Ilya Egorov
School for Advanced Studies in the Humanities, RANEPA, Moscow; i.m.jegorow@gmail.com

Basic vocabulary of closely related languages in contact: case study of Turkic languages on the Crimean Peninsula

The present paper provides two case studies of the basic vocabulary of the Turkic languages spoken on the Crimea Peninsula. Its aim is to illuminate the issues that a historical linguist, and in particular a phylogeneticist, faces when analyzing the basic vocabulary of closely related languages in a situation of intensive contact. The first case study is dedicated to the onomasiological reconstruction of the Proto-Karaim Swadesh list. The main problem here is detection of the West Oghuz loans and especially of contact-induced archaization (fake archaisms) in Crimean Karaim. The objective of the second case study is to identify the genealogical affiliation of the Crimean Tatar dialects. Both the manual analysis of the innovations in the basic vocabulary and the computational lexicostatistics (Bayesian approach, Neighbor-joining, Maximum Parsimony Analysis) confirm the traditional view that the Coastal dialect belongs to the Oghuz subgroup, the Orta dialect – to the West Kipchak subgroup, and the Steppe dialect – to the Nogai Kipchak subgroup. Such affiliations fully fit the documented ethnic history. The correct genealogical affiliation of the dialects in question became possible only after exclusion of all the loans, which has not been done in previous lexicostatistical studies of Crimean Tatar. Both cases show that careful elimination of areal influences is crucial for semantic (onomasiological) reconstruction and phylogenetic studies.

Keywords: phylogeny; semantic reconstruction; lexical borrowings; Karaim language; Crimean Tatar language; Turkic languages.

1. Introduction

The procedure of reconstruction in comparative-historical linguistics implies being able to distinguish between inherited and loaned items and patterns. This statement is true for phonologic, morphologic, and semantic reconstruction. Usually, when a word violates regular sound correspondences, it is treated as a borrowing unless it can be explained as the result of an analogical or another occasional change. Of course, extra sets of sound correspondences can appear between remotely related or unrelated languages as a result of phonological adaptation as well. However, as a rule, such borrowings can be revealed relatively simply, based on their distribution in the contacting subgroup. Various specific problems arise in the case of borrowings from a genetically related language, cf. for example the so-called "etymological nativisation", described in detail by Ante Aikio (2007) for Finnish loans in the Northern Saami.

Problems caused by contacts between closely related languages are relevant not only for traditional historical-linguistic studies, but also for linguistic phylogeny. The issue of homoplasy and especially horizontal transfer has been redefined in the last decades, see Nakhleh, Ringe & Warnow 2005; Nelson-Sathi et al. 2011 and Kassian 2017. These works make linguists aware of the problem and propose methods to uncover and eliminate it. Early criticisms of Moris Swadesh's lexicostatistic and glottochronological methods were caused mostly by incorrect interpretation of loans. One of the most known critical works is Knut Bergsland and Hans Vogt's paper (1962), where it was argued that literary Norwegian (Riksmål) demonstrates a drastically longer distance from Old Norse than Icelandic. The problem was that both contact-

induced and autonomous replacements in the basic vocabulary were considered valid for measuring genealogical distance, whereas in reality the effect of the first group is highly dependent on the specific sociolinguistic situation. Revisiting this case, Sergei Starostin (2000: 230) has shown that 16 of 20 innovations in Riksmål are loans: 11 from Danish, 3 from Swedish and 2 from German. Hence, if they are excluded, the percentage of innovations more or less equals that in the other Scandinavian languages. Nowadays, the detection of loans when reconstructing phylogeny has become an obligatory requirement at least in the Moscow school of comparative linguistics.

However, such drawbacks still arise in more recent phylogenetic studies applying lexicostatistics. For instance, confounding true cognates and borrowings, Russell Gray and Quentin Atkinson (2003) and then Remco Bouckaert et al. (2012) have inferred such a structure for the Slavic group in which Polish forms one clade with Belarusian, Ukrainian, and Russian. This contradicts the existing consensus which assumes a trifurcation of Proto-Slavic into the following subgroups: [Polish, Czech, Slovak, Sorbian], [Slovenian, Serbian/Croatian/Bosnian, Bulgarian, Macedonian], [Belarusian, Ukrainian, Russian]. Such affiliation of Polish is caused by undetected Polish loans in the Belarusian wordlist used in the forenamed works (see the linguistic supplement in Kushniarevich et al. 2015 for more detailed criticism). In section 4.6, I address identical problems in a recent work on Turkic phylogeny.

In the present paper, I intend to discuss two cases which illustrate the problems with the basic vocabulary of the languages undergoing intensive influence on the part of their close relatives. In my investigation of the Turkic languages of Crimea, I attempt to show the challenges they pose to a historical linguist, when the new method of onomasiological reconstruction is applied to identify the genealogical affiliation of a language. The Turkic languages spoken on the Crimean Peninsula provide suitable material for discussion of these issues for the following reasons: (a) they are related to each other approximately at the same depth as Riksmål and its Scandinavian relatives from the canonical example cited above; (b) the tree structure and historical phonology of the Turkic family are known well enough for the purposes of our research; (c) the ethnic, sociolinguistic, and political history of Crimea is well documented.

The remainder of this paper is structured in the following way. Section 2 contains basic information on the Turkic languages of Crimea, their traditional genealogical affiliation, and the sociolinguistic situation in the region, along with a short annotated bibliography. Sections 3 and 4 deal with semantic (onomasiological) reconstruction of the Proto-Karaim wordlist and with revision of the genealogical affiliation of the Turkic varieties spoken in Crimea respectively. Each section contains its own introductory, methodological, analytical subsections and discussions of the results. Section 5 summarizes what can be learned from the considered cases.

2. Turkic languages of Crimea

In this section I provide the most important information on the sociolinguistic situation in the Crimean Peninsula, traditional genealogical affiliation of the languages, dictionaries, grammar and other sources used in the present study.

2.1. Crimean Tatar

The group of dialects traditionally referred to as the Crimean Tatar language actually represents a paraphyletic formation (Sevortyan 1966). It includes (1) Coastal dialect, which is genetically an Oghuz language most closely related to Turkish and Gagauz, (2) Orta (also called

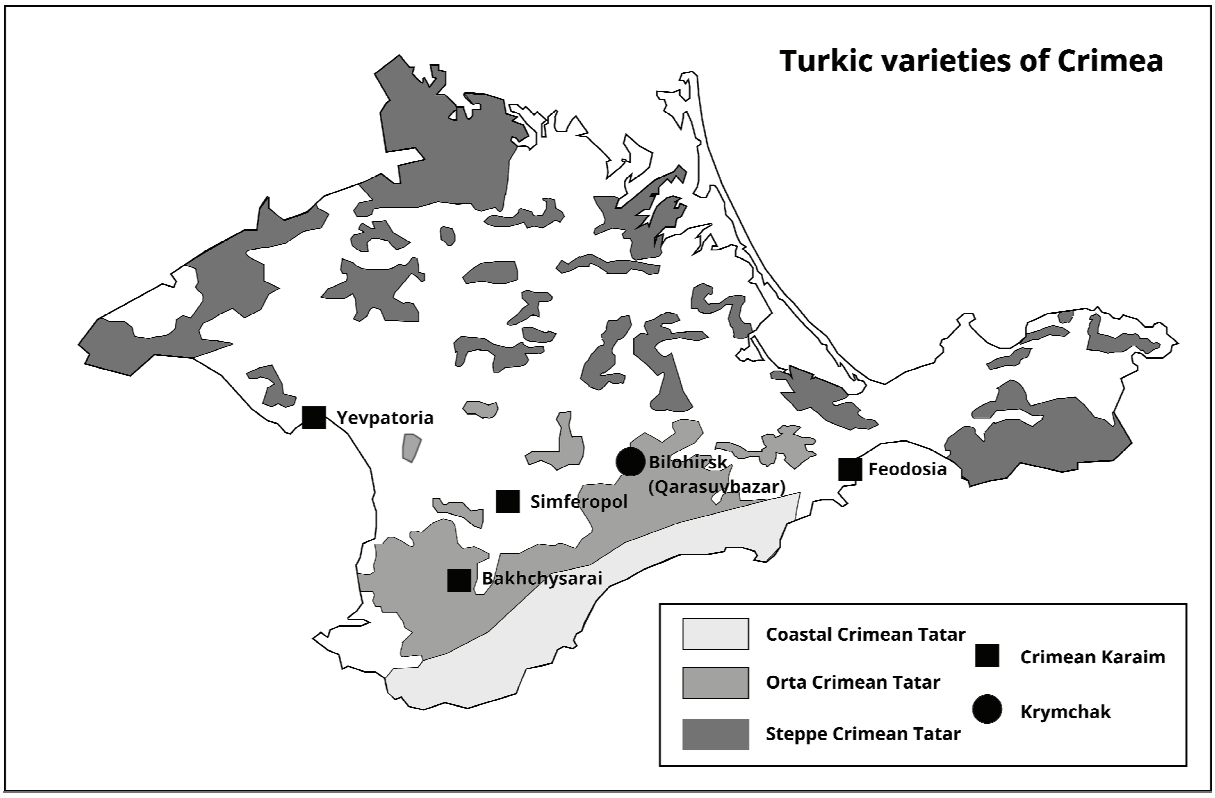

Figure 1. Turkic varieties of Crimea. The map has been drawn on the basis of the Soviet ethnographic map of Crimea 1926; Filonenko 1931 and Radloff 1896: xiv–xvi. The border between the Coastal and Orta dialects is somewhat arbitrary.

Central or Middle) and (3) Steppe dialects, both belonging to different Kipchak subgroups. The Coastal dialect is sometimes named Crimean Turkish; such term reflects its genealogical affiliation exactly. It became the dominant language in the Crimean Khanate, which was a vassal state of the Ottoman Empire. Modern literary Crimean Tatar is based on the Orta dialect.

Dictionaries: Useinov 2007 – dictionary of the Literary Crimean Tatar which is based on the Middle dialect.

Grammars: Sevortyan 1966; Izidinova 1996 – short grammar sketches; Kavitskaya 2010 – grammar based on the field notes from the early 2000s.

Other materials and studies: Polinsky 1992 – 100-wordlists for three Crimean Tatar dialects. I also use the wordlist recently collected after my own initiative, which can be found in Supplement 1.

2.2. Karaim

Karaim (also called Karaite) is a subgroup of Kipchak languages consisting of three dialects, sometimes treated as three separate languages. Only one of them was spoken by Jewish Karaite community in Crimea until recently. Two other dialects (Trakai and Halich) appeared as the result of the migration of Karaims from Crimean Khanate to the Grand Duchy of Lithuania. The migration to Trakai started in 1397, to Halich in 1407–1409 and continued up to the 15th and 16th centuries (Musaev 2010: 205–206). Karaim is traditionally classified together with Karachay-Balkar, Kumyk and with the Middle dialect of Crimean Tatar as a West Kipchak language (Johanson 1998: 82). Maria Polinsky treats Crimean Karaim as an “ethnolect” of Crimean Tatar belonging to the Oghuz subgroup.

Dictionaries: Baskakov, Szapszał & Zajączkowski 1974 remains the most reputable source on the lexicon of all dialects; Aqtay & Jankowski 2015 deals with Crimean Karaim, includes all Crimean materials from Baskakov, Szapszał & Zajączkowski 1974 and from other written sources.

Grammars: Musaev 1964 deals with the Trakai and Halich dialect; Musaev 2010 contains information on Crimean Karaim as well; Prik 1976 – grammar sketch of the Crimean Karaim.

Other materials and studies: Kocaoğlu 2006 – texts in the Trakai dialect with brief grammar sketch and vocabulary; Polinsky 1992 – 100-wordlist for the Crimean dialect; 110-wordlist for Trakai dialect, speaker's self-recording made in 2019.

2.3. Krymchak

Krymchak is the language of the other Jewish community, which survived until the end of 20th century in the town of Qarasuvbazar (Ukr. *Bilohirsk*). In the late 20th and early 21st century, some attempts at revitalization were undertaken. However, now this language is extinct. In a number of works, Krymchak is treated as a Kipchak language. Polinsky names it (as well as Crimean Karaim) an "ethnolect" of Crimean Tatar, i.e. an Oghuz language.

Dictionaries: Rebi 2004 – the dictionary created by language activists; Ianbay 2016.

Other materials and studies: Polinsky 1992 – 100-item wordlist and short grammar sketch; Polinsky 1991 – text sample; Jankowski 2017 – overview of grammar and major sources.

3. Reconstructing the Swadesh wordlist for Proto-Karaim

3.1. Introductory remarks

Traditional reconstruction of lexical semantics remains extremely arbitrary. A typical meaning of a reconstructed root or even a lexeme is 'a kind of tree' or 'to stack, to collect, to dump, to put in order, to build up'. Such definitions are the results of two wrong methods of semantic reconstruction: 1) reduction of all meanings attested in the daughter languages to a wide semantic component; 2) extrapolation of all attested meanings onto the proto-language. Sometimes this results in openly ridiculous situations: thus, according to Dybo 1996: 18, about 70% of the Proto-Indo-European verbal roots in Julius Pokorny's dictionary (1959) mean 'to bloat, to swell' or 'to bend'. For further criticism of the traditional semantic reconstructions see Burlak & Starostin 2005: 248. To solve this problem, the method of onomasiological reconstruction has been elaborated in the recent years. I will discuss it in the next section.

Why do we need reconstructed Swadesh lists? It seems reasonable to use reconstructed wordlists for commonly accepted low-level groups when investigating the tree structure of a deeper family. The principle of step-by-step reconstruction is a commonly accepted standard in comparative-historical studies. It is widely applied for phonological reconstruction. For instance, if one introduces a Germanic word into Indo-European comparison, methodologically it is more correct to use a reconstructed Proto-Germanic form instead of Gothic, Old High German, Old North and Old English, since each of them demonstrates innovations that are irrelevant to external comparison. Similarly, more correct is the use of a Proto-Germanic wordlist when reconstructing the Indo-European tree. This exact approach was recently used in Kassian et al. forthcoming. The reconstructed Proto-Karaim wordlist can be used when revising the topology of Turkic family and reconstructing the Proto-Turkic wordlist for further comparison.

