunçuyuŋ* in the second line in Table 10). The second level is marked by text in brackets. It represents text, that is slightly damaged, but still readable, and has a reading that is agreed upon. The third level comprises text that suffered moderate to heavy damage, and either has some interpretation or not. The fourth level represents parts of the stone, that are missing altogether, and is used mainly for marking the length of the missing part.

1	Clearly readable, no damage	𐰴𐰆𐰣𐰲𐰆𐰖𐰭
2	Slightly damaged, consensus on reading	(𐰴𐰆𐰣𐰲𐰆𐰖𐰭)
3	Damaged, not clear	???????
4	Missing parts of stone, no traces of text	-------

Table 11: Intensity of damage.

The substituent signs, representing damaged or missing characters of a text, vary from edition to edition. Amongst the most frequently used ones are dots and slashes, but I consider this an ill-advised solution. The reason behind the choice of the question mark and hyphen in the Orkhon corpus is, that when the text is rendered on the screen, these characters have the length of an average runiform character, and are thus more illustrative, than dots and slashes would be.

There are two frequently used options for encoding parts of texts, that are missing due to damage. Either one can state the approximate count of missing characters or is stated, or the length of the missing text in a metric unit. I consider the character count more important than the actual measurement of the missing part for the usability of the corpus. The reason is that the second option, even though objectively more accurate, has the disadvantage of leaving the character count to the user of the corpus and it is indeed the number of characters missing which the user can use as a valuable piece of information for his study of texts, much unlike the information about the physical measurements.

Characters, that are omitted by scribe, are not marked in the original level. In return they are noted in the commentary of the transcription section. There is also a special sign # used to mark severed part of the inscription (where the texts splits in two parts). I adopt the usage of this sign from Ōsawa (2011: 163).

Often in runiform texts collocations (often fossilized) and parallelism can help with interpretation. If one part of a collocation or parallelism appear in the text, then it is usually easy to supply the proper reading of the damaged part. See the Table 12:

Structure			**Original runiform**			**Transcription**	
name	**side**	**#**	**full reading**	**damage**	**c.**	**transcription**	**c.**
KT	N	3	𐰋𐰃𐰤𐰯			binip	
KT	N	4	4				
KT	N	4	𐰆𐰯𐰞𐰖𐰆	--???		oplayu	
KT	N	4	:	-		:	
KT	N	4	𐱅𐰏𐰓𐰃	𐰃?--		tegdi	

Table 12: Supplying the reading - parallelism.

In Table 12 we can see that the words *oplayu tegdi* are severely damaged, and five letters are missing altogether, the only clearly readable letter being 𐰃 I. But as the words *binip oplayu tegdi* ‘mounted (a horse) and attacked’ appear five times in a short sequence (KT N 2-5), we can safely assume, that *oplayu tegdi* is the correct reading of the damaged part.

The marking of damage is currently a problematic part of the corpus. It needs to be reworked based on a close analysis of the visual material. The early rubbings and visual material should be digitalized and aligned to the Orkhon corpus as they represent the best preserved stage of the inscriptions.

4.3.2 Transcription level - how do we read it

The transcription section is designed to represent the phonological reading of the runiform characters. I have already stated my arguments for the choice of transcription style used in Ölmez 2015 in section 2.3.4. For the same reasons I altered the system slightly and united the transcription of different /k/ characters into *k* (Ölmez uses superscribed vowels). For example, I render the words *k*^{ı}*ılıntım*, and *k*^{ö}*örür* (T W 1) as *kılıntım*, and *körür*. The reason is that as far as we know, there was no phonological difference between the five characters used to mark /k/-like sounds, and thus the marking of the superscripts is actually a marking of an orthographical, not phonological feature.

As I already noted multiple times, the orthography is volatile. The overall motto of the transcription is the following: if there is no counterevidence, transcribe the same semantic words in the same way. For example there are two different transcriptions of the word *toñukuk* in the edition Ölmez 2015. In the transcription of T inscription (Ölmez 2015: 181-187) the rendering of the name is always *toñukuk* (written as 𐱃𐰆𐰪𐰸𐰸), while in BK S 14 (Ölmez 2015: 145) the transcription is *tonyukuk* (written as 𐱃𐰆𐰣𐰖𐰸𐰸). This most likely stems from the fact, that the single occurrence of the word Tonyukuk in Bilge Kagan inscription has a different ortography than in the T inscription. But again there is no evidence that the pronunciation of those two differently written words was different, and therefore the motto applies.

Another example, but a bit more complicated is the word *eçüm / eçim* ‘my ancestor’. In Ölmez’ edition it is possible to count four different transcriptions *ėçüm* (BK E 3), *eçüm* (KT E 1), *ėçim* (BKN 9), and *eçim* (BK E 35) in multiple places. I believe that it is justified to preserve the distinction of /ü/ and /i/, but not the distinction of /ė/ and /e/. My reason for this is that there is no evidence whatsoever, that there is any difference in the first vowel (since it is never written). But there is a reason to believe that there was a rounded and unrounded version of the word ‘ancestor’ *ėçüm* and *ėçim*. This is a situation similar to the variation in some words in English, like the word *often*, that can be pronounced with silent [t] as [ˈɒfən] or with the [t] sound [ˈɒftən] while retaining the same meaning.

Similarly the /i/-/ü/ variation is not unknown in Turkic languages. Interestingly enough there is one word in Old Turkic, that we can spectate at the beginning of its long history of assimilation from /ü/ to /i/. The word *üçün* ‘for, because of’ is in the runiform script usually written as 𐰇𐰲𐰤 Üçn (KT E 6), or 𐰇𐰲𐰇𐰤 ÜçÜn (KT S 9). But there is one instance where the shape of the word is 𐰃𐰲𐰃𐰤 IçIn (KT S 12). This word takes various forms in modern Oguz languages - Turkish *için*, Azeri *üçün*, Turkmen *üçin*, and Ottoman *içün*. There is a reason to believe, that it had two different pronunciations already in the Old Turkic language. This is the kind of distinction (if there is evidence fo it) which is worth keeping in the transcription level.

Another exception from the motto are cases, where certain differences between dialects had already appeared in the Old Turkic language, for example, the difference between *ben* ‘I’ as found in most of the inscriptions and *men* ‘I’ in T inscription.

A large number of the words that we find in the inscriptions are hapax legomena. Until other discoveries are made, there will be always doubt cast over their meaning and transcription. Therefore, researchers always need to be aware of danger of working with dictionaries as there is a chance, that the translation one will find is based only on a single occurence.

In cases where were words added to the transcription level by some scholars (either for syntactic, semantic purposes or due to an error of the scribe), I ignored those and included them only in the transcriptions of particular scholars (e.g. KT E22 - anča)

4.3.3 Analytical part - what does it mean

The aim of the analytical part of the thesis is the segmentation and glossing of the texts. In order to align various levels of the corpus, the segmentation on word, and sentence level is necessary. Individual words are then segmented into moprhemes and glossed.

