Chapter 9

A Bayesian approach to the classification of the Turkic languages

Alexander Savelyev

Abstract

Despite more than 150 years of research, the internal structure of the Turkic language family remains a controversial issue. In this study, I employ the Bayesian phylogenetic approach in order to provide an independent verification of the contemporary views on the Turkic linguistic history. The data underlying the study are Turkic basic vocabularies, which are resistant to replacement and likely to reflect the genealogical relationships among the Turkic languages. The method tested in the chapter is based on the strict clock model of evolution, which assumes that relevant changes occur at the same rate at every branch of the family. This study supports the widespread view that the binary split between Bulgharic and Common Turkic was the earliest split in the Turkic family. The model further replicates most of the conventional sub-groups within the Common Turkic branch. Based on a Bayesian analysis, the time-depth of Proto-Turkic is estimated to be around 2119 years BP, which is in accordance with the traditional estimates of 2000–2500 years BP.

Keywords: Turkic languages, genealogical classification, Bayesian phylogenetic approach, basic vocabulary, the history of Turkic

9.1 Introduction

The Turkic language family represents a promising but rather puzzling case of genealogical classification. From a cross-linguistic perspective, it is a relatively young family, with a time-depth of approximately 2000–2500 years according to different estimates (see Robbeets et al., this volume: Chapter 43). At the same time, it stands out by a high level of internal diversity, numbering more than 40 languages and peculiar dialects that are accessible to historical linguistic analysis. Much of this diversity seems to be due to the fact that, since the late Proto-Turkic period and until

very recent times, most Turkic-speaking populations belonged, or at least had close ties, to the world of nomadic pastoralism (Golden 1998; Robbeets et al., this volume: Chapter 43; Hudson, this volume: Chapter 47). This same nomadic lifestyle may have contributed to one remarkable thing about the Turkic family, which is relevant for establishing its internal structure, namely that different branches and sub-branches of Turkic are known to have maintained extensive mutual contact throughout their history. Due to convergence phenomena, the original relationships between the Turkic languages are not easy to disentangle, and approaching the problem is a truly challenging task.

Numerous problematic nodes in the phylogeny of the Turkic family are sometimes considered as having their roots in what Aharon Dolgopolsky once called "the genealogical stump of the Turkic languages" as referred to by Belikov (2009: 53). Although somewhat misleading in the sense that it rejects fundamental applicability of the tree model to the Turkic family, this term wittily points to the methodological obstacles facing attempts to reveal its original structure. One practical consequence of such skepticism is that many of the traditional classifications of the Turkic languages bypass problematic issues, combining areal clusters and clear-cut genealogical nodes in the same classification, depending on which better fits the particular case (see Johanson, this volume: Chapter 8, for an overview of the previous classifications of Turkic).

The opposite point of departure, namely that it is essentially feasible to build a genealogical tree of the Turkic family by gradual elimination of the effects of intra-family contact in all controversial cases, is taken in the models based on a quantitative approach to language classification. Among several classifications of this type, those by Djačok (2001) and Dybo (2006, 2013) are based on lexicostatistics, a distance-based method relying on pairwise comparison of basic vocabularies and building a matrix of shared cognates. Both authors used the Swadesh list as a source of basic vocabulary meanings and S. Starostin's (1989) formula that excludes borrowings from the calculations. A more experimental approach was proposed by Mudrak (2009) who adapted basic assumptions of lexicostatistics to measure distances in the domain of Turkic historical morphology

and phonology. In general, the models using a distance-based approach to the classification of the Turkic languages show a high degree of overlap with those based on the method of shared innovations (Baskakov 1960: 228–229, 1981: 18–20; Menges 1995: 59–66; Johanson, this volume: Chapter 8), although with some relevant discrepancies.

Taking into account different classifications of the Turkic family proposed so far, the state of the art in the field can be summarized as follows. There is a wide consensus that the earliest split in the family was the split between the Bulgharic ("Oghuric") branch, which is represented by the only living language Chuvash (Savelyev, this volume: Chapter 27), and the Common Turkic branch including all the other Turkic languages spoken on the vast area from the Mediterranean to Northeastern Siberia. From the other end, most lower-level subgroups of Common Turkic, such as Kipchak (Northwestern Turkic), Oghuz (Southwestern Turkic) and Karluk (Southeastern Turkic), are usually considered as non-controversial. Yet, the genealogical relations between these three groupings, that is, whether they are three sister branches or rather any two of them stand in a closer relationship, remain subject to discussion. The status of the fourth large unit within Common Turkic, the Siberian Turkic (Northeastern Turkic) languages, as a homegeneous group is generally debated. Moreover, while the reliability of North Siberian (Yakut–Dolgan) as a separate grouping is beyond doubt, there is no consensus on whether the two sub-branches spoken in Southern Siberia, Sayan Turkic (Tuvan–Tofa) and Khakassic (non-Sayan-Turkic), can indeed be traced to a single ("South Siberian") genealogical node (Johanson 1998a: 83). In particular, recent studies based on a quantitative approach (Mudrak 2009: 179; Dybo 2013: 18) suggest that the Sayan Turkic branch should, rather, be paired together with the North Siberian Turkic languages. Last but not least, there are individual languages or, sociolinguistically, particular dialects whose exact position in the Turkic family tree remains controversial. These include *inter alia* different varieties of Old Turkic, an early form of Common Turkic known due to inscriptions from as early as the 7th century AD, and Khalaj, the language of a Turkic minority group populating Western Iran. Although Old Turkic is often referred to as the ancestor of (at least most) Common Turkic languages, different studies

may highlight its specific closeness to individual subgroups of Common Turkic, such as Oghuz (Johanson 1998a; Dybo 2006), Karluk (Mudrak 2009), or Siberian Turkic (Dybo 2016). Tekin (1990) and Johanson (1998a; this volume: Chapter 8), following Doerfer (1971b), consider Khalaj as a separate branch of Turkic that split off immediately after the Bulgharic branch, while Ščerbak (1997) and Dybo (2016) emphasize its affinities with the Oghuz branch, and Mudrak (2009) interprets Khalaj as an early offshoot of the Karluk branch. Most of the other Turkic languages whose place in the classification may look unclear are in some way connected to the Siberian Turkic area and seem to be severely affected by secondary convergence phenomena. That is definitely the case for at least some of Siberian Tatar varieties as well as North Altay and South Altay dialects, which are characterized by an interaction of both indigenous (Northeastern) and Kipchak (Northwestern) components; Saryg Yugur (West Yugur) as a language of South Siberian origin that has been influenced by the Karluk branch; Kirghiz as a language that shares numerous isoglosses with the Kipchak, in particular South Kipchak, languages and, at the same time, with South Altay dialects. An interesting exception in this respect is Salar, an Oghuz language strongly affected by the Karluk branch.