3.2. Methods

A relatively strict method of onomasiological reconstruction was recently developed by the representatives of the Moscow School of comparative linguistics (see Kassian, Starostin & Zhivlov 2015: 304–306; Starostin 2016). It involves tracing a way from the meaning to its optimal exponent in the protolanguage, i.e., determining which word was used for a given concept in a protolanguage. Selection of the optimal candidate is guided by five principles, which are very similar to the ones used for detecting archaisms and innovations when reconstructing phonology. The basic principle is topological (1); others (2–5) are used in competitive situations, i.e. when tree topology allows no unambiguous judgment on the candidates. Here I only provide a brief synopsis; for strict definitions, further explanations and examples see the abovementioned works:

1) tree topology: the root attested in different branches is preferable;
2) external etymology: the root is preferable if its external cognates preserve the same Swadesh meaning;
3) internal derivability: the primary root (as opposed to polymorphemic derivates) is a preferable candidate;
4) typology of semantic shifts: the typologically frequent direction of semantic shifts to be assumed when reconstructing scenario of semantic changes of the potential candidates;
5) areal effect exclusion: contact-induced innovations to be excluded.

It must be noted that these principles are different from the criteria for synchronic wordlists: terms for synchronic Swadesh wordlists are selected based on its frequency and stylistic neutrality.

Another advantage of onomasiological reconstruction is that it allows the elimination of uncertainties and mistakes in a synchronic list. Thus, if the quality of data on one language is not irreproachable, this can be compensated for with data on its relatives. In the process of reconstruction, an incorrectly selected term for one of the languages is likely to be seen as an innovation in this particular idiom. If a wordlist is overcrowded with inappropriate archaisms, the situation becomes more difficult. The solution to this problem is proposed in Section 3.3. Although the probability that one inaccurate list will influence the structure of the phylogenetic tree is lower and onomasiological reconstruction helps fix some defects in the data, the motto “garbage in – garbage out” remains fully true.

Next, I will concentrate on detecting contact innovations in Crimean Karaim, which make relatively shallow Proto-Karaim reconstruction difficult. Since the phonological inventory of Crimean Karaim and Coastal Crimean Tatar is very similar to each other, the phonological criterion is not particularly helpful in this case. Obviously, the reflexes of Proto-Turkic stems differ in these languages and sometimes these differences point to the Oghuz origin of the word. However, this criterion cannot be applied to every word. I will mostly use the distributional criterion, which can be described as follows. For instance, four languages (L1, L2, L3, L4) related to each other among which L1 is an outgroup and L2, L3, L4 form a separate clade are taken into consideration. The lexeme A is the basic term for the meaning ‘M’ in L1. In L2, lexemes A and B are synonyms; in L3, L4, ‘M’ is denoted only by the lexeme B (see Figure 2). If A is a primary root whereas B is a transparent derivative or semantic innovation, this constitutes strong evidence for reconstructing *A for the meaning ‘M’ in the Proto-L2–4. However, if L1 influences L2, A should be regarded as a borrowing. The coexistence of A and B in L2 is an additional argument for such a solution.

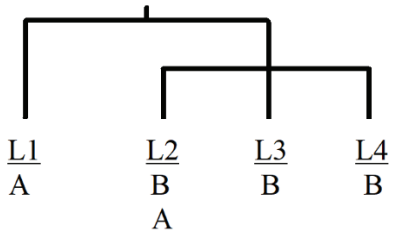

Figure 2. Tree structure and distribution of lexemes with meaning 'M' which hint at horizontal transfer of the stem A.

Extrapolating this scheme to the case in question, I will regard a word as an Oghuz borrowing in Karaim if it is widely spread in the Oghuz or at least in the West Oghuz languages and is uncommon or completely absent in Kipchak.

In addition to being subject to horizontal transfer (MAT-borrowings in Jeanette Sakel's (2007: 15) terminology), some stems can also undergo contact-induced semantic shifts (PAT-borrowings in Sakel's terminology). The last phenomenon is also known as loan meaning extensions (see Haspelmath 2009). In the case of closely related languages, this is driven by obvious, naïve logic: "if words sound similarly they must have similar meaning".

3.3. Contact innovations in Crimean Karaim: clear cases

More evident borrowings will be considered at first. All of them are loaned from the Coastal dialect of Crimean Tatar. Some of these loans were in turn borrowed into West Oghuz languages from Persian and Arabic. Theoretically, one can assume that Oghuz-like lexemes in Crimean Karaim are inherited and Kipchak-like ones are borrowed. However, such assumption faces more difficulties, since it is hardly possible to find the source of potential Kipchak borrowings which occurred in the Trakai and Halich dialects and sporadically in the Crimean dialect as well.

In citing examples, I first give the number of the concept in the 110-item Swadesh list following Kassian et al. 2010, then list the language material with references. The abbreviation for the source used in Baskakov, Szapszał & Zajączkowski 1974 and Aqtay & Jankowski 2015 is given in brackets for Crimean Karaim forms. For the full reconstructed Proto-Karaim Swadesh wordlist see Supplement 1. For the transcription and transliteration of the examples, I use the Unified Transcription System applied in the Global Lexicostatistical Database (https://starling.rinet.ru/new100/UTS.htm).

5. **big** – CrKar. *balaban* (Sz) 'big, huge' (Baskakov, Szapszał & Zajączkowski 1974: 100). It is difficult to define whether *balaban* is appropriate even for the synchronic Crimean Karaim Swadesh list. There are two other candidates that will be considered below. CrKar. *balaban* is a clear Oghuz borrowing, cf. Tur. *balaban* 'huge', Gag. *balaban* 'high', CoCrTat. *balaban* 'big'. The root is extremely rare beyond the Oghuz languages. Only two *comparanda* are mentioned in Dybo 2013: 128–131: Tat. dial. *balban* 'fat, overweight, stout', Kirg. *balpay-* 'to seem big, bulky, clumsy'. Details of its etymology remain obscure (see the cited work for the review of existing hypotheses), however, the innovative nature of the meaning 'big, huge' is quite obvious.

CrKar., TrKar., HKar. *biyik* (Sz, Par. 101 v. 1) 'big, high, great' (Baskakov, Szapszał & Zajączkowski 1974: 115) is another stem which could be treated as a contact-induced innovation. This stem with the meaning 'high' is widespread across Turkic languages (see Dybo 2013: 123) whereas the more general meaning 'big' is limited to the Oghuz subgroup (Tur. *büyük*, Gag. *bü:k*, Az. *böyük*), Karakhanid Uyghur, Old Uyghur, and Sary Yugur. Such distribution theoretically can be an indication of the antiquity of the meaning 'big' (cf. Clauson 1972: 302 for an in-

terpretation), but **ulug* is the better candidate for both Proto-Karaim and Proto-Turkic 'big'. All Karaim dialects demonstrate its reflexes: CrKar. *ulu* (Sz) ~ *ulɨ* 'great, big', TrKar. *ullu* 'big, great, important' ~ *uŋlu* 'great, big, elder' ~ *ullux* 'big, great' (with additional suffix), HKar. *ullu* 'big, great, important' (Baskakov, Szapszał & Zajączkowski 1974: 577, 579). At least modern speakers of the Trakai dialect use *ullu* as the basic word for 'big', according to our data; it is also confirmed with materials published in Kocaoğlu 2006. The semantic shift 'big' > 'great' is typical for the world's languages, the synchronic polysemy is also common in the Turkic family (Dybo 2013: 120–121) and cross-linguistically (Rzymski et al. 2019). The direction of the shift 'big (of a physical object)' > 'great (high status)' is more probable than vice versa due to the common tendency of the development from concrete meanings to more abstract ones (Campbell 2013: 237). In sum, the old term for 'big' in the Proto-Karaim subgroup is **ullu*; the semantic shift 'high' > 'big' of the stem **biyik* should have been triggered by contact with the Oghuz dialect of Crimea before the start of the migration to the Grand Duchy of Lithuania. Note also the form *büyük* in the Crimean dialect, which looks as if it was recently borrowed from Turkish. Aqtay and Jankowski (2015: 88, 100) list the latter form with the gloss 'great, big' whereas *biyik* is glossed as 'high'.

It is difficult to make a choice between three candidates with the meaning 'big' for synchronic Crimean Karaim based on existing sources, which partly contradict each other. Provisionally, I assume that *ulu ~ ulɨ* has the more abstract meaning 'great' whereas *biyik ~ büyük* and *balaban* compete with each other in the basic meaning 'big'. Both are loans, the meaning 'big' by the lexeme *biyik* is borrowed from Oghuz; *büyük* and *balaban* are MAT-borrowings. The choice is much simpler for Proto-Karaim. Thus, this case illustrates an important advantage of onomasiological reconstruction: uncertainty in the data on one of the languages does not influence the final list.

67. **red** – CrKar. *qɨrmɨzɨ ~ qɨrɨmzɨ* 'red'. This is an Arabic loan common in Oghuz languages, cf. Tur., Gag. *qɨrmɨzɨ* 'red', Az., Turkm. *gɨrmɨzɨ* 'red'. In the Trakai and Halich dialects, *kɨrmɨzɨ* denotes a specific shade 'purple, magenta' (Baskakov, Szapszał & Zajączkowski 1974: 381, 387). The archaic stem *kɨzɨl*, reflecting PT **Kɨrʸɨl* 'red', has been found with the meaning 'red, orange' for these dialects; it should be the basic term for 'red' (Baskakov, Szapszał & Zajączkowski 1974: 383). The reflexes of **Kɨrʸɨl* are not attested for Crimean Karaim in Baskakov, Szapszał & Zajączkowski 1974, but Aqtay and Jankowski (2015: 309) cite it with the gloss 'red, ruddy'. The stem **qɨzɨl* can be reconstructed for Proto-Karaim 'red' with complete certainty. It is one of the most stable Turkic stems.

69. **root** – CrKar. *kök* 'root' (Baskakov, Szapszał & Zajączkowski 1974: 337). This stem is widely attested with the meaning 'root', however, its basic meaning in the majority of the Turkic languages is more abstract ('basis'), it develops various metaphorical meanings as well. As the basic term for 'root' this stem is attested in the Oghuz languages, cf. Tur., Gag. *kök,* from which it has been borrowed into Crimean Karaim. The stem **dạmor* > **tamur* is a better candidate for Proto-Turkic and Proto-Karaim 'root'. It is preserved all over the Turkic-speaking area including the languages in question and their numerous Kipchak relatives, cf. Karaim reflexes: CrKar., *tamur ~ tamar* 'root, vein', HKar. *tamar ~ tamur* 'root, vein', TrKar *tamur* 'root, vein' (Baskakov, Szapszał & Zajączkowski 1974: 509–510).

70. **round** – CrKar. *müdever ~ mudever* 'round', *yuvarlaq* 'round, globular', *tomalaq* 'round, full, plump', *yumalaq* 'globular, round' (Aqtay & Jankowski 2015: 247, 407, 472, 474), *tögerek* 'round' (Baskakov, Szapszał & Zajączkowski 1974: 541). The first term has an Arabic origin and is borrowed via Turkish, cf. Tur. *müdevver* 'circular, round'. The second one is an Oghuz borrowing as well, cf. Tur., Gag. *yuvarlaq* 'round'. The stems *tomalaq* and *yumalaq* have been attested in other Turkic languages:

Uzb. *yumalɔq* ‘round (sphere & circle)’, Karak. *žumalaq* ‘round (circle)’, for other derivatives with the Proto-Turkic bound root **yum-* see Dybo 2013: 441–442;

Gag. *tombarlaq* ‘round’, Uzb. *dumalɔq* ‘round (sphere & cylinder)’, Uyg. *domlaq* ‘round (sphere)’, Bash. *tumalaq* ‘round (sphere & circle)’, Nog. *tɨmalaq* ‘circle, sphere (n.)’, Kaz. *domalaq* ‘sphere (n.)’, Karak. *dumalaq* ‘round (sphere)’ – note that this set of phonetically similar forms demonstrates suspiciously irregular sound correspondences!

However, none of the Karaim stems listed above have been sufficiently confirmed by the Trakai and Halich data; the meaning ‘round’ is insufficiently documented in the existing sources. Only the stem *tögerek* has a Trakai cognate. According to the recently collected wordlist for the Trakai dialect, either the collocation *galgan kibik* literally means ‘circle-like’ or *tägere ~ tegäräk* ‘round, circle’ is used as an adjective ‘round’. CrKar. *tögerek* and TrKar. *tägere ~ tegäräk* are treated as a Mongolian reborrowing and an inherited stem respectively in Dybo 2013: 238–239. However, I believe that it is reasonable to consider them true cognates and to reconstruct **tögerek* for Proto-Karaim with the meaning ‘round (circle)’.

Until there is a corpus for Karaim, the choice for the synchronic basic term is difficult both for the Crimean and Trakai dialects. The stem *tägere ~ tegäräk* is a single candidate for Trakai ‘round’, Crimean *tögerek* can apply at least for ‘round 2D’. Further details in synchronic dialects remain obscure.

79. **smoke** – CrKar., TrKar. *tüt-sü* (Sz) ‘smoke, incense’ (Baskakov, Szapszał & Zajączkowski 1974: 555). This is an old contact innovation shared by the Crimean and Trakai dialects. The substantives from the verb PT **tüt-* formed with the not especially productive suffix **-süg* have been found only in Oghuz languages, cf. *tütsü* Tur. ‘incense’, Az. *tüstü* ‘smoke’ with metathesis inside the consonant cluster, Turkm. *tüsse* ‘smoke’, Sal. *tɨssɨ* ‘smoke’, for this suffix see Räsänen 1957: 141. All other Turkic languages, including even Chuvash, demonstrate the suffix **-ün* (Dybo 2013: 479). So we treat CrKar. and TrKar. *tüt-sü* ‘smoke, incense’ as a borrowing which occurred before the migration from Crimea to the Grand Duchy of Lithuania. The inherited forms with **-ün* have been found in all Karaim dialects as well: CrKar. *tütün* ‘smoke, tobacco’, TrKar. *tʸutʸunʸ* ‘smoke, tobacco’, HKar. *titin* ‘smoke’ (Baskakov, Szapszał & Zajączkowski 1974: 532, 555, 571). Hence, **tütün* must be reconstructed for Proto-Karaim ‘smoke’.

103. **near** – CrKar. *yaqɨn* is attested in the Crimean dialect (Baskakov, Szapszał & Zajączkowski 1974: 220) beside CrKar. *yuwuq*, TrCar. *yuwux*, HCar. *yuwuk* ‘near’ (Baskakov, Szapszał & Zajączkowski 1974: 253–254). I consider the first stem an Oghuz borrowing: Tur. and Gag. *yaqɨn* ‘near’, Az. *yaχɨn* ‘near’, Turkm. *yaqɨ:n* ‘near’. Both stems are widely spread across Nuclear Turkic languages. However, the narrow distribution in the Karaim dialects allows us to treat **yaqɨn* as a borrowing. Its competitor, **yuwuq*, which can be found in all Karaim dialects, is definitely the better candidate for Proto-Karaim ‘near’. The distribution of its external cognates points to the stem discussed above as to the main exponent of the meaning ‘near’ not only in Proto-Kipchak and even in Proto-Turkic, see Dybo 2013: 539–540.