As has been discussed earlier, the punctuation sign (:) is used in various places. Generally it marks a word boundary, or a boundary of two syntactic phrases. As I am not a specialist in Old Turkic, I tried to resolve the problem of word boundaries in accord with contemporary editions of the texts and dictionaries. Segmentation of morphemes was tackled with help of grammars and other publications (Erdal 1991, Tekin 1995, Erdal 2004). Another source of inspiration were the solutions of morpheme segmentation present in the VATEC corpus (discussed in section 3.1.1, 4.3.5).

4.3.4 Segmentation

In the segmentation section of the spreadsheet the words from transcription section are segmented into morphemes. The goal is to create conditions for morpemic translation. The first column of the segmentation section is labeled *lexical root,* while the rest are subsequent suffixes. Old Turkic is suffixing-only language. Therefore, there is no need for accounting for prefixes. Besides that the lack of fusional morphology in Old Turkic allows for a neat segmentation of morphemes in majority of cases. The morpheme boundaries are marked in the corpus by a hyphen '-', placed at the end of the first morpheme. The morpheme boundary between a morpheme and a clitic is noted by equal sign '=' (Examples 12 in section 4.1.2; 19 below).

(19) 𐰇𐰞𐰼𐱅𐰲𐰃𐰚 (T W 11)
ÜlrtçIk
ölür-teçi=k
to.kill-FUT.PTCP=EMP
'(he) certainly will kill'

In very few cases two words are fused together creating a compound. For example the word *bödke* ‘at this time’ (KT S 11) formed from the demonstrative pronoun *bo* and the word *üdke* ‘at time’ (cf. Erdal 2004: 126). In those cases I assume, that they are already lexicalised and therefore historical segmentation is indicated only in the commentary.

Transcription		**Segmentation**				**Glossing**			
transcription	**c.**	**lexical root**	**suffix 1**	**suffix 2**	**c.**	**root glossed**	**suffix 1 glossed**	**suffix 2 glossed**	**c.**
üze		üze				above			
teŋrike		teŋri-	ke			heaven-	dat		
asra		asra				below			
yėrke		yėr-	ke			ground-	dat		
:		:							
yüküntüküm		yükün-	tük-	üm		worship-	obj.ptcp-	poss.1sg	
bar		bar				exist			
erti		er-	ti			to.be-	pst.3		
yaŋıltokum		yaŋıl-	tok-	um		to.err-	obj.ptcp-	poss.1sg	
yok		yok				exist.neg			

Table 13: Bombogor inscription. Segmentation and glossing.

The only exception to the segmentation rules is due to some derivative morphemes. Compared to inflection morphology, where every morpheme is easily labeled, the case of derivation morphemes is indeed more complex. Derivation often changes the meaning of the word to the extent, that they become lexicalized (especially with adjectivizers, and causative). In these cases the root-suffix segmentation would lead to having to make clumsy decisions, because either the editor would have to invent a meaning for the segmented lexical root, or some of the information would be lost in the process (e.g. there are verbs with causative morphemes, that are lexicalized without having any bases attested, cf. Erdal 2004: 299). On Examples 20, 21 I want to illustrate the solution I propose.

There are two possible ways to segment the word *başlıgıg*. The first is to segment the derivation morpheme and gloss the lexical root *baş* as ‘head’ (Example 20). The second is to keep the segment in one piece as *başlıg*, consider it the lexical root, and gloss it as ‘proud’. I incline to the second solution, as there are cases, where attempting to arrive at the diachronic segmenation would be difficult (e.g. *tonlug* ‘clothed, rich’, for which there is no attested nominal root *ton*-, and one would have to reconstruct it). One also does not have to tackle the problem, how to mark various derivation morphemes (they tend not to have special names in grammars). Instead I propose to insert the information concerning the derivative suffix in a separate column (more in section 4.3.6).

(20) *baş-lıg-ıg* (KT E 15)
head-ADJVZR?-ACC

(21) *başlıg-ıg* (KT E 15)
proud-ACC

4.3.5 Glossing

The morphemic segmentation as such can only then fully be of actual use to the corpus user, when the morphemes are glossed (Table 13). The glossing gives the user information about the grammatical and semantic properties of the lexical roots and suffixes. It strives to provide a morpheme-to-morpheme translation between the object/target language (Old Turkic) and the metalanguage (English). The glosses, and glossing system of the Orkhon runiform corpus is designed in accord with the Leipzig Glossing Rules (Comrie, Haspelmath & Bickel 2008).

As has been already noted in previous section, morpheme boundaries of regular morphemes are marked by hyphen, that is placed at the end of the first morpheme. One-to-many correspondences (for example the *ölür-* glossed as *to.kill-* in example XXX in segmenation section), are marked by dots in between the words of the metalanguage.

The reversed case of one-to-many correspondences, where multiple words of the object language correspond to one word in English, also occur sometimes. It is mostly the case of petrified collocations, that are translated as one word in English. These cases usually do not pose a problem, as usually each of the words can have its own glossing, and the overall meaning of the collocation is mentioned in the commentary, e.g. *otça borça* ‘clustered’ is glossed separately as follows:

(22) *ot-ça* *bor-ça* (KT E 37)
fire-EQT lightning-EQT
‘clustered’

The following two tables (14, 15) are list of glosses, that I used for preparation of data for the Orkhon Runiform Corpus. Cases of some morphemes, that are attested in Orkhon inscriptions only rarely *-gInçA*, or are restricted to single words, that are grammaticalized *-yIn* in *tėyin* ‘saying, in order, for’, are excluded from this list. Part of the two tables are examples from Orkhon runiform texts. The VATEC glossing column are glosses used in VATEC Corpus. Most of the forms are adopted for the use in Orkhon Corpus, but some are changed along the list of standard abbreviations in Leipzig Glossing Rules (Comrie, Haspelmath, Bickel 2008). The entries in VATEC function column are grammatical functions of the corresponding morphemes.

The column labeled VATEC morpheme represents an archimorpheme, theoretical form of the morpheme, before it is affected by vowel harmony and assimilation processes. In cases where there is not a dedicated gloss in the VATEC glossing column, the VATEC morpheme is supplied by the author. I will shortly explain the notation. Letter *X* is represents any vowel. Letters *A*, *I*, *U*, and *O* represents their respective front and back realisations /a/-/e/, /i/-/ı/, /ü/-/u/, and /ö/-/o/. Letters *D*, and *G* are are the voiced and unvoiced consonants with the same place of articulation /d/-/t/, and /g/-/k/. Letters enclosed in brackets are rendered only in some positions. For informations about the suffixes I redirect an interested reader to two publications of Erdal (1991, 2004), and to *Grammar of Orkhon Turkic* by Tekin (1997, or 2003), that is accessible either in English or Turkish.

