Journal of Language Evolution, 2020, 39–53
doi: 10.1093/jole/lzz010
Advance Access Publication Date: 14 February 2020
Research article

Bayesian phylolinguistics infers the internal structure and the time-depth of the Turkic language family

Alexander Savelyev *,†,‡ **and Martine Robbeets**†

†Eurasia3angle, Max Planck Institute for the Science of Human History, Kahlaische Strasse 10, 07745 Jena, Germany and ‡Institute of Linguistics, Russian Academy of Sciences, Bolshoy Kislovsky pereulok 1/1, 125009 Moscow, Russia

*Corresponding author: a.savelyev@iling-ran.ru

Abstract

Despite more than 200 years of research, the internal structure of the Turkic language family remains subject to debate. Classifications of Turkic so far are based on both classical historical–comparative linguistic and distance-based quantitative approaches. Although these studies yield an internal structure of the Turkic family, they cannot give us an understanding of the statistical robustness of the proposed branches, nor are they capable of reliably inferring absolute divergence dates, without assuming constant rates of change. Here we use computational Bayesian phylogenetic methods to build a phylogeny of the Turkic languages, express the reliability of the proposed branches in terms of probability, and estimate the time-depth of the family within credibility intervals. To this end, we collect a new dataset of 254 basic vocabulary items for thirty-two Turkic language varieties based on the recently introduced Leipzig–Jakarta list. Our application of Bayesian phylogenetic inference on lexical data of the Turkic languages is unprecedented. The resulting phylogenetic tree supports a binary structure for Turkic and replicates most of the conventional sub-branches in the Common Turkic branch. We calculate the robustness of the inferences for subgroups and individual languages whose position in the tree seems to be debatable. We infer the time-depth of the Turkic family at around 2100 years before present, thus providing a reliable quantitative basis for previous estimates based on classical historical linguistics and lexicostatistics.

Key words: Turkic languages; genealogical classification; Proto-Turkic; Bayesian phylogenetic linguistics.

1. Introduction

The Turkic family is represented by several dozen languages spoken in a vast area stretching from Northeast Siberia and North China in the east to the Mediterranean, Poland, and Lithuania in the west. Figure 1 shows the distribution of the thirty contemporary Turkic languages and dialects underlying our study.

From a sociolinguistic perspective, the Turkic languages range from those spoken by tens of millions of speakers and having an official status in sovereign states, such as Turkish, Azeri, and Uzbek, to almost extinct languages with only a few speakers, such as Middle Chulym in Southern Siberia. Many of the Turkic languages are to some extent mutually intelligible due to a shared cultural history and—what is more crucial—a relatively shallow time-depth of most nodes in the family. It is traditionally believed that the primary split in Turkic should have taken place somewhere between the

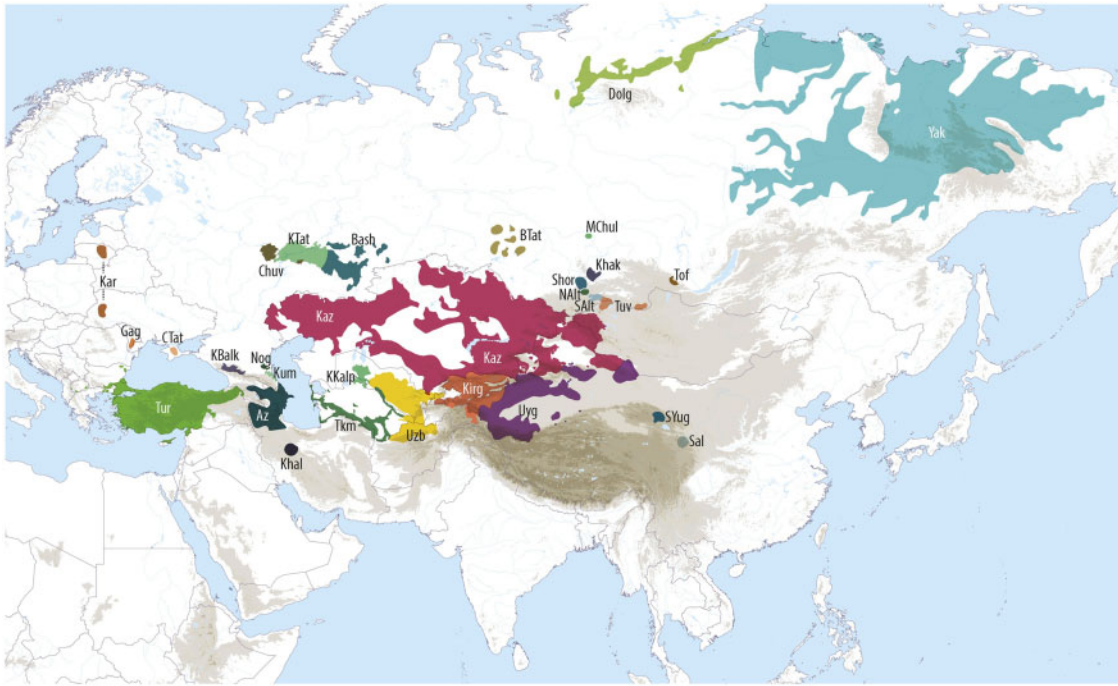

Figure 1 The distribution of the Turkic languages. (Abbreviations for languages are explained in the list of abbreviations).

fifth century BC and the turn of the eras (Róna-Tas 1998: 68; Mudrak 2009), which puts the family on approximately the same chronological level as, for example, Germanic. The high level of internal diversity (around forty different doculects that are available to historical–comparative analysis) and the wide distribution of the Turkic family as compared to its relatively shallow time-depth may be attributed to the fact that most Turkic-speaking populations practiced nomadic pastoralism from the Proto-Turkic period until recent times (Golden 1998). This nomadic lifestyle may have contributed to continuous mutual contact between different branches and sub-branches of Turkic throughout their history, a fact that should be taken into account when inferring the internal structure of the Turkic family.

Another caveat is that most of the Turkic languages are rather poorly attested in historical sources as compared to other language families such as Indo-European. The varieties attested in Old Turkic inscriptions and Middle Turkic texts from the early and late medieval periods are not particularly diverse against the overall background of the Turkic language family. There is a considerable lack of written attestations for the most divergent Turkic languages, such as Yakut and Chuvash—the languages that play a crucial role in reconstructing the early history of the Turkic family.

In sum, due to convergence phenomena and a lack of ancient written sources, the internal classification of the Turkic language family remains a challenging task, even after 200 years of research. Aharon Dolgopolsky used the expression 'the genealogical stump of the Turkic languages' (as witnessed by Belikov 2009: 53) to indicate the difficulty in imposing a tree-like classification on them. Many traditional approaches tried to bypass this problem by combining genealogical and areal criteria in setting up a classification of the Turkic languages in a somewhat arbitrary way (Bogorodickij 1934: 6–13; Benzing 1959: 2–5; Johanson 1998).

In this article, we intend to provide a more methodic, quantitative basis to classify the Turkic languages by applying Bayesian analysis to lexical data from thirty-two Turkic languages in order to trace their genealogical relationships and estimate the time-depth of the family. In the following section, we search for points of consensus and define the unresolved issues in previous classifications. In Section 3, we introduce our dataset and describe our principles of data selection and coding. After explaining our Bayesian approach in Section 4, we present our results in Section 5. This is followed by a

discussion of the tree topology, the time-depth of the Turkic language family, and the relevance of our study for the field of Turkic historical linguistics in Section 6. Finally, we conclude our article in Section 7.

2. Unresolved issues in previous classifications

The overall number of classifications of the Turkic languages was estimated to be around sixty by Doerfer (1978), and now it is likely to be close to one hundred. Among the most frequently cited classifications of Turkic are those by Samojlovič (1922), Menges (1959), Baskakov (1960), Tekin (1990), Schönig (1997–1998), and Johanson (1998). For a recent overview of classifications based on classical historical–comparative approaches, we refer to Jankowski (2017). These approaches make use of a character-based method in order to generate trees. More specifically, they rely on the parsimony method, which seeks a tree that explains a dataset by minimizing the number of evolutionary changes required to produce the observed state (Dunn 2015).