3.4. Contact archaization in Crimean Karaim: fake archaisms

In this section, I consider the most curious cases. There are some stems which can seem archaic at first sight, but in reality turn out to be loanwords. For such cases, I suggest the term ‘fake archaisms’. Revealing this kind of borrowings is crucial for onomasiological reconstruction. Fake archaism must be suspected when principles of tree topology, external etymology, and internal derivability come in conflict with the principle of areal effect exclusion. The semantic plausibility principle, i.e. the typology of semantic shifts, theoretically, can also contradict the principle of areal effect exclusion but such cases have not been attested in our material.

Thus, fake archaisms can be successfully detected when areal distribution and the direction of influence are taken into account. In Crimean Karaim, four examples of fake archaisms have been found.

22. **to eat** – there are two candidates for filling this slot:

1) CrKar. *ye-* (Sz) 'to eat' (Baskakov, Szapszał & Zajączkowski 1974: 268);

2) CrKar., TrKar. *aš-a-* (Sz) 'to eat', HKar. *as-a-* 'to eat' (Baskakov, Szapszał & Zajączkowski 1974: 79, 91).

The root **ye-* should have an advantage due to the principle of external etymology. It is found not only in numerous non-Kipchak languages but even in Chuvash, whereas **aš-a-* is limited to the Nuclear Turkic languages. In many of them, it is often a marked polite term ≈ Rus. *kušat'*. The principle of internal derivability also speaks for the primary root **ye-*, since the verbal stem **aš-a-* can be analyzed as a synchronic derivative from **aš-* 'food'. However, **ye-* is limited only to the Karaim dialect that was under intense influence on the part of Oghuz. The Oghuz languages preserve **ye-* as the basic exponent of 'to eat' (Tur., Az. *ye-*, Gag. *i-*, Turkm. *iy-*, Sal. *yɨ-*). Thus, one can simply consider CrKar. *ye-* a borrowing. A probable situation is that **aš-a-* already becomes the basic term for 'to eat' in Proto-Karaim, but archaic **ye-* as a marginal term still remains in Proto-Karaim. Under foreign influence **ye-* could become the basic term again, i.e. we deal with a semantic backformation.

83. **sun** – three words glossed in this way have been found in Crimean Karaim, and two of them can apply for the status of the basic term in Proto-Karaim.

1) CrKar. *kün* 'sun, day' (Sz, R) ~ *gun* 'day' (Par 77 v. 11), cf. TrKar. *kʸunʸ* 'day', HKar. *kin* 'day' (Baskakov, Szapszał & Zajączkowski 1974: 167, 320, 353, 396);

2) CrKar. *küneš* ~ *güneš* 'sun' (Baskakov, Szapszał & Zajączkowski 1974: 354; Aqtay & Jankowski 2015: 169, 225);

3) CrKar. *quyaš*, TrKar. *kuyaš*, HKar. *kuyas* 'sun' (Baskakov, Szapszał & Zajączkowski 1974: 344, 372).

The first item attested mostly with the meaning 'day' demonstrates also the meaning 'sun' in Crimean Karaim. It is a reflex of the stable Proto-Turkic stem **gün* 'day, sun' which retained this meaning across the whole area of the Turkic languages. The second stem was derived from the first one with a not quite clear suffix. It occurs sporadically in various languages, cf. Tur., Kum. *güneš* 'sun', OT *küneš* 'sun'. The Crimean Karaim form *küneš ~ güneš* is a transparent Western Oghuz loan due to the initial voiced consonant. The last stem, *quyaš*, is a result of the semantic shift 'heat' > 'sun', which should independently occur in a couple of Turkic languages. Thus, when one chooses between **kün* and **quyaš*, the external etymology principle strongly points to the first stem as the better candidate for filling the slot 'sun' in Proto-Karaim. However, in light of Tur., Gag. *gün* 'sun, day' (attested simultaneously with *güneš* in Turkish), it is reasonable to regard the meaning 'sun' of CrKar. *kün* as a result of backformation. Hence, the slot 'sun' must be filled by the stem **quyaš* in Proto-Karaim. The retention of the stem **kün* in Halich and Trakai collocations *kʸunʸ batɨš* 'sunset' (lit. 'sun diving'), *kʸunʸ tuvuš* 'sunrise' (lit. 'sun appearing') can prove its antiquity in this meaning and, hence, the existence of **kün* 'sun' in Pre-Proto-Karaim, but it remains questionable whether this evidence is sufficient to reconstruct Proto-Karaim **kün* as the basic term for 'sun'.

84. **to swim** – two candidates for this slot have been found:

1) CrKar. *yüz-* (Sz) ~ *üz-* (Sz) 'to swim' (Baskakov, Szapszał & Zajączkowski 1974: 261, 588);

2) CrKar. *čöm-* (Sz) 'to swim, to dip', TrKar. *čom-* 'to swim, to dip', HKar. *com-* 'to swim, to flow' (Baskakov, Szapszał & Zajączkowski 1974: 614, 632, 639).

The first one reflects the relatively stable Proto-Turkic stem **yürʸ-* 'to swim', cf. Oghuz comparanda: Tur., Turkm. *yüz-* 'to swim', Gag., Az. *üz-* 'to swim', it is also common beyond the Oghuz subgroup (Dybo 2013: 490). This stem can be safely reconstructed for Proto-Turkic

'to swim'. The second candidate is a transparent innovation. It reflects the semantic shift 'to dive, to dip' > 'to swim'. The original meaning is confirmed by a number of languages:

Tuv. *šɨm-ɨn-* (refl.) 'to dip, to dive', OUyg., KarakhUyg., Chag., Uyg. *čom-* 'to dip, to dive', Uzb. *čọm-* 'to dip, to dive', Tat. *čum-* 'to dip, to dive', Chuv. *čʊm-* 'to dip, to dive' (Tat. borrowing?), Bash. *sumɨ-* 'to dip, to dive' (Dybo 2013: 491).

Note that the polysemy 'to swim, to dive' is attested in the Karaim dialects as well. Based on the external etymology principle, one could reconstruct **yüz-* for Proto-Karaim 'to swim'. However, this stem must be regarded as a borrowing since it is limited to the one dialect in intimate contact with West Oghuz, while the archaic stem, on the contrary, is retained in Karaim.

86. **that** / 87. **this** – the system of demonstrative pronouns in the Crimean Karaim has been influenced on the part of Oghuz languages.

	Crimean	Trakai		Halich	
proximal	*bu* 'this'	*bu* 'this'	*ušpu* 'this here'	*bu* 'this'	*uspu* 'this here'
medial	*šu* 'this, that'	—	—	—	—
distal	*ol* 'that'	*ol* that'	*ošol* 'that there'	*ol* that'	*osol* 'that there'

Table 1. The subsystems of the demonstrative pronouns in the Karaim languages.

The three-way deictic system, like in Crimean Karaim, can be potentially treated as archaic. Proto-Nuclear-Turkic **šu* functions as a medial deictic pronoun in several languages:

Gag. *šu* 'this, that (medial deixis)', Turkm. *šu* 'this, that (medial deixis)', Uzb. *šu* 'this, that (medial deixis)', Kum. *šu* 'this, that (medial deixis)', Kirg. *šu* 'this, that (medial deixis)'.

Theoretically this could confirm the antiquity of CrKar. *šu*. The systems with bare **šu* as a medial deictic pronoun are common in the Oghuz languages (Tenishev & Dybo 2002: 145–156), but not typical for other Turkic subgroups. Outside Oghuz, the Proto-Turkic pronominal root **šu* is more frequently attested with various extensions:

Chuv. *ʃavз*, *ʃakз* 'this', *leš* 'that', Yak. *sol* 'that', Turkm. *šol* 'that', Bash. *ošo* 'this', *šul* 'that', Tat. *šul* 'that', Nog. *sosɨ* 'this', *sol* 'that', Kaz. *osɨ* 'this', *sol* 'that', Karak. *usɨ* 'this', *sol* 'that', Kir. *ušu* 'that'.

The fact that Crimean Karaim, Kumyk, and Kirgiz feature simply **šu* sets them apart from other Kipchak languages. Therefore, it may be suspected that Crimean Karaim demonstrates another Oghuz loan. Thus, only **bu* 'this' and **ol* 'that' can be reconstructed for Proto-Karaim with certainty. In fact, **šu* must not be a deictic pronoun but rather a deictic particle, see Dybo 2013: 497–498.

3.5. Phonological variation in Crimean Karaim

Another result of strong Oghuz influence on Crimean Karaim is the presence of phonological doublets which reflect both Kipchak and Oghuz development of the same Proto-Turkic root. The Oghuz-like counterparts are borrowings. Due to the fact that Oghuz looks more archaic than Kipchak in some parameters, these cases, considered in Sections 3.4.1 and 3.4.2, can also be regarded as fake archaisms.

3.5.1. Reflexes of PT *g

To the basic distinctions between Oghuz and Kipchak languages belong the reflexes of **g* after a low central vowel. The Oghuz languages demonstrate an uvular consonant whereas the west

majority of the Kipchak languages change the velar to a labial. A school-book example is the reflex of Proto-Tukic *da:g 'mountain':

Oghuz: Tur. *da:* (dial. *daʁ*), Az. *daʁ*, Turkm. *da:ɢ*, Sal. *da:ʁ*;

Kipchak: Kum. *taw*, K.-B. *taw*, Tat. *taw*, Bash. *tau*, Nog. *taw*, Kaz. *taw*.

Crimean Karaim demonstrates both *taw* 'forest' (Sz) and *taʁ* 'mountain' (Sz); the third variant is *daʁ* 'mountain' (ZR 45, 3). These forms contrast with TrKar., HKar. *taw* 'mountain' (Baskakov, Szapszał & Zajączkowski 1974: 168, 503, 505). The Crimean Karaim form *taʁ* must be treated as phonologically adopted. Voiced *d* was substituted with voiceless *t*, since only voiceless dentals are possible in word onset in the inherited vocabulary. The final velar does not undergo the adaptation since there is no general restriction on *ʁ* after vowels at least in the non-final position, cf. *alžaʁɨm* 'I will take', *qartnɨŋ tayaʁɨ* 'old man's stick'. A simultaneous occurrence of adopted (to various degrees) and non-adopted items is typical for the situation of intensive influence, cf. Russian loans in Kazym Khanty and Finish loans in Northern Saami:

Khant. *ăškola ~ škola* 'school' < Rus. *škola* 'school';

Khant. *wɵntər ~ andrey ~ andrʸey* 'a male personal name' < Rus. *Andrey* 'a male personal name';

SaaN. *hirbmat ~ harbmat* 'horrible' < Fin. *hirmu* 'horror' (Aikio 2007: 28–29);

SaaN. *hapmu* 'craving (for a particular food)' ~ *hipmu* 'lust, desire' < Fin. *himo* 'carving, desire' (Aikio 2007: 28–29).

Another example for *-*ag* in the final position found in the Karaim Swadesh list:

CrKar. *yaw* (Sz) ~ *yaʁ* (Sz) 'fat', cf. TrKar., HKar. *yaw* 'fat' (Baskakov, Szapszał & Zajączkowski 1974: 214–215).

Reflexes of the Proto-Turkic vocalic-consonantal cluster *-*agɨ*- are a special case. In Kipchak, not only does **g* become a labial consonant, but **ɨ* also becomes a rounded vowel. Oghuz demonstrates here an uvular consonant and an unrounded vowel.

CrKar. *awur* (Sz) ~ *aʁɨr* (Sz, R) 'heavy', cf. TrKar., HKar. *awur* 'heavy' (Baskakov, Szapszał & Zajączkowski 1974: 42, 44);

CrKar. *awuz* (Sz) ~ *awɨz* (Sz) ~ *aʁɨz* (Par 84 v. 9) 'mouth', cf. TrKar., HKar. *awuz* 'mouth' (Baskakov, Szapszał & Zajączkowski 1974: 42, 44);

CrKar. *bawur* 'liver' (Sz) ~ *baʁɨr* (Sz) 'chest, liver', cf. TrKar., HKar. *bawur* 'liver', TrKar. *bawɨr* 'liver' (Baskakov, Szapszał & Zajączkowski 1974: 94, 96).

In Tenishev & Dybo 2006: 72–73, the double reflexes of **ag#* and **agɨ* have been postulated for Karaim, i.e. *aw ~ aγ* and *awu ~ aγɨ*. It seems more reasonable to regard the reflexes with *γ* as a result of Oghuz influence. If they are eliminated, Karaim will not differ from other Kipchak languages in its reflexes of **ag#* and **agɨ*. In the opposite case, the Karaim data would require reconstructing velar (or rather uvular) consonants for Proto-Kipchak in these clusters.

3.5.2. Initial voiced dental and velar consonants

Turkish and Gagauz reflect the Proto-Turkic distinction of initial voiced and voiceless dental and velar stops. For velars the opposition can be reconstructed only in roots with front vowels. The reconstruction of the initial Proto-Turkic voiced stops and some modifications which occurred in the Oghuz languages are described in all details in Tenishev & Dybo 2002: 68–83 (see also Dybo 2007 for further details and discussion). The majority of the Kipchak languages neutralize these oppositions in favor of the voiceless series. Crimean Tatar demonstrates contact-induced variation.