Nominal suffixes	**VATEC glossing**	**VATEC function**	**VATEC morpheme**	**example**	**Orkhon C. glossing**
Accusative	ACC	case	(X)g, nI	kagan-ıg	acc
Genitive	GEN	case	(n)Aŋ, (n)Xŋ	kaganım-in	gen
Dative	DAT	case	kA	yış-ka	dat
Locative	LOC	case	DA	balık-da	loc
Ablative	ABL	case	DXn, DAn	kan-dan	abl
Equative	EQT	case	čA	ot-ča	eqt
Instrumental	INST	case	(X)n, In	kaganıŋ-ın	ins
Vocative	VOC	case	A	beglerim=a	voc
Plural	PL	plural	lAr	beg-ler	pl
Possession 1sg	POSS1	possesor	(X)m	kan-ım	poss.1sg
Possession 2sg	POSS2	possesor	(X)ŋ	kagan-ıŋ	poss.2sg
Possession 3	POSS3	possesor	(s)I(n)	kövürge-si	poss.3
Possession 1pl	POSS.1PL	possesor	(X)mXz	ėç-imiz	poss.1pl
Possession 2pl	POSS.2PL	possesor	(X)ŋIz	oglan-ıŋız-da	poss.2pl
Possession 3sg & accusative	POSS.3SG.ACC	case	(s)In	kümüş-in	poss.3sg.acc
Ordinal numeral	ORD	num	(X)nč	üç-ünç	ord
Collective	-	-	AgUn	tay-agun-uŋuz	col
Privative	PRIV	adjvzr	sXz	buŋ-sız	priv

Table 14: List of nominal glosses.

Verbal suffixes	VATEC glossing	VATEC function	VATEC morpheme	example	Orkhon C. glossing
Aorist	AOR	tense	Ir, Ur, yUr, Ar	kelür-ür	aor
Negated aorist	AOR.NEG	tense	mAz	bil-mez	aor.neg
Past	PST	tense	D	er-t-i	pst
Inferential	INFR	tense	mIš	teg-miş	infr
Negated perfect / inferential ptcp.	INFR.NEG	tense	mAdOk	kılın-madok	infr.neg
Negation	NEG	negation	mA	kork-ma-dımız	neg
Volition / Imperative 1sg	IMP.1SG	mood	(A)yIn	yoglat-ayın	imp.1sg
Volition / Imperative 2sg	IMP.2SG	mood	0, (X)ŋ	öl	imp.2sg
Volition / Imperative 3sg	IMP.3	mood	zUn	bolma-zun	imp.3
Volition / Imperative 1pl	IMP.1PL	mood	(A)lIm	basın-alım	imp.1pl
Volition / Imperative 2pl	IMP.2PL	mood	(X)ŋ	bil-iŋ	imp.2pl
Conditional converb	COND	gerund	sAr	er-ser	cond
Consecutive converb	GER1	gerund	(X)p, (X)pAn	tut-up	cvb.con
Simultaneous converb	GERA	gerund	yU, U	ula-yu	cvb.sim
Negative converb	GER.NEG	gerund	mAtI(n)	udı-matı	cvb.neg
Purpose converb	PURP.GER	gerund	GAlI	al-galı	purp.cvb
Participle	PART	part	gAn	kara-gan	ptcp
Perfect / Inferential participle	PF.PART1	part	mIš	bol-mış	pf.ptcp
Negated perfect / inferential ptcp.	PF.PART1.NEG	part	mAdOk	kılın-madok	pf.ptcp.neg
Object participle	OBJ.PART	part	DOk	tegür-tök	obj.ptcp
Necessitative / Future ptcp.	OBLG.PART2	part	sXk	tug-sık-ıŋa	nec.ptcp
Future participle	AG.PART1	part	DAčI	er-teçi	fut.ptcp
Negated future participle	-	-	mAçI	yara-maçı	fut.ptcp.neg
Agentive participle	AG.PART2	part	gUçI	ay-guçı	ag.ptcp2
Agentive participle	AG.PART3	part	(X)glI	ö-gli	ag.ptcp3
Agentive participle	AG.PART4	part	(X)gmA	aytı-gma	ag.ptcp4
Expectation participle	EXP.PART	part	gUlXk	bil-gülük	exp.ptcp
Emphasis	-	-	(O)k	ölörteçi=k	emp

Table 15: List of verbal glosses.

4.3.6 Further annotation

Orkhon Runiform Corpus will be furthermore annotated for parts of speech in the future. Part of speech annotation/tagging is one of the useful features for linguistic research. It provides information about the word and its syntactical neighbours, and the distribution of various parts of speech in a clause can affect various linguistic phenomena, i.e. possible morphological suffixes.

The annotation process can be semi-automatic and can be based on the the glosses of suffixes. Finite verb forms, converbs, participles, and cases are unique markers for their respective parts of speech. The list of pos tags is available in Table 16. There are two categories that stand out from the standard list of parts of speech - converbs, and participles. They are categorised as parts of speech on their own, because their affiliation to other classes is problematic.

noun	n
verb	v
adjective	a
adverb	adv
pronoun	pro
numeral	num
postposition	pp
converb	cvb
participle	ptcp
particle	ptcl
punctuation	i

Table 16: List of parts of speech tags.

In Table 9 (section 4.2.2) I presented two columns headed as *special morphology* and *special semantics*. These two columns are used for marking the categories and informations, that are better outside the rest of the system. I stated the reasons about the special morphology already in the section 4.3.4. This column is used for marking causative and passive derivational morphemes as well as nominal derivation with the exception of privative *sIz*, that has straightforward meaning, and tends not to be lexicalized (cf. Table 14). In Example 23 the word *yüküntürmiş* ‘(he) subjugated’ is segmented and glossed without segmentation of the causative -*tür*- morpheme. The causative is instead noted in the special morphology column. More detailed description of the derivation process, or lexicalization of the particular word can be provided in the commentary section.

(23) *yüküntür-miş* (KT E 2)

to.subjugate-INFR

The column special semantics is used for marking personal names, place names, and titles. Those three groups of words are better accompanied with encyclopaedic information that encompasses the actual knowledge about the person, the place or the title (cf. Ölmez 2015b). The placenames should be accompanied by their geographic location.

4.3.7 Metadata

Metadata are informations about the individual inscriptions. They play a key role in organizing the corpus in a way that enhances the processing of the data. Metadata should not aim to be a substitution of a proper description. Their importance lays in enabling the user of the corpus to filter through various texts, and eventually create a subcorpus designed for a particular enquiry. One example of such use of metadata would be for example to create a subcorpus of Orkhon inscriptions, that would encompass all the inscriptions written after the year 742 CE. Another use of metadata would be to filter out all the funerary inscriptions. When creating a subcorpus, the option to combine multiple criteria should be possible as well. So the question is, what data should be included as metadata, so we can benefit from them?

One of the features, that can be considered also as metadata is the information about the location of a word on the inscription. Inscriptions are traditionally split into lines and sides. For example the western side of the first stone of the T inscription has 8 lines. By this practice it is easy to locate a word, and reference to it. All the structural metadata is taken over from classical editions and are marked in the first group of columns (see Table 12).