New prospects for inferring the genealogical relationships of the Turkic languages have opened up after the novel Bayesian phylogenetic method has been implemented into a historical linguistic framework. Bayesian analysis is a probabilistic approach that allows to integrate different forms of prior knowledge, such as dating extinct languages and non-controversial nodes in the tree structure, based on proper linguistic or interdisciplinary (e.g., archaeological) evidence. Another advantage of this method is that, rather than producing a single optimal tree, it offers a distribution of trees sampled in proportion to their posterior probability given the data and the model. This allows the level of support for each grouping to be quantified. The models themselves are statistically comparable via the Bayes factors (Dunn 2015; Bowern and Atkinson 2012). In the recent decades, the Bayesian phylogenetic method has been applied to trace the origin of different language families, such as Indo-European (Gray and Atkinson 2003), Austronesian (Gray et al. 2009), and

Semitic (Kitchen et al. 2009).

Previously, Bayesian inference has been applied to the Turkic data by Hruschka et al. (2014). The study is focused on general theoretical aspects of linguistic evolution, namely on analyzing regular sound changes as events of concerted evolution. The phonetic evidence provided by 26 Turkic languages (adapted from Starostin et al. 2003) is used to test the authors' hypothesis about the nature of concerted changes. One of the obtained phylogenetic trees—the one presenting the evolution of the Turkic languages through historical events of regular sound change—replicates many of the non-controversial nodes in the family, although the position of some languages in the tree contradicts the conventional understanding of the Turkic linguistic history.

In this chapter, I present an application of Bayesian inference to lexical data from 30 Turkic languages in order to trace their genealogical relations and estimate the time-depth of the family. In Section 9.2, I introduce the data and the method of the study. Section 9.3 provides the results of Bayesian analysis. In Section 9.4, I propology a historical interpretation of the obtained tree, regarding its topology and time-depth. In conclusion (Section 9.5), I summarize the results obtained and discuss future prospects for studying the internal structure of the Turkic family.

9.2 Data and method

9.2.1 Languages

This study is based on lexical evidence from 32 Turkic languages. The 30 modern languages include Chuvash (based on the more archaic Viryal dialect), Yakut (Sakha), Dolgan, Khalaj, Karaim (based on the dialects of Halich and Trakai and not including the highly distinct Crimean dialect), Kumyk, Karachay-Balkar, Uzbek, Modern Uyghur, Kazan Tatar, Siberian Tatar (based on the Baraba variety), Bashkir, Nogai, Kirghiz, Kazakh, Karakalpak, Salar, Turkmen, Crimean Tatar, Azeri, Turkish, Gagauz, Saryg Yugur (West Yugur, Yellow Uyghur), North Altay (based on the Chelkan dialect) and South Altay (based on the Altay-Kiži dialect), Tuvan, Tofa, Middle Chulym, Shor, and Khakas (based on the Kacha dialect). Extinct Turkic languages are represented by Old

Turkic and Cuman. For the sake of uniformity, Old Turkic data are restricted as far as possible to the evidence of Old Uyghur texts from the 9th century AD. The label “Cuman” is applied to a Middle Kipchak variety attested in the *Codex Cumanicus* manuscript, dating from the early 14th century AD. The languages in the dataset represent all essential groupings of the Turkic family obtained by the previous studies, including both controversial and non-controversial entities.

9.2.2 Basic vocabulary list

In this study, I compared Turkic basic vocabularies, that is, sets of lexical items that are, cross-linguistically, particularly resistant to replacement, be it for internal (semantic shift or lexical replacement) or external (borrowing) reasons. The underlying set of basic vocabulary meanings is the 200-Leipzig-Jakarta, reduced to 195 items due to a merge of several meanings. The Leipzig-Jakarta list has certain advantages as compared to the more traditional Swadesh list, e.g., the former is based on a quantitative comparison of most stable words in languages across the world while the latter is based mainly on intuition (Haspelmath and Tadmor 2009a). I supplemented it further by the 200-Jena list, which for the most part replicates the 200-Swadesh list and is currently applied to a comparison of Indo-European languages in the CoBL (Cognacy in Basic Lexicon) project (Anderson and Heggarty n.d.). Given the large overlap between the two lists, the final list has included 254 meanings given in Appendix 1. Although some of these meanings appeared to be rather unstable as applied specifically to the Turkic family, the list was further kept intact in order to avoid cherry-picking, to maintain consistency with Bayesian classifications that have been developed in parallel for the other branches of Transeurasian (see Section 6.6 in Whaley and Oskolskaya, this volume: Chapter 6, for a classification of Tungusic) and to provide a uniform benchmark for a future Bayesian classification of the whole Transeurasian family. The meanings were precisely defined in order to avoid possible ambiguity, partially based on previous attempts at basic vocabulary semantic specification, such as Kassian et al. (2010), Dybo (2013), Starostin (2013), and the CoBL documentation.