CrKar. *keča* 'night' (Par 83 v. 3) ~ *keče* 'night' (Sz, Man 3a, 8a) ~ *geže* 'evening' (Kž III–IV, 81) ~ *geče* 'night' (ZR 52, 26, Q 9) ~ *gečа* 'night' (ZR 52, 20), cf. TrKar. *kʸečʸa* 'night', HKar. *kece*

'night' (Baskakov, Szapszał & Zajączkowski 1974: 159, 167, 311–312, 394–395; Aqtay & Jankowski 2015: 164, 202) < PT *ge:če;

CrKar. *kel-* 'to come' (Sz, Cam, Dan 1:1, Man 2a) ~ *gel-* 'to come' (Man 3a, Q34), cf. TrKar., HKar. *kel-* 'to come' (Baskakov, Szapszał & Zajączkowski 1974: 301–302, 390; Aqtay & Jankowski 2015: 164, 204) < PT *gel-;

CrKar. *köz* 'eye' (Sz, Cam, Psa 10:1, Man 5a) ~ *göz-* 'eye' (Par 82 v. 1, ZR 79, 15, ZR 95, 30, Man 5a, Q 4), cf. TrKar. *kʸozʸ-*, HKar. *kez-* 'eye' (Baskakov, Szapszał & Zajączkowski 1974: 161, 300, 312, 336; Aqtay & Jankowski 2015: 168, 221) < PT *görʸ;

CrKar. *kör-* 'to see' (Par 83 v. 5, Man 1a, Q 38) ~ *gör* 'to see' (Q 36, 49), cf. TrKar. *kʸor-*, HKar. *ker-* 'to see' (Baskakov, Szapszał & Zajączkowski 1974: 306, 314, 339; Aqtay & Jankowski 2015: 167–168, 218) < PT *gör-;

CrKar. *kün* 'sun, day' (Sz, R, Man 3a) ~ *gun* 'day' (Par 77 v. 11, Man 2b, Q 73), cf. TrKar. *kʸunʸ* 'day', HKar. *kin* 'day' (Baskakov, Szapszał & Zajączkowski 1974: 167, 320, 353, 396; Aqtay & Jankowski 2015: 169, 224) < PT *gün;

CrKar. *taš* 'stone' (Sz, Fil 7, 120, Q 81) ~ *daš* 'stone' (Par 83 v. 12, Q 21), cf. TrKar. *taš* 'stone', HKar. *tas* 'stone' (Baskakov, Szapszał & Zajączkowski 1974: 170, 516, 518; Aqtay & Jankowski 2015: 132, 386) < PT *dia:λ;

CrKar. *taw ~ taʁ ~ daʁ* < PT *da:g (details see above);

CrKar. *tamar* 'vein, root' (Sz) ~ *tamur* 'vein, root' (Sz, Q 431) ~ *damar* 'vein' (ZR 78, 18), cf. TrKar., HKar. *tamur* 'vein, root', HKar. *tamar* 'vein, root' (Baskakov, Szapszał & Zajączkowski 1974: 169, 509–510; Aqtay & Jankowski 2015: 131, 381) < PT *dąmor;

CrKar. *terek* 'tree' (Man 10a) ~ *teraq* 'tree' (Par 83 v. 14) ~ *derek* 'tree' (Fil 8, 150, Q 58) ~ *direk* 'tree' (Sz, ZR 44, 30) ~ *diraq* 'post, column' (ZR 16, 21), cf. TrKar. *tʸerʸak* 'fruit tree', HKar. *terek* 'id' (Baskakov, Szapszał & Zajączkowski 1974: 178, 185, 522, 565, 567; Aqtay & Jankowski 2015: 136, 396) < PT *dęrek;

CrKar. *tüz* 'knee' (Sz) ~ *diz* 'knee' (Q 628) ~ *düz* 'knee' (KM 61b), cf. TrKar. *tiz ~ tizʸ* 'knee', HKar. *tiz ~ kiz* 'knee' (Baskakov, Szapszał & Zajączkowski 1974: 317, 525–526; Aqtay & Jankowski 2015: 139, 144) < PT *di:rʸ;

CrKar. *tolɨ* 'full' (Sz, Par 102 v. 13) ~ *tolu* 'full' (Sz, R) ~ *dolɨ* 'full' (Q 187), cf. TrKar., HKar. *tolu* 'full' (Baskakov, Szapszał & Zajączkowski 1974: 537; Aqtay & Jankowski 2015: 140, 407) < PT *do:l-;

CrKar. *til* 'tongue' (Q 18, 49, Meq 60, 70) ~ *dil* 'tongue' (Q 223), cf. TrKar. *tilʸ*, HKar. *til ~ kil* 'tongue' (Baskakov, Szapszał & Zajączkowski 1974: 319, 528; Aqtay & Jankowski 2015: 403, 138) < PT *dɨl ~ *dil;

CrKar. *tiš* 'tooth' (Sz) ~ *čiš* 'tooth' (Sz, Q 125) ~ *diš* 'tooth' (ZR 70, 12, Q 302); cf. TrKar. *tiš* 'tooth', HKar. *tis ~ kis* 'tooth' (Baskakov, Szapszał & Zajączkowski 1974: 178, 323, 531–532, 629; Aqtay & Jankowski 2015: 124, 132, 404) < PT *di:λ;

CrKar. *tur-* 'to stand' (Sz, Man 5a) ~ *dur-* 'to stand' (Par 77 v. 12, Q 54), cf. TrKar., HKar. *tur-* 'to stand' (Baskakov, Szapszał & Zajączkowski 1974: 181, 547; Aqtay & Jankowski 2015: 142, 413) < PT *dur-.

It should be noted that not all Proto-Turkic stems with initial *d found in Swadesh list demonstrate voiced consonants in Crimean Karaim: PT *dɨrŋa-k > CrKar. *tɨrnaq* 'fingernail, claw', PT *di:λ-le- > CrKar. *tišle- ~ čišle-* 'to bite', *dęri > CrKar. *teri* 'skin', *durʸ > CrKar. *tuz* 'salt'. Such inconsistency indicates that in this case they are not regular reflexes but borrowings. The Proto-Turkic stems with initial *k and *t are found always with voiceless consonants: PT *kül > CrKar. *kül* 'ashes', PT *köń- > CrKar. *küy-* 'to burn (intr.)', PT *köp > CrKar. *köp* 'many', PT *kiλi > CrKar. *kiši* 'man (person)', PT *kičük > CrKar. *kiči* 'small', PT *kem > CrKar. *kim* 'who', PT *tü:k > CrKar. *tük* 'feather', PT *tün > CrKar. *tün* 'night', PT *tüt-ün > CrKar. *tütün* 'smoke'.

3.5.3. Other Oghuz loans

The initial consonant of PT **s(i)ač* 'hair' yields *č* or *š* in majority of Turkic languages. However, reconstruction of the initial **s* is proven by Yakut *as* (where **s- > 0-* regularly) and Oghuz reflexes with retained *s-*. Crimean Karaim demonstrates doublets with Oghuz- and Kipchak-like reflexes. The first one should be a loan, since the Halich and Trakai dialects point to Proto-Karaim **č*. This case belongs to fake phonological archaisms.

CrKar. *sač* (Par 107 v. 13) 'a hair (Rus. *volos* – hair[SG])' ~ *seč* (Sz) 'hair (Rus. *volosy* – hair-PL); tuft, crest' ~ *čač* (R) 'a hair, hair', TrKar. *čač* 'a hair, hair', HKar. *cac* 'hair, fur' (Baskakov, Szapszał & Zajączkowski 1974: 470, 500, 613, 625).

Two more Oghuz loanwords in Crimean Karaim are *ver-* 'to give' and *var-* 'to go', which reflect an Oghuz innovation. Although these words are not fake archaisms, I include them here since they additionally confirm the direction of borrowings in the pair Crimean Karaim < Coastal Crimean Tatar / Turkish. These stems reflect the shift of initial **b* to *v* in monosyllabic stems with *r* in the coda.

CrKar. *ber-* (Sz, R) ~ *ver-* (Par 77 v. 19, ZR 32, 27) 'to give', cf. TrKar. *bʸer-* 'to give', HKar. *ber-* 'to give' (Baskakov, Szapszał & Zajączkowski 1974: 112, 151, 158);

CrKar. *var-* (Q 4) ~ *bar-* (Sz) 'to go', TrKar., HKar. *bar-* 'to go' (Baskakov, Szapszał & Zajączkowski 1974: 102; Aqtay & Jankowski 2015: 436).

3.6. Preliminary conclusions

The onomasiological reconstruction of the Proto-Karaim Swadesh list is complicated mainly by the set of fake archaisms. Fake archaisms are a particular type of homoplastic development, namely MAT-borrowings and semantic back-formations from a sister subgroup which preserved more archaic (in the perspective of a whole family) items. In the Crimean Karaim case, Oghuz nature of the archaic-looking items is proven by the large amount of other Oghuz borrowings and by the history of the sociolinguistic situation.

Examination of sources used in Baskakov, Szapszał & Zajączkowski 1974 and in Aqtay & Jankowski 2015 shows that some of them are more "Oghuzized" than others. From our data it is clear that Par, ZR, and Q contain many more Oghuz forms than Sz. Apparently, they demonstrate language shift to Coastal Crimea Tatar. Aqtay and Jankowski's more detailed study of the lexicon (2015: 9) confirms this statement. Data from these sources are inappropriate for phylogenetic studies.

Consistent detection of all borrowed elements allows mostly trivial reconstruction of the Swadesh list for the not particularly deep Proto-Karaim taxon.

4. Classifying languages of Crimea

In this section, I address the discussion of the genealogical affiliation of Crimean Karaim, each of three Crimean Tatar dialects and Krymchak. My goal within the scope of this section is not to build a complete phylogenetic tree but only to define the closest relatives of the idioms in question. Needless to say, disclosure of borrowings plays a crucial role in this procedure. Before comparing wordlists all loans, including inter-Turkic ones, must be excluded. Although this statement may seem trivial, in Section 4.6 it will be shown that even recent phylogenetic research still continues to be affected by undetected borrowings.

4.1. Previous research

Beginning with Radloff, the language of Crimean Karaims is fully identified with Crimean Tatar or seen as one of its dialects. This opinion is shared by Zajączkowski, Doerfer, and Polinsky. At the same time, Crimean Karaim (whatever the term means) is included in the Karaim-Russian-Polish dictionary (Baskakov, Szapszał & Zajączkowski 1974). In his earlier grammar (Musaev 1964: 36–37), Musaev maintains that the Crimean Karaits' variety is not distinguishable from Crimean Tatar and must not be included in the notion *Karaim language*. Information on the Crimean dialect was later included in his sketch of Karaim dialectology (Musaev 2010) by the editors. The discussion is summarized in the work by Jankowski (2003: 109–112), who attempts to show that Crimean Karaim is different from Crimean Tatar, involving phonological, syntactical, onomastic and lexical arguments. Basic vocabulary remains beyond his interest. The Swadesh list was examined in Polinsky 1992. She comes to the conclusion that Crimean Karaim together with Krymchak language is very close to the Orta and Coastal Crimean Tatar dialects and, hence, belongs to the Oghuz subgroup. Polinsky does not distinguish borrowed and inherited vocabulary when calculating lexicostatistical matches, therefore her conclusions can be called into question. To be fair, it must be noted that, to the best of my knowledge, the requirement to exclude contact innovations was yet to be explicitly formulated in 1992.

Currently, no detailed descriptions of the Crimean Tatar dialects exist and they are unlikely to appear in the future. Commonly accepted is Ervand Sevortyan's (1966) dialectal classification, which distinguishes three dialects of Crimean Tatar (Steppe, Coastal and Central) highlighting their heterogeneous origin. According to Sevortyan's classifications, the Steppe dialect belongs to the Nogai Kipchak subgroup; Orta is Cuman Kipchak, i.e. West Kipchak; Coastal belongs to the Oghuz group. The original dialectal differentiation was violated as the result of Soviet deportation of Crimean Tatars to Uzbekistan in 1944. After the return to Crimea in the early 1990s, most families were not able to settle in their native villages. This provides further dialectal mixture. Already during Darya Kavitskaya's fieldwork in 2002–2003 and 2009, only older speakers had "clear dialect affiliation" (Kavitskaya 2010: 3). Dialectal mixture is quite visible both in my and Polinsky's data. See Normanskaya 2019 on dialectal mixture in literary Crimean Tatar.

4.2. Methods

To define the genealogical affiliation of the Turkic languages of Crimea, I first apply manual subgrouping based on lexical innovation and then compare the obtained results with the inference of the computational lexicostatistical algorithms. I use three approaches which are currently most widespread in linguistic phylogeny: Maximum Parsimony Analysis, MCMC Bayesian approach, Neighbor-joining algorithm.

An important advantage of the manual subgrouping applied in the present paper is that it fits the commonly accepted requirement to build genealogical classification based on innovations (Campbell 2013: 175). This requirement is ignored by the lexicostatistical framework, where every match, whether it is an innovation or a retention, has similar value. The principle of subgrouping sufficient for our purposes is drastically trivial. Languages A and B are regarded as specifically related to each other if this pair demonstrates the highest amount of shared non-contact-induced innovations. This method was used by Leonid Kogan (2015) for the classification of the Semitic languages. When reconstructing the phylogeny of a whole fam-

ily from scratch, this method leads to a vicious circle, since tree topology must be already known for most cases to distinguish between innovations and retentions. However, if the goal is merely to find the positions of newly involved taxa on a previously constructed tree, such a method is applicable.

The most important technical details on the applied computational lexicostatistical algorithms are summarized in Table 2.

Algorithm	Software	Basic settings
Maximum Parsimony	TNT v. 1.5 (Goloboff & Catalano 2016)	Implicit enumeration Collapse trees after search Outgroup: Yakut

Bayesian MCMC	MrBayes 3.2.7a x86_64
	(Huelsenbeck & Ronquist 2001;
	Ronquist et al. 2012)
		covariation F81 model;
		datatype = restriction;
		coding = noabsencesites;
		rates = gamma
		covariation = yes
		brlenspr = clock:fossilization
		clockvarpr = TK02

Neighbor-joining	Starling v. 2.7.0–42f0a13 (Starostin 2007a)	Method: Experimental Replacement rate: 4.88 (default value)

Table 2. Information on the software and basic settings applied for the lexicostatistical analyses.

Based on trees obtained as the result of Maximum Parsimony, a strict consensus tree was produced. The settings for Bayesian MCMC are adopted from Kassian et al. forthcoming. The full dataset and output files can be found in Supplement 2. Cognate encoding has been done within Starling software and then converted into the Nexus file with a binary matrix. The derivational drift free dataset has been used; on the principles of the derivational drift elimination see Kassian et al. forthcoming.

I have compared six wordlists of the Turkic idioms spoken until recently on the Crimean Peninsula (see Table 3) with Halich and Trakai Karaim, Turkish, Gagauz, Proto-Nogai, Proto-Kazakh-Karakalpak, Proto-Kumyk, Proto-Karachay-Balkar. Since Maximum Parsimony analysis requires an outgroup taxon, I included the Proto-Yakut list, which clearly belongs neither to the Oghuz nor to the Kipchak clade. Lists of the proto-languages are reconstructed (using methodology described in Section 3.1) by me in collaboration with Anna Dybo and Alexei Kassian within the framework of an ongoing project devoted to revision of the Turkic phylogenetic tree structure. All Karaim wordlists were collected from the sources mentioned in Section 2.1; Turkish and Gagauz ones are based on the dictionary sources as well (Parker 2008; Bogochanskaya & Torgashova 2009; Gaydarzhi et al. 1973; Sesli Sözlük; Rajki 2007). Collecting the lists, I was guided by the semantic specification proposed in Kassian et al. 2010.