For the rest of the metadata (called descriptive metadata) I follow practices from the corpora discussed in section 3.1. Considering the amount of knowledge about the inscription the following list of metadata is proposed:

Dating of text. One of the most important criteria for linguistic research is knowledge of the time, when the text was written. It enables the researcher to keep traces of how the language might have changed.

Place of discovery. Every text should provide for its provenance. It is an important aspect for exposing patterns of dialectological variation (the data might be combined with knowledge about the location of particular tribes). The data about the location should be sufficiently accurate and should include modern-day administrative units and GPS location.

Text type / Genre. Various linguistic features are dependent on the genre, and text type. We can expect difference in the lexicon, grammar and syntax between different text types like grafitti or epitaphs as the first might have been produced by a lost wanderer, while the second might be classified as literary language of political elites. It is also possible to establish more fine-grained distinctions by adding more levels of the text type taxonomy.

Length of text. Another criteria for variability in the language might be length of text. The information should include number of characters, another option is to divide texts in groups as has been indicated in section 4.1.

Language variety affiliation. The language of the texts written during the Uyghur Kaghanate is sometimes called Old Uyghur. One of the distinctions that presents itself is to divide texts to Orkhon Turkic texts, and Old Uyghur texts.

5 TOWARDS THE CREATION OF SEARCHABLE CORPUS

In this section I will describe the basic functions of the corpus, including query language, structure of concordance list, and export of results. This part of the thesis is unfortunately still in the planning phase and I will thus not be able to provide detailed information about how the project will develop in the future.

5.1 Query language and search engine

The goal of any corpus is to allow the user to search words, morphemes, or any other information, that is annotated in the corpus. Query language is generally a name for any language/notation system, that is used by the user in order to be able to retrieve information from a database. The query language is designed in dependence on the markup and structure of the data, in our case the spreadsheet. It enables the search engine to look for matching data in the corpus and the matching data are then simply copied to the results screen. A good query language enables the user to pose complicated queries, including specific information about any of the marked categories, syntax, or by allowing the user to use regular expressions.

There are two ways to prepare data in a database, that is searched by the search engine. The first option is fulltext database, a single file, that contains all the data in form of a vertical. The search engine is looking for matches in the vertical, and saves them as a result. This approach is very simple, does not need any further programming, and it is suitable for smaller corpora. The second option is to index (collect, parse, and store) data in the file called the index. It represents a file, where answers for a set of queries are already processed. The search engine then finds the matching answer, that includes references to the location of the matching data. This approach facilitates the retrieval of information, and lowers the computational load of more complicated queries. Considering the size of the Orkhon Runiform Corpus, that will have approximately 50 000 runiform characters, if all the currently discovered inscriptions are processed, there is no need for indexing of the corpus.

5.2 Results

The data structure of the Orkhon Runiform Corpus has multiple *levels* (section 4.2), that are mutually aligned. The user of the corpus should have the option to search through all the levels, that are part of the corpus (original runiform, transcription, glossing, commentaries, etc.). Additionally the user of the corpus should have more options, when designing the structure of results. i.e. let us consider a researcher, that is not interested in the Old Turkic runiform letters at all, but wants to find an example of transitive veb construction for his typologically oriented linguistic research. This researcher should have the option to disable the original runiform level in the results, and display only the levels he/she considers useful.

The output of the whole procedure are results, that are displayed on the results screen. They consists of a list of concordances, that are evaluated as matching the query by the search engine. The concordance is an excerpt from the corpus, that consists of a string of words, and that is centered around the KeyWord In Context (KWIC). The multiple level feature of the corpus is manifested by the option to display multiple levels in alignment to the KWIC.

5.3 Other functions

Standard function of any modern corpus is exporting results to various formats (.xlsx, .xml, .ods, .txt, .csv). Export of data is useful for example in cases, when a user wants to use the data as part of his/her work, or to continue working on the data analysis offline. Another useful format of data export might be exporting concordances in format proposed in Leipzig Glossing Rules.

As has been already mentioned in section 4.3.7, creating and managing subcorpora is an essential part of corpus data analysis. The user should be able to choose precisely the texts he wants to work with in the subcorpus, and filter out any unwanted data.

6 CONCLUSION REMARKS AND OUTLOOK ON THE FUTURE OF THE ORKHON RUNIFORM CORPUS

In the previous chapters I aimed to describe the process of creating a corpus of Orkhon runiform inscriptions and preparation of data. In Chapter 2 I provided a short summary of the history and society of Turkic and Uyghur Kaghanates, Orkhon inscriptions, their language, runiform script, and various transcriptions of the Old Turkic language. In Chapter 3 I reviewed other projects that digitalized Old Turkic texts and commented on the technical solutions - especially encoding, fonts, and keyboard layouts - in order to be able to work with the runiform script on a computer. In Chapter 4 that constitutes the most essential part of the thesis I described the process of building the Orkhon Runiform Corpus. The sections included in this chapter focus on the choice of the initial set of inscriptions, design of the spreadsheet data structure, alignment, marking of damage, metadata, and the overall operationalization of the language data into the corpus. In the previous chapter I proposed, how should the end-product corpus work.

I believe, that online accessible electronic corpus of Orkhon runiform inscriptions will prove itself useful in the future. Although there is still much work to be done to introduce the full list of Orkhon inscriptions into the corpus, further options present themselves just behind the horizon. First and foremost imperative of the Orkhon Runiform Corpus should be providing access to photographs, rubbings, and other visual material, that can help to confront the edited text with the original monument. Because the jury is still out on reading of some of the words, if these visual materials would be parsed and aligned to texts, it would help tremendously to point out inconsistencies and emend the text.

Another option pending on the hypothetical to-do list is to publish the corpus as a electronic text edition (in order to see how such an project might look like, cf. Kytö, Grund, Walker 2011). There is also the possibility to automatically compile Orkhon Turkic dictionary, including the English meanings and location of the words in texts. The value of this enterprise is rising with every word, that is added to the volume of the corpus.

What the Orkhon Runiform Corpus project should definitely do in the future is to close the gap between the spreadsheet format, that has been designed as the format to store data, and the TEI format (discussed in section 4.2.1). The reason to not use TEI markup language against its obvious advantages, can be to a certain extent paraphrased as using a sledgehammer to crack a nut. Not to underestimate the corpus, the stage of the corpus is indeed in the situation when TEI is a tool yet too strong for the job. The spreadsheet format is only a temporary solution, that posed less complications during the annotation process, but certainly will be more problematic, when part of the searchable electronic corpus.

Problems with some of the editions of the Orkhon runiform texts are, that often there is a missing commentary of problematic part of the inscription. My hope with the Orkhon Runiform Corpus is, that exposing texts on one website can provide a shared platform, or a kind of shared workbench, that will eventually help to focus work on problematic parts of Orkhon inscriptions. I believe that leveraging the power of computers for study of texts will make working with the language more interactive and even more appealing.

7 REFERENCES

Aalto, P., 1958. Materialen zu den alttürkischen Inschriften der Mongolei, Journal de la Société Finno-Ougrienne, LX. Helsinki.