9.2.3 Sources and data selection

I relied on a multidimensional understanding of a word's "basicness", that is, its suitability for inclusion in the basic vocabulary list, using a wide range of criteria. In case of considerable synonymy, preference was given to more generic, more frequent, stylistically neutral and morphologically simple terms. It can be argued that the most appropriate sources in view of these criteria, reflecting the pragmatic aspect of basicness, would be those based on direct elicitation from informants or comprehensive bilingual corpora. However, as such options are not easily accessible for most of the Turkic languages, the primary source of evidence on basic vocabulary were, inevitably, bilingual dictionaries or, in more problematic cases, wordlists extracted from grammar descriptions. If available, I still double-checked the basicness of a word through parallel texts, such as included in textbooks and phrasebooks. For two languages in the sample, Chuvash and Middle Chulym, my fieldnotes from 2011–2015 and 2015, respectively, were the main source of basic vocabulary. Given the geographical distribution of the Turkic languages and the historical context of their documentation, English as the language of basic vocabulary meanings and the target Turkic languages were in most cases mediated by Russian. The full list of the sources used in this study for collecting Turkic basic vocabularies is given in Appendix 2.

Selection of two lexemes as synonyms for one basic vocabulary meaning was allowed within reasonable bounds. The basic principle of dealing with synonymy can be formulated as follows: a word was included in the list unless there was positive evidence that it was less basic than at least one of its synonyms (e.g., based on semantic nuances or frequency in available texts). An important restriction was that singletons—that is, words that were present in a given basic vocabulary meaning only in one language—were removed from the dataset in case they had at least one non-singleton synonym that fit the criteria for basic status. The reason was that, for a number of languages in the dataset, the only sources suitable for the study's purposes were dictionaries providing a set of synonyms as a translation for a given basic vocabulary meaning, without

commenting on the difference in use. Due to a lack of other dictionary-like or corpus-like sources on these languages, it was hardly feasible to rank such synonyms by basicness. Being included in the dataset based on a lack of evidence on their "non-basicness", singleton synonyms would be likely to affect critically the position of the corresponding language in the tree—all because of the quality of the sources, rather than linguistic features. Therefore, excluding singleton synonyms from the dataset seemed to be a way to counter the uneven quality of sources for different languages.

Another major constraint was that all borrowings that have been identified using the comparative method (see Section 9.2.4 for more detail) were excluded from the dataset in order to provide a clearer phylogenetic signal. The lexical items that neither could be reliably identified as borrowings, nor had a plausible Turkic etymology, were preserved in order to avoid the loss of relevant data.

The quality of the resulting dataset can be described as follows. Almost all languages in the sample are rather well documented, and lexical data are evenly distributed across the family. The mean amount of missing data due to gaps in documentation or borrowings in basic vocabulary is around 7 %. Several languages in the dataset are clearly undersampled against the background of the other Turkic languages. Khalaj, Salar and Baraba Tatar as well as the extinct Cuman are rather limitedly documented languages. In addition, Khalaj is drastically affected by Persian and, to a lesser extent, South Azeri varieties, including its basic vocabulary. The same refers to Chinese borrowings in Salar basic vocabulary. Due to these factors, the amount of missing data is 30 % for Khalaj, 21 % for Cuman, 18 % for Salar and 17 % for Baraba Tatar. For the other languages in the sample, this figure varies from 11 % for Dolgan to 1 % for Nogai and Karaim. The amount of present cognates is 22 % for Khalaj; for the other languages, it varies from 25 % (Cuman, Salar, and Dolgan) to 32 % (Nogai and Modern Uyghur).

9.2.4 Cognate coding

For each word in the dataset, a thorough etymological analysis was done in order to establish cognacy classes and exclude borrowings. To this end, I used the latest comparative etymological

dictionaries of the Turkic family (Sevortjan et al. 1974–2003; Tenišev et al. 2001; Dybo 2013; the Turkic part of Starostin et al. 2003) as well as dictionaries of individual languages that include etymological information, such as (Clauson 1972; Stachowski 1993; Fedotov 1996; Tatarincev 2000–2008). Cognacy classes were established based on regular sound correspondences. Each cognacy class was coded as present (1) or absent (0) in each language. In order to process cognates, I used the EDICTOR software tool (List 2017). The resulting matrix of binary characters included 910 cognacy classes, covering 254 basic vocabulary meanings across 32 Turkic languages. An additional all-zero column has been added to the dataset as a correction for ascertainment bias in order to compensate for missing data.

9.2.5 Bayesian analysis

I applied a Bayesian approach as implemented in BEAST v. 2.4.8 (Bouckaert et al. 2014). The Bayesian analysis adopted in BEAST uses Markov chain Monte Carlo (MCMC) algorithm to sample the posterior probability distribution of tree topologies. MCMC chains were run for 50 million generations, sampled every 1000 generations. The first 5 million iterations were discarded as burn-in, and post-run analysis as implemented in the Tracer v. 1.6 component of the BEAST package (Rambaut et al. 2014) revealed that runs had reached convergence by the end of the burn-in period. I used then the TreeAnnotator tool in BEAST to achieve the maximum clade credibility tree. The analysis was conducted using the Fossilized Birth-Death model, which is most appropriate for the data that contain extinct languages (Stadler et al. 2018). I further compared the fit of three probable models of cognate evolution: the simple binary model (Gray and Atkinson 2003), the binary covarion model, allowing cognates to be in either a “fast” or “slow” state (Gray et al. 2009), and the stochastic Dollo model, which assumes that cognates can be gained once but lost multiple times (Nicholls and Gray 2006). In this study, I focused on testing these models as combined with the strict clock model of evolution, assuming that every branch in the tree evolves according to the same evolutionary rate. To calibrate the clock, I relied on the sampling dates for the two extinct

languages in the dataset, 1150 years BP for Old Turkic and 700 years BP for Cuman. No monophyletic constraints were introduced on the branches.

9.3 Results

I used the model comparison option as implemented in Tracer v. 1.6 in order to compare the fit of different cognate evolution models to the data in combination with the strict clock. The best fit was shown by the covarion model. Below I report the results for this model only.