Since some of Crimean Tatar wordlists do not meet all the modern requirements to Swadesh lists (see Section 4.3 for details), computational lexicostatistics can only play a secondary role in the present research. However, I believe that, despite somewhat faulty data, application of three different computational approaches still has some relevance.

4.3. Data

Table 3 shows which wordlists of idioms spoken in Crimea were used in the present paper.

Idiom	Comments and references
Crimean Karaim	The material of the list was collected from Baskakov, Szapszał & Zajączkowski 1974 and Aqtay & Jankowski 2015. All Oghuz loans discussed in section 3 have been excluded, so that the list reflects sources with minor Oghuz influence.
Coastal Crimean Tatar	The list published in Polinsky 1992.
Orta Crimean Tatar	The list published in Polinsky 1992.
Steppe Crimean Tatar	The list published in Polinsky 1992.
Crimean Tatar (dialect not defined)	The list was collected in 2020 from two speakers. I avoid labeling it with any dialectal affiliation due to the reasons described below.
Krymchak	The list published in Polinsky 1992, revisited and extended based on Ianbay 2016 and Rebi 2004.
Turkish	The material of the list was taken from Parker 2008; SesliSözlük; Bogochanskaya & Torgashova 2009.
Gagauz	The material of the list was taken from Gaydarzhi et al. 1973; Rajki 2007.
Proto-Kumyk	The list is reconstructed by the author in collaboration with Anna Dybo and Alexei Kassian on the basis of Bammatov & Gadzhiakhmedov 2011 and recently collected dialectal data.
Proto-Karachay-Balkar	The list is reconstructed by the author in collaboration with Anna Dybo and Alexei Kassian on the basis of Gochiyayeva & Suyunchev 1989 and recently collected dialectal wordlists.
Proto-Nogai	The list is reconstructed by the author in collaboration with Anna Dybo and Alexei Kassian on the basis of Baskakov 1956 and recently collected dialectal wordlists.
Proto-Kazakh-Karakalpak	The list is reconstructed by the author in collaboration with Anna Dybo and Alexei Kassian on the basis of dictionary sources (Bektaev 1996; Bekturov & Bekturova 2001; Syzdykova & Khusaiyn 2008; Baskakov 1967) and recently collected dialectal wordlists.
Proto-Yakut	The list is reconstructed by the author in collaboration with Anna Dybo and Alexei Kassian on the basis of dictionary sources (Pekarskiy 1959; Pekarskiy 1916; Stachowski 1993; Stachowski 1998) and recently collected dialectal wordlists.

Table 3. 110-item Swadesh lists for the Turkic varieties of Crimea.

It should be noted that three Crimean Tatar wordlists used in this research had been collected long before the semantic specification of the Swadesh list was undertaken in Kassian et al. 2010. Therefore, they are not fully compatible with my data. Three important discrepancies have been found; 'all (omnes)', 'to burn (intr.)', 'to go' are used now instead of traditional 'all (totus)', 'to burn (tr.)', 'to walk'. Incompatible items are marked as not attested in Polinsky's data; moreover, the original Swadesh 100-wordlist has been extended with 10 items absent from the older record. To make my lists more compatible with older ones, I deviate from Kassian et al. 2010 on three points: taking 'earth (ground)' instead of 'earth (soil)', 'round (2D)' without 'round 3D' as a synonym and not accepting medial deictic pronouns.

I now avoid labeling my data on Crimean Tatar with any terms from the traditional classification. It has been collected from two informants, a couple, about 60 years old, who were interviewed independently. They were born in Uzbekistan, now they live in the Bakhchisaray district. The wife's parents come from Duvanköy (Ukr. *Verxnjosadove*), the traditional territory of the Orta dialect; the husband's parents come from the nearest suburbs of Gurzuf, the traditional territory of the Coastal dialect. The two obtained wordlists correspond with each other fully as far as lexical items are concerned; at the same time, they considerably differ in phonology. The woman's idiolect lacks labial harmony and the distinction of voiced and voiceless stops in the word-onset; the voiceless uvular is a stop. So, it should be regarded as a Kipchak-

based dialect. Thus, she pronounces 'fat' with the uvular consonant as *yaʁ*, but 'forest' with a labial as *taw*, cf. section 3.4.1. Her husband's idiolect demonstrates consistent labial harmony, sporadic preservation of the PT initial **d*, and the fricativization of the voiceless uvular; these features are typical for the Coastal (Oghuz-based) dialect. The word *auz* 'mouth' has been found to have the typically Kipchak reflex of **agɨ* in the both idiolects, but *aʁɨr* 'heavy' with the Oghuz reflex. Polinsky's lists contain some inconsistencies in phonology as well; the most glaring is a sporadic **y* > *ǯ* before *a* in the Coastal dialect, which is expected only in Steppe Crimean Tatar, and, at the same time, the lack of this shift in some lexemes from the Steppe dialect. The distribution of reflexes of PT **g* across dialects confounds expectations as well: cf. *aʁɨz* in all dialects, CoCrTat. *tau*, *yav*, OrCrTat *yaʁ*, *daʁ* (with *d* instead of expected *t*!). The phonology represented in Polinsky's data is completely inconsistent with existing description of the Crimean Tatar dialects. As is shown below, basic vocabulary allows us to make some clearer conclusions about the original genealogical affiliation of the dialects documented in the wordlists under consideration and about the direction of their development.

4.4. Innovations in the basic vocabulary

Turkic languages of the Crimea demonstrate discrepancies in 25 slots out of the 110-item wordlist. Cases in which the variation is caused by Persian and Arabic borrowings are excluded, i.e. only potentially autonomous innovations are taken into account. Table 4 below presents the genetically relevant features. Archaisms, borrowings, and items innovative from the Common Turkic perspective but still not informative for the current question (i.e. innovations shared by both Kipchak and Oghuz languages) are underlined. The Oghuz loans revealed above and phonological variants have been excluded from the Crimean Karaim wordlist. Indexes in the superscript identify the subgroup in which cognates of the word are found, an exclamation mark labels singletons. These indexes are somewhat rough, since I ignore the fact that some of the considered words can occur sporadically in the other Turkic languages which cannot be applied to the closest relative of the studied Crimean varieties. These inaccuracies are partially clarified in the commentary immediately after the table. I use mostly *Kip* and *Ogh* meaning primarily West Kipchak and West Oghuz respectively; when possible, I refer to the low level subgroups instead of Kipchak. The full lists for the Turkic idioms of Crimea and for languages they have been compared with can be found in Supplement 2.

1. **all** – CrTat., Krym. *epsi*, cf. Tur., Gag. *hepsi* 'all (omnes)' – the word is limited to the mentioned languages; no fully acceptable Turkic or foreign etymology (Dybo 2013: 66), the root has a probable Persian origin (cf. Räsänen 1969: 158), but this hypothesis faces some phonological difficulties.

4. **belly** – CrKar., CrTat., Krym. *qursaq* – assuming the stem as a basic term is innovative as a result of the elevation 'paunch' > 'belly'; widespread in the Kipchak subgroup (Dybo 2013: 106–108).

10. **bone** – CrTat., Krym. *kemik*, cf. Tur., Gag. *kemik* 'bone' – result of generalization 'spongy bone' > 'bone'; as a basic term, limited to the mentioned languages (Dybo 2013: 173–174).

11. **breast** – Strictly speaking, both stems are archaic: CrTat., Krym. *göküs*, *köküs*, *kokus*, *koks* reflect PT **göküŕ*; CrKar. *kökrek*, CrTat. *kokrek* reflect the same root with the fossilized diminutive suffix **gökŕek* (Dybo 2013: 178). Since the derivational connection between these stems has been erased a long time ago, I believe that their distribution is informative for genealogical classification. However, in agreement with the principles of derivational drift elimination, I mark the simplex and diminutive form with one index in the lexicostatistical dataset.

	Crimean Karaim	Crimean Tatar (author's data)	Steppe Crimean Tatar	Middle Crimean Tatar	Coastal Crimean Tatar	Krymchak
			(Polinsky 1992)			
all	*barɨ* *barča*	*epsi*^{Ogh}	not attested			*epsi*^{Ogh}
belly	*qursaq* ^{Kip}	*qursaq~χursaχ* ^{Kip}	*xursax* ^{Kip}	*qursaq* ^{Kip}	*qarɨn*	*qursaq* ^{Kip}
big	*biyik*^{Ogh}	*balaban*^{Ogh}	*bijk*^{Ogh}	*buyuk*^{Ogh}	*büyük*^{Ogh}	*balaban*^{Ogh} *buyuk*^{Ogh}
bone	*süvek*	*kemik*^{Ogh}	*kemik*^{Ogh}	*süyek*	*süyek*	*kemik*^{Ogh}
breast	*kökrek* ^{Kip}	*kokrek* ^{Kip}	*koks*^{Ogh}	*kokus*^{Ogh}	*gökü̈s*^{Ogh}	*kokus*^{Ogh}
to burn	*küydür-* ^{Kar}	*yaq-*	not attested			
dog	*it*	*kopek*^{Ogh}	*it*	*kopek*^{Ogh}	*köpek*^{Ogh}	*kopek*^{Ogh}
dry	*quru*	*qurɨ*	*χatɨ*$^{!}$	*quru*	*quru*	*quru*
fat	*yav*	*yaʁ*	*may*^{Nog}	*yaʁ*	*yaw*	*yaʁ*
hand	*qol* ^{Kip}	*qol* ^{Kip}	*qol* ^{Kip}	*el*^{Ogh}	*el*^{Ogh}	*qol* ^{Kip}
feather	*yun* ^{Kar}	*qanat~χanat*	*qušin*^{Nog}	*quš-qanat*	*lelek*^{Ogh}	*pux*
man	*er* *erkak*	*aχay~aqay*$^{!}$	*erkek*	*er*	*marda*	*erkek*
many	*köp*	*čoq*^{Ogh}	*köp*	*čok*^{Ogh}	*čok*^{Ogh}	*čok*^{Ogh}
mountain	*taw*	*bayɨr*^{Ogh}	*bair*^{Ogh}	*daʁ*	*tau*	*daʁ*
to sleep	*yuqla-*^{Kip}	*yuqla-~yuχla-* ^{Kip}	*ǯat-*$^{!}$	*yuχla-* ^{Kip}	*yuχla-* ^{Kip}	*yuχla-* ^{Kip}
small	*kiči*	*ifaq*^{Ogh}	*kišik*	*yufaq*^{Ogh}	*yufaq*^{Ogh}	*kičkene*
smoke	*tütün*		*duman*^{Ogh}	*tutun*	*duman*^{Ogh}	*tutun*
sun	*quyaš* ^{Kar}	*kuneš*^{Ogh}	*kuneš*^{Ogh}	*küneš*^{Ogh}	*güneš*^{Ogh}	*küneš*^{Ogh}
to swim	*čom-*^{Kar}	*yalta-*^{Nog}	*ǯalda-*^{Nog}	*yalda-*^{Nog}	*yüz-*	*yalda-*^{Nog}
tree	*aʁač*	*derek~terek* ^{Kip}	*derek* ^{Kip}	*terek* ^{Kip}	*aʁač*	*terek* ^{Kip}
	teraq ^{Kip}					
to go	*bar-*	*kit-*^{Ogh}	not attested			*bar-~var-* *kit-*^{Ogh}
warm	*issi* *yɨllɨ*	*sɨžaχ*^{Ogh}	*sižaχ*^{Ogh}	*čilli*	*yilli*	*sɨžaχ*^{Ogh}
woman	*qatɨn*	*apay*$^{!}$	*χɨsayaχlɨ*	*qadɨn*	*qadɨn*	*qadɨn*
far	*yɨraq*	*uzaq ~ uzaχ*	not attested			*uzaχ*
near	*yuwuq*	*yaqɨn ~ yaχɨn*	not attested			*yaqɨn*

Table 4. Innovations in the basic vocabulary of the Turkic varieties of Crimea.

12. **to burn** – CrKar. *küydür-* – causative from Proto-Turkic **köń-* 'to burn'; cognates have been found in other Karaim dialects, Karachay-Balkar, Kazan Tatar and in some other languages, mainly in the Middle Asian area (Dybo 2013: 189).

18. **dog** – CrTat., Krym. *köpek*, Tur., Gag. *köpek* 'dog' – result of generalization 'hound' > 'dog'; as a basic term, limited to the mentioned languages (Dybo 2013: 189).

20. **dry** – SCrTat. *χatɨ* – result of the semantic shift 'solid' > 'dry' (Räsänen 1969: 241).

26. **fat** – SCrTat. *may* – result of the semantic shift 'butter, suet' > 'fat'; as a basic term, common in Volga and Middle Asian Kipchak languages and in Nogai (Dybo 2013: 249–250).

28. **feather** – CrKar. *yun*, cf. HKar., TrKar., yun ‘feather / down’ –result of the semantic shift ‘fur / down’ > ‘feather / dawn’; with original meaning, common in the Kipchak languages (Dybo 2013: 259–260).

SCrTat. *quš-in* – etymology is somewhat obscure; the stem should be derived from *quš* ‘bird’, *-in* can be an old instrumental affix; exclusive isoglosses with the Nogai subgroup. The etymology proposed here is more probable than the hypothetical Persian loan mentioned in Dybo 2013: 261. Anna Dybo has as of now rejected the etymology involving Pers. *kuč* ‘fish scale’ (personal communication).

CoCrTat. *lelek*, cf. Tur. *yelek* ‘feather’ – derived from PT **ye:l* ‘mane’; attested in all Oghuz languages, including Salar, and in Khalaj (Dybo 2013: 259).

37. **hand** – CrKar., CrTat. *qol* – result of the semantic shift ‘arm’ > ‘hand’; extremely widespread across Nuclear Turkic languages, particularly in all languages belonging to the Kipchak subgroup. Oghuz and some other languages preserve reflexes of the stem **elg* ‘hand’ (Dybo 2013: 300–307).

51. **man** – CrTat. *aqay ~ aχay* – fossilized vocative form from **aqa* ‘uncle, older relative’ (Sevortyan et al. 1974: 121). The stem looks rather like a borrowing from baby-talk (cf. Russian baby-talk words *d’ad’a* ‘man’ < ‘uncle’ and *t’ot’a* ‘woman’ < ‘aunt’), so its relevance for the genealogical classification is questionable, cf. ‘woman’.

52. **many** – CrTat., Krym. *čoq*, cf. Tur., Gag. *čoq*, Az. *čoχ* ‘many’ – reconstruction of the original semantics is somewhat difficult but its innovative nature is obvious; as a basic term, attested only in the mentioned languages (Dybo 2013: 371).