There are also a number of quantitative approaches to the classification of Turkic but all are distance-based analyses. Apart from a Levenstein distance analysis by Mennecier et al. (2016), restricted to four languages from the same region (i.e., Karakalpak, Kazakh, Kyrgyz, and Uzbek), these attempts have been mainly based on lexicostatistics. They estimate the relationship between two languages by measuring the amount of difference in shared cognate proportion between them, using the Swadesh list as a source of basic vocabulary concepts in addition to the mathematical algorithm proposed by Starostin (1989). The first classification of this type, presented in a short paper by Djačok (2001), was quite undetailed, classifying all the languages except for Chuvash, North Siberian (Yakut-Dolgan), and Sayan (Tuvan-Tofa) simply as 'West Turkic'. However, Dybo (2006, 2013) further refined the lexicostatistic classification of Turkic, reaching a more detailed subbranching. Mudrak (2009) took a more experimental approach called 'morphostatistics', measuring the amount of difference in shared cognate proportion in the domain of historical morphology and phonology rather than in the basic vocabulary. In general, lexicostatistic approaches resulted in tree structures that replicate most of the branching inferred in classical historical–comparative linguistic approaches.

In the different classifications proposed so far, there is a wide consensus that the earliest split in the family was between the Bulgharic (also known as 'Oghuric') branch, which today only survives in Chuvash, and the Common Turkic branch, which is ancestral to all other contemporary Turkic languages. Three lower level subgroups of Common Turkic, that is, Kipchak (Northwestern Turkic), Oghuz (Southwestern Turkic), and Karluk (Southeastern Turkic), are usually considered uncontroversial, but the status of the fourth subgroup of Common Turkic, the Siberian Turkic (Northeastern Turkic) languages, as a monophyletic group is generally debated. Moreover, the genealogical relations between and within these subgroups remain subject to discussion. For Kipchak, Oghuz, and Karluk, there is some controversy as to whether they are three sister branches or whether any two of them are more closely related to each other than to the third. For Siberian Turkic, it is debated whether the two Southern Siberian sub-branches Sayan Turkic (Tuvan–Tofa) and Khakassic derive from a single 'South Siberian' genealogical node (Johanson 1998: 83). In particular, lexicostatistic studies (Mudrak 2009: 179; Dybo 2013: 18) suggest that the Sayan Turkic branch should, rather, be paired together with the North Siberian Turkic languages, that is, Yakut and Dolgan.

Finally, there are individual languages and dialects of widely dispersed languages such as Old Turkic and Khalaj, whose exact position in the Turkic family tree remains controversial. Old Turkic is the written language underlying three corpora, that is, the runic inscriptions from the eighth to the tenth century CE, the Old Uyghur manuscripts from the ninth to the thirteenth century CE, and the eleventh-century texts from the Karakhanid state. Although it is often referred to as the ancestor of Kipchak, Oghuz, Karluk, and Siberian Turkic, different studies highlight a certain closeness to individual subgroups of Common Turkic, such as Oghuz (Johanson 1998; Dybo 2006), Karluk (Mudrak 2009), or Siberian Turkic (Dybo 2016).

Khalaj is the language of a Turkic minority group in Western Iran. Tekin (1990) and Johanson (1998), following Doerfer (1971), consider it a separate branch of Turkic that separated immediately after the Bulgharic split, while Ščerbak (1997: 471) and Dybo (2016: 87) emphasize its affinities with the Oghuz branch, and Mudrak (2009) interprets it as an early offshoot of the Karluk branch.

Most of the other Turkic languages whose place in the classification is unclear are, or were in the past, spoken in the Siberian Turkic area and seem to be severely affected by secondary convergence phenomena, that is, intra-family horizontal transmission in phylogenetic terms. That is the case for some Siberian Tatar varieties as well as for North Altay and South Altay dialects,

which are characterized by an interaction of both indigenous (Northeastern) and Kipchak (Northwestern) components: Saryg Yugur (West Yugur), a language of South Siberian origin that has been influenced by the Karluk branch; and Kirghiz, a language that shares numerous isoglosses with South Kipchak languages in addition to South Altay dialects.

There are various estimates of the time-depth of the Turkic language family, based on a wide range of approaches, such as combining cultural reconstruction with information from archaeology, contact studies, inferences from historical records, and quantitative methods. Róna-Tas (1998: 68–9) dates the lower limit of Proto-Turkic to around the middle of the first millennium BC, which is based on his interpretation of Proto-Turkic and Proto-Bulgharic contact history. A similar view on the early contact relations of Turkic underlies a less specific estimate by Janhunen (2010: 290), who discusses the period between 2,500 and 2,000 years ago as the most likely time of the first split in the family. Different glottochronological studies on the time-depth of Proto-Turkic have led to very similar results: 120–0 BC according to Mudrak's (2009: 181) calculation based on phonological and morphological isoglosses and 100–0 BC according to Dybo's (2007: 66) application of lexicostatistics. The time estimates inferred from these quantitative methods are used to support the rather controversial idea that the Xiongnu of Old Chinese chronicles were at least in part the speakers of Proto-Turkic.[1] The reasoning is that the primary split in the Turkic family should be associated with the initial stages of the disintegration of the Xiongnu tribal confederation around the first century BC.

In our paper, we aim to resolve some of the pending issues with regard to the classification of Turkic, in addition to providing a time estimate for the root and the nodes in the language tree.

3. Data

3.1 Languages

This study is based on lexical evidence from thirty-two Turkic languages. The contemporary Turkic languages include Azeri, Baraba Tatar (a variety of Siberian Tatar), Bashkir, Crimean Tatar, Dolgan, Chuvash (the Viryal dialect, which is more archaic as compared to standard Chuvash based on the Anatri dialect), Gagauz, Karachay-Balkar, Karaim (the dialects of Halych and Trakai, not including the highly distinct dialect of Crimea), Karakalpak, Kazan Tatar, Kazakh, Khakas (the Kacha dialect), Khalaj, Kirghiz, Kumyk, Middle Chulym, Nogai, North Altay (the Chelkan dialect), Salar, Saryg Yugur (West Yugur, Yellow Uyghur), Shor, South Altay (the Altay-Kiži dialect), Tofa, Turkish, Turkmen, Tuvan, Modern Uyghur, Uzbek, and Yakut (Sakha); see Fig. 1 for their geographical distribution. The historical varieties included in our study are Old Turkic and Cuman. For the sake of uniformity, Old Turkic data are restricted to the evidence of Old Uyghur texts from the ninth century AD. The label 'Cuman' is applied to a Middle Kipchak variety attested in the *Codex Cumanicus* manuscript, dating from the early fourteenth century AD. The languages in the dataset represent all essential groupings of the Turkic family contained in previous studies, whether these are considered to be controversial or not.

3.2 Wordlists

In this study, we compare basic vocabulary lists across the Turkic languages. A basic vocabulary list is a compilation of concepts that are relatively independent of cultural context and available across the languages of the world and, therefore, cross-linguistically, and are particularly resistant to replacement. Words with basic meanings not only tend to resist borrowing more successfully than random lexical items, but they are also more resistant to internal change such as semantic shift or lexical replacement.

Our basic vocabulary list merges the Leipzig–Jakarta 200 list (Haspelmath and Tadmor 2009) with the Jena 200 list (Heggarty and Anderson 2019) and contains 254 different concepts. The underlying Leipzig–Jakarta 200 list is reduced to 195 items due to the merger of a few meanings (e.g. 'breast' and 'chest'). It is supplemented by the Jena 200 list, which is an updated version of the Swadesh 200 list, currently applied to a comparison of Indo-European languages in the CoBL (Cognacy in Basic Lexicon) project by Anderson and Heggarty. Given the large overlap between the two lists, the final list amounts to 254 concepts (available as Appendix 1).