Figure 9.1 shows the DensiTree representation of tree topologies for the Turkic family in the posterior probability distribution. The DensiTree program provides an overview of the areas that agree with each other along with the areas of topological uncertainty (Bouckaert 2010).

<Insert Figure 9.1 here>

Figure 9.1 A DensiTree for the Turkic family

Figure 9.2 shows the maximum clade credibility tree as produced by the TreeAnnotator tool using the mean heights option. The node labels show the posterior probability of the given node, that is, the number of trees supporting this node as weighted to the total number of trees. The scale axis below is a time scale, with "1" for one thousand years.

<Insert Figure 9.2 here>

Figure 9.2 The maximum credibility tree for the Turkic family

The conflicting signal in DensiTree is reflected in low posterior probabilities as presented in the maximum credibility tree. Basically, such nodes correlate with the cases where the historical signal remained weak due to undetected borrowings, or because of lower-quality data. In the obtained tree, the nodes with a lower posterior probability are: the position of Saryg Yugur as the first separated

branch of South Siberian Turkic (0.58); the node linking the Oghuz languages to the Kipchak and Karluk languages (0.62); and the node linking Khakas and Shor together and leaving Middle Chulym as an outlier (0.75). For the other 28 nodes in the tree, the posterior probability is higher than 0.8, and in most cases it is rather close or equal to 1.

With regard to topology, the obtained tree divides the modern Turkic languages into six principle sub-branches (in the order of their divergence): Bulgharic, North Siberian, South Siberian, Khalaj-Salar, Oghuz, and Kipchak-Karluk ("Macro-Kipchak"). The time-depth of the Turkic family on the maximum credibility tree is estimated to be around 2119 years BP, with a 95 % highest posterior density between 1541 and 2793 years BP. The topology and age of the obtained tree are discussed in detail below in Section 9.4.

9.4 Discussion

9.4.1 Topology

In general, the obtained tree structure seems to be quite compatible with the contemporary understanding of the Turkic linguistic history, as presented in (Johanson, this volume: Chapter 8). The early split between the Bulgharic branch and the Common Turkic languages shapes the Turkic family as having a clear-cut binary structure. This agrees with most of the previous classifications of the Turkic family, be they based on the arbitrary method of shared innovations or a distance-based quantitative approach (Menges 1995: 60–61; Johanson 1998a: 81–83; this volume, Chapter 8; Dybo 2006: 766–817, 2013: 18; Mudrak 2009: 172–179). There is no support for the original division between "Eastern" (Karluk and Siberian Turkic) and "Western" (Bulgharic, Kipchak and Oghuz) Turkic languages as proposed by Baskakov (1960: 228–229; 1981: 18–20).

In accordance with most of the contemporary classifications, the obtained tree does not support "Siberian Turkic" as a valid genealogical node. Instead, it nominates the North Siberian (Yakut–Dolgan) branch as the second earliest offshoot from the Turkic family and the earliest breakaway group in Common Turkic. Other Siberian Turkic languages form a separate group comprising the

following subgroups, in the order of branching: Saryg Yugur, Altay (including North and South Altay languages), Sayan (Tuvan–Tofa) and Khakassic (Khakas, Shor, and Middle Chulym). Thus, our model supports the hypothesis that the South Siberian Turkic languages evolved from a common ancestor rather than due to areal convergence. At the same time, the exact relations between individual South Siberian languages and language clusters remain somewhat controversial. In particular, the status of Saryg Yugur as the first language to diverge from the lineage has a low posterior probability and contradicts the widely accepted view on Saryg Yugur as having a Khakassic origin. Its current position in the tree can be provisionally attributed to extensive—and not always clearly detectable—lexical influence from the Karluk branch of Turkic it has strong areal connections to (Johanson 1998a: 83; Johanson, this volume: Chapter 8). Another point of interest is the relationship between Middle Chulym, Khakas, and Shor. Middle Chulym shares many phonological isoglosses with Shor and is sometimes regarded as its dialect. However, the analysis of basic vocabularies supports the conclusions by Mudrak (2009), who assumes, based on archaic morphological isoglosses, that Khakas and Shor are more closely related to each other than to Middle Chulym. This controversy, which may point to a kind of dialect continuum between the three languages, is reflected in the relatively low statistical support for the Khakas–Shor node and the conflicting signal linking Shor to Middle Chulym.

Old Turkic appears in the tree as a possible ancestor of all Turkic languages except for Chuvash and Yakut–Dolgan. The other extinct language, Cuman, forms its own sub-branch as the model does not reflect its specific relatedness to the Kipchak languages.

The next sub-branch is represented by the two idiosyncratic languages that are often discussed in connection with the Oghuz languages, Khalaj (Ščerbak 1997: 471; Dybo 2016: 87) and Salar (Dwyer 2017). Tekin (1990) assumed that both languages should be considered as branch-level isolates in the Turkic family. From a historical linguistic perspective, the presumed link between Khalaj and Salar could probably be explained by the fact that Salar is a language of Oghuz origin with significant Karluk elements in its basic vocabulary (Dwyer 2017), while Khalaj can be

considered a language of Karluk origin (Mudrak 2009) whose basic vocabulary includes some Oghuz (South Azeri) elements (Doerfer and Tezcan 1980). In both cases, some of the borrowings in basic vocabularies may have remained unrecognized due to phonological similarities between Karluk and Oghuz languages. From a phylogenetic viewpoint, the link between Khalaj and Salar could be also attributed to the fact that these two languages do not exhibit the innovations that the other related languages share because of unequal data coverage.