Ajdarov, G., 1966. Jazyk orxonskogo pamjatnika Bil’ge-kagana. Alma-Ata.

Alyılmaz, C., 2000. Bilge Tonyukuk yazıtları üzerine birkaç düzeltme. TDA10: 103-112.

Alyılmaz, C., 2003. Bugut Yazıtlı ve Anıt Mezar Külliyesi Üzerine. Selçuk Universitesi Türkiyat Araştırmaları Dergisi, 13: 11-22.

Alyılmaz, C., 2003. Moğolistanda eski Türk kültür ve medeniyetine ait bazı eserler ve bulundukları yerler. Türkiyat Araştırmaları Enstitüsü Dergisi, 21. Erzurum: Atatürk Üniversitesi. pp. 181-199.

Alyılmaz, C., 2005. Orhun Yazıtlarının Bugünkü Durumu. Ankara: Kurmay.

Alyılmaz, C., Yakar, M., Yılmaz, H.M., 2010. Drawing of petroglyphs in Mongolia by close range photogrammetry. Scientific Research and Essays Vol. 5(11), pp. 1216-1222.

Amanžolov, A.S., 2003. Istorija i teorija drevnetjurkskogo pis'ma. Almaty: Mektep.

Aspelin, J.R., 1889. Inscriptions de l’Iénissei. Recueillies et publiées par la Société Finlandaise d’Archéologie. Helsinki: Société Finlandaise d’Archéologie.

Aydın, E., 2007. Şine Usu Yazıtı. Çorum: KaraM.

Aydın, E., 2008. Ongi yazıtı üzerine incelemeler. İlmî Araştırmalar, 25. pp. 21-38.

Aydın, E., 2011a. Yenisey Yazıtlarında Geçen Unvanlar ve Unvan Niteleyecekleri. ??? 2 belleten.

Aydın, E., 2011b. Uygur Kağanlığı Yazıtları. Konya: Kömen.

Aydın, E., Alimov, R., Yıldırım, F., 2013. Yenisey - Kırgızistan Yazırları ve Irk Bitig. Ankara: BilgeSu.

Aydın, E., 2014. Orhon Yazıtları (Köl Tegin, Bilge Kağan, Tonyukuk, Ongi, Küli Çor). Konya: Kömen.

Aydın, E., 2015. Yenisey Yazıtları. Konya: Kömen.

Aydın, E., 2016. Eski Türk Yer Adları. İstanbul: Bilge Kültür Sanat.

Bahry, S., 2016. Language Ecology: Understanding Central Asian Multilingualism. In: Ahn, E.S., Smagulova, J., (eds.). Language Change in Central Asia. Berlin: De Gruyter Mouton.

Batmanov, I.A., Kunaa, A.Č., 1963. Pamjatniki drevnetjurkskoj pis’mennosti Tuvi. Kizil.

Battulga, Ts., 2005. Mongolin runi bichgiyn baga dursgaluud. Ulaanbaatar: Corpus Scriptorum.

Bazin, L., 1964. La littérature épigraphique turc ancienne. In: Philologiae Turcicae Fundamenta, tom 2. Wiesbaden: ???. pp. 192–211.

Berta, Á., 2004. Szavaimat jól halljátok… A türk és ujgur rovásírásos emlékek kritikai kiadása. Szeged: JATEPress.

Berta, Á., 2010. Sözlerimi İyi Dinleyin… Türk ve Uygur Runik Yazıtlarının Karşılaştırmalı Yayını. Ankara: TDK Yayınları.

Bold, L., 1990. BNMAU-in nutag dahʻhadni bičees. Ulaanbaatar: Ulsin Hevlelijn Gazar. ???

Caferoğlu, A., 1968. Eski Uygur Türkçesi Sözlüğü. Istanbul: Edebiyat Fakültesi Basımevi

Cengiz, M., 2012. Katalog Drevnetyurskix Runiçeskix Pamyatnikov. Türkiyat Araştırmaları Dergisi, 2012 Bahar 16. Ankara: Hacettepe Üniversitesi yayınları. pp. 259-268

Clauson, Sir G., 1957. The Ongin inscription. Journal of the Royal Asiatic Society. pp. 177-192.

Clauson, Sir G., 1967. "Eski Türkçe Üzerine Üç Not". Translated into Turkish by A. Levendoğlu, Türk Dili Araştırmaları Yıllığı Belleten 1966, 19-37.

Clauson, Sir G., 1970. The Origin of the Turkish "Runic" alphabet. AO32. pp. 51-76.

Clauson, Sir G., Tryjarski, E., 1971. The inscription at Ikhe Khushotu. Rocznik Orientalistyczny 34. pp. 1-33.

Clauson, Sir G., 1972. An Etymological Dictionary of Pre-Thirteenth-Century Turkish. Oxford: Clarendon.

Clauson, Sir G., 2002. Studies in Turkic and Mongolic Linguistics. (2. ed.) London & New York: Routledge.

Comrie, B., Haspelmath, M., Bickel, B., 2008. The Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses. [online] Availible at: <https://www.eva.mpg.de/lingua/resources/glossing-rules.php> (accessed 30 Jul 2017)

Doerfer, G., 1983. Die Lanze und der alttürkische Genitiv. IV. MATK. Istanbul.

Doerfer, G., 1993. Bemerkungen zur Transkription des Runentürkischen.Journal of Turcology 1: 7-22.

Džumagulov, Č., 1971. Jazyk siro-tjurkskix (nestorjanskix) pamjatnikov Kirgizii. Frunze: Ilim.

Eraslan, K., 1980. Eski Türkçede isim-fiiller. İstanbul: İstanbul Üniversitesi Edebiyat Fakültesi.

Eraslan, K., 2012. Eski Uygur Türkçesi Grameri. Ankara: TDK Yayınları.

Erdal, M., 1979. The chronological classification of Old Turkish texts. CAJ 23. pp.151-175.

Erdal, M., 1991. Old Turkic word formation. A functional approach to the lexicon, volume I-II. Wiesbaden: Otto Harrassowitz.

Erdal, M., 1998. Old Turkic. In: Johanson, L., Csató, E., (eds.). The Turkic Languages. London: Routledge. pp.138-157.

Erdal, M., Gippert, J., Röhrborn, K. & Zieme, P., 2003. VATEC: Vorislamische Alttürkische Texte: Elektronisches Corpus. [online] Available at: <http://vatec2.fkidg1.uni-frankfurt.de/> (accessed 29 Jul 2017)

Erdal, M., 2004. Grammar of old Turkic. Leiden, Boston, Köln: Brill.

Erdal, M., 2010. Ongin Yazıtı. In: Şavk, Ü.Ç., (ed.). III. Uluslararası Türkiyat Ataştırmaları Sempozyumu 26-29 Mayıs Bildiriler Kitabı. Ankara: Research Institute for Turkish Studies of Haccetepe University. pp.363–372.