Compared to the traditional Swadesh list, which is mainly based on intuition, the Leipzig–Jakarta list takes a more systematic and empirical approach to the basic vocabulary because it is based on a quantitative comparison of stable words in the languages across the world. The strength of basic vocabulary is not in the stability of a single concept, but in the overall stability of the body of concepts as a whole. Even if certain concepts may appear to be rather unstable as applied specifically to the Turkic language family, the list should be kept intact in order to avoid cherry-picking. Merging previous proposals about the semantic specification of basic

vocabulary, such as Kassian et al. (2010), Dybo (2013), Starostin (2013), and the CoBL documentation, we defined the basic concepts by way of precise descriptions to avoid possible ambiguity. These definitions are given in Supplementary data SI1.

3.3 Sources and data selection

The primary sources from which our Turkic basic vocabulary is extracted are bilingual dictionaries. In cases where these are not available, we use wordlists given in grammar descriptions. For two languages in the sample, Chuvash and Middle Chulym, Savelyev's fieldnotes from 2011 to 2015 and 2015, respectively, were the main source of basic vocabulary. In problematic cases, we examine the occurrence of a particular word in written texts, including textbooks and phrasebooks. Given the geographical distribution of the Turkic languages and the historical context of their documentation, the English translation of our Turkic target data is often mediated by Russian. The full list of the sources underlying our collection of Turkic basic vocabulary is given in Supplementary data SI2.

In order to decide whether a certain Turkic word is indeed the most 'basic' one and thus suitable for inclusion in our dataset, we use the following criteria. In case of synonymy, we prefer the more generic, frequent, stylistically neutral, and morphologically simple terms. Ideally, given the pragmatic aspect of 'basicness', this decision should be based on direct elicitation from informants or comprehensive bilingual corpora, but unfortunately this is currently not feasible for practical reasons.

We deal with cases of synonymy in that we allow more than one word with a certain basic meaning in our dataset unless there is evidence that it is less basic than one of its synonyms. Singletons—that is, words that are present in a given basic meaning only in one language—are removed from the dataset in case they have a non-singleton synonym that fits the criteria for basic status.

Borrowings that can be identified using clear-cut historical–comparative criteria are excluded from the dataset to provide a clearer phylogenetic signal. Prototypical characteristics used for loanword identification include the attestation of a plausible model form outside the Turkic family, the contradiction of regular sound correspondence, morphological complexity of a word in one Turkic language but unsegmentability in the other, and the restriction of shared semantics to a secondarily developed or cultural meaning (Robbeets 2016). Lexical items that cannot be reliably identified as borrowings on the basis of these criteria, even if they lack a plausible Turkic etymology because they are singletons or poorly distributed across the Turkic languages, will nevertheless be preserved in the dataset in order to avoid the loss of relevant data. As singletons and poorly distributed words can represent newly arisen cognate classes, removing them might result in more recent nodes appearing shallower than they are.

As far as the quality of the resulting dataset is concerned, the majority of languages in our sample are well-documented and lexical data are evenly distributed across the family. The average amount of missing data due to gaps in documentation or borrowings in basic vocabulary is around 7%. A few languages in the dataset such as Khalaj, Salar, Baraba Tatar and the ancient Cuman language are undersampled because of their limited documentation. In addition, Khalaj basic vocabulary is drastically influenced by Persian and, to a lesser extent, South Azeri varieties. This is also the case for Chinese borrowings in Salar basic vocabulary. Therefore, the amount of missing data is 30% for Khalaj, 21% for Cuman, 18% for Salar, and 17% for Baraba Tatar.

3.4 Cognate coding

For each word in the dataset, we provide an etymological analysis to establish cognacy classes and exclude borrowings. To this end, we use comparative etymological dictionaries of the Turkic language family (Sevortjan et al. 1974–2003; Tenišev et al. 2001; Dybo 2013; the Turkic part of Starostin, Dybo, and Mudrak 2003) in addition to dictionaries of individual languages that include etymological information, such as Clauson (1972), Stachowski (1993), Fedotov (1996), and Tatarincev (2000–2008). Cognacy classes are established on the basis of regular sound correspondences (available as Supplementary data SI5). Each cognacy class is coded as present (1) or absent (0) for all languages in the dataset. If evidence for inheritance of a given basic vocabulary item is lacking, this is coded as a gap.

In order to process cognates, we made our datasets available in excel format (Supplementary data SI3) as well as in cross-linguistic data formats (CLDF) format (Forkel et al. 2018), using the EDICTOR software tool (List 2017). The resulting matrix of binary characters includes 905 cognacy classes, covering 254 basic vocabulary meanings across thirty-two 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. An example of our coding strategy as illustrated by the terms for 'nose' in Bashkir,

Table 1. Basic terms for 'nose' in Bashkir, Chuvash, South Altai, Turkish, and Tuvan.

Proto-Turkic	Bashkir	Chuvash	South Altay	Turkish	Tuvan	Notes
		sə̂msa				Likely borrowed from Mongolian *samsaɤa* 'wing of nose' (Dybo 2013: 415) and excluded from the dataset.
	tanaw					Likely borrowed from Mongolian *tanaɤa* 'wing of nose' (Dybo 2013: 414) and excluded from the dataset.
***kaŋ-**					*χ*aay	In Tuvan, both words should be regarded as basic.
***dumčuk**			tumčuḳ		dumčuḳ	
***burun**				burun		

Chuvash, South Altai, Turkish, and Tuvan is given in Table 1.

After exclusion of borrowings, we obtain the following character sequences for the three inherited roots: <???> for Bashkir, <???> for Chuvash, <010> for South Altai, <001> for Turkish, and <110> for Tuvan. The same procedure is repeated throughout our dataset; see a wider sample of this coding procedure involving ten basic vocabulary meanings in all thirty-two languages and the overall outcome of our coding in Supplementary data (SI4a and SI4b).

4. Methods

Due to some serious methodological issues, such as the assumption of a constant rate of language change, historical linguists have justly criticized lexicostatistic methods (McMahon and McMahon 2006; Campbell and Poser 2008; Greenhill 2015). Unfortunately, this criticism has led to a certain aversion amongst linguists toward quantitative methods in general. Meanwhile, however, evolutionary biologists have developed new methods for building and dating trees, that can account for rate variation in molecular evolution, using the assumption of a 'molecular clock'. These methodological advances in biology have been applied rather successfully to language, Russell Gray being a pioneer in the field of Bayesian phylolinguistics (Gray and Atkinson 2003; Gray, Drummond, and Greenhill 2009). The Bayesian method is a character-based approach, which seeks to explain a set of observed data by quantifying how likely it is that they have been produced by a certain model of the evolution of cognates along a tree. It is a probabilistic approach that allows the integration of different forms of prior knowledge, such as information about the time-depth of ancient language varieties and uncontroversial nodes in the tree structure. 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 (Bowern and Atkinson 2012; Dunn 2015).

In recent decades, the Bayesian phylogenetic method has been applied to build and date trees for language families, such as Indo-European (Gray and Atkinson 2003), Austronesian (Gray, Drummond, and Greenhill 2009), and Semitic (Kitchen et al. 2009). Although Hruschka et al. (2014) take a Bayesian approach to the establishment of regular sound changes across twenty-six Turkic languages and twenty-three Turkic languages were included in the study of the Transeurasian languages in Robbeets and Bouckaert (2018), Bayesian phylogenetic inference has never been applied to a lexical dataset to build and date a detailed Turkic family tree before.