The (Core) Oghuz branch appears as a clear-cut grouping, which stands apart from the Karluk-Kipchak ("Macro–Kipchak") cluster, although the posterior probability for the node linking these two branches together is rather low. In line with most of the contemporary classifications, the model argues for Turkmen as an early offshoot of the Oghuz branch, representing "East Oghuz" in Johanson's terminology, whilst Azeri, Turkish, and Gagauz form the "West Oghuz" node. The position of Crimean Tatar among the Oghuz languages is expected to be controversial in view of its heterogeneous origin as a mix of both Oghuz and Kipchak dialects along with drastic sociolinguistic consequences of the Crimean Tatar's deportation (1944) and return (since 1989 onwards). Yet, based on the recent dictionary of standard Crimean Tatar (Useinov 2007) and additional materials available, the modern Crimean Tatar should be clearly classified as an Oghuz variety that separated from the lineage following Turkmen.

The last major node in the tree brings together what is traditionally labeled as Karluk (Southeastern) and Kipchak (Northwestern) languages. It turns out that South Kipchak (Kazakh, Karakalpak, Kirghiz, and Nogai) and North Kipchak, or Volga-Ural Kipchak (Kazan Tatar, Bashkir, and Siberian Tatar represented by the Baraba variety) languages share more similarities in basic vocabulary with the Karluk branch than with the West Kipchak languages (Kumyk, Karachay-Balkar, and more remotely related Karaim). Thus, the two Karluk languages, Uzbek and Modern Uyghur, appear in the tree as a part of the larger "Macro-Kipchak" node. This contradicts the almost universally accepted view on Kipchak and Karluk as clear-cut sister branches and, at the same time, replicates the tree topology achieved by a distance-based comparison of Turkic 110-Swadesh lists

(Dybo 2006). That is, different quantitative approaches and datasets agree that the two branches are not distinguishable based solely on the evidence from basic vocabulary. The question remains open as to what extent the current structure is caused by transmission of basic vocabulary items among "standard Turkic" languages, which may be difficult to detect based on purely phonological or morphological grounds. The original relationships between the Karluk and Kipchak languages might be obscured, *inter alia*, by two factors. First, there are well-known areal relationships between the West Kipchak and Oghuz languages (note the label "Kipchak-Oghuz" used for the West Kipchak languages by Baskakov (1952: 127–128)). It can be argued that some undected Oghuz words entered the West Kipchak basic vocabulary lists and pulled them off the other Kipchak languages. Second, many Kipchak and Karluk languages shared the same literary tradition (Chagatai) since the late medieval period and until very recently. It has caused extensive lexical convergence, which might affect the domain of basic vocabulary as well.

At the lowest level, the sub-branches of Kipchak are structured in a predictable way. One caveat should be made with regard to the position of Kirghiz. This language shares some important isoglosses, mainly phonological and morphological, with the South Siberian Turkic languages, and particularly with the dialects of Altay. These isoglosses are usually interpreted as pointing to a South Siberian origin of Kirghiz (e.g., Mudrak 2009: 179–180; Johanson, this volume: Chapter 8). While this interpretation seems to be correct, it is clear that, based on the evidence from basic vocabulary, the modern Kirghiz should be placed among South Kipchak languages, along with Kazakh, Karakalpak, and more remotely related Nogai. Another point of interest is the position of Baraba Tatar among the North Kipchak languages, Kazan Tatar and Bashkir. The model supports Mudrak's interpretation (2009: 177), according to which Siberian Tatar dialects have a Volga Kipchak origin, while Northeastern Turkic elements in their structures are due to contact phenomena.

9.4.2 Time-depth

The age of 2119 years BP lies within the bounds traditionally discussed as the probable time-depth of the Turkic family (2000–2500 years BP). In general, this figure is quite compatible with the traditional association of the Proto-Turkic speakers, as well as their immediate descendants (primarily, Proto-Bulgharic groups) with the Xiongnu of Old Chinese sources (Robbeets et al., this volume: Chapter 43). Accurate dating for more shallow nodes is far less approachable due to the limits of the tested strict clock model; therefore, divergence dates for the sub-branches of Common Turkic remain beyond the scope of this chapter.

9.5 Conclusion

In this chapter, I presented results of a study on the internal structure of the Turkic language family, using the Bayesian phylogenetic approach. The tested covarion model of cognate evolution, combined with a strict clock model, replicates most of the traditional groupings in the Turkic family. It also points in a plausible way to the original genealogical affiliation for a number of Turkic languages that have been severely affected by intra-family contact phenomena. The overall dating of the primary split in the Turkic family is consistent with the previous chronological estimates obtained by quantitative linguistic studies and supported by general historical considerations. Detailed dating of individual nodes in the family requires further elaboration.

Several avenues can be taken for the study's future development. There are still Turkic varieties that are rather poorly documented, and acquiring more information on their basic vocabularies would definitely improve the quality of the dataset. Some peculiar dialects that are not covered in this study might help in clarifying the nodes with a less reliable status. Further work on contact relations of the Turkic languages would allow to exclude the borrowings that so far remain undetected in order to get a clearer signal in controversial cases. Last but not least, it seems to be promising to explore in detail some of the more sophisticated models of Bayesian analysis, including adaptation of the relaxed clock model to the Turkic phylogenetic data, that might give a better fit to the data.

Acknowledgements

The research leading to these results has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 646612) granted to Martine Robbeets. I am very grateful to Remco Bouckaert, Simon Greenhill, and Johann-Mattis List for their invaluable methodological support, which made this study possible. I have also greatly benefited from the Bayesian mini-school that was held in Jena, Germany, on November 6–9, 2017, and would like to thank the tutors Ezequiel Koile, Nataliia Hübler and Annemarie Verkerk.