55. **mountain** – CrTat. *bayır*, cf. Gag. *bayır* ‘mountain’ – result of the semantic shift ‘hill’ > ‘mountain’; attested as a basic term only in the mentioned languages (Dybo 2013: 380).

76. **to sleep** – CrKar., CrTat., Krym. *yuq-la-* – derived from **uykı* ‘sleep’; common in Kipchak languages (Dybo 2013: 473–474).

77. **small** – CrTat. *yufaq*, *ifaq*, cf. Tur. *ufaq* ‘small, little’ – result of the generalization ‘small, fine (of pebble, crumb, powder etc.)’ > ‘small, little’; attested in most languages with more specific meanings (Sevortyan et al. 1974: 560–561).

76. **smoke** – CrTat. *duman*, cf, Tur., Gag. *duman* ‘smoke’ – result of the semantic shift ‘fog’ > ‘smoke’; this meaning is found only in the mentioned languages (Dybo 2013: 481).

82. **sun** – CrTat., Krym. *kuneš*, *güneš*, *küneš* – formal innovation, root extension is not quite clear, apparently constitutes an analogical rhyme with *quyaš* ‘heat, blazing sun’; with this extension, this root is attested mainly in Oghuz languages but occurs beyond this subgroup as well (Dybo 2013: 488–489). Since no meaning shift has occurred, I mark reflexes **gün* and **güneš* with the same index in the lexicostatistical dataset.

83. **to swim** – CrTat., Krym. *yalta-*, *yalda-*, *žalda-* – derived from *yal* ‘horse mane’ with further semantic shift ‘to swim or to cross a river holding the horse’s mane’ > ‘to swim’; the verb with its original meaning occurs in Kipchak languages; as the basic term for ‘to swim’, only in Nogai (Dybo 2013: 492; Sevortyan et al. 1974: 93–94).

90. **tree** – CrKar., CrTat., Krym. *terek*, *derek* – result of the generalization ‘poplar’ > ‘tree’; attested as the basic term for ‘tree’ in West Kipchak and Sary Yugur (Dybo 2013: 510).

92. **to go** – CrTat., Krym. *kit-*, cf. Tur. Gag. *git-* – result of the semantic shift ‘to go away’ > ‘to go’; attested as the basic term for ‘to go’ in the Oghuz subgroup (Dybo 2013: 515).

93. **warm** – CrTat., Krym. *sıžaχ*, cf. Gag. *sıžaq* ‘warm’ – a formal innovation, PT **ısıg* ‘warm, hot (?)’ extended with a diminutive suffix; found with this suffix only in the mentioned languages (Dybo 2013: 518; Räsänen 1969: 173–174).

99. **woman** – CrTat. *apay* – fossilized vocative form from **apa* ‘elder sister’ (Sevortyan et al. 1974: 159). The stem looks rather like a borrowing from baby-talk (cf. Russian baby-talk words

d'ad'a 'man' < 'uncle' and *t'ot'a* 'woman' < 'aunt'), so its relevance for genealogical classification is questionable.

SCrTat. *χɨsayaχlɨ* – the compound **qɨz-ayal-kɨ* 'girl-woman-nominal suffix', metathesis of the cluster under the analogical influence of the more productive suffix *-lɨ* (etymology proposed by Anna Dybo, p.c.). Due to the fact that the main component *-ayal-* is an Arabic loan, I mark the whole stem as a borrowing.

101. **far** – CrTat., Krym. *uzaq ~ uzaχ* – result of the shift 'far (adj.)' > 'far (adv.)'; found in numerous languages. The isogloss is not particularly informative for affiliation of Crimean Tatar and Krymchak, however, it opposes Karaim to other Turkic varieties of Crimean (Dybo 2013: 534); the same is true for 'near'.

103. **near** – CrTat., Krym. *yaqɨn ~ yaχɨn* – found in numerous languages, however, the distribution points to PT **yaguk* (Kar. *yuwuq*) as a more archaic term for 'near (adv.)' from the Common Turkic perspective (Dybo 2013: 539).

Table 5 shows the amount of the innovations shared by each of the investigated wordlists with other Turkic subgroups.

	Crimean Karaim	Crimean Tatar (author's data)	Steppe Crimean Tatar	Middle Crimean Tatar	Coastal Crimean Tatar	Krymchak
Oghuz	1	10	7	6	8	10
Kipchak	4	4	2	3	1	3
Nogai	0	1	3	1	0	1
Karaim	3	0	0	0	0	0

Table 5. The amount of the innovations (including both inherited items and inter-Turkic borrowings) shared with other Turkic subgroups.

Four innovations ('to burn', 'feather', 'sun', 'to swim') clearly connect Crimean Karaim with the Halich and Trakai dialect and oppose it to other studied languages. Since it demonstrates only one Kipchak-looking stem, Coastal Crimean Tatar must be classified as Oghuz language. The situation with other idioms is not so transparent. Oghuz-looking innovations are predominant in all wordlists, yet at the same time they demonstrate some typically Kipchak lexemes. Two possible interpretations of this situation may be offered: 1) the Oghuz innovations can be identified as inherited ones and Kipchak as substrate loans; 2) the Oghuz lexemes can be regarded as loans from the dominant language of the region and Kipchak innovations as inherited ones. Such ambiguity indicates that there has been a language shift. The range of borrowings can be easily identified in Crimean Tatar (Steppe, Orta and my data) and in Krymchak based on the same distributional criterion which has already been applied to the Crimean Karaim data. All lexemes with the meanings mentioned above are widespread in Oghuz languages but not in Kipchak. Taking into account the prestige status of Crimean Turkish, i.e. Coastal Crimean Tatar, extensive borrowing from it is very probable. Such interpretation is supported by the fact that the lexemes suspected to be loanwords concentrate predominantly in the less stable part of the Swadesh list: 'small' – 110; 'mountain' – 107; 'many' – 106; 'big' – 101; 'far' – 100; 'near' – 95; 'warm' – 90; 'to go' – 89; 'all' – 84; 'to swim' – 78; 'breast' – 49; 'smoke' – 40; 'sun' – 39; 'bone' – 38; 'dog' – 16. The number after the concept indicates its position in the stability ranking (the higher the number, the less stable the concept), according to Sergei Starostin's estimations (2007b). Avoiding the discussion of every particular case, I would like to draw attention only to OrCrTat. *el* 'hand' which is a classic example of a fake archaism.

Other lexemes are more likely to be borrowed West Oghuz innovations. It must be also emphasized that at least in some cases we should deal not with MAT-borrowings (like *epsi* 'all'; *küneš* 'sun'; *sɨžaχ* 'warm'; *kopek* 'dog' < 'hound') but rather with PAT-borrowings, i.e. with contact-induced semantic shifts (*kemik* 'bone' < 'spongy bone'; *bayɨr* 'mountain' < 'hill'; *ifaq* 'small, little' < 'small, fine (of pebble, crumb, powder etc.)'; *kit-* 'to go' < 'to go away').

Considering all the Oghuz-looking lexemes as borrowings, hence, irrelevant for genealogical classification, I come to the conclusion that Crimean Karaim does indeed belong to the same subgroup with other Karaim dialects; Steppe Crimean Tatar – to the Nogai subgroup; Crimean Tatar (based on my data), Middle Crimean Tatar (Polinsky's data) and Krymchak are close to West Kipchak. Any further conclusions about their proximity to a particular subgroup within Kipchak cannot be made with enough certainty.

The proximity of Steppe Crimean Tatar to the Nogai subgroup is proven by three non-trivial innovations ('fat', 'feather', 'to swim') which are not attested in any language from potential candidates for the closest relatives. Two of these innovations (*qušɨn* 'feather' and *yalda-* 'to swim') do not occur elsewhere in the Turkic languages. Orta Crimean Tatar, Crimean Tatar (based on my data), and Krymchak *yalda-* must be analyzed as a borrowing from the Steppe dialect. Despite the fact that the Steppe dialect was not a dominant idiom, some influence on its part cannot be excluded. Otherwise, such non-trivial ('to swim or to cross a river holding horse's mane' > 'to swim') innovations can be regarded only as a signal of relatedness. This is less probable, since there are no other facts pointing to the specific proximity of Orta Crimean Tatar and of the dialect reflected in my data to the Nogai subgroup.

To sum up, I assume the following subgrouping based on innovations in the basic vocabulary: [Turkish, Gagauz, Coastal Crimean Tatar], [[Nogai, Steppe Crimean Tatar], [Halich Karaim, Trakai Karaim, Crimean Karaim], [Orta Crimean Tatar, Crimean Tatar (my data), Krymchak, Kumyk, Karachay-Balkar].

The alternative approach, i.e. consider Oghuz lexemes as inherited and Kipchak lexemes as borrowings, leads to difficulties. Must we mark as borrowings only lexemes looking similar to Kipchak innovations or must we regard typical Kipchak retentions as borrowings too, i.e. as fake archaisms? This question does not have a satisfactory answer. Had Kipchak innovations been considered borrowings, we would have to deal with a suspiciously archaic Oghuz idiom simultaneously overflowing with Kipchak loans. It should be noted that these fictitious Kipchak loans would have concentrated in the somewhat more stable part of the lists than the real loans considered above, cf. 'belly' – 109; 'fat' – 81; 'to sleep' – 73; 'breast' – 49; 'tree' – 42; 'hand' – 11. If Kipchak retentions such as *süyek* 'bone', *it* 'dog', *bar-* 'to go' etc. had been fake archaisms in the tentative Oghuz language, the mass of borrowings would have strongly contradicted the direction of influence proven by sociolinguistic factors. Thus, it seems reasonable to reject such a decision.

4.5 Results of lexicostatistical analysis

All expectations based on innovations in the basic vocabulary are confirmed by formal computational methods. Trees inferred by three applied algorithms differ only in some details. They all agree on the following points: (a) Coastal Crimean Tatar belongs to the same clade as Turkish and Gagauz; (b) other languages are included in the Kipchak subgroup; (c) Steppe Crimean Tatar is combined with Nogai and Kazakh-Karakalpak; (d) Crimean Karaim forms a clade with other Karaim dialects.

As for the internal structure of the Kipchak clade, the applied analyses are in minor disagreement with each other. Neighbor-joining and Bayesian MCMC suggests a first split into Nogai-Kazakh-Karakalpak-Steppe-Crimean-Tatar and remaining languages, followed by a split

into Karachay-Balkar-Kumyk and Karaim-Orta-Crimean-Tatar-Krymchak. Both algorithms have established that Krymchak is the closest to Crimean Tatar (according to my recently collected data) and the two idioms are related to the Orta dialect. The strict consensus tree build by Maximum Parsimony analysis shows multifurcation of the Kipchak clade into the following taxons: [Halich, Trakai, and Crimean Karaim], [Nogai, Steppe Crimean Tatar, Kazakh-Karakalpak], Orta Crimean Tatar, Crimean Tatar (my data), Krymchak, Kumyk, Karachay-Balkar. Such a structure for the tree fully fits all my assumptions made on the basis of common innovations (Section 4.4).

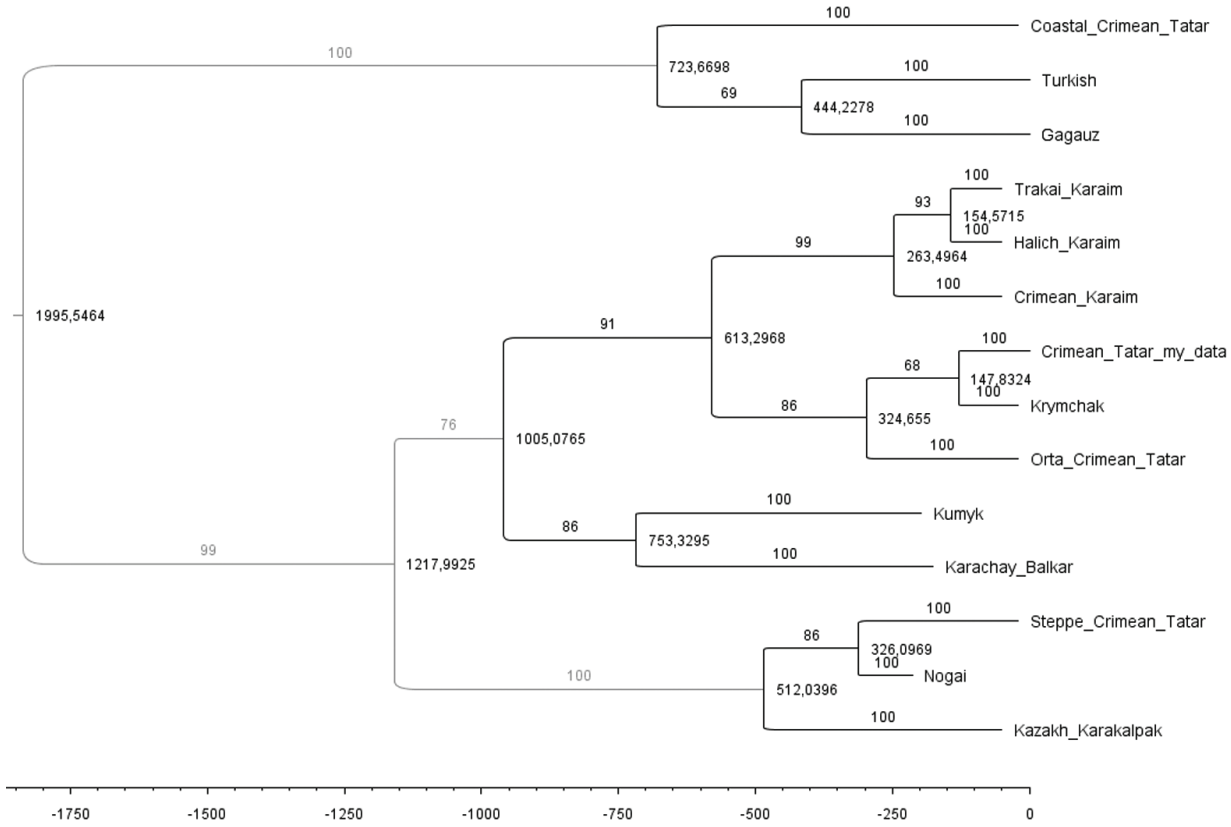

Figure 3. Tree constructed with Bayesian MCMC algorithm in MrBayes software visualized in FigTree software. Numbers near the nodes define mean age; numbers near branches define their probability in percent.