First, we used SPLITS TREE 4 (Huson and Bryant 2006), a program that calculates NeighborNets and displays split graphs to estimate the tree-likeness of the Turkic phylogeny. We obtained the average delta score = 0.34 and Q-residual score = 0.01, both indicating that the evolution of the Turkic basic vocabulary is reasonably tree-like, and hence suitable for phylogenetic analysis. By way of comparison, the tree-likeness scores calculated for a subset of twelve Indo-European languages (Gray, Bryant and Greenhill 2010) have similar scores as our Turkic languages with an average delta score = 0.23 and Q-residual score = 0.03. The scores range between 0 and 1, but the closer they are to 0, the more tree-like the data are. These observations provide us with confidence that our dataset is sufficiently tree-like and carries a historical signal.

To infer the internal structure of the Turkic family, we applied a Bayesian phylogenetic approach as implemented in BEAST 2.5.1 (Bouckaert et al. 2014). The Bayesian analysis adopted in BEAST uses a Markov

Table 2. Marginal log-likelihoods of simple CTMC, covarion, and SDollo cognate evolution models combined with the strict and relaxed clock models.

Clock	Cognate evolution model	CTMC	Covarion	SDollo
Strict		–8,451	**–8,200**	–10,172
Relaxed		–8,558	–8,482	–10,206

Note: The best fit is indicated in bold.

chain Monte Carlo (MCMC) algorithm to sample the posterior probability distribution of tree topologies. MCMC chains were run for 50 million generations, sampled every 1,000 generations and resulted in a sample of 50,000 trees. A posterior sample of 45,000 trees was left after the first 5 million iterations were discarded as a burn-in. Post-run analysis involving convergence assessment by comparing the posteriors was made using the Tracer v. 1.6 component of the BEAST package (Rambaut et al. 2014). Multiple runs were attempted, reaching convergence by the end of the burn-in period. The estimated sample size (ESS) was well over 200 for the posterior and all the other important parameters, including the prior, the likelihood, and the tree height. We then used the TreeAnnotator tool in BEAST to achieve the maximum clade credibility (MCC) tree.

The analysis was conducted using the Fossilized Birth-Death model, which is most appropriate for data that contain ancient language varieties (Stadler et al. 2018). We tested three models of cognate gain and loss: the simple two-state CTMC model (Gray and Atkinson 2003), the binary covarion model, allowing cognates to be in either a 'fast' or 'slow' state (Gray, Drummond, and Greenhill 2009), and the stochastic Dollo model, which assumes that cognates can be gained once but lost multiple times (Nicholls and Gray 2006). We further tested two different clock models to account for rate variation across branches. The strict clock model of evolution assumes that every branch in the tree evolves according to the same evolutionary rate, while the uncorrelated log-normal relaxed clock allows for variations in rates between branches. All models, including the CTMC, had gamma-distributed rate heterogeneity (four categories).

To calibrate the clock, we relied on the sampling dates (implemented as dated tips) for the two ancient language varieties in the dataset: 1,150 years BP for Old Turkic and 700 years BP for Cuman (the language of *Codex Cumanicus*), with 2,000 taken as the date for the present. No time constraints were put on the shallowest nodes in the Turkic tree as inferring absolute dates for recent events in the Turkic linguistic history was beyond the scope of our research. Given that the principle objective of the study was to verify both conventional and controversial nodes in the Turkic family, no monophyletic constraints were introduced on the branches.

To compare different models and reveal the one that shows the best fit to the data, we estimated marginal log-likelihoods using the path sampling procedure as implemented in BEAST (Baele et al. 2012), with fifty steps and 1 million samples per step.

5. Results

Table 2 summarizes the results from the model selection procedure. The best fit is shown by the covarion model with strict clock (marginal log-likelihood = −8,200). Below we report results for this model only.

Figure 2 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).

Figure 3 presents the MCC as produced by the TreeAnnotator tool using the median 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 obtained subgroups are labeled with the traditional terms if conventional; for the few unconventional groupings, *ad hoc* terminology is used. The scale axis below is a time scale, with '1' for 1,000 years.

The conflicting signal in the 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 missing 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 (0.62); the node linking the Oghuz languages to the Kipchak and Karluk languages (0.71); and the node linking Khakas and Shor together and leaving Middle Chulym as an outlier (0.77). For the other twenty-eight 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 principal subbranches (in the order of their divergence): Bulgharic, North Siberian, South Siberian, Khalaj-Salar, Oghuz, and Kipchak-Karluk ('Macro-Kipchak'). The time-depth

Figure 2. A DensiTree for the Turkic family.

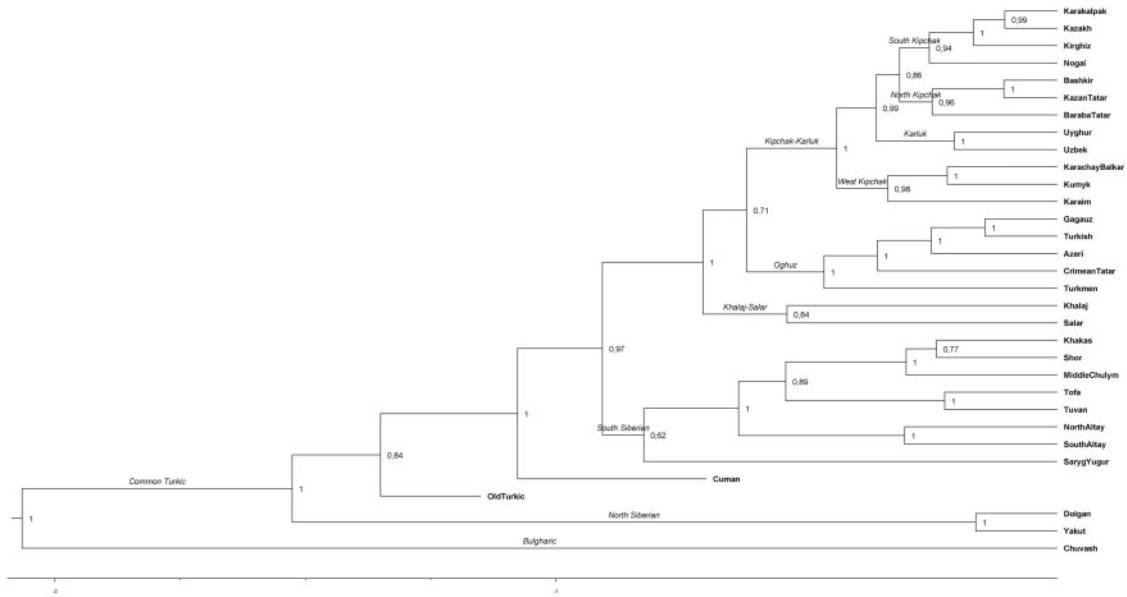

Figure 3. The maximum credibility tree for the Turkic family.

of the Turkic family on the maximum credibility tree is estimated to be around 2,066 years BP (median height of the node), with a 95% highest posterior density between 1,517 and 2,755 years BP. The topology and the age of the obtained tree are discussed in further detail in Section 6.

6. Discussion

6.1 Topology

In general, the obtained tree structure seems to be quite compatible with the contemporary understanding of the Turkic linguistic history as presented in different classifications from the last decades.

The early split between the Bulgharic branch and the Common Turkic languages shapes the Turkic language family as a clear-cut binary structure. This agrees with most of the previous classifications of the Turkic language family, whether they are based on the historical–comparative or lexicostatistic approaches (Tekin 1990: 16; Menges 1995: 60–1; Johanson 1998: 81–3; Dybo 2006: 766–817, 2013: 18; Mudrak 2009: 172–79). 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: 228–9, 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 depicts the North Siberian (Yakut–Dolgan) branch as the second earliest offshoot from the Turkic language family and the earliest breakaway group in Common Turkic. The 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.

The status of Saryg Yugur as a separate—and most divergent—sub-branch of South Siberian Turkic is not universally accepted: it belongs to the Khakassic subgroup according to Baskakov (1952: 132) and Mudrak (2009: 179). On the other hand, its current position in the tree replicates the lexicostatistic results obtained by Dybo (2006: 770–1). The two Altay languages belong to different groups of Turkic according to Baskakov (1952: 134), but our results are again compatible with the cited previous quantitative studies.