Appendix 1. 254-basic vocabulary list

195-LJ list items (in order from most basic to least basic): 'fire', 'nose', 'to go', 'water', 'mouth', 'tongue', 'blood', 'bone', 2SG personal pronoun ('thou'), 'root', 'come', 'breast / chest', 'rain', 1SG personal pronoun ('I'), 'name', 'louse', 'wing', 'meat', 'arm / hand', 'fly' (n.), 'night', 'ear', 'neck', 'far', 'to do / make', 'house', 'stone', 'bitter', 'to say', 'tooth', 'hair', 'big', 'one', 'who?', 3SG personal pronoun ('he / she / it'), 'to hit / beat', 'leg / foot', 'horn', proximal demonstrative ('this'), 'fish', 'yesterday', 'to drink', 'black', 'navel', 'to stand', 'to bite', 'back' (n.), 'wind', 'smoke', 'what?', 'child' (as a kin term), 'egg', 'to give', 'new', 'to burn' (intransitive), verbal negation (in indicative) ('not'), 'good', 'to know', 'knee', 'sand', 'laugh', 'to hear', 'soil / earth', 'leaf', 'red', 'liver', 'to hide' (transitive), 'skin / hide', 'to suck', 'to carry', 'ant', 'heavy', 'to take', 'old', 'to eat', 'thigh', 'thick', 'long', 'to blow', 'wood', 'to run', 'to fall', 'eye', 'ash', 'tail', 'dog', 'to cry / weep', 'to tie', 'to see', 'sweet', 'rope', 'shade / shadow', 'bird', 'salt', 'small', 'wide', 'star', 'in (≈ inside)', 'hard', 'to crush / grind', 'mountain', 'to sit', 'fingernail', 'to throw', 'three', 'right', 'to wash', 'to grasp', 'branch', 'man', 'raw', 'tomorrow', 'two', 'bottom', 'to lie (down)', 'snake', 'cloud', 'year', 'tear', 'to ask', 'to weave', 'at' (≈ locative), 'edge', 'chin', 'to play', 'cheek', 'pus', 'to fly', 'hole', 'to grow', 'head', 'belly', 'shoulder', 'claw', 'which?', 'to dig', 'to

pull’, ‘hot’, ‘firewood’, ‘to remain’, ‘cold’, ‘feather’, ‘to cough’, ‘thin’, ‘grass’, ‘foam’, ‘sour’, ‘full’, ‘day’, ‘sleep’, ‘month’, ‘white’, ‘to sew’, ‘to kill’, ‘to jump’, ‘throat’, ‘woods / forest’, ‘there’, ‘to find’, ‘to flow’, ‘many’, ‘to chew’, ‘to swallow’, ‘wet’, ‘four’, ‘soft’, ‘to look’, ‘nasal mucus’, ‘that’, ‘to cut’, ‘mother’, ‘to scratch’, ‘sun’, ‘to look for’, ‘brain’, ‘warm’, ‘to cover’, ‘woman’, ‘deep’, ‘above’, ‘female (of an animal)’, ‘to put on’, ‘other’, ‘forehead’, ‘left’, ‘to rise’, ‘dry’, ‘how?’, ‘to break’, ‘where?’, ‘to spin’, ‘to ripe’, ‘to lick’, ‘to open’, ‘tall’.

Additional Jena list items (in alphabetical order): 'bad’, ‘bark’, ‘to breathe’, ‘to count’, ‘to die’, ‘dirty’, ‘dust’, ‘fat’ (n.), ‘father’, ‘to fear’, ‘to fight’, ‘five’, ‘flower’, ‘fog’, ‘to freeze’, ‘fruit’, ‘green’, ‘guts’, ‘heart’, ‘here’, ‘to hunt’, ‘ice’, ‘lake’, ‘to live’, ‘moon’, ‘narrow’, ‘near’, ‘person’, ‘to push’, ‘river’, ‘to be(come) rotten’, ‘round’, ‘sea’, ‘seed’, ‘sharp’, ‘short’, ‘to sing’, ‘sky’, ‘to smell’ (intransitive inactive), ‘smooth’, ‘snow’, ‘to spit’, ‘stick’, ‘straight’, ‘to swell’, ‘to swim’, ‘they’, ‘to think’, ‘tree’, ‘true’, ‘to turn’ (transitive), ‘to vomit’, ‘to walk’, 1 PL personal pronoun (‘we’), ‘when?’, ‘with’ (comitative), ‘worm’, ‘yellow’, 2 PL personal pronoun (‘you’).

Appendix 2. Sources of basic vocabulary items used in the study

1. Comparative etymological dictionaries

Dybo, Anna V. (2013). *Etimologičeskij slovar’ bazisnoj leksiki tjurkskix jazykov* [An etymological dictionary of Turkic basic vocabularies]. (Etimologičeskij slovar’ tjurkskix jazykov 9.) Astana: TOO “Prosper Print”.

Sevortjan, Ervand V., et al. (1974–2003). *Etimologiceskij slovar’ tjurkskich jazykov* [An etymological dictionary of the Turkic languages]. Moskva: Nauka.

Starostin, Sergej A., Anna V. Dybo, and Oleg A. Mudrak (2003). *Etymological Dictionary of the Altayc Languages*. Leiden: Brill.

Tenišev, Edxäm R., et al. (2001). *Sravnitel’no-istoričeskaja grammatika tjurkskix jazykov. Leksika* [A historical comparative grammar of the Turkic languages. Lexicon]. Moscow: Nauka.

2. Azeri

Gusejnov, Gejdar N. (1941). *Azerbajdzhansko-russkij slovar'* [Azeri-Russian Dictionary]. Baku: 1941

Širəliyev, Məmmədağa Š. (1951). *Rusca-azərbaycanca-lüğət* [Russian-Azeri dictionary]. Baku: Azerbaycan SSR elmlər Akademiyası nəşriyatı.

Tağiyev, Məmməd T., et al. (2006). *Azərbaycanca-rusca lüğət* [Azeri-Russian dictionary]. Baku: 2006.

3. Baraba Tatar

Dmitrijeva, Ljudmila V. (1981). *Jazyk barabinskix tatar* [Baraba Tatar]. Leningrad: 1981

4. Bashkir

Uraksin, Zinnur G. (1996). *Baškirsko-russkij slovar'* [Bashkir-Russian dictionary]. Moscow: Digora, Russkij jazyk.

Uraksin, Zinnur G. (2005). *Russko-baškirskij slovar'* [Russian-Bashkir Dictionary]. Ufa: Baškirskaja enciklopedija.