Node	Mean	Median	95% HPD
West Oghuz, i.e. [CoCrTat., [Tur., Gag.]]	723,6698	679,4849	[297,8038, 1238,5883]
[Tur., Gag]	444,2278	415,7284	[173,811, 774,879]
West Kipchak, i.e. [[CrKar., [TrKar., HKar.]], [[OCrTat. [CrTat.(my data), Krym.]], [Kum., KB]]]	1005,0765	959,0648	[581,0544, 1545,5003]
[[OCrTat. [CrTat.(my data), Krym.]], [CrKar., [TrKar., HKar.]]]	613,2968	578,6829	[281,4897, 1026,826]
[CrKar., [TrKar., HKar.]]	263,4964	246,5246	[118,5255, 446,7021]
[TrKar., HKar.]	154,5715	143,146	[66,1245, 269,8982]
[OCrTat. [CrTat.(my data), Krym.]]	324,655	297,9322	[100,4207, 605,93]
[CrTat.(my data), Krym.]	147,8324	129,064	[24,7837, 314,6661]
[Kum., KB]	753,3295	717,8587	[418,5504, 1164,1166]
South Kipchak, i.e. [Kaz-Karak., [Nog., SCrTat.]]	512,0396	484,1897	[242,3641, 844,1445]
[Nog., SCrTat.]	326,0969	311,1897	[149,6245, 521,9043]

Table 6. Mean age, median age and 95% HPD age of the law level nodes according to Bayesian MCMC.

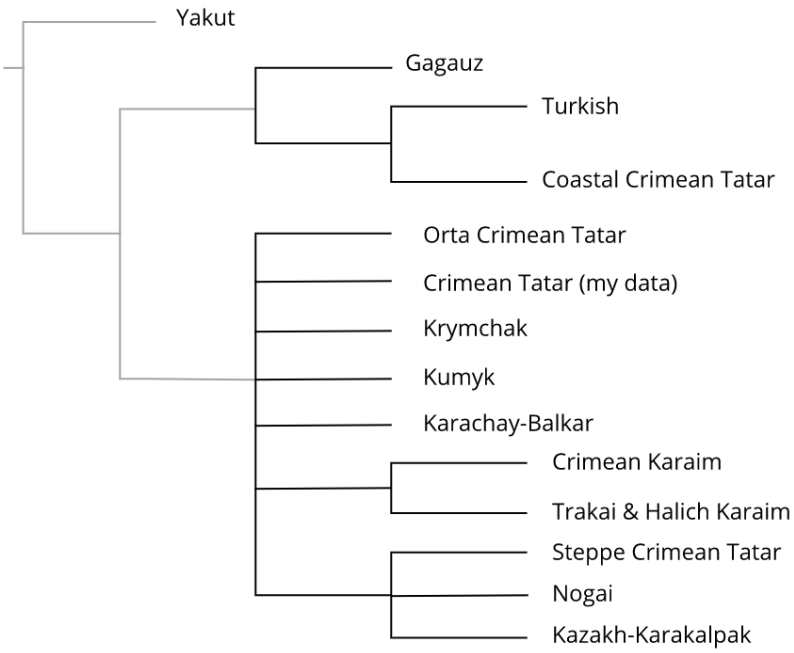

Figure 4. Manually redrawn strict consensus tree constructed with Maximum Parsimony algorithm.

The advantage of Bayesian MCMC and neighbor-joining is that they suggest a clade joining all the languages which had originated on the Crimea Peninsula. The ancestors of the Karaim and Orta Crimean Tatar speaking people came to this region when the West Kipchak languages should have already been (insignificantly) diversified. The invasion of Tatars into Crimea dates back to the 1220s (Fisher 1978: 2); around this time the adherents of Karaite Judaism migrated here from Byzantium (Jankowski 2017: 452–453) and adopted the local Kipchak language. Then the language common for Tatars and Karaims should have split due to the closeness of the Karaim community. The Krymchak speakers, groups of Rabbinic Jews heterogeneous in their origin, adopted Orta Crimean Tatar. Thus, Polinsky is right calling Krymchak an ethnolect of Crimean Tatar (see Polinsky 1992: 173–176). However, another of Polinsky's statements must be revisited. She classifies the Orta dialect together with Krymchak and even Crimean Karaim as Oghuz languages. But even if one admits it is methodologically tolerable not to exclude inter-Turkic borrowings, such an affiliation actually reflects a later language shift. Polinsky's data at least on Krymchak and Orta Crimean Tatars allows the reconstruction of their original genealogical affiliation. The identification of the borrowings plays here a crucial role. My study confirms Sevartyan's (1966) view on the Crimean Tatar dialects as three genealogically distinct items. The early separation of Nogai-Kazakh-Karakakpak-Steppe-Crimean-Tatar from remaining Kipchak languages in question corresponds to the traditional opinion that Nogai does not belong to the West Kipchak subgroup. The speakers of the Steppe dialect have massively settled in Crimea only in the beginning of the 17th century; this was the result of Nogai migration from Lower Volga Steppes which had started a century before (see Trepavlov 2014).

A recent attempt at another revision of the Turkic classification fell victim to undetected loans as well. Martine Robbeets and Alexander Savelyev (2020) include Crimean Tatar into the Oghuz subgroup. They discuss this contradiction with previous classifications and correctly explain it by a lot of Oghuz elements in the wordlist. However, the authors do not try to exclude them despite careful elimination of all non-Turkic borrowings. Robbeets and Savelyev's wordlist of Literary Crimean Tatar (based on the Orta dialect) is compiled on the basis of Useinov 2007. 13 lexemes from this list can be treated as Oghuz loans and contact-influenced semantic innovations: *el* 'hand', *qarın* 'belly', *kemik* 'bone', *köküs* 'breast', *köpek* 'dog', *uzaq* 'far',

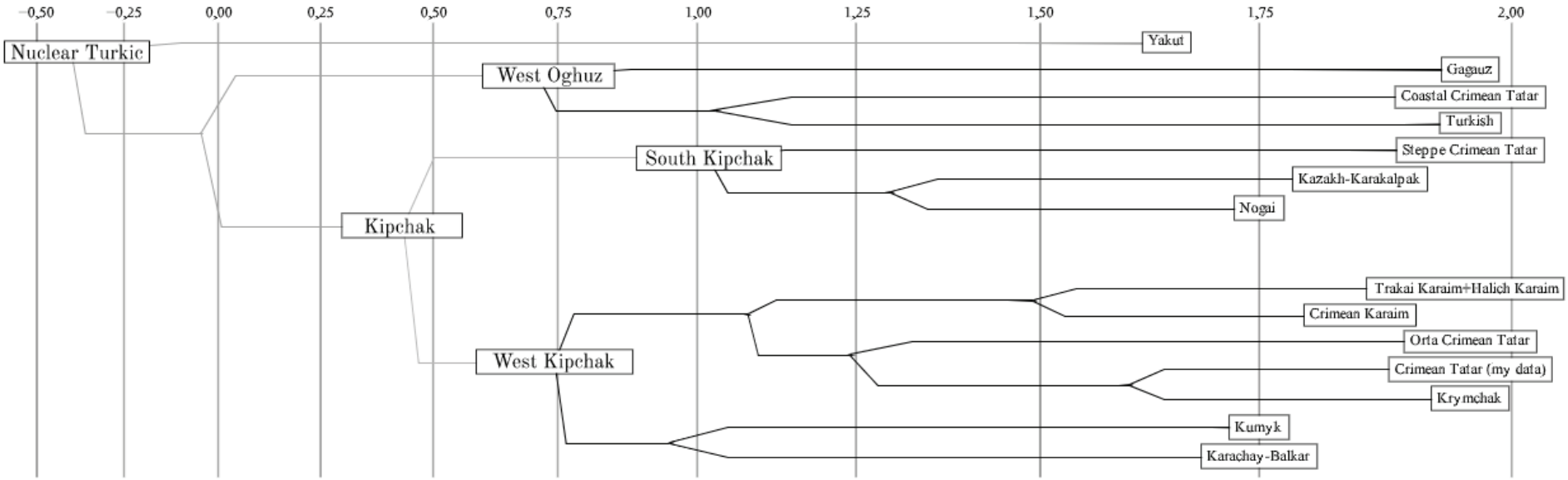

Figure 5. Tree constructed with neighbor-joining algorithm in Starling software.

ket- ‘to go’, *eyi* ‘good’, *čoq* ‘many’, *ufačiq* ‘small’, *küneš* ‘sun’, *sižaq* ‘warm’, *sivalčan* ‘worm’. At the same time, the list contains the following lexemes typical for the Kipchak subgroup: *qursaq* ‘belly’, *aša-* ‘to eat’, *ayt-* ‘to say’, *yuqla-* ‘to sleep’, *terek* ‘tree’. My remark made on Polinsky’s results is true for this study as well. If an effort is made to exclude 13 items from the dataset, we should be able to identify the original genealogical affiliation of the Orta Crimean Tatar language. That a language should contain so many borrowings from one source in its basic vocabulary is not a frequent case, but hardly a unique exception either (the abovementioned case of Riksmål is another example of a language with a similar amount of borrowings in the basic vocabulary).

5. Conclusions

Detection of borrowings is a necessary procedure for the purposes of phylogenetic studies. A historical linguist should be attentive not only to external but also to intra-family loans. When closely related languages are in intimate contact with each other, the areal criterion becomes more important than the phonological one. The similar phonotactics, phonological inventories, and the minor differences in the phonological shape of words driven by historical phonological processes often make application of phonological criteria difficult or even completely impossible. Loans from a genetically related language can sometimes look like archaic items. They can be both MAT- and PAT-borrowings. These fake archaisms can be revealed based on the distributional criterion if the general direction of influence is known.

Borrowings from a closely related dominant language can strongly influence the basic vocabulary of less prestigious idioms and make the genealogical classification of the latter quite difficult. However, careful elimination of all borrowings makes it possible to identify the subgroup to which the language in question belongs. Both manual analysis of isoglosses and computational lexicostatistics give acceptable results if the dataset is free from borrowings.

Acknowledgments

I thank Anna Dybo and Alexei Kassian for their consultations at all stages of preparation of the present paper. I also acknowledge Maria Polinsky, Christopher Straughn, Mark Zimin, Eldar Idrisov, and Svetlana Egorova for making available to me the necessary materials and for their help in data collection. Last but not least, I would like to thank my friends Tatiana Kliuchnikova and Nikolay Kim for our informal discussions which helped me make much clearer the general idea and the actual text of the paper. All mistakes are, of course, mine.

All supplementary materials mentioned in the paper are archived online at: www.jolr.ru//jlr18/egorov.zip.

Abbreviations of sources used in Baskakov, Szapszał & Zajączkowski 1974 and Aqtay & Jankowski 2015

Cam – Cambridge manuscript of the whole Bible except Chronicles, vol i-iv, Cambridge University Library, classmark BSMS 288.

Fil – Filonenko, V. I. 1929. Atalar sozy: karaim idioms. Proceedings of the Tauride society of history, archeology and ethnography. Vol. 1. Simferopol.

Man – Manchester manuscript of some portions of the Bible. The Rylans Library, classmark Gaster H 170.

Meq – Meqabbeç, a prayer book printed in 1734.

R – Radloff 1896.
Par. – Karaim manuscript from French National Library, signature Hebr. 666.
Sz – Card files collected for Szapszał's dictionary of Crimean Karaim, manuscript.
ZR – Zeḳer rav, published in 1831 in Istanbul by Joseph Shelomo Lucki, see edition Poznański 1913.
Q – Qılcı's Mejuma, edition Aqtay 2009.

Abbreviations for names of languages and dialects

Az. – Azerbaijani
Bash. – Bashkir
Chag. – Chagatai
Chuv. – Chuvash
CoCrTat. – Coastal Crimean Tatar
CrKar. – Crimean Karaim
CrTat. – Crimean Tatar
dial. – dialectal
Fin. – Finnish
HKar. – Halich Karaim
K.-B. – Karachay-Balkar
Karak. – Karakalpak
KarakhUyg. – Karakhanid Uyghur
Kaz. – Kazakh
Khant. – Khanty
Kip. – Kipchak
Kirg. – Kirgiz
Krym. – Krymchak
Kum. – Kumyk
Nog. – Nogai
Ogh. – Oghuz
OrCrTat. – Orta Crimean Tatar
OT – Old Turkic
OUyg. – Old Uyghur
PT – Proto-Turkic
Rus. – Russian
SaaN. – Northern Saami
Sal. – Salar
SCrTat. – Steppe Crimean Tatar
Tat. – Tatar
TrKar. – Trakai Karaim
Tur. – Turkish
Turkm. – Turkmen
Turkm. – Turkmen
Tuv. – Tuvinian
Ukr. – Ukrainian
Uyg. – Uyghur
Uzb. – Uzbek
Yak. – Yakut

References

Aikio, Ante. 2007. Etymological nativization of loanwords: A case study of Saami and Finnish. In: Ida Toivonen, Diane Nelson (eds.). *Saami Linguistics* (Current Issues in Linguistic Theory 288): 17–52. Amsterdam, Philadelphia: John Benjamins Pub. Co.

Aqtay, Gulayhan. 2009. *Eliyahu Ben Yosef Qılcı's Anthology of Crimean Karaim and Turkish Literature*. Istanbul: Yıldız Dil ve Edebiyat Dizisi 8.

Aqtay, Gulayhan & Henryk Jankowski. 2015. *A Crimean Karaim-English dictionary* (Prace Karaimoznawcze 2). Poznań.

Bammatov, Burgan G., Nurmagomed E. Gadzhiakhmedov. 2011. *Kumyksko-russkij slovar' [Kumyk-Russian dictionary]*. Makhachkala.

Baskakov, Nikolay A. (ed.). 1956. *Russko-Nogajskij slovar' [Russian-Nogai dictionary]*. Moscow: Gosudarstvennoe izdatel'stvo innostrannyx i nacyonal'nyx slovarej.

Baskakov, Nikolay A. 1967. *Russko-karakalpakskij slovar' [Russian-karakalpak dictionary]*. Moscow: Sovetskaya entsiklopediya.

Baskakov, Nikolay A., Seraja Ben Mordechai Szapszał, Ananiasz Zajączkowski. 1974. *Słownik karaimsko-rosyjsko-polski [Karaim-Polish-Russian dictionary]*. Moscow: Russkij Jazyk.

Bektaev, Kaldybay. 1996. *Bol'šoj kazaxsko-russkij russko-kazaxskij slovar' [Large Kazakh-Russian Russian-Kazakh dictionary]*. Kazaxsij proekt razvitija gosudarstvennogo jazyka.

Bekturov, Shabken & Ardak Bekturova. 2001. *Kazaxsko-russkij slovar' [Kazakh-Russian dictionary]*. Astana: Foliant.