Another point of interest is the relationship between Middle Chulym, Khakas, and Shor. Middle Chulym shares numerous phonological isoglosses with Shor and, therefore, can be regarded as the closest relative of the latter. However, on the basis of archaic morphological isoglosses, Mudrak (2009: 179) concluded that Khakas and Shor are more closely related to each other than to Middle Chulym. This contradiction is reflected in the relatively low statistical support for the Khakas–Shor node and the conflicting signal linking Shor to Middle Chulym and may point to a certain dialect continuum between the three languages.

The position of the two ancient varieties, Old Turkic and Cuman in the tree diverges from mainstream thinking, although statistically it is supported by high posterior probabilities (0.84 and 1, respectively). Both languages appear in the tree as branch-level isolates that separated from the tree right after Chuvash and Yakut–Dolgan. Therefore, the alleged close connection of Old Turkic to any specific branch, be it Oghuz, Karluk, or Siberian Turkic is not supported by the data. More surprisingly, the model fails to reveal specific affinities between Cuman and contemporary Kipchak, or the West Kipchak languages in particular, a connection that enjoys broad support among Turkologists (Baskakov 1952; Čečenov 1997: 110).[2]

In relation to the disagreement among Turkologists about the precise position of Khalaj in the Turkic language family, our tree displays a noteworthy 'Khalaj-Salar' node, which recalls a proposal by Tekin (1990: 16–8) to regard both languages as branch-level isolates. The link between Khalaj and Salar may be explained by the observation that Salar is a language of Oghuz origin with significant Karluk elements in its basic vocabulary (Dwyer 2007), while Khalaj can be considered a language of Karluk origin (Mudrak 2009) whose basic vocabulary includes Oghuz (South Azeri) elements (many of which were recognized as such in Doerfer and Tezcan 1980). Thus, in this case, borrowings in the basic vocabulary may have been misinterpreted as cognates. An alternative explanation for the unexpected link between Khalaj and Salar could be unequal data coverage, both languages having significant gaps in the datasets (Section 3.2), potentially preventing common innovations with other related languages from being detected.

The Oghuz branch excluding Salar appears as a clear-cut grouping, which stands apart from the Karluk-Kipchak ('Macro–Kipchak') clade, even if 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 Core Oghuz, representing 'East Oghuz' in Johanson's (1998: 82) terminology, while Azeri, Turkish, and Gagauz form the 'West Oghuz' node.

Our analysis classifies Crimean Tatar as an Oghuz variety that separated from the branch after the separation of Turkmen. Such a result is unexpected as Literary Crimean Tatar is conventionally considered a Kipchak variety, based mainly on phonological and morphological classificatory criteria (Johanson 1998: 82–3). However, Crimean Tatar is heavily influenced by Oghuz Turkic, especially as regards its lexical stock (Berta 1998: 301). Moreover, according to some accounts, one of the Crimean Tatar varieties, the one traditionally labeled as its Southern dialect, should be classified as Oghuz Turkic in origin (Izdinova 1997; Dybo 2002). Interestingly, the author of the recent dictionary of Crimean Tatar (Useinov 2007) that we used

for analysis was born in a village located in the Southern dialect zone, and it is likely that he was a native speaker of the Southern dialect. Therefore, it cannot be excluded that the author's native variety affected severely the selection of lexemes in his dictionary (even if it was conceived as a 'standard' one) and, consequently, the position of Crimean Tatar in our phylogenetic tree.

The last major node in the tree brings together what is traditionally labeled as Karluk (Southeastern Turkic) and Kipchak (Northwestern Turkic) languages. Our analysis indicates 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 general agreement on Kipchak and Karluk as clear-cut sister branches but confirms the tree topology achieved by a lexicostatistic comparison of Turkic 110-Swadesh lists (Dybo 2006; Dybo 2013). This implies that both quantitative approaches cannot distinguish these branches on the basis of basic vocabulary evidence alone. There are two possible reasons for this. First, it could be because of the well-known areal relationships between the West Kipchak and Oghuz languages, reminiscent of Baskakov's (1952: 127–8) use of the label 'Kipchak-Oghuz' for West Kipchak. Under this scenario, some undetected Oghuz loanwords may have entered the West Kipchak basic vocabulary and set it apart from the other Kipchak languages. The second reason may be that many Kipchak and Karluk languages shared the same literary tradition (Khorezmian and Chagatai) from the late medieval period until recently, which may have caused borrowing in the basic vocabulary.

At the lowest level, the sub-branches of Kipchak are structured in a predictable way, but some caution is due 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 South Altay dialects. These isoglosses are usually interpreted as pointing to its South Siberian origin (Mudrak 2009: 179–80). At the same time, evidence from basic vocabulary place Kirghiz among the South Kipchak languages, along with Kazakh, Karakalpak, and the more remotely related Nogai. Another point of interest is the position of Baraba Tatar among the North Kipchak languages, Kazan Tatar and Bashkir. Here, our model supports Mudrak's (2009: 177) interpretation, according to which Siberian Tatar dialects have a Volga Kipchak origin, and that Northeastern Turkic elements in their structures are due to contact phenomena.

Table 3. Dating and maximum credibility intervals for the root and the deepest nodes in the Turkic tree.

Node separating from Turkic	Dating	Lower bound	Upper bound
Bulgharic (the root)	66 BC	483 AD	755 BC
North Siberian Turkic	474 AD	809 AD	7 BC
Old Turkic	650 AD	804 AD	427 AD

In short, although our estimated phylogeny largely confirms the basic tree structure of the Turkic family as proposed in mainstream classifications, we nevertheless find a few instances of poorly supported nodes or unexpected relationships. Among other reasons, these may be explained by undetected borrowings. Although we made significant efforts to eliminate loanwords from our dataset, we only excluded those words that answered to clear-cut phonological, morphological, or semantic criteria that identified them as non-inherited forms. Lexical items, for which a poor distribution across the Turkic languages suggested borrowing but could not be reliably identified as such, were preserved in the dataset. In some cases of intensive language contact, these forms may have accumulated into a residue of unidentified borrowings, skewing the actual genealogical relationships. However, poorly supported nodes or unexpected relationships may also be accounted for by explanations other than borrowing such as cognate attrition from deep nodes, insufficiently observable losses on short nodes and poor data coverage preventing common innovations with other related languages from being detected.

6.2 Dating

Table 3 presents the dating for the root and the earliest nodes in the Turkic family in combination with the maximum credibility intervals.

The age of 2,066 years BP (66 BC) lies within the bounds traditionally discussed as the probable time-depth of the Turkic language family (2,000–2,500 years BP). However, the size of the credible intervals exceeds one thousand years for the root of the family, allowing for any time between 755 BC and 483 AD. The large size of the potential time span for proto-Turkic leaves the current controversy among Turkologists, whether to associate the dominant language spoken during the Xiongnu empire (209 BC–100 AD) with Proto-Turkic (Menges 1995: 17; Dybo 2007: 75–115; Mudrak 2009:

181) or with Proto-Bulgharic (Janhunen 2010) unsettled. An earlier date for the separation of proto-Turkic, preceding 209 BC would support the identification of Xiongnu language with proto-Bulgharic or one of its subgroups, while a later date of separation would make its association with proto-Turkic more plausible.

The date of the first split in Common Turkic, with the North Siberian Turkic languages branching off, is estimated to be around 1,526 years BP (474 AD), which also seems reasonable in view of the possible historical affiliation of the Yakuts and the Dolgans. Their ancestors are often associated with the historical Kurykan tribes that inhabited the Lake Baikal area in the middle of the first millennium BC (Kormušin 2002: 602).