5. Chuvash

Ašmarin, Nikolaj I. (1928–1950). *Thesaurus Linguae Tschuvaschorum*. Vol. 1–17. Kazan: Krasnyj Pečatnik.

Savelyev, Alexander (2011–2015). *Fieldnotes on Chuvash*.

6. Crimean Tatar

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

7. Cuman

Codex Cumanicus bibliothecæ ad templum Divi Marci Venetiarum / primum ex integro ed. prolegomenis notis et compluribus glossariis instruxit Comes Géza Kuun. B.udapestini : Scient. Acad. Hung., 1880.

8. Dolgan

Aksenova, Jevdokija, et al. (1992) *Slovar' dolgansko-russkij i russko-dolganskij* [Dolgan-Russian and Russian-Dolgan Dictionary]. Sankt-Peterburg: Prosveščenije.

Stachowski, Marek (1993). *Dolganischer Wortschatz*. Kraków: Uniwersytet Jagielloński.

9. Saryg Yugur

Malov, Sergei Je. (1957). *Jazyk želtyx ujgurov* [The language of the Yellow Uyghurs]. Alma-Ata: Izdatel'stvo Akademii nauk Kazaxskoj SSR.

Tenišev, Edxäm R., and Buljaš X. Todajeva (1966). *Jazyk želtyx ujgurov* [The language of the Yellow Uyghurs]. Moscow: Nauka.

10. Gagauz

Baskakov, Nikolaj A. (1973). *Gagauzsko-russko-moldavskij slovar'* [Gagauz-Russian-Moldovan dictionary]. Moscow: Sovetskaja enciklopedija.

Radova-Karanastas, Olga, and Stepan Kuroglo (2016). Rusça-gagauzça sözleşmäk kiyadı [Russian-Gagauz phrasebook]. Komrat: Gagauziya M.V. Maruneviç adına Bilim-aaraştırma merkezi.

11. Karačay-Balkar

Tenišev, Edhjam R., and Sujunčev, Xanafij I. (1989). *Karačajevo-balkarsko-russkij slovar'* [Karachay-Balkar-Russian dictionary]. Moscow: Russkij jazyk.

Sujunčev, Xanafij I., and Ibragim X.-M. Urusbijev (1965). *Russko-karačajevo-balkarskij slovar'*

[Russian-Karachay-Balkar dictionary]. Moscow: Sovetskaja enciklopedija.

12. Karaim

Baskakov, Nikolaj A., Ananiasz Zajaczkowski, and Seraja Szapszał (1974). *Karaimsko-russko-pol'skij slovar'* [Karaim-Russian-Polish dictionary]. Moscow: Russkij jazyk.

13. Karakalpak

Baskakov, Nikolaj A. (1958). *Karakalpaksko-russkij slovar'* [Karakalpak-Russian dictionary]. Moscow: Gosudarstvennoe izdatel'stvo inostrannyx i nacional'nyx slovarej.

Baskakov, Nikolaj A. (1967). *Russko-karakalpakskij slovar'* [Russian-Karakalpak dictionary]. Moscow: Sovetskaja enciklopedija.

14. Kazakh

Bektaev, Qaldybaj (1995). *Bol'šoj kazaxsko-russkij i russko-kazaxskij slovar'* [Big Kazakh-Russian and Russian-Kazakh dictionary]. Almaty: Kazaxstanskij projekt razvitija gosudarstvennogo jazyka.

Bekturov, Šabken, and Ardak Bekturova (2001). *Kazaxsko-russkij slovar'* [Kazakh-Russian dictionary]. Astana: Foliant.

Sauranbaev, Nigmet T. (1954). *Russko-kazaxskij slovar'* [Russian-Kazakh dictionary]. Moscow: Gosudarstvennoe izdatel'stvo inostrannyx i nacional'nyx slovarej.

15. Khakas

Baskakov, Nikolaj A. (1953). *Xakassko-russkij slovar'* [Khakas-Russian dictionary]. Moscow: Gosudarstvennoe izdatel'stvo inostrannyx i nacional'nyx slovarej.

Čankov, Dmitrij I. (1961). *Russko-xakasskij slovar'* [Russian-Khakas dictionary]. Moscow: Gosudarstvennoe izdatel'stvo inostrannyx i nacional'nyx slovarej.

Subrakova, Ol'ga V. (2006). *Xakassko-russkij slovar'* [Khakas-Russian dictionary]. Novosibirsk:

Nauka.

16. Khalaj

Doerfer, Gerhard, and Semih Tezcan (1980). *Wörterbuch des Chaladsch (dialekt von Charrab)*. Budapest: Akadémiai Kiadó.

17. Kirghiz

Judakhin, Konstantin K. (1985). *Kirgizsko-russkij slovar'* [Kirghiz-Russian dictionary]. 2 vol. Moscow: Sovetskaja enciklopedija.

Judakhin, Konstantin K. (1957). *Russko-kirgizskij slovar'* [Russian-Kirghiz dictionary]. Moscow: Moscow: Gosudarstvennoe izdatel'stvo inostrannyx i nacional'nyx slovarej.

18. Kumyk

Adžijev, Abdulakim M., and Kalsyn S. Kadyradžijev (1992). *Rusča-qumuqča qïlawuz* [Russian-Kumyk phrasebook]. Makhachkala: Institut jazyka, literatury i iskusstva imeni Gamzata Cadasy DNC RAN.

Bammatov, Zajnal Z. (1960). *Russko-kumykskij slovar'* [Russian-Kumyk dictionary]. Moscow: Gosudarstvennoe izdatel'stvo inostrannyx i nacional'nyx slovarej.

Bammatov, Zajnal Z. (1969). *Kumyksko-russkij slovar'* [Kumyk-Russian dictionary]. Moscow: Sovetskaja enciklopedija.

19. Middle Chulym

Savelyev, Alexander (2015). *Fieldnotes on Chulym*.