Ergin, M., 1984. Orhun Abideleri. Hisar. [online] Available at: <https://www.academia.edu/3376912/Orhun_%C3%A2bideleri> (accessed 29 Jul 2017)

Everson, M., 2008. Proposal for encoding the Old Turkic script in the SMP of the UCS. [online] Available at: <http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3357.pdf> (accessed 29 Jul 2017)

von Le Coq, A.A., 1909. "Köktürkisches aus Turfan (Manuskriptfragmente in köktürkischen 'Runen' aus Toyoq und Idiqut-Schähri [Oase von Turfan])". Sitzungsberichte der Preußischen Akademie der Wissenschaften, Phil.-hist. Klasse, 1909/41: 1047-1061.

von Gabain, A., 1941. Alttürkische Grammatik. Wiesbaden: Harrassowitz. Leipzig: Porta Linguarum Orientalium: 23.

von Gabain, A., 1963. Zenlralasiatische türkische Literaturen. I. Vorislamische alttürkische Literatur. In: Handbuch der Orientalistik, l, Abteilung, V. Band, l. Abschnitt, Turkologie. Leiden-Köln. pp.207-228.

von Gabain, A., 1964. Die alttürkische Literatur. In: Philologiae Turcicae Fundamenta, tom 2. Wiesbaden. pp.211-243.

von Gabain, A., 1974. Alttürkische Grammatik (3rd edition). Wiesbaden: Harrassowitz.

Gül, B., 2006. Moğolistan'daki Türk Yazıtları Üzerine Yeni Bir Eser. Modern Türklük Araştırmaları Dergisi (3,4). Ankara: Ankara Üniverstesi.

Golden, P. B., 2011. Studies on the Peoples and Cultures of the Eurasian Steppes. Bucharest-Braila: Editura Academiei Române - Muzeul Brâilei Editura Istros.

Hacıeminoğlu, N., 1996. Karahanlı Türkçesi Grameri. Ankara: TDK Yayınları.

Harmatta, J., (ed.), 1999. History of civilizations of Central Asia. Volume II. The development of sedentary and nomadic civilizations: 700 B.C. to A.D. 250. Delhi: Motilal Banarsidass.

Hayashi, T., Osawa, T., 1999. Site of Ikh-Khoshoot and Küli Čor inscription. In: Moriyasu, T., Ochir, A., eds. Provisional report of researches on historical sites and inscriptions in Mongolia from 1996 to 1998. The Society of Central Eurasian Studies. Tokyo. pp. 148-157.

Heikel, A.O., 1892. Les monuments prés de l´Orkhon. Inscriptions de l´Orkhon. Helsingfors.

Hovdhaugen, E., 1979. The structure and origin of the Turkish runic alphabet. I. MATK Tebliğler. 2. Türk Dili ve Edebiyatı. Istanbul: Istanbul Üniversitesi Edebiyat Fakültesi Türkiyat Enstitüsü. pp. 470-478.

Jadrincev, N. M., 1889. Predvaritel´nyj otchet o poezdke s archeologicheskoy celyu v Severnuyu Mongoliyu i verschiny Orkhona, Izvestiya Vostochnosibirskogo otdeleniya Russkogo Geograficheskogo Obshchestva 20, No 4, str. 1 n.

Janhunen, J. & Rybatzki, V., eds. 1999. Writing in the Altaic World. Studia Orientalia 87. Helsinki: Finnish Oriental Society.

Jísl, L., 1960. Výzkum Külteginova pamatníku v Mongolské Lidové Republice. Archeologické Rozhledy, 12-1, str.86-115.

Johanson, L., 1995. On Turkic converb clauses. In: Haspelmath, M. & König, E., eds. Converbs in Cross-Linguistic perspective. Structure and Meaning of Adverbial Verb Forms – Adverbial Participles, Gerunds. Berlin: Mouton de Gruyter. pp.313-347.

Johanson, L. & Csató, É. Á., eds. 1998. The Turkic Languages. London: Routledge.

Károly, L., Rentzsch, J., (eds.). A Database of Turkic Runiform Insciptions. [online] Available at: <http://www.runiform.lingfil.uu.se/> (accessed 29 Jul 2017)

Katayama, A., 1999. Tariat inscription. In: Moriyasu, T., Ochir, A., eds. Provisional report of researches on historical sites and inscriptions in Mongolia from 1996 to 1998. The Society of Central Eurasian Studies. Tokyo. pp. 168-176.

Kljaštornyj, S.G., Livšic, V.A., 1972. The Sogdian inscription of Bugut revised. Acta Orientalia Academiae Scientiarum Hungaricae, Vol. 26, No. 1. Akadémiai Kiadó. pp. 69-102.

Kljaštornyj, S.G., 1994. The Royal Clan of the Turks and the Problem of Early Turkic-Iranian Contacts. In: Acta Orientalia Academiae Scientiarum Hungaricae, XLVII. pp. 445-447.

Kondrat‘ev, V.G., 1981. Grammaticheskij stroj jazyka pamjatnikov drevnetjurkskoj pismennosti VIII-XI vv. Leningrad: Izdatelstvo Leningradskogo Universiteta.

Kononov, A.N. 1980. Grammatika jazyka tjurkskix runičeskix pamjatnikov. Leningrad: Nauka.

Kytö, M., Grund, P.J., Walker, T., 2011. Testifying to Language and Life in Early Modern England (Including CD-ROM An Electronic Text Edition of Depositions 1560-1760 (ETED)). Amsterdam: John Benjamins.

Kyzlasov, I.L., 2002. Pamjatniki runičeskoj pis’mennosti Gornogo Altaja (Učebnoe posobie). Čast’ pervaja: Pamjatniki jenisejskogo pis’ma. Gorno-Altajsk: RAN – Institut Arxeologii, et al.

Mallitskij, N.G., 1897. O svjazi tjurkskih tamga s orhonskimi pismenami. In: Protok. sozed. i soobšč. Turkestanskogo kružka ljub. arheologiji. (quoted from Caferoğlu 1968)

Malov, S.J., 1951. Pamjatniki drevnjetjurskoj pismennosti. Teksty i issledovanija. Moskva: Akademija nauk SSSR.

Malov, S.J., 1952. Enisejskaja pis‘mennost‘ tjurkov. Moskva-Leningrad: Akademija Nauk SSSR.

Malov, S.J., 1959. Pamjatniki drevnjetjurskoj pismennosti v Mongolii i Kirgizii. Moskva-Leningrad: Akademija Nauk SSSR.

Maue, D., 1996. Alttürkische Handschriften. Teil 1. Dokumente in Brahmi und tibetischer Schrift. Beschrieben und herausgegeben von D.M. In: Verzeichnis der Orientalischen Handschriften in Deutschland XIII 9. Stuttgart: Steiner.

Mau-Tsai, L., 1958. Die chinesischen Nachrichten zur Geschichte der Ost-Türken (T’u-küe), 2 Bde., Wiesbaden 1958 (Göttinger Asiatische Forschungen, Bd. 10).