If Old Turkic indeed represents a separate branch of Common Turkic, rather than its ancestral stage, its separation from the ancestor of all contemporary Turkic languages, except Chuvash, Yakut, and Dolgan, is dated to 1,350 BP (650 AD). This dating is convincing as it precedes the first reliably dated Old Turkic records, notably the inscriptions on stone steels in present-day Mongolia's Orkhon Valley, the earliest of which date to the eighth century.

Divergence dates for more recent splits in Common Turkic seems to be less reliable. In particular at the lowest level, some dates appear to be implausibly shallow. Provisionally, this could be attributed to the fact that closely related Oghuz, Kipchak, and Karluk languages tend to form dialect continua, with parallels lexical and semantic developments that make the original time-depth of their genealogical unity look less remote than it really is.

Greenhill, Currie, and Gray (2009) argued that removal of borrowed items may be problematic in dating cognate data sets including ancient languages because the latter may contain undetectable borrowing whereas loans are excluded from modern languages. As the rate of change is mainly inferred from modern and medieval languages, the measured rate of change would appear lower than the real rate of change. By consequence, the measured time-depth would appear deeper than the real-time depth. This would cause the root of the tree to be placed farther in the past when borrowed items are removed from modern languages. In practice, however, the chronological effect of removing loanwords seems to be rather small. Chang et al. (2015: 220) proposed a median root age of 5,950 BP for Indo-European without loanword extraction vis-à-vis 6,050 BP when loanwords are excluded. The ancient Turkic languages included in our study are considerably less remote in time than Sanskrit or Ancient Greek and, therefore, still allow for the identification of a limited amount of loanwords. As such the chronological effect of excluding loanwords is expected to be even lower than in the Indo-European case. As a matter of fact, conducting rough analyses leaving the loanwords in place and using different settings and variables at different stages of our research, the root always remained within the limits of 500 BC–0 AD, usually around 200–100 BC.

6.3 Relevance for Turkic historical linguistics

Our study contributes to the field of Turkic historical linguistics in the following ways.

First, we collected a new dataset of 254 basic vocabulary items for thirty contemporary Turkic languages and dialects and two ancient language varieties. The compilation of such a comprehensive basic vocabulary based on the recently introduced Leipzig–Jakarta concepts is unprecedented for the Turkic languages.

Second, we here for the first time applied Bayesian phylogenetic inference to a lexical dataset to verify and date previously proposed classifications of the Turkic family. Our study shows that Bayesian-based quantitative methods are, in principle, applicable to the Turkic phylogeny, being able to replicate most of the conventional nodes in the family tree. Of particular interest are the overlaps between our model and the previous quantitative-based—that is, lexicostatistic or morphostatistic—studies on the Turkic phylogeny (e.g. the ultimate Volga Kypchak affiliation of Siberian Tatar). As these studies are based on different datasets and apply different methods, the matches between the models seem to support the reliability and compatibility of different quantitative approaches.

Third, we were able to date our classification, by expressing the uncertainty encountered in previously proposed dating in terms of probability. By yielding maximum credibility intervals for the time depth of the root as well as for the nodes in the family, our Bayesian approach gives statistical underpinning for traditionally discussed time bounds.

Finally, discrepancies between our results and the conventional views on the Turkic phylogeny help us to pinpoint specific problems of Turkic historical linguistics that need to be prioritized in future research. This is, for instance, the case for possible intra-family loans in Turkic basic vocabularies, such as tentative Oghuz loans in West Kipchak or South Kipchak loans in Kirghiz, which remain undetected up to date. Another issue that is raised by our study is the complex relationship between literary Turkic languages and vernaculars. This is the case for Kipchak and Oghuz Turkic varieties spoken in Crimea, which are all labeled as Crimean Tatar

because of using Literary Crimean Tatar as a standard language. As the case of Crimean Tatar shows, the discrepancy between sociolinguistics and genealogy may affect the position of a language in the phylogenetic tree. In order to resolve these issues, it will be necessary to focus future research on lexical borrowing between closely related Turkic varieties and engage in fieldwork to generate basic vocabularies of critical varieties rather than relying on standard dictionaries.

7. Conclusion

In this article, we inferred the internal structure and the time-depth of the Turkic language family, using a Bayesian phylogenetic approach. Our research suggests a binary topology, that is, a topology without polytomies, which replicates most of the traditionally proposed sub-branches of the Turkic tree. In addition, we trace the root of the family back to 66 BC, with credibility intervals between 755 BC and 483 AD. Both the family structure and the dating largely confirm previously made proposals on the basis of qualitative and quantitative linguistic research and are consistent with general historical considerations.

Our tree structure answers the question of whether Siberian Turkic is a valid genealogical node in the negative, but nevertheless supports South Siberian Turkic as a genealogical unity. The ongoing debate on the position of Khalaj in the family seems to be settled in favor of the idea that it does not represent an early branch in the tree. Another thought-provoking finding is the position of Old Turkic. First, it is not ancestral to the Yakut-Dolgan branch as is sometimes thought and, second, there is no reason to believe that it is particularly closer to any of the individual subgroups of Common Turkic, Oghuz, Karluk, or South Siberian Turkic than to the other ones. In addition to verifying the internal structure and time-depth of Old Turkic, our results move the field forward in that they provide a quantitative basis on which to test various competing hypotheses and give us an estimate of the likelihood of the proposed branches. Beyond the field of Turkic linguistics, our insights in the timing and taxonomy of these languages may be relevant for our understanding of broader historical connections between Turkic and the Transeurasian languages.

Future studies in this area may benefit from better documentation of some poorly described varieties of Turkic. Adding their basic vocabularies would improve the quality of the dataset and may lead to clarifying nodes with a less reliable status. In addition, more detailed research in the contact relations among the Turkic languages could help us to exclude undetected borrowings and yield a clearer historical signal.

Abbreviations

Az.	Azeri
Bash.	Bashkir
BTat.	Baraba Tatar
Chuv.	Chuvash
CTat.	Crimean Tatar
Dolg.	Dolgan
Gag.	Gagauz
Kar.	Karaim
Kaz.	Kazakh
KBalk.	Karachay-Balkar
Khak.	Khakas
Khal.	Khalaj
Kirg.	Kirghiz
KKalp.	Karakalpak
KTat.	Kazan Tatar
Kum.	Kumyk
MChul.	Middle Chulym
NAlt.	North Altay
Nog.	Nogai
Sal.	Salar
SAlt.	South Altay
SYug.	Saryg Yugur
Tkm.	Turkmen
Tof.	Tofa
Tur.	Turkish
Tuv.	Tuvan
Uyg.	Modern Uyghur
Uzb.	Uzbek
Yak.	Yakut

Notes

1. Traditional Chinese historiography connects the early Turkic speakers to the Xiongnu tribes who dominated the area north and northwest of China between approximately the third century BC and the second century AD. Evidence on the Xiongnu language is scarce, comprising only several dozen—and often controversial—attestations in Old Chinese chronicles. Yet the assumption that at least some groups in the multiethnic and multilingual Xiongnu confederation were Turkic-speaking has gained some acceptance among historical linguists (Ramstedt 1922: 30–1; Bazin 1948; Gabain 1949; Doerfer 1973; Menges 1995: 17; Dybo 2007: 75–115; Mudrak 2009: 181; Janhunen 2010).
2. Note that the standard Fossilized Birth-Death prior in BEAST 2 does not permit ancient languages to be inferred as direct ancestors to modern languages, so

our analysis cannot unequivocally answer the question of whether Cuman and Old Turkic are ancestral or off on their own branch, although the latter seems the case. Undetected ancient borrowing is unlikely to account for Old Turkic and Cuman clustering together because Cuman was geographically isolated from Old Turkic.

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 program (grant agreement No 646612) granted to Martine Robbeets. We are grateful to Simon Greenhill, Remco Bouckaert, and Nataliia Hübler for their methodological advice and to Johann-Mattis List for his computational support in preparing the data for the analysis. We also greatly benefited from the Bayesian mini-school for Transeurasian linguists that was held in Jena, Germany, on 6–9 November 2017, and would like to especially thank our tutors at that event. Finally, our gratitude goes to Michelle O'Reilly, who helped designing the map of the Turkic language family.