20. Nogai

Baskakov, Nikolaj A. (1963). *Nogajsko-russkij slovar'* [Nogai-Russian dictionary]. Moscow:

Gosudarstvennoe izdatel’stvo inostrannyx i nacional’nyx slovarej.

Baskakov, Nikolaj A. (1956). *Russko-nogajskij slovar’* [Russian-Nogai dictionary]. Moscow: Gosudarstvennoe izdatel’stvo inostrannyx i nacional’nyx slovarej.

Kapajev, Isa S., and Kel’dikhan I. Kumratova (2007). *Russko-nogajskij razgovornik* [Russian-Nogai phrasebook]. Stavropol’: Jurkit.

21. North Altay (Chelkan)

Baskakov, Nikolaj A. (1985). *Dialekt lebedinskix tatar-čalkancev (kuu-kiži). Grammatičeskij očerk, teksty, perevody, slovar’* [The dialect of Chelkan (Quu-Kiži). Grammar sketch, texts, translations, dictionary]. Moscow: Nauka.

22. Old Turkic

Clauson, Gerard (1972). *An etymological dictionary of pre-thirteenth-century Turkish*. Oxford: Clarendon Press.

23. Salar

Tenišev, Edxäm R. (1976). *Stroj salarskogo jazyka* [The system of Salar]. Moscow: Nauka.

24. Shor

Kurpeško-Tannagaševa, Nadežda N., and Fedor Ja. Apon'kin (1993). *Šorsko-russkij i russko-šorskij slovar’* [Shor-Russian and Russian-Shor dictionary]. Kemerovo: Kemerovskoje knižnoje izdatel’stvo.

25. South Altay (Altay-Kiži)

Baskakov, Nikolaj A., and Taisija M. Toščakova (1947). *Ojrotsko-russkij slovar’* [An Oyrot-Russian dictionary]. Moscow: Gosudarstvennoje izdatel’stvo inostrannyx i nacional’nyx slovarej.

26. Tatar

Dmitrijev, Nikolaj K., et al. (1955). *Russko-tatarskij slovar'* [Russian-Tatar dictionary]. Kazan: Tatknigoizdat.

Osmanov, M. M., et al. (1966). *Tatarsko-russkij slovar'* [Tatar-Russian dictionary]. Moscow: Sovetskaja enciklopedija.

27. Tofa

Rassadin, Valentin I. (1971). *Fonetika i leksika tofalarskogo jazyka* [The phonology and vocabulary of Tofa]. Ulan-Ude: Burjatskoje knižnoje izdatel'stvo.

Rassadin, Valentin I. (2005). *Tofalarsko-russkij i russko-tofalarskij slovar'* [Tofa-Russian and Russian-Tofa dictionary]. Saint-Petersburg: Drofa.

28. Turkish

Baskakov, Nikolaj A. (1977). *Turecko-russkij slovar'* [Turkish-Russian dictionary]. Moscow: Russkij jazyk.

Iz, Fahiz, and Henry C. Hony. (1992). T*he Oxford Turkish dictionary*. Oxford: Oxford University Press.

29. Turkmen

Baskakov, Nikolaj A. (1968). *Turkmensko-russkij slovar'* [Turkmen-Russian dictionary]. Moscow: Sovetskaja enciklopedija.

Khamzajev, Mašan Ja., and S. Altajev (1968). *Kratkij russko-turkmenskij slovar'* [A concise Russian-Turkmen dictionary]. Ashgabat: Ylym.

30. Tuvan

Harrison, K. David, and Gregory D. S. Anderson (2002). *Tuvan-English, English-Tuvan Dictionary*. Kyzyl: Tipografija Goskomiteta po pečati i informacii RT.

Monguš, Dorug-ool A. (1988). *Russko-tuvinskij učebnyj slovar'* [Learner's Russian-Tuvan dictionary]. Moscow: Russkij jazyk.

Tenišev, Edhjam R. (1968). *Tuvinsko-russkij slovar'* [Tuvan-Russian dictionary]. Moscow: Sovetskaja enciklopedija.

31. Uyghur

Nadžip, Emir N. (1968). *Ujgursko-russkij slovar'* [Uyghur-Russian dictionary]. Moscow: Sovetskaja enciklopedija.

32. Uzbek

Awde, Nicholas, William Dirks, and Umida Hikmatullajeva (2002). *Uzbek-English / English-Uzbek Dictionary and Phrasebook: Romanized*. New York: Hippocrene Books.

Kary-Nijazov, Tašmuxamed N., and Aleksandr K. Borovkov (1941). *Uzbeksko-russkij slovar'* [Uzbek-Russian dictionary]. Taškent: Izdatel'stvo Uzbekistanskogo filiala akademii nauk SSSR.

Koščanov, Mat'akub K., et al. (1983–1984). *Russko-uzbekskij slovar'* [Russian-Uzbek dictionary]. Taškent: Glavnaja redakcija Uzbekskoj Sovetskoj Enciklopedii.

33. Yakut

Afanasjev, Pjotr S., and Luka N. Xaritonov (1968). *Russko-jakutskij slovar'* [Russian-Yakut dictionary]. Moscow: Sovetskaja enciklopedija.

Pekarskij, Eduard K. (1907–1930) *Slovar' jakutskogo jazyka* [A dictionary of Yakut]. S.-Peterburg.

Sleptsov, Pjotr A. (1972). *Jakutsko-russkij slovar'* [Yakut-Russian dictionary]. Moscow: Sovetskaja enciklopedija.

This is a draft version of a chapter that appears in Robbeets, M. and A. Savelyev (eds). *The Oxford Guide to the Transeurasian Languages* (OUP, 2020),
see https://global.oup.com/academic/product/the-oxford-guide-to-the-transeurasian-languages-9780198804628. The research leading to these results has received funding from the European Research Council under the Horizon 2020 Program/ ERC Grant Agreement n. 646612 granted to Martine Robbeets.