McEnery, T., Wilson, A., 1996. Corpus Linguistics. Edinburgh: Edinburgh University Press.

Mert, O., 2009. Ötüken Uygur Dönemi Yazıtlarından Tes, Tariat, Şine Us. Ankara: Belen.

Moriyasu, T. & Ochir, A., eds. 1999. Provisional Report of Researches on Historical Sites and Inscriptions in Mongolia from 1996 to 1998. Society of Central Eurasian Studies.

Nadeljaev, V.M. et al., 1969. Drevnetjurkskij Slovar’. Leningrad: Nauka.

Nevskaya, I., 2011. Some paleographic and ortographic features of Altay Runic inscriptions. In: Şavk, Ü.Ç. (ed.). Orhon Yazıtlarının Bulunuşundan 120 Yıl sonra. Proceedings of the 3rd Runic symposium. Ankara. pp. 589-599.

Novák, Ľ., 2016. Babylónské zmatení písem v předislámské Střední Asii. In: Nekvapil, L., Ed. 2016. Kultura psaní v dějinách. Pardubice: Filozofická fakulta Univerzity Pardubice, pp.42-57.

Ölmez, M., 1998. Eski Türk Yazıtları ve Bugünkü Durumu. Çağdaş Türk Dili. 02/1998, 120.

Ölmez, M., 2010. Runik harfli Eski Türk Yazıtları / Old Turkic Runic Inscriptions. İBB Kültür ve Sosyal İşler Daire Başkanlığı: İstanbul.

Ölmez, M., 2011. “Eski Uygur ve Çin Kaynakları Işığında Orhon Yazıtlarında Geçen Yer ve Kişi Adları”, Orhon Yazıtlarının Bulunuşundan 120 Yıl Sonra Türklük Bilimi ve 21. Yüzyıl // 3. Uluslararası Türkiyat Araştırmaları Sempozyumu, 26-29 Mayıs 2010, Hacettepe Üniversitesi-Ankara.

Ölmez, M., 2015. Orhon-Uygur Hanlığı Dönemi. Moğolistan‘daki Eski Türk Yazıtları (Metin-Çeviri-Sözlük). Ankara: BilgeSu.

Ölmez, M., 2015b. What Should a New Edition of the Old Turkic Inscriptions Look Like? In: Taishan, Y. & Jinxiu, L. (eds.). International Journal of Eurasian Studies 2. pp.80-93.

Orkun, H. N., 1936. Eski Türk Yazıtları I. Istanbul: TDK.
- 1938. Eski Türk Yazıtları II. Istanbul: TDK.
- 1940. Eski Türk Yazıtları III. Istanbul: TDK.

Ōsawa, T., 2000. Moğolistan‘daki Eski Türk Anıt ve Yazıtları Üzerine Yeni Araştırmalar. 1996-1998 Japon-Moğol Ortak Çalışmalarının Ön Raporu. In: Türk Dilleri Araştırmaları 10 (2000). pp.191-204.

Ōsawa, T., 2008. Site and Inscription of Ongi revised. In: Türk Dilleri Araştırmaları, 18. pp.253-322.

Ōsawa, T., 2011. Revisiting the Ongi inscription of Mongolia from the Second Turkic Qaγanate on the basis of rubbings by G.J.Ramstedt. In: SUSA/JSFOu, 93. pp. 147-203.

Pritsak, O., 1963. Das Alttürkische. Handbuch der Orientalistik, 1. Abt., Bd.5, Abschn. 1. Leiden: Brill.

Pritsak, O., 1980. Turkology and the comparative study of the Altaic languages. The system of the Old Runic script. JTS4: 83-100.

Radlov, V.V., 1892-1899. Atlas drevnostey Mongolii. St. Peterburg.
Radloff, W., 1892-1899. Atlas der Altertümer der Mongolei. St. Petersburg.
1894. "I" Die Alttürkischen inschriften der Mongolei, Erste Lieferung. St. Petersburg.
1897. "II" Die Alttürkischen inschriften der Mongolei, Neue Folge. St. Petersburg.
1899. "II" Die Alttürkischen inschriften der Mongolei, Zweite Folge. St. Petersburg.

Radloff, V., 1897. Grammatische Skizze der alttürkischen Sprache, Neue Folge. St. Petersburg.

Radlov, V.V., Melioranskij, P.M., 1897. Drevnye-tyurkskie pamyatniki v Kosho-Tsaydame. Sbornik trudov Orkhonskoy ekspedicii IV. St. Petersburg.

Räsänen, M., 1957. Materialien zur Morphologie der türkischen Sprachen. SOF 21. Helsinki.

Räjäbov, A., Mämmädov, J., 1993. Orhon-Jenisej abidäläri. Baku: Jazichi.

Reynolds, J., Roueché, Ch., Bodard, G., 2007. Inscriptions of Aphrodisias. [online] Available at: <http://insaph.kcl.ac.uk/iaph2007>

Rinčen, B., 1968. Mongol nutag dax‘ xadnı bičees, gerelt xöšöönii züil. Ulaanbaatar: Corpus scriptorum Mongolorum.

Röhrborn, K., 1977-1988. Uigurisches Wörterbuch. Sprachmaterial der vorislamischen türkischen Texte aus Zentralasien. Lfg. 1-6. Wiesbaden: Steiner.

Röhrborn, K., 1996. Zur Suffixklassifikation im Alttürkischen. UAJb N.F. 14: 176-186.

Róna-Tas, A., 1987. On the development and origin of the East Turkic “Runic” script. AOH41: 7-14.

Róna-Tas, A., 1998a. The reconstruction of Proto-Turkic and the genetic question. In: Johanson, L., & Csató, E., (eds.). The Turkic Languages. London: Routledge. pp.67-80.

Róna-Tas, A., 1998b. Turkic writing systems. In: Johanson, L., & Csató, E., (eds.). The Turkic Languages. London: Routledge. pp.126-137.

Rybatzki, V., 1997. Die Toňukuk-Inschrift. In: Studia Uralo-Altaica, 40. Szeged.

Rybatzki, V., 1999. Punctuation rules in the Tonyuquq inscription?. In: Janhunen & Rybatzki, eds. Writing in the Altaic World. Studia Orientalia 87. Helsinki: Finnish Oriental Society. pp.207-225.

Rybatzki, V., 2011. Between East and West: Central Asian writing systems. In: Ölmez, M. & Yıldırım, F., eds. Orta Asya‘dan Anadolu‘ya Alfabeler. Istanbul: Eren Yayıncılık.

Sertkaya, O.F., 1985. Fragmente in altturkürkischer Runenschrift aus den Turfan-Funden. In: Röhrborn, K., Veenker, W. (eds.) Runen, Tamgas und Graffiti aus Asien und Osteuropa. Wiesbaden: Otto Harrassowitz.

Sertkaya, O.F., Alyılmaz, C., Battulga, T., 2001. Moğolistan‘daki Türk Anıtları Projesi Albümü. Ankara: Tika Yayını.