Supplementary data

Supplementary data is available at *Journal of Language Evolution* online. The supplementary data contains the data as it was edited in the EDICTOR tool, an additional Excel Spreadsheet for convenient comparison of the data, as well as the file fed to BEAST in the original nexus format and in the XML format required by BEAST. The data are archived with Zenodo, where they can be downloaded from https://doi.org/10.5281/zenodo.3556954. The data have also been converted to the formats supported by the CLDF initiative (https://cldf.clld.org), they are hosted on GitHub (https://github.com/lexibank/savelyevturkic) and have also been archived with Zenodo (https://doi.org/10.5281/zenodo.3556518).

Conflict of interest statement. None declared.

Appendix 1: 254 basic vocabulary list

The 195 basic concepts taken from the Leipzig–Jakarta list in order from most basic to least basic include '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', and 'tall'.

Additional basic concepts on the Jena list, not present on the Leipzig–Jakarta 200 list include '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', and 2 PL personal pronoun ('you').

References

Baele, G. et al. (2012) 'Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty', *Molecular Biology and Evolution*, 29/9: 2157–67.

Baskakov, N. A. (1952) 'K voprosu o klassifikacii tjurkskix jazykov [On the Classification of the Turkic Languages]', *Izvestija Akademii Nauk SSSR, Otdelenije Literatury i Jazyka*, 11.2: 121–34.

——— (1960) *Tjurkskie jazyki [the Turkic languages]*. Moscow: Izdatel'stvo vostočnoj literatury.

——— (1981) *Altajskaja semja jazykov i ee izučenije [The Altaic language family and its study]*. Moscow: Nauka.

Bazin, L. (1948) 'Un texte proto-turc du IVe siecle: le distique Hiong-nou du *Tsin-chou*', *Oriens*, 1/2: 208–19.

Belikov, V. I. (2009) 'Jazykovyje kontakty i genealogičeskaja klassifikacija [Language Contact and Genealogical Classification]', *Journal of Language Relationship*, 1: 49–68.

Benzing, J. (1959) 'Classification of the Turkic Languages', in Deny, J., Grønbech, K., Scheel, H., and Togan, Z. V. (eds) *Philologiae Turcicae Fundamenta*, pp. 1–5. Wiesbaden: Steiner.

Berta, Á. (1998) 'West Kipchak Languages', in Johanson L. and Csató É. A. (eds) *The Turkic Languages*, pp. 301–17. London: Routledge.

Bogorodickij, V. A. (1934) *Vvedenije v tatarskoje jazykoznanije v svjazi s drugimi tjurkskimi jazykami [An Introduction to Tatar Linguistics as Seen in the Context of Other Turkic Languages]*. Kazan: Tatgosizdat.

Bouckaert, R. (2010) 'DensiTree: Making Sense of Sets of Phylogenetic Trees', *Bioinformatics*, 26/10: 1372–3.

——— et al. (2014) 'BEAST 2: A Software Platform for Bayesian Evolutionary Analysis', *PLoS Computational Biology*, 10/4: e1003537.

Bowern, C., and Atkinson, Q. (2012) 'Computational Phylogenetics and the Internal Structure of Pama-Nyungan', *Language*, 88/4: 817–45.

Campbell, L., and Poser, W. J. (2008) *Language classification: History and method*. Cambridge, UK: Cambridge University Press.

Čečenov, A. A., (1997) 'Poloveckij jazyk [Cuman]', in Tenišev E. R. et al. (eds) *Tjurkskije jazyki*, pp. 110–16. Biškek: Kyrgyzstan.

Chang, W. et al. (2015) 'Ancestry-constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis', *Language*, 91/1: 194–244.

Clauson, G. (1972) *An Etymological Dictionary of Pre-Thirteenth-Century Turkish*. Oxford: Clarendon Press.

Djačok, M. T. (2001) 'Glottoxronologija tjurkskix jazykov (predvaritel'nyj analiz) [The Glottochronology of the Turkic Languages: A Preliminary Analysis]', in *Nauka. Universitet. 2001. Materialy Vtoroj naučnoj konferencii, pp. 14–16*. Novosibirsk: Novyj Sibirskij universitet.

Doerfer, G. (1971) *Grammatik des Chaladsch. (Turcologica 4.)*. Wiesbaden: Harrassowitz.

——— (1973) 'Zur Sprache der Hunnen', *Central Asiatic Journal*, 17: 1–50.

——— (1978) 'Zur Stellung der Chaladsh im Kreise der Türksprachen', *Rocznik Orientalistyczny*, 39/2: 15–31.

———, and Tezcan, S. (1980) *Wörterbuch Des Chaladsch (Dialekt Von Charrab)*. Budapest: Akadémiai Kiadó.

Dunn, M. (2015) 'Language Phylogenies', in Bowern C. and Evans B. (eds) *The Routledge Handbook of Historical Linguistics*, pp. 190–211. London: Routledge.

Dwyer, A. M. (2007) *Salar: A Study in Inner Asian Language Contact Processes*. Wiesbaden: Harrassowitz.

Dybo, A. V. (2007) *Lingvističeskije kontakty rannix tjurkov. Leksičeskij fond. Pratjurkskij period [Linguistic contacts of the early Turks. Lexical stock. Proto-Turkic period]*. Moscow: Vostočnaja literatura.

——— (2002) 'Oguzskaja gruppa [Oghuz Group]', in Tenišev E. R. (ed.) *Sravnitel'no-istoričeskaja grammatika tjurkskix jazykov. Regional'nyje rekonstrukcii*, pp. 7–215. Moscow: Nauka.

——— (2006) 'Xronologija tjurkskix jazykov i lingvističeskije kontakty rannix tjurkov [The Chronology of the Turkic Languages and Contact Relations of Early Turkic Speakers]', in Tenišev, E. R. and Dybo, A. V. (eds) *Sravnitel'no-istoričeskaja grammatika tjurkskix jazykov. Pratjurkskij jazyk-osnova. Kartina mira pratjurkskogo etnosa po dannym jazyka*. Moscow: Nauka, pp. 766–817.

——— (2013) *Etimologičeskij slovar' bazisnoj leksiki tjurkskix jazykov [An etymological dictionary of Turkic basic vocabularies]. (Etimologičeskij slovar' tjurkskix jazykov 9.)*. Astana: TOO "Prosper Print".

Dybo, A. (2016) 'New Trends in European Studies on the Altaic Problem', *Journal of Language Relationship*, 14/2: 71–106.

Fedotov, M. R. (1996) *Etimologičeskij slovar' čuvašskogo jazyka [An Etymological Dictionary of Chuvash]*. Cheboksary: ČGIGN.

Forkel, R. et al. (2018) 'Cross-linguistic Data Formats, Advancing Data Sharing and Reuse in Comparative Linguistics', *Scientific Data*, 5, DOI:10.1038/sdata.2018.205.

Gabain, A. von (1949) 'Review of Louis Bazin. "Un text proto-turc due IVe siecle"', *Der Islam*, 29: 244–6.

Golden, P. (1998) 'The Turkic Peoples: A Historical Sketch', in Johanson, L. and Csató, É. A. (eds) *The Turkic Languages*, pp. 16–29. London: Routledge.

Gray, R. D., and Atkinson, Q. D. (2003) 'Language-tree Divergence Times Support the Anatolian Theory of Indo-European origin', *Nature*, 426: 435–9.

———, Drummond, A. J., and Greenhill, S. J. (2009) 'Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement', *Science*, 323: 479–83.

———, Bryant, D., and ——— (2010) 'On the Shape and Fabric of Human History', *Philosophical Transactions of the Royal Society B*, 365: 3923–33.

Greenhill, S. (2015) An Online Database of New Guinea Languages. *PLoS ONE* 10/10: e0141563.