Bergsland, Knut, Hans Vogt. 1962. On the Validity of Glottochronology. *Current Anthropology* 3(2): 115–153.

Bogochanskaya, Nina N., Anna S. Torgashova. 2009. *Bol'šoj russko-tureckij slovar' [Large Russian-Turkish dictionary]*. Moscow: Dom Slavjanskoj Knigi.

Bouckaert, R., P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard, Q. D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337: 957–960. doi:10.1126/science.1219669.

Burlak, Svetlana A., Sergei A. Starostin. 2005. *Sravnitel'no-istoričeskoe jazykoznanie [Comparative-historical linguistics]*. Moscow: Academia.

Campbell, Lyle. 2013. *Historical linguistics: an introduction*. 3rd edn. Edinburgh University Press.

Clauson, Gerard. 1972. *An etymological dictionary of Pre-Thirteenth-Century Turkish*. Oxford University Press.

Dybo, Anna V. 1996. *Semantičeskaja rekonstrukcija v altajskoj étimologii: somatičeskie terminy (plečevoj pojas) [Semantic reconstruction in the Altaic etymology: somatic terms (shoulder girdle)]*. Moscow.

Dybo, Anna V. 2007. *Lingvističeskie kontakty rannix tjurkov. Leksičeskij fond. Pratjurkskij period [Language contacts of the early Turkic peoples: vocabulary: Proto-Turkic epoch]*. Moscow: Vostochnaya Literatura.

Dybo, Anna V. 2013. *Ètimologičeskij slovar' tjurkskix jazykov [Etymological dictionary of the Turkic languages]. Vol. 9: Ètimologičeskij slovar' bazisnoj leksiki tjurkskix jazykov [Etymological dictionary of the Turkic basic vocabulary]*. Astana: Prosper Print.

Étnografičeskaja karta Kryma [Ethnographic map of Crimea]. 1926. GosKrymIzdat.

Filonenko, W. J. 1931. *Ethnografische Karte der Krimer autonomen sozialistischen Sowjetrepublik* [Ethnographic map of the Crimean Autonomous Soviet Socialist Republic]. Wien.

Fisher, Alan W. 1978. *Crimean Tatars* (Studies of Nationalities in the USSR). Stanford, Calif.: Hoover Institution Press.

Gaydarzhi, Gavril A., Ludmila A. Pokrovskaya, Boris P. Tukan, Elena K. Koltsa. 1973. *Gagauzsko-russko-moldavskij slovar' [Gagauz-Russian-Moldovan dictionary]*. (Ed.) Nikolay A. Baskakov. Moscow: Sovetskaja entsiklopedija.

Gochiyayeva, Sofya A., Kh. I. Suyunchev. 1989. *Karačaevo-balkarsko-russkij slovar' [Karachay-Balkar-Russian dictionary]*. Moscow: Russkij Jazyk.

Goloboff, Pablo A., Santiago A. Catalano. 2016. TNT version 1.5, including a full implementation of phylogenetic morphometrics. *Cladistics* 32(3): 221–238. doi:10.1111/cla.12160.

Gray, Russell D., Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426: 435–439. doi:10.1038/nature02029.

Haspelmath, Martin. 2009. Lexical borrowing: concepts and issues. In: Uri Tadmor, Martin Haspelmath (eds.). *Loanwords in the World's Languages: A Comparative Handbook*: 35–54. Berlin / New York: Walter de Gruyter. doi:10.1515/9783110218442.

Huelsenbeck, J. P., F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17: 754–755.

Ianbay, Iala. 2016. *Krimchak Dictionary*. Wiesbaden: Harrassowitz Verlag.

Izidinova, Seveli R. 1996. Krymskotatarskij jazyk [Crimean Tatar language]. In: Edkhyam R. Tenishev (ed.). *Tjurkskie jazyki [Turkic languages]* (Jazyki Mira [World's Languages]): 298–309. Moscow.

Jankowski, Henryk. 2003. On the Language Varieties of Karaims in the Crimea. *Studia Orientalia* 95: 109–130.

Jankowski, Henryk. 2017. Karaim and Krymchak. In: Lily Kahn, Aaron D. Rubin (eds.). *Handbook of Jewish Languages*: 452–489. Leiden: Brill. doi:10.1163/9789004359543_015.

Johanson, Lars. 1998. The history of Turkic. In: Lars Johanson, Éva Á. Csató (eds.). *The Turkic languages* (Routledge Language Family Descriptions): 81–125. London: Routledge.

Kassian, Alexei S., George Starostin, Anna Dybo, Vasily Chernov. 2010. The Swadesh wordlist. An attempt at semantic specification. *Journal of Language Relationship* 4: 46–89.

Kassian, Alexei S., George S. Starostin, Mikhail A. Zhivlov. 2015. Proto-Indo-European-Uralic comparison from the probabilistic point of view. *Journal of Indo-European Studies* 43(3–4): 301–347.

Kassian, Alexei S., Mikhail A. Zhivlov, George S. Starostin, Artem A. Trofimov, Petr A. Kocharov, Anna Kuritsyna, Mikhail N. Saenko. Forthcoming. *Rapid radiation of the Inner Indo-European languages: an advanced approach to Indo-European lexicostatistics*. https://www.academia.edu/39903804/.

Kavitskaya, Darya. 2010. *Crimean Tatar* (Languages of the World/Materials 477). München: LINCOM EUROPA.

Kocaoğlu, Timur. 2006. *Karay: the Trakai dialect* (Languages of the World/Materials 458). München: LINCOM EUROPA.

Kogan, Leonid E. 2015. *Genealogical classification of Semitic: the lexical isoglosses*. Boston, Berlin: De Gruyter.

Musaev, Kenesbai M. 1964. *Grammatika karaimskogo jazyka: fonetika i morfologija [Grammar of the Karaim langauge: phonetics and morphology]*. Moscow: Nauka.

Musaev, Kenesbai M. 2010. Dialekty karaimskogo jazyka [Dialects of the Karaim language]. In: Anna V. Dybo (ed.). *Dialekty tjurkskix jazykov [Dialects of the Turkic languages]*: 205–235. Moscow: Vostochnaya literatura.

Nakhleh, Luay, Donald A. Ringe, Tandy Warnow. 2005. Perfect Phylogenetic Networks: A New Methodology for Reconstructing the Evolutionary History of Natural Languages. *Language* 81(2): 382–420. doi:10.1353/lan.2005.0078.

Parker, Philip M. 2008. *Webster's Turkish-English thesaurus dictionary*. San Diego: ICON Classics.

Pekarskiy, Eduard K. 1916. *Kratkij russko-jakutskij slovar' [Concise Russian-Yakut dictionary]*. Saint Petersburg: Tipografija imperatorskoj akademii nauk.

Pekarskiy, Eduard K. 1959. *Slovar' jakutskogo jazyka [Dictionary of the Yakut language]*. Leningrad: Izdatelstvo Akademii Nauk SSSR.

Pokorny, Julius. 1959. *Indogermanisches etymologisches Wörterbuch [Indo-European etymological dictionary]*. 2 vols. Bern / Munich: A. Francke.

Polinsky, Maria. 1991. The Krymchaks: History and Texts. *Ural-Altaic Yearbook* 63: 123–154.

Polinsky, Maria. 1992. Crimean Tatar and Krymchak: classification and description. In: Howard I. Aronson (ed.). *The Non-Slavic languages of the USSR: linguistic studies*: 157–188. Chicago Linguistic Society, University of Chicago.

Poznański, Samuel. 1913. Karäisch-tatarische Literatur. *KeletiSzemle* 13: 37–47.

Radloff, Wasilij. 1896. *Der Volksliteratur der nördlichen türkischen Stämme [The folk literature of the northern Turkish tribes]*. Vol. 7: Die Mundarten der Krym. Saint Petersburg: Tipografija imperatorskoj akademii nauk.

Rajki, András. 2007. *A concise Gagauz dictionary*. Budapest.

Räsänen, Martti. 1957. *Materialien zur Morphologie der türkischen Sprachen* (Studia Orientalia 21). Helsinki: Societas Orientalis Fennica.

Räsänen, Martti. 1969. *Versuch eines etymologisches Wörterbuch der Türksprachen [Attempt at etymological dictionary of Turkic languages]* (Lexica Societatis Fenno-Ugricae). Helsinki: Suomalais-Ugrilainen Seura.

Rebi, David. 2004. *Krymčakskij jazyk: krymčaksko-russkij slovar' [Krymchak language: Krymchak-Russian dictionary]*. Simferopol: Dolya.

Ronquist, Fredrik, Maxim Teslenko, Paul van der Mark, Daniel L. Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A. Suchard, John P. Huelsenbeck. 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* 61(3): 539–542. doi:10.1093/sysbio/sys029.

Rzymski, Christoph, Tiago Tresoldi, Johann-Mattis List, Simon Greenhill, Robert Forkel. 2019. *The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies*. https://clics.clld.org/.

Sakel, Jeanette. 2007. Types of loan: matter and pattern. In: Jeanette Sakel & Yaron Matras (eds.). *Grammatical borrowing in cross-linguistic perspective* (Empirical Approaches to Language Typology 38): 15–29. Berlin, New York: De Gruyter Mouton.

Savelyev, Alexander, Martine I. Robbeets. 2020. Bayesian phylolinguistics infers the internal structure and the time-depth of the Turkic language family. *Journal of Language Evolution* 5(1): 39–53. doi:10.1093/jole/lzz010.

SesliSözlük. *SesliSözlük. Online English-Turkish and Multilingual Dictionary*. https://www.seslisozluk.net/ (29 July, 2020).

Sevortyan, Ervand V. 1966. Krymsko-tatarskij jazyk [Crimean Tatar language]. In: Viktor V. Vinogradov (ed.). *Jazyki narodov SSSR, vol. 2: Tyurkskie jazyki* [Turkic languages]: 234–259. Moscow: Nauka.

Sevortyan, Ervand V., Liya S. Levitskaya, Anna V. Dybo, Valentin I. Rassadin, Galina F. Blagova, Dmitriy M. Nasilov. 1974. *Ètimologičeskij slovar' tjurkskix jazykov [Etymological dictionary of the Turkic languages]*. 9 vols. Moscow: Nauka.

Stachowski, Marek. 1993. *Dolganischer Wortschatz* (Zeszyty naukowe Uniwersytetu Jagiellońskiego, Prace językoznawcze 1086. zesz. 114). Kraków: Nakl. Uniewersytetu Jagiellońskiego.

Stachowski, Marek. 1998. *Dolganischer Wortschatz: Supplementband*. Kraków: Księgarnia Akademicka.

Starostin, George S. 2016. From wordlists to proto-wordlists: reconstruction as 'optimal selection.' *Faits de langues* 47(1): 177–200. doi:10.3726/432492_177.

Starostin, Sergei A. 2000. Comparative-historical linguistics and lexicostatistics. In: Colin Renfrew, April McMahon & Larry Trask (eds.). *Time depth in historical linguistics, vol. 1*: 223–265. Cambridge, England: The McDonald Institute for Archaeological Research.

Starostin, Sergei A. 2007a. Rabočaja sreda dlja lingvista [Linguist's workspace]. In: Sergei A. Starostin. *Trudy po jazykoznaniju [Works on linguistics]*: 481–496. Moscow: Jazyki slavjanskix kul'tur.

Starostin, Sergei A. 2007b. Opredelenie ustojčivosti bazisnoj leksiki [Defining the stability of basic lexicon]. In: Sergei A. Starostin. *Trudy po jazykoznaniju [Works on linguistics]*: 827–839. Moscow: Jazyki slavjanskix kul'tur.

Syzdykova, Rabiga G., K. Sh. Khusaiyn. 2008. *Kazaxsko-russkiy slovar' [Kazakh-Russian dictionary]*. Alma-Ata.

Tenishev, Edkhyam R., Anna V. Dybo (eds.). 2002. *Sravnitel'no-istoričeskaja grammatika tjurkskix jazykov: regional'nye rekonstrukcii [Comparative grammar of Turkic languages: regional reconstructions]*. Moscow: Nauka.

Tenishev, Edkhyam R., Anna V. Dybo (eds.). 2006. *Sravnitel'no-istoričeskaja grammatika tjurkskix jazykov: pratjurkskij jazyk-osnova. Kartina mira pratjurkskogo ètnosa po dannym jazyka [Comparative grammar of Turkic languages: the Proto-Turkic language. A world-view of the Proto-Turkic people according to language data]*. Moscow: Nauka.

Trepavlov, Vadim V. 2014. Nogajskaja orda [Nogai Horde]. In: Rafael S. Khakimov (ed.). *Istorija tatar s drevnejšix vremen [History of Tatars from antiquity]*, vol. 4: 223–252. Kazan.

Useinov, Seyran M. 2007. *Russko-krymskotatarskij, krymskotatarsko-russkij slovar' [Russian-Crimean Tatar, Crimean Tatar-Russian dictionary]*. Simferopol: Tezis.

И. М. Егоров. Базисная лексика близкородственных языков в ситуации языкового контакта: тюркские языки Крымского полуострова

Настоящая статья объединяет два разыскания в области базисной лексики тюркских языков Крымского полуострова. Ее цель — заострить внимание на проблемах, с которыми сталкиваются лингвисты при диахроническом — и в особенности филогенетическом — анализе интенсивно контактирующих друг с другом близкородственных языков. Первое исследование посвящено ономасиологической реконструкции пракараимского списка Сводеша. Основная рассматриваемая здесь проблема — выявление западно-огузских заимствований и, в первую очередь, контактно обусловленных архаизмов (fake archaisms) в крымском диалекте караимского. Задача второго исследования — определение генеалогической принадлежности крымскотатарских диалектов. Ручной анализ инноваций в базисной лексике и алгоритмы вычислительной филогенетики (байесовский метод, метод ближайших соседей, метод максимальной бережливости) подтверждают традиционное мнение о том, что береговой диалект принадлежит к огузской группе, средний — к западно-кыпчакской, а восточный — к ногайско-кыпчакской группе. Такой результат полностью подтверждается данными по этнической истории. Установить правильную генеалогическую аффилиацию рассматриваемых диалектов удалось только после выявления всех заимствований, чего не делалось в предыдущих лексикостатистических исследованиях по крымско-татарскому языку. Оба изученных кейса показывают, что элиминация ареальных влияний принципиально важна и для семантической (ономасиологической) реконструкции, и для филогенетических исследований.

Ключевые слова: филогенетика; семантическая реконструкция; заимствования; караимский язык; крымскотатарский язык; тюркские языки.