Sertkaya, O.F., 2008. Göktürk (Runik) Harfli Yazıtların Envanter, Alfabe ve Bibliyografya Problemleri Üzerine. Dil Araştırmaları Dergisi 2/2008.

Scharlipp, W.-E., 1992. Die Frühen Türken in Zentralasien. Darmstadt: Wissenschaftliche Buchgesellschaft.

Scharlipp, W.-E., 1994. Introduction to the Old Turkish Runic Inscriptions. Nicosia: University of Cyprus.

Scharlipp, W.-E., 2004. The Decipherment of the Turkish Runic Inscriptions and its Effects on Turkology in East and West. Journal of Turkish Civilization Studies, 2004, Issue 1, pp. 303-318.

Schulz, P., 1978. Verbalnomina und Konverbien als adverbiale Ergänzungen im Alttürkishen. PhD. Thesis, Giessen University.

Sims-Williams, N., 1975. 'Notes on Sogdian Palaeography.' In: Bulletin of the School of Oriental and African Studies, University of London, XXXVIII/I, 132-9.

Sims-Williams, N., Hamilton, J., 1990. Documents turco-sogdiens du IX^{e}-X^{e} siècle de Touen-houang. Corpus Inscriptionum Iranicarum: Part II Inscriptions of the Seleucid and Parthian Periods and of Eastern Iran and Central Asia. Vol. III Sogdian. London: School of Oriental and African Studies.

Sinor, D., 1990. The Cambridge History of Early Inner Asia. Cambridge University Press.

Subaşı Uzun, L., 1995. Orhon yazıtlarının dilbilimsel yapısı. In: Türk Diller ve Araştırmaları 7. Ankara: Simurg.

Suzuki, K., 2010. Newly Found Turkic Inscription from Bömbögör: On the Conflict for the Hegemony in Mongolia from the Qarluqs‘ Viewpoint. In: Arakawa, S., Takai, Y., Watanabe, K., (eds.). New Trends in Studies on Liao, Jin and Xi-Xia (3). Tokyo. pp. 1-30.

Ščerbak, A.M., 1961. Grammatičeskij očerk jazyka tjurkskix tekstov X-XIII vv. iz vostočnogo Turkestana. Moskva-Leningrad: Akademija nauk SSSR.

Šmahelová, L., 2014. Kül-Tegin monument. Turkic khaganate and research of the First Czechoslovak- Mongolian expedition in Khöshöö Tsaidam 1958. Disertační práce. FF UK UPRAV. https://is.cuni.cz/webapps/zzp/detail/104639/ (text & příloha)

Tekin, T., 1988. Orhon Yazıtları. Ankara: TDK.

Tekin, T., 1993. Irk Bitig. The book of omens. Wiesbaden: Harrassowitz.

Tekin, T., 1995. Les inscriptions de l'Orkhon. Istanbul: Simurg.

Tekin, T., 1997. A Grammar of Orkhon Turkic. The Hague: Mouton & Co.

Tekin, T., 2003. Orhon Türkçesi Grameri. Ankara: Sanat Kitabevi.

Tenišev, E.R., 1976. Otraženije dialektov v tjurkskix runicheskix i ujgurskix pamjatnikov. ST 1976/1. pp. 27-33.

Tezcan, S., 1990. Gibt es einen Namen Kök-türk wirklich? In: Baldauf, I., Kreiser, K., Tezcan, S., (eds.). Türkische Sprachen und Literaturen, Materialen der ersten Deutschen Turkologen Konferenz. Wiesbaden. pp. 357-76.

Thomsen, V., 1893. Déchiffrement des inscriptions de l'Orkhon et de l'ienissei, Notice préliminare. Kopenhag: Bulletin de l'Académie Royale des Sciences et des Lettres de Danemark. pp.185-299.

Thomsen, V., 1896. Les inscriptions de l'Orkhon déchifrées. Helsingfors: Société de littérature

Thomsen, V., 1922. L‘alphabet runiforme turc. Samlede Afhandlinger III. Kobenhavn. pp.27-82.

Tryjarski, E., 1966. The Present State of Preservation of Old Turkic Relics in Mongolia and the Need for their Conservation. UAJb 37. UAJb 38. pp. 158-173.

Tryjarski, E., 1981. Die alttürkischen Runeninschriften in den Arbeiten der letzten Jahre. Befunde und kritische Übersicht. AOF VIII, 339-352.

Tybykova, L. N., Nevskaya, I. A., & Erdal, M., 2012. Katalog drevnetjurkskix runičeskix pjamjatnikov Respubliki Gornyj Altaj. Gorno-Altajsk: Gorno-Altajskoe knižnoe izdatel’stvo.

Tybykova, L., Nevksaya, I., 2013. Pamjatniki runičeskogo pisma Gornogo Altaja. [online] Availible at: <http://www.altay.uni-frankfurt.de/> (accessed 29 Jul 2017)

User, H.Ş., 2009. Köktürk ve Ötüken Uygur Kağanlığı Yazıtları. Söz Varlığı İncelemesi. Konya: Kömen Yayınevi.

User, H.Ş., 2011. Runik Türk Yazıtları Çerçevesinde katun ve kunçuy. In: Ölmez, M., Aydın, E., Zieme, P., Kaçalin, M. S., (eds.). Ötüken‘den İstanbul‘a Türkçenin 1290. Yılı (720-2010) Sempozyumu, Bildiriler. İstanbul: İstanbul Büyükşehir Belediyesi. pp. 281-294.

User, H.Ş., 2015. Bombogor Inscription: Tombstone of a Turkic Qunčuy („Princess“). Journal of the Royal Asiatic Society, November 2015. pp. 1-9.

Vasil'ev, D.D., 1983. Inventory of graphic monuments of Türkic runic writing in the Asia area. (Attempt of systematizing). Moskva.

Vasil'ev, D.D., 1983. Korpus tjurkskix runičeskix pamjatnikov bassejna Jeniseja. Leningrad: Akademija Nauk SSSR.

Vovin, A., 2005. The End of the Altaic Controversy (review article of Sergei Starostin, Anna Dybo, and Oleg Mudrak, 2003. Etymological dictionary of the Altaic Languages. Leiden: Brill.). Central Asiatic Journal, 49.1. pp. 71-132.

Yoshida, Y., Moriyasu, T., 1999. Bugut Inscription. In: Moriyasu, T. & Ochir, A., (eds.), 1999. Provisional Report of Researches on Historical Sites and Inscriptions in Mongolia from 1996 to 1998. Society of Central Eurasian Studies. pp. 122-124.

Yoshida, Y., 2013. When Did Sogdians Begin to Write Vertically?. In: Linguistic Papers 33, pp. 375-391.

8 APPENDIX: LIST OF STANDARD ABBREVIATIONS

1	first person
2	second person
3	third person
ACC	accusative
ADJVZR	adjectivizer
AOR	aorist
CVB	converb
EMP	emphatic
EQT	equative
INFR	inferential
POSS	possessive
PST	past
FUT.PTCP	future participle
SG	singular