Greenhill, S. J., Currie, T. E., and Gray, R. D. (2009) 'Does Horizontal Transmission Invalidate Cultural Phylogenies?', *Proceedings of the Royal Society B*, 276: 2299–306.

Haspelmath, M., and Tadmor, U. (2009) *Loanwords in the World's Languages: A Comparative Handbook*. Berlin: Mouton de Gruyter.

Heggarty, P., and Anderson, C. (eds) (2019) *Cognacy in Basic Lexicon (CoBL)*. Jena: Max Planck Institute for the Science of Human History.

Hruschka, D. J. et al. (2014) 'Detecting Regular Sound Changes in Linguistics as Events of Concerted Evolution', *Current Biology*, 25: 1–9.

Huson, D., and Bryant, D. (2006) 'Application of Phylogenetic Networks in Evolutionary Studies', *Molecular Biology and Evolution*, 23: 254–67.

Izdinova, S. R. (1997) 'Krymsko-tatarskij jazyk [Crimean Tatar]', in Tenišev E. R. et al. (eds) *Jazyki mira: Tjurkskije jazyki*, pp. 298–309. Biškek: Kyrgyzstan.

Janhunen, J. (2010) 'Reconstructing the Language Map of Prehistorical Northeast Asia', *Studia Orientalia*, 108: 281–303.

Jankowski, H. (2017) 'Some Notes on Talat Tekin's Classification of Turkic Languages', in Sertkaya, O. F. et al. (eds), *Prof. Dr. Talât Tekin Hatıra Kitabı*, pp. 461–76. İstanbul: Çantay.

Johanson, L. (1998) 'The History of Turkic', in Johanson L. and Csató É. A. (eds) *The Turkic Languages*, pp. 81–125. London: Routledge.

Kassian, A. et al. (2010) 'The Swadesh Wordlist. An Attempt at Semantic Specification', *Journal of Language Relationship*, 4: 46–89.

Kitchen, A. et al. (2009) 'Bayesian Phylogenetic Analysis of Semitic Languages Identifies an Early Bronze Age Origin of Semitic in the near East', *Proceedings of the Royal Society B: Biological Sciences*, 276/1668: 2703–10.

Kormušin, I. V. (2002) 'Tobasskaja gruppa [The Toba Group]', in Tenišev, E. R. (ed.) *Sravnitel'no-istoričeskaja grammatika tjurkskix jazykov. Regional'nyje rekonstrukcii*, pp. 600–60. Moscow: Nauka.

List, J.-M. (2017) 'A Web-based Interactive Tool for Creating, Inspecting, Editing, and Publishing Etymological Datasets', *Proceedings of the EACL 2017 Software Demonstrations, Valencia, Spain, 3–7 April 2017*, pp. 9–12. Valencia: Association for Computational Linguistics.

McMahon, A., and McMahon, R. (2006). Why linguists don't do dates: evidence from Indo-European and Australian languages. In Forster, P. and Renfrew, C. (eds) *Phylogenetic methods and the prehistory of languages*, pp. 153–160. Cambridge, UK: McDonald Institute for Archaeological Research.

Mennecier, P. et al. (2016) 'A Central Asian Language Survey', *Language Dynamics and Change*, 6/1: 57–98.

Menges, K. H. (1959) 'Classification of the Turkic Languages', in Deny, J. et al. (eds) *Philologiae turcicae fundamenta*, pp. 5–8. Wiesbaden: Steiner.

——— (1995) *The Turkic Languages and Peoples. An Introduction to Turkic Studies*. Wiesbaden: Harrassowitz.

Mudrak, O. A. (2009) *Klassifikacija tjurkskix jazykov i dialektov s pomoščju metodov glottoxronologii na osnove voprosov po morfologii i istoričeskoj fonetike [A Glottochronological Classification of the Turkic Languages and Dialects Based on a Questionnaire on Morphology and Historical Phonology]*. Moskva: RGGU.

Nicholls, G. K., and Gray, R. D. (2006) 'Quantifying Uncertainty in a Stochastic Dollo Model of Vocabulary Evolution', in Forster, P. and Renfrew, C. (eds) *Phylogenetic Methods and the Prehistory of Languages*, pp. 161–72. Cambridge: The McDonald Institute for Archaeological Research.

Rambaut, A. et al. (2014) *Tracer v. 1.6*. Institute of Evolutionary Biology, University of Edinburgh.

Ramstedt, G. J. (1922) *Zur Frage nach der Stellung des Tschuwassischen. (Journal de la Société Finno-Ougrienne XXXVIII, 1)*. Helsinki: Suomalais-Ugrilainen Seura.

Robbeets, M. (2016) 'Transeurasian Basic Verbs: Copy or Cognate?', in Csató, É., Karakoç, B. and Menz, A. (eds) *The Uppsala Meeting. Proceedings of the 13th International Conference on Turkish Linguistics*, pp. 199–212. Wiesbaden: Harrassowitz.

———, and Bouckaert, R. (2018) 'Bayesian Phylolinguistics Reveals the Internal Structure of the Transeurasian Family', *Journal of Language Evolution*, 3/2: 145–62.

Róna-Tas, A. (1998) 'The Reconstruction of Proto-Turkic and the Genetic Question', in Johanson, L. and Csató, É. Á. (eds) *The Turkic Languages*, pp. 67–80. London: Routledge.

Samojlovič, A. (1922) *Nekotoryje dopolnenija k klassifikacii tureckix jazykov [Some Addenda to the Classification of the Turkic Languages]*. Petrograd: Rossijskaja Gosudarstvennaja Akademičeskaja Tipografija.

Ščerbak, A. M. (1997) 'Xaladžskij jazyk [Khalaj]', in Tenišev, E. R. et al. (eds) *Jazyki Mira: Tjurkskije Jazyki*, pp. 470–76. Biškek: Kyrgyzstan.

Schönig, K. (1997–1998) 'A New Attempt to Classify the Turkic Languages (1-3)', *Turkic Languages* 1, 1 (1997): 117–33; 1, 2 (1997): 262–77; 2, 1 (1998): 130–51.

Sevortjan, E. V. et al. (1974–2003) *Etimologičeskij slovar' tjurkskix jazykov [An Etymological Dictionary of the Turkic Languages]*. Moscow: Nauka.

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

Stadler, T. et al. (2018) 'The Fossilized Birth-death Model for the Analysis of Stratigraphic Range Data under Different Speciation Modes', *Journal of Theoretical Biology*, 447: 41–55.

Starostin, G. S. (2013) *Jazyki Afriki. Opyt postrojenija leksikostatističeskoj klassifikacii. T. 1: Metodologija. Kojsanskije jazyki [The Languages of Africa. An Attempt at Lexicostatistic Classification. Vol. 1: Methodology. Khoisan Languages]*. Moscow: Jazyki slavjanskoj kul'tury.

Starostin, S. A. (1989) 'Sravnitel'no-istoričeskoje jazykoznanije i leksikostatistika [Historical comparative linguistics and lexicostatistics]', in Kullanda, S. V. et al. (eds) *Lingvističeskaja rekonstrukcija i drevnejšaja istorija Vostoka*, pp. 3–39. Moscow: Nauka.

Starostin, S., A., Dybo, A. V., and Mudrak, O. A. (2003) *Etymological Dictionary of the Altaic Languages*. Leiden: Brill.

Tatarincev, B. I. (2000–2008) *Etimologičeskij slovar' tuvinskogo jazyka [An Etymological Dictionary of Tuvan]*. Vol. 1–5. Novosibirsk: Nauka.

Tekin, T. (1990) 'A New Classification of the Turkic Languages', *Türk dilleri araştırmaları*, 1: 5–18.

Tenišev, E. R. et al. (2001) *Sravnitel'no-istoričeskaja grammatika tjurkskix jazykov. Leksika [A Historical Comparative Grammar of the Turkic Languages. Lexicon]*. Moscow: Nauka.

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